Ein Computerlinguistisches Lexikon als komplexes System

Von der Philosophisch-Historischen Fakultät der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Arne Fitschen

aus Hamburg

Hauptberichter: Prof. Dr. Christian Rohrer
Mitberichter: HD Dr. Ulrich Heid
Mitberichterin: Prof. Dr. Anke Lüdeling

Tag der mündlichen Prüfung: 29. September 2004

Institut für maschinelle Sprachverarbeitung
Universität Stuttgart

2004

D93

Danksagung

Diese Dissertationsschrift entstand während meiner Arbeit am Institut für Maschinelle Sprachverarbeitung (IMS) an der Universität Stuttgart. Sie wäre ohne die offene und freundliche Umgebung, die das IMS bietet, und die kompetente Unterstützung durch die Kollegen nicht möglich gewesen. Mein besonderer Dank gilt hierbei meinem Hauptberichter Christian Rohrer. Sein großes Interesse am Lexikon sorgte dafür, dass ich alle Unterstützung erhielt, die ich mir wünschen konnte, und dass er stets ein offenes Ohr für meine Fragen hatte. Ebenfalls herzlich bedanken möchte ich mich bei meinem Mitberichter Ulrich Heid, der meine Arbeit von Anfang an mit großem Engagement begleitet hat und mich durch seine zahlreichen kompetenten und kritischen Kommentare von manchem Irrweg abbrachte.

Ein herzlicher Dank geht an meine Mitberichterin Anke Lüdeling, die mein Interesse an der Morphologie des Deutschen geweckt hat. Ohne ihre Konzeption des DeKo-Lexikons (gemeinsam mit Tanja Schmid und anderen) hätte es für diese Arbeit keine Grundlage gegeben. Anke hat einen besonderen Beitrag zu dieser Arbeit geleistet, weil sie sich in meiner Phase des Zweifelns viel Zeit genommen hat, mich wieder auf den richtigen Weg zu bringen.

Weiterhin möchte ich mich bei Esther König bedanken, die mir in der Anfangsphase als Ansprechpartnerin zur Seite stand und die mir half, viele Ideen zu entwickeln. Dank ihrer Unterstützung konnte ich Projektarbeit und Dissertation so miteinander verknüpfen, dass sie wechselseitig voneinander profitierten.

Für das Anlegen und Auffüllen eines groß angelegten computerlinguistischen Lexikons bedarf es der Unterstützung durch studentische Hilfskräfte, auf deren Arbeit man sich verlassen kann. Ich möchte mich an dieser Stelle besonders bei André Blessing bedanken, der die graphische Oberfläche für den Lexikonzugriff und das Werkzeug für die automatische Umwandlung der Lexikondaten in eine relationale Datenbank programmiert hat, und bei Stefanie Anstein und Gerhard Kremer, die mit hoher Sprachkompetenz das Lexikon erweitert haben.

Schließlich danke ich den Kollegen und Freunden, die mir fachliche Unterstützung boten, mich die Arbeit aber auch einmal vergessen lassen konnten, allen voran den beiden besten Kollegen der Welt, Stefanie Dipper und Wolfgang Lezius. Für die nötige Ablenkung sorgten auch die tägliche Mensa-Runde, der Stammtisch, das Laufen und Schwimmen. Vielen Dank, Heike Zinsmeister, Jonas Kuhn, Sabine Schulte im Walde, Jasmin Saric, Stefan Evert, Bettina Säuberlich, Arndt Riester, Beate Dorow, Piklu Gupta und Ciprian Gerstenberger!

Ich danke meinen Eltern dafür, dass sie mir all dies ermöglicht haben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Ein Desideratum für die maschinelle Verarbeitung geschriebener Sprache ist ein Verfahren, das die einzelnen Elemente dieser Sprache eindeutig zu identifizieren und zu klassifizieren vermag. Ein solches Verfahren existiert noch nicht und kann in absehbarer Zeit auch nicht erwartet werden, da zum einen durch den produktiven Prozess der Wortbildung die Menge der Elemente nicht endlich ist, also nie vollständig aufgezählt werden kann, zum anderen die Sprache selbst mit vielen Mehrdeutigkeiten aufwartet, die oftmals gar nicht disambiguiert werden sollen. Für den Menschen ist es ein Leichtes, Fehler zu verarbeiten, seien es Tippfehler in Texten, fehlende oder unbekannte Wörter, die durch den Kontext leicht verstanden werden können. Der Computer hingegen kann weder auf ein mentales Lexikon noch auf langfristig gelerntes Weltwissen zugreifen.

Eine notwendige Voraussetzung für ein Werkzeug, das auch dem Computer die Analyse von Sprache ermöglicht, ist eine Ressource, die Informationen zu den Elementen der Sprache enthält. Es sind für einen möglichst großen Teil der in geschriebenen und gesprochenen Texten vorkommenden Einheiten Informationen zu Morphologie, Syntax, Semantik und Phonetik zu speichern, um z.B. die syntaktische Zerlegung der Texte zu ermöglichen (Parsing; benötigt Morphosyntax), die automatische Zusammenfassung oder das schnelle Auffinden bestimmter Informationen zu erleichtern (Information Retrieval; benötigt Morphologie, Semantik) oder das Aussprechen eines Textes durch einen Computer vornehmen zu lassen (Sprachsynthese; benötigt Morphologie, Phonetik).

Zur Zeit sind keine sog. **maschinenlesbaren Wörterbücher** für das Deutsche verfügbar, die ein Werkzeug der skizzierten Art unterstützen. Überhaupt sind nur sehr wenige maschinell gespeicherte Wörterbücher zu finden, die über detaillierte Informationen für eine große Anzahl von in Texten vorkommenden Wörtern verfügen und anspruchsvolle NLP-Anwendungen unterstützen. Die Gründe dafür sind vielfältig: Zum einen muss ein hoher Aufwand getrieben werden, die geschätzten mehreren zehntausend Elemente, die sich durch die Regeln der Wortbildung und Flexion zu einigen Millionen verschiedenen Wort-

formen kombinieren lassen, aufzulisten und mit den benötigten Informationen zu versehen. Zum anderen verbessert sich ein Verfahren zur Analyse von Wortformen nicht automatisch mit zunehmender Lexikongröße. Schließlich herrscht bei einigen morphologischen Phänomenen Uneinigkeit über deren Status bzw. Behandlung (z.B. Konversion: Vgl. die Wortarten von RECHT oder ESSEN in *du hast* RECHT*, lass uns mal* ESSEN *gehen*).

Neben der Lexikonressource selbst stellt sich die Frage nach einer **Datenquelle**, aus der das Material für den Lexikonaufbau oder die Lexikonerweiterung hergenommen wird bzw. anhand derer das Verfahren überprüft werden kann. Schon seit über hundert Jahren beziehen sich Forscher dabei auf *Korpora* geschriebener Texte, die anfangs noch von Hand durchgesehen wurden – Kaeding erstellte schon 1897 für seine Forschungen zu Vorkommenshäufigkeiten von Wortformen in deutschsprachigen Texten ein Korpus mit 11 Millionen laufenden Wörtern –, heutzutage jedoch in elektronisch gespeicherter Form vorliegen. Für die vorliegende Arbeit wird als **empirische Basis** ein Textkorpus aus 200 Millionen Wortformen verwendet, welches überwiegend Zeitungstexte aus den Jahren 1988 bis 1994 umfasst.

Am Institut für Maschinelle Sprachverarbeitung wird seit einigen Jahren am Aufbau eines umfangreichen maschinenlesbaren Wörterbuches gearbeitet, das mit Beginn des DeKo-Projektes (vgl. Schmid et al. (2001)) systematisch um die genannten Informationen erweitert wird. Mit dieser Dissertation wird versucht,

- den Aufbau und die Konzeption dieses Lexikons[1] zu erläutern,
- das Zusammenspiel zwischen der Ressource und der **morphologischen Analyse** der Wortformen aus der Datenquelle zu veranschaulichen und
- darzulegen, wie die Komplexität auf verschiedenen Ebenen (sowohl in Verbindungen innerhalb einzelner Einträge wie auch zwischen Einträgen) gehandhabt werden kann, ohne bei der Qualität der Ressource als Ganzes Abstriche machen zu müssen.

Die aus dem Software-Engineering bekannten Prinzipien der Transparenz und Modularität sorgen für die Erweiter- und Skalierbarkeit der Ressource.

1.1 Motivation: Ein Lexikon für die morphologische Analyse

Bei der maschinellen **morphologischen Analyse** von geschriebener Sprache wird die Morphologiekomponente bei jedem neuen Text mit vorher ungese-

[1] Da es in dieser Arbeit ausschließlich um maschinenlesbare Wörterbücher geht, also die Gefahr einer Verwechslung nicht gegeben ist, wird im Verlauf der gesamten Arbeit der Begriff des *Lexikons* in der Bedeutung *maschinenlesbares Wörterbuch* verwendet.

henen Wortformen konfrontiert. Aufgrund der vielfältigen Möglichkeiten der Wortbildung ist die Menge der potentiell auftretenden Wortformen theoretisch unendlich groß. Aus *denken* entsteht durch Ableitung *bedenken*, daraus *bedenklich*, *unbedenklich* und schließlich *Unbedenklichkeit*. Analog kann *klären* zu *erklären, Erklärung* erweitert werden. Die Substantive, die bei den beiden Ableitungen entstanden sind, lassen sich wiederum zusammenfügen zu einer Wortform *Unbedenklichkeitserklärung*. Daneben gibt es noch *Unbedenklichkeitsbescheinigung, Unbedenklichkeitszeugnis, Unbedenklichkeitsnachweis, Unbedenklichkeitsgutachten* usw. Allen diesen Bildungen ist gemein, dass sie nach bestimmten **Regeln** ablaufen, also von jedem Sprecher[2] des Deutschen problemlos gebildet und auch verstanden werden können. Die beiden Haupt-Wortbildungsmuster im Deutschen, Komposition und Derivation, sind sehr **produktiv**, d.h., sie sind für einen großen Teil der ungesehenen Wortformen verantwortlich. Eine Morphologiekomponente, die **regelbasiert** arbeitet, kann also auch Wortformen analysieren, die nicht als Ganzes in ihrem internen Lexikon verzeichnet sind.

Allerdings gibt es einige Faktoren, die die automatische morphologische Analyse erschweren. Zum einen können bei der Wortbildung morphologische Prozesse stattfinden, die eine Formveränderung der beteiligten Elemente hervorrufen. Im Deutschen sind dies Umlautung, Fugung und Tilgung. So ist *Öfchen* eine Ableitung von *Ofen*, *Häusermeer* ist eine Zusammensetzung von *Haus* und *Meer*. Es muss also nicht nur der Wortbildungstyp ermittelt werden, sondern die Bestandteile müssen einer möglichen Grundform zugeordnet werden. Dies gilt ebenso für Formen, die erkennbar regelhaft gebildet werden (*Biologe, Biologie, biologisch* und *Geologe, Geologie, geologisch*; *Politbüro, Politprofi, Politskandal,* ...), deren vordere Bestandteile sich aber nicht so leicht einer existierenden Form zuordnen lassen. Hier die relevanten Muster und Prozesse zu identifizieren, ist Aufgabe einer **morphologischen Theorie**. Die Bestandteile schließlich einer Morphologiekomponente zur Verfügung zu stellen, um diese in der morphologischen Analyse von Wortformen zu unterstützen, ist die Aufgabe des **Lexikons**. Ein Modell, wie beides miteinander in Einklang zu bringen ist, wird im folgenden Abschnitt vorgestellt.

1.2 Anforderungen an das Lexikon eines Morphologiesystems

Zur Feststellung der Anforderungen an das Lexikon eines Morphologiesystems ist zunächst zu untersuchen, wie eine morphologische Analyse abläuft.

[2] ... und jeder Sprecherin: Das grammatische Geschlecht 'Maskulinum' ist in dieser Arbeit bei Personenbezeichnungen nicht mit dem tatsächlichen Geschlecht zu verwechseln, sondern wird nicht-diskriminierend für beide Geschlechter verwendet.

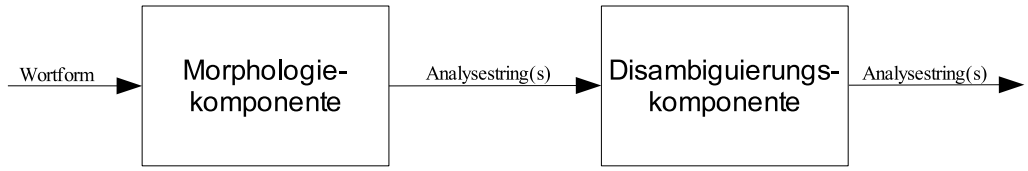

Abbildung 1.1: Morphologische Analyse – Datenfluss

In Abbildung 1.1 ist der Datenfluss bei der morphologischen Analyse (von links nach rechts) wiedergegeben: Ein Modul zur Durchführung der morphologischen Analyse (in der Abbildung und im weiteren Verlauf dieser Arbeit abkürzend **Morphologiekomponente** oder **Morphologiesystem** genannt) erhält als Eingabe eine zu analysierende Wortform und gibt null oder mehr Analysestrings aus. Diese enthalten Angaben zur Wortart, zur Grundform und zu den morphologischen Merkmalen der Eingabewortform. Die Analysestrings durchlaufen fakultativ (sie können auch einfach 'durchgereicht' werden) ein Filtermodul, in dem eine **Disambiguierung** durchgeführt wird. Hier werden nicht gewollte Analysen herausgefiltert. Die verbleibenden Analysestrings (im Idealfall: genau einer) werden an die nachfolgende Verarbeitungseinheit weitergegeben, z.B. eine Syntaxkomponente.

Die Realisierung dieses Datenflusses setzt die Existenz von zwei Komponenten voraus: eine für die morphologische Analyse und eine für die Bewertung und Disambiguierung der Analyseergebnisse. Morphologiesysteme sind vorhanden und bilden die Flexion des Deutschen und teilweise die Wortbildung ab. Eine Komponente zur automatischen korrekten Disambiguierung der Analyseergebnisse hingegen exisitiert noch nicht. Die zentralen Fragen nach der **Korrektheit** einer morphologischen Analyse und nach der **Vollständigkeit** der korrekten Lösungen sind ungeklärt. Dies hängt damit zusammen, dass beide Aspekte nur im Rahmen der morphologischen Theorie einer Sprache betrachtet werden können, dass es aber für das Deutsche keine alle morphologischen Phänomene umfassende allgemein akzeptierte Theorie gibt. Darüber hinaus erzeugen gängige Morphologiesysteme für viele Wortformen überhaupt keine Analyse. Vor der Disambiguierung muss also zunächst einmal eine Untersuchung stattfinden, die die Qualität der Analyseergebnisse bewertet.

In Abbildung 1.2 ist ein Modell vorgestellt, das diese Untersuchung skizziert. In diesem Modell werden die Analysestrings, die von der Morphologiekomponente ausgegeben werden, einer intellektuellen (nicht automatischen) **Bewertung** unterzogen. Die Bewertung richtet sich nach der Behandlung morphologischer Phänomene und der Definition morphologischer Einheiten.

Eine bei den Analysestrings fehlende korrekte Analyse wird als schlechter Fall bewertet und liefert den Anstoß für eine Anpassung des Lexikons. Diese Anpassung wiederum führt i.A. zu einer Verbesserung der Morphologiekompo-

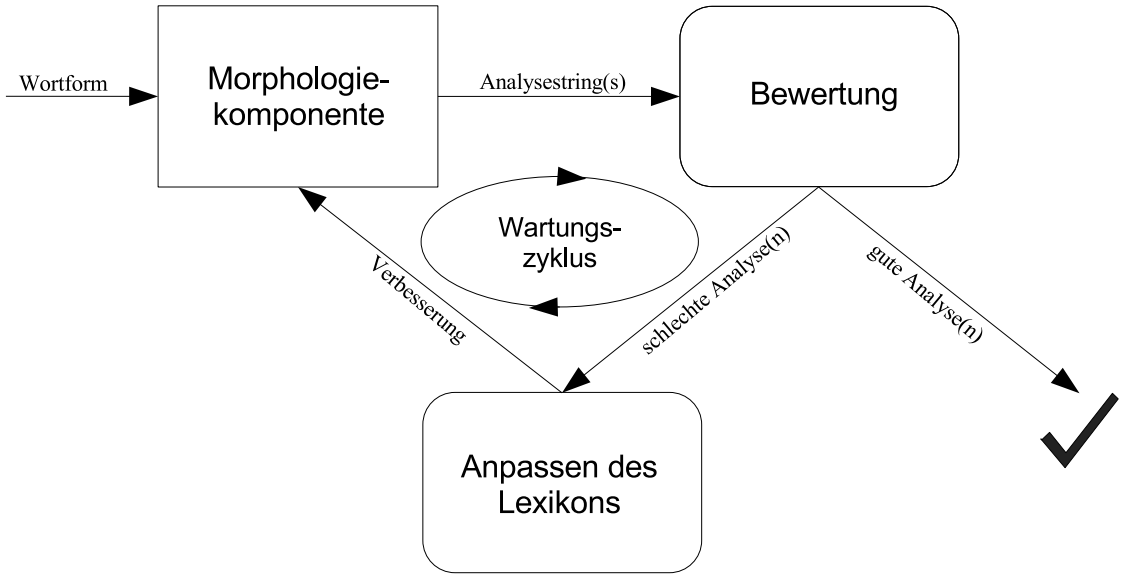

Abbildung 1.2: Wartungszyklus von morphologischer Analyse und Lexikon

nente. Kann z.B. eine einfache Wortform nicht analysiert werden, so wird sie ins Lexikon eingetragen und kann im nächsten Durchlauf des Wartungszyklus dann von der Morphologiekomponente erkannt werden.

Durch die Beschränkung der Menge möglicher Eingabewortformen (z.B. auf alle Wortformen aus einem Korpus) ist es theoretisch möglich, nach mehreren Durchläufen des Wartungszyklus für alle Wortformen 'gute' Analysen zu erhalten. In der Praxis ist dies unwahrscheinlich, da in Textkorpora viele Phänomene vorkommen, die außerhalb des Bereiches regulärer deutscher Morphologie liegen: Tippfehler, Tokenisierungsfehler (z.B. bei Zeilenumbrüchen getrennte Wörter, die nicht wieder zusammengefügt wurden), fremdsprachliches Material, etc.

Neben der Frage nach der Erkennung der 'guten' Analysen stellt sich die Frage nach der Vermeidung der 'schlechten'. Da es noch keine automatische Unterscheidung gibt, kann zunächst nur versucht werden, die automatisch erkennbar ungewollten Analysen aus der Resultatsmenge herauszufiltern. Die im Lexikon gespeicherten Einheiten und die in der Morphologiekomponente enthaltenen Wortbildungsregeln sind für die meisten der in den Analysestrings enthaltenen Ambiguitäten verantwortlich. Daher ist eine Disambiguierung der Analysestrings ohne Kenntnis des Zusammenspiels von Morphologie und Lexikon nicht praktikabel.

Dem Modell liegt eine Theorie der Morphologie zugrunde, die für jedes Phänomen eine adäquate Behandlung vorsieht. Die Realisierung der Theorie verteilt sich auf die Module Morphologiekomponente und Lexikon: Erstere enthält die Wortbildungs- und Flexionsregeln, letzteres speichert die Einheiten, auf de-

nen die Regeln operieren. Wenn für den größten Teil der zu erwartenden Phänomene Regeln aufgestellt sind, wird jedes Durchlaufen eines Wartungszyklus aufgrund fehlender Analysen zu einer Erweiterung oder Änderung des Lexikons führen. Gesetzt den Fall, die Morphologiekomponente weist der Wortform *Feuerwehr* eine Analyse *Feuer=Wehr+NN.Neut* zu (wegen der Simplexform *(das) Wehr* im Lexikon), so wird dies in der Bewertung als eine falsche Analyse erkannt. Um eine richtige Analyse zu erhalten, muss entweder *(die) Feuerwehr* oder *(die) Wehr* ergänzt werden. Aufgrund des Vorkommens anderer Komposita mit *(die) Wehr* als Kopf (*Bürgerwehr, Bundeswehr*) wird die Simplexform eingetragen, so dass zusätzlich zur falschen Analyse nun auch eine Analyse *Feuer=Wehr+NN.Fem* erzielt wird.

1.3 Abgrenzung von verwandten Arbeiten

Ein in Umfang und Zielsetzung dem in dieser Arbeit beschriebenen Lexikon ähnliches computerlinguistisches Lexikon stellt das CISLEX dar (vgl. Langer et al. (1996), Maier-Meyer (1995)). Auch dieses verfolgt das Ziel, die morphologische Analyse (bei CISLEX als **Lemmatisierung** bezeichnet) von Zeitungskorpora mit einer möglichst hohen Abdeckung zu unterstützen. In der vorliegenden Arbeit wird allerdings zum einen der Aspekt der Wortbildung wesentlich stärker betont. (Derivation ist in CISLEX nur für den Bereich "häufiger Suffixe" vorgesehen, die als "spezielle Kategorien" im Lexikon für einfache Formen aufgenommen werden, vgl. Maier-Meyer (1995), S. 32.) Zum anderen liegt der Fokus dieser Arbeit auf der **Struktur** bzw. **Repräsentation** eines Lexikons zur Unterstützung der morphologischen Analyse. Die interne Struktur des CISLEX hingegen wird in den CISLEX-Publikationen nicht weiter aufgeschlüsselt.

Das WordManager-System (vgl. Domenig und ten Hacken (1992)) ist als Entwicklungsumgebung für computerlinguistische Lexika konzipiert worden. Es steht jedoch auch als System zur morphologischen Analyse im Internet zur Verfügung (vgl. CANOO (o.J.)). Dies ist das einzige mir bekannte System, das Komposition und Derivation für das Deutsche umfassend behandelt (also nicht nur für eine Handvoll Suffixe Lösungsansätze bereithält) **und** auch Phänomene der neoklassischen Wortbildung berücksichtigt. Die Internet-Version ist allerings in dieser Hinsicht eingeschränkt: Während die Wortform *Thermohose* gefunden und analysiert wird, ist die Wortform *Thermojacke* dem System unbekannt (CANOO (o.J.) am 1.6.2004). Auch für dieses System gilt allerdings, dass die Struktur und Repräsentation des zugrundeliegenden Lexikons nicht weiter beschrieben wird.

Neben CISLEX und WordManager existieren eine Reihe von Morphologiesystemen für Deutsch, die in Hausser (1996) beschrieben sind. Bei diesen ist oft das Lexikon mit dem Prozessierungssystem verwoben, also z.B. als Prolog-Datenbank oder in einer Lisp-Struktur abgelegt, so dass die Interaktion mit

dem Morphologieprogramm einfacher ist. Die Abhängigkeiten der Module untereinander erschweren in diesen Systemen jedoch die Lexikonerweiterung und verhindern Transparenz.

1.4 Ziele der Dissertation

Die zentrale Fragestellung in dieser Dissertation lautet:

> Wie muss ein computerlinguistisches Lexikon beschaffen sein, um die maschinelle morphologische Analyse optimal zu unterstützen?

Zunächst geht es darum, die Einheiten zu identifizieren, die für die regelbasierte Behandlung morphologischer Phänomene benötigt werden. Die Zusammenhänge zwischen den Einheiten müssen erkannt und mit den Einheiten repräsentiert werden. Erst wenn ein Format gefunden ist, in dem sich eintragsübergreifende Zusammenhänge repräsentieren lassen, ist das Lexikon unter Wahrung der Konsistenz pfleg- bzw. erweiterbar.

Der Beitrag dieser Dissertation für die Forschung liegt in der Vorstellung einer flexiblen Lexikonstruktur, der ein Modell für die Behandlung der in deutschen Gegenwartstexten vorkommenden morphologischen Phänomene zugrundeliegt. Während in der Wortbildungsliteratur seit langer Zeit die Phänomene (kontrovers) beschrieben werden, aber nicht in einem realisierten System auf ihre Praxistauglichkeit hin überprüft werden können, werden in den vorhandenen Morphologiesystemen für das Deutsche Phänomene wie Derivation und neoklassische Wortbildung entweder stiefmütterlich oder gar nicht behandelt. In dieser Arbeit werden sowohl die Phänomene benannt als auch die Möglichkeiten ihrer praktischen Behandlung vorgestellt.

1.5 Empirische Basis

Als Datenquelle wird in dieser Arbeit ein großes Zeitungskorpus verwendet, das geschriebene deutschsprachige Texte aus den Jahren 1988 bis 1994 umfasst (vgl. Tabelle 1.3). Es wird fortan in dieser Arbeit mit dem Kürzel **HGC** (*Huge German Corpus*) bezeichnet. Sofern es nicht anders erwähnt wird, beziehen sich alle Beispielwortformen und Angaben zu ihren Vorkommenshäufigkeiten auf das HGC.

In Tabelle 1.3 sind die Bestandteile das HGC mitsamt der Anzahl der in ihnen enthaltenen **Tokens** aufgelistet. Das HGC umfasst 204 Millionen Tokens, die sich auf 3,2 Millionen verschiedene Typen (**Types**) verteilen. Für die weiteren Untersuchungen wird die Menge der betrachteten Tokens auf diejenigen

Zeitung	Jahrgänge	Korpusgröße in Tokens
Frankfurter Rundschau (FR)	1992 - 1993	40,6 Millionen
Stuttgarter Zeitung (STZ)	1991 - 1993	36,2 Millionen
VDI-Nachrichten (VDI)	1990 - 1991	0,2 Millionen
tageszeitung (TAZ)	1988 - 1994	111,3 Millionen
German Law (Gesetzestexte)	k.A.	5,7 Millionen
Donaukurier (DK)	1992 - 1993	8,4 Millionen
Computerzeitung (CZ)	1993 - 1994	2,1 Millionen
gesamt: HGC	1988 - 1994	204,5 Millionen

Abbildung 1.3: Die Bestandteile des HGC

eingeschränkt, die sich aus den Buchstaben des deutschen Alphabets mit Hinzunahme des französischen *é*[3] zusammensetzen, also *a-z, ä, ö, ü* in Klein- und Großschreibung sowie *ß* und *é*.[4] Von den 204 Millionen Tokens sind dies 166 Millionen oder 81% aller Tokens des HGC. Diese verteilen sich auf 2,1 Millionen Types. Somit umfasst die Datengrundlage für das in dieser Arbeit beschriebene Lexikon- und Morphologiesystem 2,1 Millionen Types.[5]

Für Textkorpora gilt der Grundsatz der *large number of rare events*. Dieser besagt, dass eine große Anzahl Types sehr geringe Tokenhäufigkeiten aufweist, während einige wenige Types sehr hohe Tokenfrequenzen haben. Nach Zipfs Gesetz gilt, dass der Rang einer Wortform in einer nach Häufigkeit absteigend sortierten Liste umgekehrt proportional seiner Vorkommenshäufigkeit ist (vgl. Zipf (1949), Baayen (2001)). Im HGC äußert sich das dergestalt, dass 1,1 Millionen (53%) der 2,1 Millionen Types **Hapax Legomena** (griechisch für 'einmal Gesagtes') oder abkürzend **Hapaxe** sind, also Types, die nur genau einmal im Korpus vorkommen. Dies verweist noch einmal auf die oben angegebene Motivation: Bei diesen 1,1 Millionen Wortformen handelt es sich zum größten Teil um morphologisch komplexe Wörter.

Die Liste der 2,1 Millionen Wortformen und Häufigkeiten, nach Häufigkeiten absteigend sortiert, wird im weiteren Verlauf dieser Arbeit als **HGC-Wortliste** bezeichnet.

[3] *é* ist das einzige fremde Zeichen, das in eingedeutschten Fremdwörtern wie *Café, Varieté* häufiger vorkommt.

[4] Eine ausführliche Analyse von "Sonderformen" und "Mischformen", also solchen, die sich nicht ausschließlich aus Buchstaben zusammensetzen, wird in Maier-Meyer (1995), S. 109ff., vorgenommen.

[5] Unter den 2,1 Millionen Wortformen kommen sehr viele Schreibfehler, fremdsprachiges Material und textsortenspezifische Formen (in Zeitungstexten z.B. Korrespondentenkürzel) vor, deren Behandlung nicht Gegenstand der Morphologie ist. Die Datengrundlage umfasst dennoch weit mehr als eine Million Wortformen, die jeweils eine Analyse erhalten sollen.

1.6 Aufbau der Dissertation

Der Aufbau der Arbeit richtet sich nach dem in Abschnitt 1.2 vorgegebenen Modell (vgl. Abbildung 1.2). Demnach wird zunächst beschrieben, was die morphologische Analyse ist und was ihre Ergebnisse sind (vgl. Kapitel 2). Im Anschluss daran werden die Methoden beschrieben, nach denen die morphologische Analyse durchgeführt wird, und es werden Morphologiesysteme vorgestellt (vgl. Kapitel 3). Als nächstes folgt eine Bewertung der Ergebnisse. Da die Korrektheit und Vollständigkeit von morphologischen Analysen immer nur im Bezug auf die zugrundeliegende Theorie der Morphologie überprüft werden kann, wird hier darauf eingegangen, welche Phänomene auftreten und wie sie behandelt werden sollten (vgl. Kapitel 4). Es wird eine grundlegende Unterteilung nach zwei bekannten Morphologiemodellen vorgenommen. Nachdem die Einheiten eingeführt wurden, wird beschrieben, wie sie in einem Lexikon repräsentiert werden, und es werden einige Lexikonsysteme vorgestellt (vgl. Kapitel 5). Aus der Betrachtung der Lexikonsysteme resultiert die Konzeption des IMSLEX (vgl. Kapitel 6). Die Realisierung des IMSLEX, die Frage danach, wie es mit Informationen gefüllt wird und auf welche Weise die Daten eingesehen und geändert werden können, ist Thema in Kapitel 7. Den Abschluss des Lexikonteils bildet die Frage, wie die Informationen aus dem Lexikon wieder der Morphologiekomponente zugute kommen können (vgl. Kapitel 8). Damit schließt sich der Kreis von morphologischer Analyse, Bewertung der Ergebnisse, Anpassung bzw. Erweiterung des Lexikons und erneuter morphologischer Analyse.

1.7 Notationskonventionen in dieser Arbeit

In Abbildung 1.4 sind Beispiele für die in dieser Arbeit verwendeten Notationskonventionen angegeben. Beispiele für Wortformen, Morpheme, Lexeme etc. sind grundsätzlich *kursiv* gesetzt. Verschiedene Wortbildungsarten können durch die in den Beispielen verwendeten Morphemgrenzenmarkierungen unterschieden werden. Ausgaben von Computerprogrammen werden in `'text type'` gesetzt.

Phänomen	Notationsbeispiele
Lexeme	*gehen*$^{\mathbb{P}}$, *grün*$^{\mathbb{P}}$
Lexeme mit Wortart	*gehen*$^{\mathbb{P}}{}_{V}$, *grün*$^{\mathbb{P}}{}_{ADJ}$
Wortformen	*schön, schönes*
Komposition	*Haus=Tür, Augen=Blick*
Derivation	*Häus·chen, un·be·deut·sam*
Morphemgrenzen allgemein	*Un·be·denk·lich·keits·be·schein·ig·ung*
Affixoide	*super≐reich, Affen≐schande*
Klammerstruktur	`[[Un [[be denk] lich]] keit]`
Klammerstruktur und Wortbildungstypen	`[[be denk]`$_{Derivation}$ `lich]`$_{Derivation}$
Analysestring für *Hauses*	`Haus+NN.Neut.Gen.Sg`
Morphologiestring	`+NN.Neut.Gen.Sg`
Korpusbeleg mit Vorkommenshäufigkeit	*allerdings*$_{(71636)}$, *Marmorkuchen*$_{(8)}$
Korpussatz mit Quelle	*Trockener Marmorkuchen ...* $_{(HGC:4962011)}$

Abbildung 1.4: Notationskonventionen in dieser Arbeit

Kapitel 2

Grundlagen der morphologischen Analyse

In diesem Kapitel wird die morphologische Analyse eingeführt. Ihre zentrale Rolle als Mittler zwischen Textwort und Lexikonwort wird herausgestellt, ihr Ziel und die Methoden vorhandener Morphologiekomponenten, um dieses Ziel zu erreichen, werden vorgestellt.

2.1 Morphosyntaktische Merkmale der Wortform

Bei der morphologischen Analyse handelt es sich um eine Prozedur, die zu einer **Wortform** das zugehörige **Lexem** (bzw. die zugehörigen Lexeme) und die passende(n) Stelle(n) innerhalb des durch das Lexem definierten **Paradigmas** ermittelt. Ein System zur maschinellen Durchführung der morphologischen Analyse wird als **Morphologiesystem** oder **Morphologiekomponente** bezeichnet. In den folgenden Abschnitten wird erklärt, was die Begriffe Wortart, Lexem und Paradigma bedeuten und wie die Entitäten in der Ausgabe der morphologischen Analyse repräsentiert werden.

2.1.1 Die Wortart

Eine zentrale Entität, die die Einteilung des Wortschatzes in definierte Klassen erlaubt, ist die **Wortart**[1]. Sie ist keineswegs unveränderbar vorgegeben: "Beispielsweise gibt es eine lange Diskussion darüber, wieviele und welche Wortarten das Deutsche hat." (Eisenberg (1994), S. 19) Schuch diskutiert ausführlich die verschiedenen Sichtweisen und ordnet sie in Typen, die sich eines oder mehrerer Kriterien aus einer Menge der syntaktischen, semantischen, morphologi-

[1]Der theoretische Status der Wortart ist in dieser Arbeit nicht von Belang. Ihre Einordnung als morphosyntaktisches Merkmal bietet den großen Vorteil, dass jede Wortform über mindestens ein solches Merkmal verfügt.

schen oder phonologischen Kriterien bedienen (vgl. Schuch (1990), S. 73ff.). Sie kommt zu dem Schluss, dass "Wortartkategoriebildung [..] der Versuch [ist], so allgemein wie möglich Gemeinsamkeiten und Unterschiede in den grammatischen Verwendungspotentialen lexikalischer Einheiten einer Sprache zu erfassen. [...] Diese Verhaltenseigenschaften manifestieren sich aber nicht als Eigenschaften bestimmter irgendwie vorgegebener Wortartkategorisierungen." (Schuch (1990), S. 79)

Eisenberg nimmt eine Unterteilung in **offene** und **geschlossene** Wortarten bzw. -klassen vor, die auch häufig in der Literatur zu finden ist: "Die grammatischen Kategorien als Wortarten sind nach Auffassung fast aller Grammatiken in zwei Gruppen zu unterteilen, nämlich die *lexikalischen* oder *offenen Kategorien* Substantiv, Verb, Adjektiv und Adverb und die *Funktionswörter* oder *abgeschlossenen Kategorien* Präposition, Partikel, Konjunktion, Artikel und Pronomen. [...] Von offenen Kategorien spricht man, weil die Zahl der Substantive, Verben, Adjektive und Adverbien groß ist und sich relativ schnell verändert." (Eisenberg (1994), S. 34)

Eine weitere Unterteilung ist möglich nach dem morphologischen Verhalten in **flektierende** und **nicht flektierende** Klassen. Bei den offenen Klassen sind die Adverbien nicht flektierend, bei den geschlossenen die Präpositionen, Partikeln und Konjunktionen. Das Flexionsverhalten einer Wortart ist relevant für die Begriffe Lexem und Paradigma.

2.1.2 Flexionsparadigma und Lexem

Grammatische Eigenschaftsklassen[2] wie Numerus, Genus, Person etc. definieren ein **Flexionsparadigma** (im Folgenden kurz **Paradigma**) für einen Vertreter einer bestimmten Wortart: Die Anzahl der möglichen Ausprägungen einer Kategorisierung, der Kategorien, bestimmt die Menge der Plätze, die in einem Paradigma zu einer Wortart für eine Kategorisierung zur Verfügung gestellt werden müssen. Gemeinsam definieren die Merkmale ein abstraktes Konstrukt, das als **Lexem** bezeichnet wird. Da es umständlich wäre, immer das komplette Paradigma anzugeben, um auf ein Lexem zu verweisen, wird ein Lexem durch eine per Konvention ausgewählte Form repräsentiert, die als **Lemma** oder **Grundform** bezeichnet wird.[3] Lexeme werden im weiteren Verlauf dieser Arbeit durch die Angabe des Lemmas mit einem hochgestellten *P* (für *Paradigma*) notiert: $Haus^{\mathbb{P}}$, $schnell^{\mathbb{P}}$, $gehen^{\mathbb{P}}$. Geht die Wortart nicht aus dem unmittelbaren Zusammenhang hervor, wird sie als Index mit angegeben: $Horst^{\mathbb{P}}{}_{NN}$, $licht^{\mathbb{P}}{}_{ADJ}$.

[2] **Kategorisierungen** nach Eisenberg: "Kategorisierungen sind Mengen von Kategorien" (Eisenberg (1994), S. 38).

[3] Die Grundform entspricht oft einer Form aus dem Paradigma. Daher wird sie in der Literatur oft mit einer Wortform gleichgestellt. Es handelt sich aber lediglich um einen Bezeichner für das Gebilde *Lexem*. Die Begriffe *Lexem* und *Lemma* werden in dieser Arbeit hingegen (wie allgemein üblich) leicht unscharf synonym zueinander verwendet.

Die grammatischen Kategorisierungen werden gewöhnlich in einen engen Zusammenhang mit der Wortart gebracht. Einige wenige Kategorisierungen sind dem Lexem inhärent. Ein Beispiel dafür ist das Genus bei Substantiven: Diese Eigenschaft ist an das Substantiv gebunden und unveränderlich. Sie lässt sich i.A. nicht an der orthographischen oder phonetischen Form festmachen (vgl. etwa *der Kutter, die Butter, das Futter*). Eisenberg bezeichnet diese inhärenten Eigenschaften als "Paradigmenkategorisierungen" (Eisenberg (1994), S. 40). Andere ergeben sich in der syntaktischen Verwendung der Wortformen, wie der Kasus und der Numerus bei Substantiven und Adjektiven.

	Sg	Pl
Nom	*Gefährt*	*Gefährte*
Gen	*Gefährts*	*Gefährte*
Dat	*Gefährt(e)*	*Gefährten*
Akk	*Gefährt*	*Gefährte*

	Sg	Pl
Nom	*Gefährte*	*Gefährten*
Gen	*Gefährten*	*Gefährten*
Dat	*Gefährten*	*Gefährten*
Akk	*Gefährten*	*Gefährten*

Abbildung 2.1: Paradigmen von *Gefährt*$^{\mathbb{P}}$ und *Gefährte*$^{\mathbb{P}}$

In Abbildung 2.1 sind die Paradigmen für zwei Substantive dargestellt, *Gefährt*$^{\mathbb{P}}$ und *Gefährte*$^{\mathbb{P}}$. Die Anzahl aller mit Wortformen zu belegenden Plätze berechnet sich also aus zwei (Kategorisierung Numerus mit Kategorien Singular und Plural) mal vier (Kategorisierung Kasus mit Kategorien Nominativ, Genitiv, Dativ und Akkusativ). Es ist zu beachten, dass es im Dativ Singular für *Gefährt*$^{\mathbb{P}}$ zwei Wortformen gibt, *(dem) Gefährt* und *(dem) Gefährte* (im Beispiel durch die runden Klammern angedeutet). Die letzte Form ist veraltet, aber beide sind grammatikalisch korrekt. In vielen Fällen unterscheiden sich die Wortformen auf verschiedenen Plätzen nicht, wie bei *Gefährt*$^{\mathbb{P}}$ am Beispiel des Nominativ Singular und Akkusativ Singular gezeigt. Dies wird als **Synkretismus** bezeichnet. Auch paradigmenübergreifend kann es zu identischen orthographischen Formen kommen (**Homonymie**).

Die Paradigmen anderer flektierender Wortarten lassen sich nicht mehr so leicht tabellarisch darstellen: Bei Adjektiven müssen neben Kasus, Numerus und Genus noch die starke, gemischte und schwache Flexion berücksichtigt werden, die sich ergeben, wenn kein Artikelwort, unbestimmter Artikel oder bestimmter Artikel vor dem Adjektiv steht, und es kommen die Steigerungsformen Positiv, Komparativ und Superlativ hinzu. Bei Verben müssen neben Person und Numerus noch die Kategorisierungen Tempus und Modus berücksichtigt werden, aber zusätzlich gibt es die infiniten Kategorien Imperativ, Partizip und Infinitiv, bei denen die genannten Kategorisierungen größtenteils irrelevant sind.

Ein Paradigma muss nicht vollständig gefüllt werden. Es gibt **defektive** Paradigmen, bei denen Formen fehlen. Dies gilt z.B. bei Verben wie *regnen*, die im

Allgemeinen nur in der dritten Person Singular mit einem expletiven *es* verwendet werden: *Es regnet.* Ebenso aus semantischen Gründen sind manche Adjektive nicht steigerbar (*tiefblau, endlos*) oder gibt es für manche Substantive keine Pluralformen (*Akribie, Durst, Tod*: Singulariatantum) bzw. keine Singularformen (*Kosten, Leute*: Pluraliatantum). Allerdings sind Sprecher des Deutschen durchaus in der Lage, die fehlenden Formen zu bilden bzw. sie zu erkennen: *?Ich regne Sterne für Dich. Ich sterbe tausend Tode.* Es erscheint also durchaus plausibel, diese defektiven Paradigmen in einem Morphologiesystem genauso wie die normalen Paradigmen zu behandeln.

2.2 Die Aufgabe der morphologischen Analyse

Die morphologische Analyse hat traditionell zweierlei Aufgaben: Zum einen muss sie für eine gegebene Wortform ein dazugehöriges oder mehrere dazugehörige Lexeme identifizieren. Zum anderen muss sie die Stellen des Paradigmas ermitteln, die der Wortform entsprechen. Bei der morphologischen Analyse handelt es sich also um eine Prozedur, die eine Wortform in die zugehörigen morphosyntaktischen Merkmale zerlegt. Die Ausgabe[4] besteht aus der Grundform, der Wortart und den weiteren syntaktischen Kategorien. Die Wortform *Hauses* beispielsweise lässt sich dem Lexem *Haus*$^{\mathbb{P}}$ in der Wortart Substantiv zuordnen, der Kasus ist Genitiv und der Numerus ist Singular. Diese Informationen werden im Folgenden in einem **Analysestring**[5] notiert (vgl. Beispiel 2.1). Die Grundlage für diese Notation bildet der Standard **STTS** (Stuttgart-Tübingen Tagset, vgl. Schiller et al. (1999)).[6]

(2.1) `Haus+NN.Neut.Gen.Sg`

Für den Fall, dass das Lemma in einem Zusammenhang nicht relevant ist, wird ein **Morphologiestring**, der nur die Informationen zur Wortart und zu den morphosyntaktischen Kategorien enthält, verwendet (vgl. Beispiel 2.2).

(2.2) `+NN.Neut.Gen.Sg`

[4]Da die morphologische Analyse i.A. von einer Morphologiekomponente durchgeführt wird, werden die Begriffe *Resultat* und *Ausgabe* der morphologischen Analyse in dieser Arbeit synonym zueinander verwendet.

[5]Die Bezeichnung *grammatisches Wort*, die hierfür in der Literatur zu finden ist, halte ich für missverständlich in seiner Konnatation zum Begriff des *Wortes*, da es doch gerade nicht um die Einheit *Wort*, sondern um die Darstellung der einer Wortform inhärenten morphosyntaktischen Merkmale geht.

[6]Kategorien lassen sich jeweils eindeutig einer Kategorisierung zuweisen. Eine Auflistung aller Kategorisierungen und der dazugehörigen Kategorien für das Deutsche findet sich in Schiller et al. (1999). Sie ist zusätzlich in Anhang B auf Seite 141 angegeben. Eine formale Beschreibung der Syntax von Analysestrings im EBNF-Format findet sich in Anhang A auf Seite 139.

Die Aufgabe der morphologischen Analyse ist es, zu einer gegebenen Wortform alle zu dieser passenden Analysestrings auszugeben. Es handelt sich bei Analyse- und Morphologiestring lediglich um eine Notationsform: Bei verschiedenen Morphologiekomponenten kann diese Ausgabe verschiedene Formen annehmen, allen gemeinsam ist allerdings, dass das Lemma, die Wortart und die grammatischen Kategorien in der Ausgabe enthalten sind.

Wortform	Analysestring	Lemma	Morphologiestring
Hauses	`Haus+NN.Neut.Gen.Sg`	*Haus*$^{\mathbb{P}}$	`+NN.Neut.Gen.Sg`
Gefährte	`Gefährte+NN.Masc.Nom.Sg`	*Gefährte*$^{\mathbb{P}}$	`+NN.Masc.Nom.Sg`
	`Gefährt+NN.Neut.Akk.Pl`	*Gefährt*$^{\mathbb{P}}$	`+NN.Neut.Akk.Pl`
	`Gefährt+NN.Neut.Gen.Pl`		`+NN.Neut.Gen.Pl`
	`Gefährt+NN.Neut.Nom.Pl`		`+NN.Neut.Nom.Pl`
	`Gefährt+NN.Neut.Dat.Sg`		`+NN.Neut.Dat.Sg`
denn	`denn+ADV`	*denn*$^{\mathbb{P}}{}_{ADV}$	`+ADV`
	`denn+KONJ.Kon`	*denn*$^{\mathbb{P}}{}_{KONJ}$	`+KONJ.Kon`

Abbildung 2.2: Wortformen und ihre morphologische Analyse (I)

In Tabelle 2.2 sind einige Beispielwortformen mitsamt ihren jeweiligen Grundformen, Analyse- und Morphologiestrings aufgelistet. Fast immer existieren mehrere Analysestrings zu einer Wortform (*Hauses* mit genau einer Analyse ist die Ausnahme im Beispiel). Auch bei nicht flektierenden Wortarten kann es mehrere Analysestrings zu einer Wortform geben, wenn etwa eine Wortform wie *denn* als Adverb oder als Konjunktion auftreten kann. Der minimale Analysestring besteht immer aus einem Lemma und einer Wortart.[7]

2.3 Der Status der Wortbildung in der morphologischen Analyse

Bis hierhin wurde ausschließlich die Flexion beschrieben. Die innere Form morphologisch komplexer Wortformen[8] ist jedoch auch Gegenstand der Morphologie. Da sich Wortbildungsprodukte aus Bestandteilen zusammenzusetzen, die selber wieder flektieren können, muss der Zusammenhang von Flexion

[7]Die Kategorie `Kon` im Beispiel der Konjunktionslesart der Wortform *denn* ist ein Beispiel für eine rein syntaktische Kategorie: Sie spezifiziert die Konjunktion als koordinierende Konjunktion. Bei einigen nicht flektierenden Wortarten sieht das STTS rein syntaktische Kategorien vor.

[8]Der Begriff *morphologisch komplexe Wortform* bezieht sich in dieser Arbeit immer auf die Wortbildung, nicht auf die Flexion. Es handelt sich immer um eine Form, die einen Wortbildungsprozess durchlaufen hat.

und Wortbildung betrachtet werden. Ein weiterer wichtiger Aspekt betrifft die **Struktur** von Wortbildungsprodukten. Beide Aspekte werden in den nachfolgenden Abschnitten behandelt.

2.3.1 Der Zusammenhang von Flexion und Komposition

Wortform	Analysestring	Lemma
Schiffskapitäns	`Schiffs=Kapitän+NN.Masc.Gen.Sg`	*Schiffskapitän*$^{\mathbb{P}}$
Kapitäns	`Kapitän+NN.Masc.Gen.Sg`	*Kapitän*$^{\mathbb{P}}$
Anzeigenadel	`Anzeigen=Adel+NN.Masc.Akk.Sg`	*Anzeigenadel*$^{\mathbb{P}}$
	`Anzeigen=Adel+NN.Masc.Nom.Sg`	
	`Anzeigen=Adel+NN.Masc.Dat.Sg`	
	`Anzeige=Nadel+NN.Fem.Akk.Sg`	*Anzeigenadel*$^{\mathbb{P}}$
	`Anzeige=Nadel+NN.Fem.Dat.Sg`	
	`Anzeige=Nadel+NN.Fem.Gen.Sg`	
	`Anzeige=Nadel+NN.Fem.Nom.Sg`	

Abbildung 2.3: Wortformen und ihre morphologische Analyse (II)

In Tabelle 2.3 sind einige morphologisch komplexe Beispielwortformen mitsamt ihren jeweiligen Analysestrings und Grundformen aufgelistet. In den Analysestrings für Komposita sind bei den Grundformen die Grenzen der Bestandteile durch ein Gleichheitszeichen markiert.[9] Es zeigt sich, dass sich die Flexionsinformation nach dem am weitesten rechts stehenden Bestandteil richtet. Andersherum ausgedrückt: Ist der Analysestring der Wortform *Kapitäns* bekannt und handelt es sich bei einer Wortform um ein Kompositum mit dem Kopf *Kapitäns*, so stimmt der Morphologiestring des Kompositums mit dem der Wortform überein. Bei der Wortform *Anzeigenadel* in Tabelle 2.3 zeigt sich eine strukturelle Mehrdeutigkeit, die Einfluss auf die Flexion besitzt: Je nachdem, ob die Wortform *Adel* oder *Nadel* den Kopf des Kompositums bilden soll, gibt es verschiedene Analysestrings, die jeweils denen von *Adel* bzw. *Nadel* entsprechen.

Der Nutzen der Kompositumszerlegung für die Durchführung der Flexionsanalyse liegt auf der Hand: Ist bekannt, dass der Kopf des Kompositums *Donaudampfschifffahrtskapitäns* die Wortform *Kapitäns* ist, so muss die Flexionsinformation übereinstimmen. Diese Tatsache ermöglicht erst die regelbasierte Durchführung der morphologischen Analyse.

[9] Strenggenommen kann also nicht mehr gesagt werden, dass im Analysestring die Grundform mit angezeigt wird. Da diese jedoch immer noch eindeutig rekonstruierbar ist, wird im Folgenden auch die mit Morphemgrenzen versehene Zeichenkette der Einfachheit halber als *Grundform* bezeichnet.

2.3.2 Die Analyse der Wortbildungsstruktur

Bei Wortbildungen gibt es zwei Aspekte, die Gegenstand einer morphologischen Analyse sind: Zum einen sind dies die Bestandteile, aus denen sich eine morphologisch komplexe Form zusammensetzt, zum anderen ist es die *Struktur* eines Wortbildungsproduktes. Die Erkennung der Struktur setzt die Kenntnis der Bestandteile voraus. Die morphologische Analyse von Wortbildungsprodukten umfasst beide Aspekte. Sie wird allgemein als **Wortbildungsanalyse** oder **erweiterte morphologische Analyse** bezeichnet.

(2.3) *Un·be·denk·lich·keits·er·klär·ung*

Die Zerlegung einer Wortform in Morpheme[10] wird als **Segmentierung** bezeichnet. Für die Wortform *Unbedenklichkeitserklärung* kann eine wie in 2.3 dargestellte Segmentierung angegeben werden. Aus der linearen Struktur, die alle Morpheme als gleichberechtigt darstellt, lässt sich allerdings weder die **hierarchische Struktur** der Zerlegung erkennen, noch die Wortbildungsart ablesen. Die hierarchische Struktur eines Wortbildungskonstrukts zeigt sich erst bei sukzessiver Zerlegung in **unmittelbare Konstituenten**. Bis auf wenige Ausnahmen können komplexe Wortformen im Deutschen in jeweils zwei unmittelbare Konstituenten zerlegt werden.[11]

(2.4) *Unbedenklichkeits=Erklärung*

Für die Wortform aus 2.3 ist eine plausible Zerlegung in unmittelbare Konstituenten in 2.4 dargestellt. Beide unmittelbaren Konstituenten sind selber wieder Wortbildungen, die sich weiter zerlegen lassen.

Eine Baumdarstellung zeigt eine mögliche Struktur der Wortbildungskonstruktion mit allen Hierarchie-Ebenen (vgl. Abbildung 2.4). Der jeweilige Kopf einer Untergliederung ist in der Baumdarstellung unterstrichen.

Sobald eine Wortbildung mehr als zwei Bestandteile aufweist, sind mehrere Baumdarstellungen möglich (vgl. Abbildung 2.5). Die linke Darstellung zeigt eine Lesart als Kompositum, bei der die Adjektive *klein* und *städtisch* zusammengesetzt werden. Die rechte Darstellung zeigt eine Lesart als Derivativum, bei dem das Substantiv *Kleinstadt* mit dem Adjektivsuffix *-isch* zu einem Adjektiv abgeleitet wird. Mit zunehmender Anzahl an Bestandteilen nimmt die Anzahl möglicher Strukturdarstellungen zu.

Auch in Abbildung 2.5 ist der jeweilige **Kopf** einer Hierarchie-Ebene unterstrichen dargestellt.[12]

[10] Zum Begriff des Morphems vgl. Abschnitt 4.2.1. Das Zeichen · kennzeichnet Morphemgrenzen innerhalb einer Wortform.

[11] Eine Ausnahme stellen Komposita dar, die mehr als zwei gleichberechtigte Bestandteile nebeneinanderstellen: *schwarzrotgold*.

[12] In dieser Arbeit wird Derivationsaffixen (wie *-isch, -lich* und *-keit*) Kopfstatus zugesprochen.

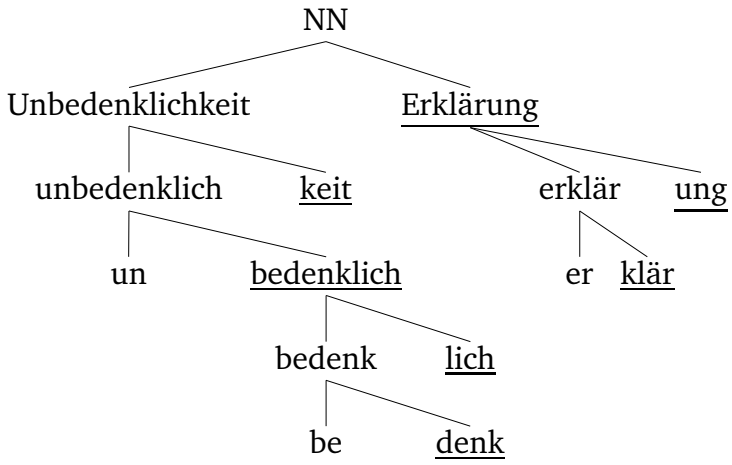

Abbildung 2.4: Struktur des Kompositums *Unbedenklichkeitserklärung*

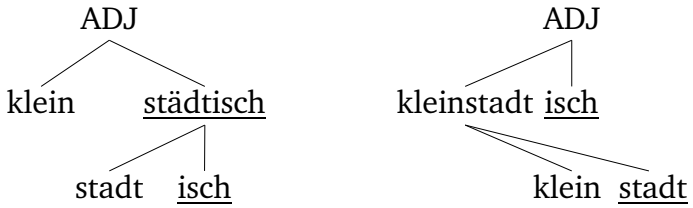

Abbildung 2.5: Strukturen der komplexen Form *kleinstädtisch*

(2.5) `[[klein stadt]`$_{Komposition}$ `isch`$_{Derivation}$`]+ADJ.Pos.Adv`

Die Ausgabe einer erweiterten morphologischen Analyse mit Angaben zur Struktur der Wortbildungskonstruktion kann wie in Beispiel 2.5 gezeigt dargestellt werden. Für die Wortform *kleinstädtisch* sind in den Analysestrings die Flexion sowie die einzelnen Wortbildungsmuster, aus denen sich die Wortform zusammensetzt, angegeben.

(2.6) `[[klein(`$klein^{\mathbb{P}}{}_{ADJ}$`) städt(`$Stadt^{\mathbb{P}}{}_{NN}$`)]`$_{Komposition}$
`isch(-`$isch^{\mathbb{P}}{}_{ADJSuff}$`)]`$_{Derivation}$`+ADJ.Pos.Adv`

In Beispiel 2.6 sind zusätzlich zu der Darstellung in Beispiel 2.5 noch die Wortbildungskomponenten genauer spezifiziert. Dieses Beispiel veranschaulicht, was eine morphologische Analyse, die Flexion und Wortbildung berücksichtigt, an Informationen ausgeben kann.

2.3.3 Die Produktivität von Wortbildung

Wortbildungsmuster wie Komposition und Derivation sind im Deutschen sehr produktiv. “Der deutsche Wortschatz besteht zum größten Teil, nämlich zu etwa

zwei Dritteln, aus Nominalkomposita." (vgl. Ortner et al. (1991), S. 3) Die **Reihenbildung** ist in der Wortbildung stark vertreten. Ist eine bestimmte Wortbildung eingeführt, so können nach demselben Muster zahlreiche Neubildungen entstehen. Zwei Beispiele sollen dies verdeutlichen.

> *argus-, beid-, blau-, blind-, braun-, dunkel-, ein-, flamm-, fremd-, frosch-, gelb-, glanz-, glotz-, glubsch-, glut-, glüh-, groß-, grün-, hell-, hohl-, kalt-, katzen-, klein-, knopf-, kuh-, kuller-, luchs-, mandel-, mond-, reh-, rot-, samt-, scharf-, schlitz-, schön-, tief-, trief-, vieläugig*

Abbildung 2.6: Wortbildungen mit *-äugig* aus dem HGC

Als erstes Beispiel sind hier die Wortbildungen auf *-äugig* aufgelistet, die im Korpus vorkommen (vgl. Abbildung 2.6).[13] Es handelt sich insofern um ein inhomogenes Muster, als die Komponenten links von der Zeichenkette *-äugig* verschiedenen Wortarten angehören. Für die Flexion der Wortformen ist dies jedoch irrelevant.

> *Politaktivist, -amateur, -apparat, -barometer, -bonze, -bühne, -büro, -clown, -elite, -freak, -funktionär, -geschäft, -größe, -kabarett, -karriere, -kaste, -kern, -kitsch, -kommissar, -krimi, -landschaft, -magazin, -management, -manager, -neuling, -offizier, -parole, -poker, -posse, -profi, -programm, -promi, -prominenz, -propaganda, -prozeß, -rentner, -satire, -sekte, -sendung, -show, -skandal, -song, -spektakel, -star, -stratege, -sumpf, -szene, -theater, -thriller, -tourismus, -tourist, -täter, -unterricht, -verein, -zirkel, -ökonom, -ökonomie, . . .*

Abbildung 2.7: Wortbildungen mit *Polit-* aus dem HGC

Als zweites Beispiel sind Wortbildungen aufgelistet, die mit *Polit-* beginnen und als zweite Komponente ein Substantiv haben (vgl. Abbildung 2.7). Es handelt sich nur um einen kleinen Ausschnitt, die vollständige Liste für das HGC umfasst einige hundert Wortformen.

2.4 Abdeckung und Korrektheit

Für das eingangs formulierte Ziel, die optimale Unterstützung der morphologischen Analyse, ergibt sich aus diesem Kapitel folgende Aussage: Die Erkennung der Bestandteile morphologisch komplexer Wortformen ist für die **Abdeckung**, die die morphologische Analyse erzielt, sehr förderlich. Tritt eine Wortform,

[13] Die Wortbildungen mit *äugig* und ihre Verarbeitung werden in Abschnitt 8.2 auf Seite 125 wieder aufgegriffen.

die morphologisch analysiert werden kann, als Kopf von Wortbildungen auf, so können diese automatisch ebenfalls morphologisch analysiert werden. Dies gilt unabhängig von der Art und Anzahl der Komponenten links vom Kopf: *Börse, Geldbörse, Ledergeldbörse* und *Rinderledergeldbörse* erhalten alle denselben Morphologiestring.[14]

Was die **Korrektheit** der Analyseergebnisse angeht, so kann die morphologische Analyse von der Erkennung der Bestandteile morphologisch komplexer Wortformen insofern profitieren, als ein einmal gefundener Fehler systematisch bei den anderen Formen mit demselben Kopf korrigiert werden kann. Die Korrektheit von Wortbildungszerlegungen und Wortbildungsmustern hingegen hängt ausschließlich davon ab, ob die zur Zerlegung notwendigen Bestandteile identifiziert werden können und ob für eine Wortbildung ein Wortbildungstyp bekannt ist. Diese Fragen nach dem Lexikon, das die potentiellen Bestandteile enthält, und einem Regelapparat, der die Wortbildungsregeln enthält, werden im nachfolgenden Kapitel behandelt.

[14]Dies gilt natürlich nur, wenn das Lexikon die Lexeme $Geld^{\mathbb{P}}$, $Leder^{\mathbb{P}}$ und $Rind^{\mathbb{P}}$ enthält.

Kapitel 3

Methoden der morphologischen Analyse

In diesem Kapitel werden Verfahren für die automatische morphologische Analyse von Wortformen vorgestellt. Nachdem im letzten Kapitel geklärt wurde, **was** die morphologische Analyse bezweckt, geht es hier darum, **wie** sie arbeitet. Zunächst wird gezeigt, wie die Aufgabe computerlinguistisch modelliert werden kann (Abschnitt 3.1). Im Anschluss daran wird ein beispielhaft ausgewähltes regelbasiert arbeitendes Morphologiesystem ausführlich vorgestellt (Abschnitt 3.2). Abschnitt 3.3 schließlich fasst die Ergebnisse aus dem vorherigen und diesem Kapitel zusammen.

3.1 Computerlinguistische Modellierung

Morphologiekomponenten wurden in dieser Arbeit bislang als 'black boxes' angesehen, die zu einer vorgegebenen Eingabe eine definierte Ausgabe erzeugen. In diesem Abschnitt wird in diese 'black boxes' hineingeschaut. Erst werden verschiedene Vorgehensweisen präsentiert, die morphologische Analyse anzugehen, danach wird exemplarisch die derzeit vorherrschende Methode maschineller morphologischer Analyse, die sich der Finite-State-Transducer bedient, vorgestellt.

3.1.1 Vollformlexikon vs. regelbasiertes System

Es gibt zwei grundsätzliche Möglichkeiten der Konstruktion eines Morphologiesystems:

(a) die simple Auflistung aller Wortformen mitsamt ihren Analysestrings oder

(b) die Verwendung eines regelgesteuerten Systems.

Variante (a), auch **Vollform(en)lexikon** genannt, ist nicht praktikabel, da ständig neue Wörter gebildet werden, die Liste also nie vollständig sein kann. "Ohne regelgesteuerte Wortzerlegung sind die Systeme auf die individuelle Erfassung der Einzellexeme angewiesen und können mit dem ständig wachsenden Lexikon der natürlichen Sprachen nicht fertig werden." (Hausser (1996), S. 19) In der Liste nicht enthaltene Wortformen werden nicht erkannt. Trotz dieser Einschränkungen existiert mit der *CELEX Lexical Database* eine solche Ressource für die Sprachen Niederländisch, Deutsch und Englisch.[1] Der Zugriff auf diese Ressource erfolgt durch simples Nachschauen, ob eine Wortform enthalten ist.

Variante (b) umfasst alle Systeme, in denen Flexion und/oder Wortbildung auf Regeln zurückgeführt werden. Die Regeln können fest mit der Programmlogik verbunden sein oder aber getrennt vom Programm explizit vorliegen. Sie operieren auf den Einheiten, die im Lexikon einer Morphologiekomponente abgelegt sind. Dies können Lexeme, Stämme oder Morpheme sein. Dazu kommen Flexionselemente. In Stammlexika sind allomorphe Stämme einzeln aufgelistet, es gibt also beispielsweise zwei Einträge *Apfel* und *Äpfel*. In Lexemlexika muss für den Eintrag *Apfel* der umgelautete Pluralstamm extra berechnet werden. In einem Morphemlexikon sind sowohl Allomorphe als auch Derivationsaffixe aufgelistet. Wenn Allomorphe im Lexikon enthalten sind, kann eine Morphologiekomponente rein **konkatenativ** arbeiten. Die Wortformen setzen sich vollständig und disjunkt aus im Lexikon gespeicherten Einheiten zusammen. Dieses Modell wird häufig als **Item and Arrangement** (IA) bezeichnet (vgl. Abschnitt 4.2). Sind keine Allomorphe im Lexikon enthalten, müssen nicht konkatenativ ablaufende morphologische Prozesse wie Umlautung und Tilgung während der morphologischen Analyse berücksichtigt werden. Dieses Modell wird als **Item and Process** (IP) bezeichnet (vgl. Abschnitt 4.3).

Beispiele für Morphologiesysteme, bei denen Allomorphe explizit im Lexikon abgelegt sind, sind *Morph* (vgl. Hanrieder (1996)) und *MPRO* (vgl. Maas (1996)). In den Systemen *Morphix* (vgl. Finkler und Lutzy (1996)) und *Morphy* (vgl. Lezius (1996)) werden allomorphe Stämme starker Verben im Lexikon gespeichert, während Umlautung als regulärer Prozess behandelt wird, also Umlaute bei der Analyse probehalber durch die ihnen zugrundeliegenden Vokale ersetzt werden.[2] Das System *LA-Morph* (vgl. Schüller und Lorenz (1996)) schließlich enthält Allomorph-Regeln, mit deren Hilfe vor Beginn der morphologischen Analyse aus dem Lexem-Lexikon alle allomorphen Formen berechnet werden.

[1] In Abschnitt 5.2 wird der deutschsprachige Teil von CELEX ausführlich vorgestellt.
[2] Im folgenden Abschnitt wird die Verarbeitung am Beispiel von *Morphy* kurz vorgestellt.

3.1.2 Methoden der regelbasierten Verarbeitung

Stemming

In einem regelbasierten Morphologiesystem gibt es i.A. zwei Methoden der Verarbeitung der Eingabe: Abarbeitung von links nach rechts oder Abarbeitung von rechts nach links. Die Methode von rechts nach links, also mit dem Ende der Wortform beginnend, ähnelt dem **Stemming**. Das ist eine linguistisch gesehen recht ungenau arbeitende Variante der Lemmatisierung, die häufig bei *Information Retrieval* bzw. *Information Extraction* eingesetzt wird. Sie erfordert lediglich ein Lexikon der Flexionsendungen einer Sprache (und ggf. der produktiven Derivationssuffixe) und ermöglicht die Rückführung von rein konkatenativ gebildeten Flexionsformen auf eine Art flexionsendungsloser Stammform. Dass diese nicht mit der morphologisch gesehen richtigen Grundform übereinstimmen muss, zeigen alle auf *-e* oder *-en* endenden Substantive, da diese Endungen als potentielle Flexionsendungen abgetrennt werden: Aus *Freude* wird *Freud*, aus *Eisen Eis* (allerdings werden die flektierten Formen *Bilds, Bildes, Bilder, Bildern* allesamt auf eine Stammform *Bild* zurückgeführt).

Ein Beispiel für ein Morphologiesystem, das die Eingabewortform von rechts nach links abarbeitet, ist *Morphy* (vgl. Lezius (1996)). Hier werden bei der Analyse einer Wortform sukzessive einzelne Zeichen abgetrennt und es wird überprüft, ob ein Stamm gefunden wurde. Zusätzlich wird in jedem Schritt versucht, morphologische Prozesse wie Umlautung und *ß/ss*-Wechsel (*Kuß/Küsse* in alter Rechtschreibung) rückgängig zu machen, um so am Ende eine Grundform und mögliche Flexionsendungen zu finden. Ist dies der Fall, wird geprüft, ob die Eingabe-Wortform aus der gefundenen Grundform generiert werden kann. Im Erfolgsfall ist eine mögliche Analyse mit Grundform, Flexionsstamm und morphologischer Information (diese wird mit den Flexiven zusammen gespeichert) ermittelt worden. Durch den Generierungsschritt wird das Problem des Stemmings, die Erkennung falscher Grundformen, umgangen.

Schritt	Eingabe	Test	Resultat
1	*Bäume*	*Bäume, Baume*	–
2	*Bäum e*	*Bäum, Baum*	*Baum*$^{\mathbb{P}}$ + Pluralendung
3	*Bäu me*	*Bäu, Bau*	– (*me* ist keine Flexionsendung)
4	*Bä ume*	*Bä, Ba*	–
5	*B äume*	*B*	–

Abbildung 3.1: Morphologische Analyse von *Bäume* in Morphy

In Abbildung 3.1 ist die Abarbeitung der Wortform *Bäume* dargestellt. In jedem Schritt wird ein weiteres Zeichen rechts abgetrennt und der Rest links davon geprüft. In diesem Beispiel wird immer sowohl nach der umgelauteten

als auch nach der nicht umgelauteten Form im Lexikon gesucht. *Baum*$^{\mathbb{P}}$ und *Bau*$^{\mathbb{P}}$ werden als einzige Formen im Lexikon gefunden, aber nur im Falle von *Baum*$^{\mathbb{P}}$ passen auch die Umlautung und die Flexionsendung. Die Abarbeitung endet, wenn das linke Ende der Eingabe erreicht ist.

Endliche Automaten

Bei der Abarbeitung der Eingabe von links nach rechts haben sich die sog. **Endlichen Automaten** durchgesetzt: "Endliche Automaten [...] sind der einfachste und zugleich verbreitetste Formalismus bei der Modellierung von morphologischen Regelsystemen." (Trommer (2001), S. 183) Da sie die Eingabe Zeichen für Zeichen abarbeiten, sind sie von linearer Komplexität: Der komputationelle Aufwand der Verarbeitung einer Wortform ist proportional zu deren Länge. "Because most morphological phenomena can be described with regular expressions the use of finite-state techniques for morphological components is common." (Trost (2003), S. 39) Allerdings sind endliche Automaten nicht in der Lage, einer Zerlegung eine hierarchische Struktur zuzuweisen. Sie verfügen weder über ein 'Gedächtnis', d.h., sie merken sich nicht, was sie bereits abgearbeitet haben, noch über eine 'Vorschaufunktion', d.h. die Möglichkeit, an einer Stelle abzuwarten und zu schauen, was noch an Eingabesymbolen kommt, um davon abhängig Entscheidungen zu treffen. Man kann sich einen endlichen Automaten vorstellen als eine Folge von Zuständen und Zustandsübergängen. Die Zustandsübergänge werden durch das nächste Zeichen in der Eingabewortform gesteuert. Ist der nach Abarbeiten der kompletten Eingabewortform erreichte Zustand ein Endzustand, so gilt die Eingabe als abgearbeitet, die Wortform als **erkannt** bzw. **akzeptiert**.

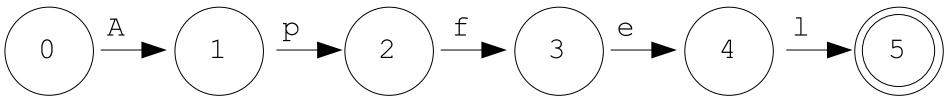

Abbildung 3.2: Ein simpler endlicher Automat

Der in Abbildung 3.2 dargestellte Automat besteht aus fünf Zuständen, die durch Kreise gekennzeichnet und durchnumeriert sind. Bei den Übergängen dazwischen wird jeweils ein Zeichen von der Eingabe gelesen. Stimmt die Eingabe mit dem jeweiligen Zeichen (im Normalfall: mit einem von mehreren angebotenen Zeichen) an der Zustandsübergangskante überein, wird in den nachfolgenden Zustand gewechselt. Ist die Eingabe abgearbeitet und ist ein Endzustand (doppelter Kreis) erreicht, so wird die Eingabe, in diesem Fall die Zeichenkette *Apfel*, akzeptiert.

Ein reiner 'Erkenner' ist für die morphologische Analyse noch nicht tauglich, da damit kein Analysestring erzeugt werden kann. Dies leistet ein **Finite-State-**

Transducer oder kurz **Transducer**, bei dem die Eingabe modifiziert wird oder Zeichenketten zusätzlich ausgegeben werden können.

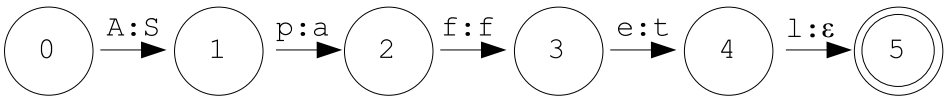

Abbildung 3.3: Ein simpler Transducer

Der in Abbildung 3.3 dargestellte Transducer liest wiederum die Eingabezeichenkette *Apfel*, vertauscht jedoch bei jedem Zustandsübergang das Eingabezeichen gegen ein anderes Zeichen.[3] Das Resultat nach Abarbeiten der Eingabe ist die Zeichenkette *Saft*.

3.1.3 Problem regelbasierter Systeme: Übergenerierung

Der am einfachsten zu realisierende Automat[4] erlaubt die beliebige Verkettung aller im Lexikon vorkommenden Morpheme. Jeder Zustand, der nach dem Lesen des letzten Zeichens eines Morphems im Lexikon auftritt, wird als Endzustand definiert, von dem aus ein Übergang in den Startzustand stattfindet. Die Eingabe wird nur dann abgelehnt, wenn sie sich nicht aus aneinandergereihten Morphemen zusammensetzt. Umgekehrt wird jede beliebige Kombination oder auch Wiederholung von Morphemen akzeptiert, also auch den Regeln der Wortbildung zuwiderlaufende Phantasieformen wie **wend·ab·keit, *keit·bar, *keit·keit·keit* usw. Dieses Phänomen wird als **Übergenerierung** bezeichnet. Die *Sprache* (die Menge aller Zeichenketten), die der Automat erkennt, umfasst zwar alle Wortformen, die als korrekt erkannt werden sollen, darüber hinaus aber auch noch einen wesentlich größeren Teil von Formen, die keine gültigen Wortformen des Deutschen sind.

Eine Übersicht über einige der möglichen Morpheme und Morphemgruppen, die in der Wortform *Auseinandersetzungen* enthalten sind (vgl. Abbildung 3.4) zeigt die Komplexität, die bereits mit der Anzahl der möglichen Zerlegungen einer Wortform verbunden ist. Während das menschliche Gehirn die meisten der als zufällige Zeichenketten in der Wortform enthaltenen Bestandteile wie den Namen *Ina*, das Substantivsuffiv *-and* oder das Substantiv *Zunge* einfach auszublenden vermag, hat ein Computer zunächst einmal keine Anhaltspunkte, aus welchen Bestandteilen sich die komplexe Wortform sinnvoll zusammensetzt. Dies funktioniert lediglich über die explizite Angabe von Re-

[3] Das Zeichen ϵ (im letzten Übergang) wird in der Ausgabe nicht dargestellt, es steht für ein leeres Zeichen.

[4] Hier und im Folgenden ist immer ein *Transducer* gemeint, wenn von einem *Automaten* gesprochen wird. Ein Transducer ist lediglich eine spezielle Art eines Automaten.

geln: Eine Wortform muss vollständig und ohne Überlappungen zusammengesetzt werden, einem Nomensuffix muss ein Nomen vorweggehen usw.

Morphem(gruppe)	Kategorie(n) in IMSLEX
au	Interjektion
aus	Adposition, Verbpartikel
auseinander	Verbpartikel
auseinandersetz	Partikelverb
auseinandersetzung	Substantiv
sein	Verb
ei	Substantiv, Nomensuffix
ein	Artikel, Verb, Verbpartikel
einander	Pronomen
in	Praefix, Adposition, Nomensuffix
ina	Name
na	Interjektion
an	Adposition, Verbpartikel
and	Nomensuffix
ander	Adjektiv
anders	Adverb, Name
de	Praefix
der	Artikel
er	Adjektivsuffix, Nomensuffix, Praefix, Pronomen
ersetz	Verb
ersetzung	Substantiv
set	Substantiv
setz	Verb
setzung	Substantiv
zu	Adposition, Partikel, Verbpartikel
zunge	Substantiv
un	Praefix
ung	Nomensuffix
ge	Praefix
gen	Adposition, Substantiv
en	Adjektivsuffix

Abbildung 3.4: Morpheme und Morphemgruppen in *Auseinandersetzung*

Übergenerierung kann vermindert werden, indem Teilautomaten für bestimmte Aufgaben vorgesehen und hintereinandergeschaltet werden. Durch separate Automaten für Präfixe, Stämme und Suffixe kann verhindert werden, dass Präfixe an Suffix- oder Stammposition auftreten. Auch hier kommt es aber zu massiver Übergenerierung, da immer noch jedes im Lexikon verzeichnete Affix an jeden Stamm treten kann, auch wenn dies unsinnige Kombinationen wie **be·ruder·keit, *ver·baum·lich* etc. ergeben kann. Wenn ein Automat eine Wortform wie *unaufhörlich* erkennen soll, wird er auch **unaufhaltlich* erkennen, da beide demselben Wortbildungsmuster folgen. Die stetige Verfeinerung des Automaten zur immer genaueren Erkennung endet irgendwann bei den

Vollformen: Jede Wortform der Sprache hat einen eigenen Teilautomaten.

In der Praxis wird ein gewisser Grad an Übergenerierung akzeptiert, da sich bestimmte produktive Muster wie die Komposition von Substantiv und Substantiv nicht einschränken lassen: Die meisten Morphologiesysteme werden für Wortformen wie *Konsumenten* und *Nachteile* neben der richtigen Analyse auch die in den seltensten Fällen intendierten Zerlegungen *Konsum=Enten* und *Nacht=Eile* erzeugen.

3.1.4 Zwei-Ebenen-Morphologie

"Die bekannteste Anwendung von Finite State-Techniken in der Morphologie ist die Two-Level-Morphologie (Zwei-Ebenen-Morphologie, TWOL)." (Heid (2000), S. 684) Der Zwei-Ebenen-Formalismus erlaubt die elegante Modellierung morphophonologischer Prozesse durch die Anwendung sogenannter *Zwei-Ebenen-Regeln* parallel zur Konkatenation morphologischer Einheiten (vgl. Koskenniemi (1983)). Die beiden Ebenen, die unterschieden werden, sind die **Oberflächenebene** und die **lexikalische Ebene**. Sie setzen eine Wortform und ihre (linguistische) Repräsentation im Lexikon miteinander in Beziehung. Bei einem Transducer geschieht dies statisch und kontextunabhängig, indem für jeden Zustandsübergang die Zeichen auf beiden Ebenen angegeben werden. Das Besondere am Zwei-Ebenen-Formalismus ist hingegen, dass Regeln angegeben werden können, die dynamisch und in Abhängigkeit des Kontextes Zeichen verändern, hinzufügen oder entfernen. Auf diese Weise können im Umfeld der eigentlich rein konkatenativ arbeitenden endlichen Automaten Prozesse modelliert werden.

Lexika und Regeln werden zusammen in einen Finite-State-Transducer kompiliert. Die Verarbeitungsgeschwindigkeit von Wortformen im kompilierten Automaten ist sehr hoch: "The TWOL program can achieve a very satisfactory speed, and the Xerox Lexical Tools allow for speeds of about 250 GB per hour with a highly compressed dictionary." (Koskeniemmi und Haapalainen (1996), S. 134)

Lexikon

Neben den Zwei-Ebenen-Regeln besteht ein Zwei-Ebenen-System aus einem Lexikon bzw. einem System von Sublexika. Hier sind für Stämme sog. **Fortsetzungsklassen** angegeben. "Fortsetzungsklassen sind Verweise auf Sublexika: verweist ein lexikalisches Zeichen auf eine Fortsetzungsklasse, so wird damit ausgedrückt, daß jedes lexikalische Zeichen dieser Fortsetzungsklasse direkt rechts an dasjenige Zeichen angehängt werden darf, von dem der Verweis ausgeht." (Heid (2000), S. 685) Dieses System ist sehr flexibel und erspart Redundanz.

LEXICON	NN_Stems	*Tag*	NMasc_es_e;
		Spiel	NNeut_es_e;
LEXICON	NMasc_es_e	*+NN.Masc:0*	N_es_e;
LEXICON	NNeut_es_e	*+NN.Neut:0*	N_es_e;
LEXICON	N_es_e		NSg_es;
		0:+e	NPl_0;
LEXICON	NSg_es	*.Nom.Sg:+*	N#;
		.Gen.Sg:+es^Gen	N#;
		.Dat.Sg:+	N#;
		.Dat.Sg:+e	N#;
		.Akk.Sg:+	N#;
LEXICON	NPl_0	*.Nom.Pl:0*	N#;
		.Gen.Pl:0	N#;
		.Dat.Pl:n	N#;
		.Akk.Pl:0	N#;

Abbildung 3.5: Lexikoneinträge in der Zwei-Ebenen-Morphologie

In Abbildung 3.5 sind einige Beispieleinträge (aus DMOR, vgl. Schiller (1996), siehe auch Abschnitt 3.2.1) aufgelistet, um die Verarbeitungsschritte nachvollziehbar zu machen. Für die beiden Lexeme *Tag*$^{\mathbb{P}}$ und *Spiel*$^{\mathbb{P}}$ sind die Fortsetzungsklassen `NMasc_es_e` und `NNeut_es_e` angegeben. Diese definieren Sublexika, die beide in eine Fortsetzungsklasse `N_es_e` weiterverzweigen. Die Information zu Wortart und Genus wird auf der Oberflächenebene verzeichnet (`+NN.Masc` bzw. `+NN.Neut`). Das Sublexikon `N_es_e` wiederum verzweigt in die Singularflexion (`NSg_es`) und die Pluralflexion (`NPl_0`). Bei der Pluralflexion geschieht dies allerdings nur dann, wenn die Eingabewortform hinter den Stämmen *Tag* oder *Spiel* ein *e* aufweist. Die Sublexika `NSg_es` und `NPl_0` bilden nun den Abschluss: In Abhängigkeit der nächsten Zeichen in der Eingabe werden weitere Informationen auf der Oberflächenebene verzeichnet und es wird in den Endzustand `N#` weiterverzweigt.[5]

In Abbildung 3.6 sind einige Analysen dargestellt, die sich aus den in Abbildung 3.5 angegebenen Stämmen und Lexika ergeben. Die Analysestrings werden beim Durchlaufen der Fortsetzungsklassen konkatenativ zusammengesetzt.

[5]Bei der Zeichenkette *^Gen*, die an die Genitivendung *-es* angehängt ist, handelt es sich um einen Trigger für eine Zwei-Ebenen-Regel, vgl. nachfolgenden Abschnitt.

Eingabe	Analysestring	Eingabe	Analysestring
Spiel	`Spiel+NN.Neut.Akk.Sg`	*Tag*	`Tag+NN.Masc.Akk.Sg`
	`Spiel+NN.Neut.Nom.Sg`		`Tag+NN.Masc.Nom.Sg`
	`Spiel+NN.Neut.Dat.Sg`		`Tag+NN.Masc.Dat.Sg`
Spieles	`Spiel+NN.Neut.Gen.Sg`	*Tages*	`Tag+NN.Masc.Gen.Sg`
Spiels	`Spiel+NN.Neut.Gen.Sg`	*Tags*	`Tag+NN.Masc.Gen.Sg`

Abbildung 3.6: Morphologische Analyse von *Spiel(es)* und *Tag(es)*

Zwei-Ebenen-Regeln

Zwei-Ebenen-Regeln dienen der Modellierung morpho-phonologischer Prozesse. Sie sind so konzipiert, dass sie das Auftreten oder Nicht-Auftreten eines Zeichens auf der Oberflächenebene in Abhängigkeit des Kontextes beeinflussen können. Im Folgenden sind zwei Beispiele angegeben, die dies anhand einer Elisions- und einer Epentheseregel illustrieren.

(3.1) `e:0 <=> _ +:0 e:e`

Die Elisionsregel in 3.1 besagt, dass das Zeichen *e* auf der lexikalischen Ebene durch ein Nullzeichen auf der Oberflächenebene realisiert wird[6], wenn es (das Zeichen wird in der Darstellung durch _ repräsentiert) vor einer Morphemgrenze steht (repräsentiert durch +), auf die wiederum ein *e* folgt. Aus

(3.2) $leise + er$

auf der lexikalischen Ebene wird

(3.3) $leiser$

in der Ausgabe. Ohne Zwei-Ebenen-Regeln ließe sich die Steigerung von auf *e* endenden Adjektiven nicht so elegant modellieren: Es müsste ein allomorpher Flexionsstamm *leis* vorgesehen werden, an den die Komparativendung *-er* angehängt werden könnte. Dies ist genau das, was die Zwei-Ebenen-Regel als Prozess realisiert, ohne dafür allerdings einen separaten Stammeintrag im Lexikon zu erfordern. Die angegebene Regel erspart also eine explizite Erzeugung von allomorphen Stämmen bei allen auf *e* endenden Adjektiven.

In Abbildung 3.7 ist die Zwei-Ebenen-Regel aus 3.1 als Transducer dargestellt. Mit Zwei-Ebenen-Regeln wird der Formalismus also nicht verlassen, sie lassen sich mit denselben Mitteln darstellen wie die Konkatenation von Zeichen oder Morphemen.

[6]Der Doppelpunkt trennt lexikalische Ebene und Oberflächenebene. Das Nullzeichen steht stellvertretend für das Zeichen ϵ. Da ϵ nicht im Standardzeichensatz vorhanden ist, wird in der Praxis gewöhnlich '0' verwendet.

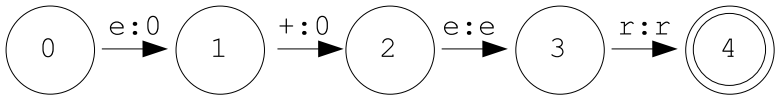

Abbildung 3.7: Ein Transducer für eine Zwei-Ebenen-Regel

(3.4) `%&:e <=> [ [c h] | [c k] | b | d | f | g | m | p ] n _`

Ein weiteres Beispiel für eine Zwei-Ebenen-Regel ist die e-Epenthese-Regel (vgl. Beispiel 3.4). Diese sorgt in bestimmten Kontexten für die Einfügung eines *e* zwischen Stamm und Flexiv, das dort aus (morpho)phonologischen Gründen benötigt wird. Hierzu wird ein Trigger `&` (das Prozentzeichen markiert ein Sonderzeichen) beim Flexiv verzeichnet, der im Normalfall auf dem Ausgabeband getilgt wird. In bestimmten Kontexten wird er jedoch in der Ausgabe durch ein *e* ersetzt.

Pronomen	Verben	Flexiv	Morphologiestring
ich	*gehe, rudere, ordne, rechne*	*-e*	`+V.1.Sg.Pres.Ind`
du	*gehst, ruderst, ordnest, rechnest*	*-&st*	`+V.2.Sg.Pres.Ind`
er/sie/es	*geht, rudert, ordnet, rechnet*	*-&t*	`+V.3.Sg.Pres.Ind`

Abbildung 3.8: Verbflexion mit e-Elision

In Abbildung 3.8 ist der für die Regel aus Beispiel 3.4 relevante Teil der Verbflexion dargestellt. Ohne die *e*-Epenthese erhielte man in der regulären Verbflexion Wortformen wie **ordnst, *ordnt*, die keine gültigen phonologischen Wörter des Deutschen darstellen. Auch hier verhindert die Zwei-Ebenen-Regel die Einführung allomorpher Flexionsstämme wie *ordne* und *rechne*.

Ein weiteres Einsatzfeld der Zwei-Ebenen-Regeln ist die Umlautbildung beim Plural und bei den Steigerungsformen der Adjektive. In der alten Rechtschreibung wird darüber hinaus die *ß/ss*-Alternation (*der Fluß, des Flusses*) gesteuert. In beiden Fällen muss allerdings, wie bei der Epenthese, ein Trigger verwendet werden, der die umlautbaren Stämme markiert. Dadurch wird ausgeschlossen, dass eine Regel an anderer, unbeabsichtigter, Stelle wirkt und so zu falschen Analysen führt. In Abbildung 3.5 (s. Seite 28) im vorhergehenden Abschnitt ist die Zeichenkette *^Gen* ein Trigger für das Wegfallen des *e* beim Genitiv: *(des) Spieles/Spiels* sind beides gültige Genitivformen im Deutschen.

Die beiden auf der *Morpholympics* vorgestellten Zwei-Ebenen-Morphologiesysteme für das Deutsche sind *Gertwol* (vom Entwickler des Zwei-Ebenen-Formalismus, Kimmo Koskeniemmi; vgl. Koskeniemmi und Haapalainen (1996)) und *DMOR*, das am IMS verwendete Morphologiesystem (vgl. Schiller (1996)). *DMOR* wird in Abschnitt 3.2.1 beschrieben.

3.2 Morphologiesysteme

Nachdem in den vorangegangenen Abschnitten beschrieben wurde, wie mit Finite-State-Transducern die morphologische Analyse vonstatten geht und wie insbesondere Zwei-Ebenen-Systeme die Vorteile der schnellen und effizienten Verarbeitung in Finite-State-Technik mit der Eleganz der linguistischen Beschreibungsmöglichkeiten morphologischer Prozesse vereinen, wird in diesem Abschnitt ein Zwei-Ebenen-Morphologiesystem beschrieben. Dafür wurde das System DMOR ausgewählt, das die Datengrundlage für das in dieser Arbeit beschriebene Lexikon IMSLEX bildet.

Unterschiede in der Behandlung von Fugenelementen oder Wortbildungsmustern ändern nichts daran, dass im Endeffekt jedes Morphologiesystem die Zerlegung einer Wortform in ihre morphosyntaktischen Merkmale vollzieht. Die Unterschiede ergeben sich im Detail, durch die Liberalität von Wortbildungsregeln und die Auswahl der Einheiten, die im internen Lexikon des Systems abgelegt werden.

Im Anschluss an die Beschreibung von DMOR werden einige Aspekte zur Bewertung der Performanz von Morphologiesystemen vorgestellt, die später in dieser Arbeit helfen werden, die Qualität des Lexikons zu bewerten.

3.2.1 DMOR – ein Zwei-Ebenen-System

Bei DMOR handelt es sich um die Implementierung einer Zwei-Ebenen-Morphologie für das Deutsche, die als Finite-State-Transducer realisiert ist (vgl. Abschnitt 3.1.2). Das System wurde in Schiller (1996) vorgestellt. Es besteht aus einem Lexikonteil und einer Sammlung von Zwei-Ebenen-Regeln, die gemeinsam in einen endlichen Automaten kompiliert werden. DMOR behandelt Flexion und Komposition. Derivation ist in DMOR nur für zwei eingeschränkte Bereiche realisiert: Movierung (*Sammler* → *Sammler<u>in</u>*; *Schwabe* → *Schwäb<u>in</u>*) und Adjektivbildung bei Städtenamen (*Stuttgart* → *Stuttgart<u>er</u>* → *Stuttgarter<u>in</u>*). Diese beiden Anwendungsbereiche sind bereits in den Flexionsklassen der Substantive markiert (Movierung durch $=in$ im Flexionsklassenbezeichner, vgl. Tabelle 3.19). Alle anderen Derivationen müssen in den Lexika aufgelistet werden.[7]

Die DMOR-Lexika sind nach Wortarten gegliedert auf Dateien verteilt. Fachsprachliche Substantive und geographische Namen werden gesondert behandelt. Die nicht-flektierenden Klassen *Adverbien, Adpositionen, Konjunktionen, Interjektionen, Partikeln* werden in einer Datei zusammengefasst.

[7] Die Gruppe der Derivationen auf *-ung, -heit, -keit, -ion, -(i)tät, -schaft* ist aufgrund ihres gleichartigen Flexionsverhaltens in einer Flexionsklasse `NFem-Deriv` versammelt, wird aber nicht weiter als Derivation gekennzeichnet. Dies gilt analog für Derivationen auf *-chen* und *-lein* mit der Flexionsklasse *NNeut-Dimin*.

Jede Datei ist weiter unterteilt in Sublexika. Für Substantive sind dies `NN_Stems`, `NN_Stems/NoHead`, `NN_Stems/NoCp` und `NN_Truncs`. Die Untergliederung dient der Steuerung des Kompositionsverhaltens der Stämme. Unter `NN_Stems/NoHead` sind jeweils einige Stämme aufgelistet, die nicht als Kompositumsköpfe auftreten dürfen (z.B. *Ei* zur Vermeidung von Falschanalysen wie *Drucker=Ei, Bücher=Ei, Ziegel=Ei, ...*). Unter `NN_Stems/NoCp` (*no compounding*) werden Stämme zusammengefasst, die generell nicht an Komposition teilnehmen dürfen (einzelne Buchstaben z.B. oder Adjektive wie *gang*, das nur in einer festen Fügung *gang und gäbe* vorkommt, und *lose*). Das `NN_Truncs`-Sublexikon ist eine Besonderheit: Hier werden Kompositionserstglieder aufgelistet, die nicht über Fortsetzungsklassen erzeugt werden können (vgl. unten S. 37). Bei Verben dienen die unterschiedlichen Sublexika der Unterscheidung der Partizipbildung: Es wird noch nach Partizipbildung mit und ohne *ge-* differenziert.

Die Sublexika werden in Teilautomaten kompiliert, die jeweils eine Funktion übernehmen: Es gibt Automaten, die für jeweils kleingeschriebene oder großgeschriebene nicht-komponierbare Formen zuständig sind, solche, die für kleingeschriebene Kompositionsköpfe (oder, bei Bindestrich-Komposita, für großgeschriebene) zuständig sind, und solche, die für klein- und großgeschriebene Kompositionserstglieder zuständig sind. Ein Grund dafür ist die Vermeidung von zu starker Übergenerierung, die entstünde, wenn alle Wortarten gleichberechtigt an Komposition teilhaben dürften (vgl. z.B. die Anzahl der möglichen Morpheme in einer Wortform in Abbildung 3.4, S. 26).

(3.5) *Sofort**h**ilfe* → *s**o**fort=**H**ilfe*

In Beispiel 3.5 wird veranschaulicht, dass in einer morphologisch komplexen Wortform Groß- und Kleinschreibung bestimmter Wortarten gerade vertauscht vorkommen können.

Zu jedem Stamm im Lexikon ist eine Fortsetzungsklasse angegeben, die wiederum ein Sublexikon definiert, das in weitere Fortsetzungsklassen verzweigt.[8] Die Fortsetzungsklassen auf der obersten Ebene, also bei den Stammeinträgen, entsprechen den Flexionsklassen von Lexemen. Implizit repräsentiert ein Paar aus Stamm und Flexionsklasse in einem Sublexikon also ein Lexem. Bei den nicht-flektierenden Klassen werden ebenfalls 'Flexionsklassen' angegeben, die dann allerdings nicht in Fortsetzungsklassen verzweigen, sondern nur die Wortart festlegen und ggf. syntaktische Informationen kodieren: bei Adverbien die Tatsache, ob es sich um Pronominaladverbien oder Frageadverbien handelt, bei Konjunktionen der Typ (koordinierend, subordinierend, vergleichend), bei Adpositionen der Kasus, der regiert wird.

Bei den geschlossenen Wortarten wie den Pronomen oder bei unregelmäßigen Paradigmen werden in DMOR die Wortformen einzeln aufgelistet und mit

[8] Eine Illustration hierzu ist in Abbildung 3.5 auf Seite 28 gegeben.

Stammeintrag	Forts.kl.	Beispiel
`alle+INDEF.attr.Fem.Nom.Sg.St:alle`	`Closed#`	*alle Mühe*
`alle+INDEF.attr.Fem.Gen.Sg.St:aller`	`Closed#`	*trotz aller Mühe*
`alle+INDEF.attr.Fem.Dat.Sg.St:aller`	`Closed#`	*mit aller Mühe*
`alle+INDEF.attr.Fem.Akk.Sg.St:alle`	`Closed#`	*für alle Mühe*

Abbildung 3.9: Vollformeneintrag in DMOR, Beispiele für *alle*$^{\mathbb{P}}$

dem jeweiligen Analysestring versehen (vgl. die Formen für Femininum Singular von *alle*$^{\mathbb{P}}$ in Abbildung 3.9). Hier entspricht im Stammeintrag der Teil vor dem Doppelpunkt genau dem Analysestring, den die Morphologiekomponente ausgibt. Hinter dem Doppelpunkt steht die Wortform. Die Fortsetzungsklasse ist grundsätzlich `Closed#`. Dies markiert einen Eintrag für eine geschlossene Klasse, die nicht weiterverzweigt.

Die Substantivlexika enthalten etwa 20 000 Appellativa[9], 2 000 Personennamen und 1 000 geographische Namen. Das Adjektivlexikon enthält etwa 7 000 Adjektive, ein Großteil davon Derivationen auf *-isch, -ig, -lich, -bar, -haft, -abel*, *un*-Präfigierungen und Komposita. Das Verblexikon enthält knapp 5 600 Verben. Zwei Drittel davon sind Präfixverben.

Flexion

Regelmäßig ablaufende Flexion wird in DMOR durch das Konzept der Fortsetzungsklassen behandelt. Unregelmäßig ablaufende Flexion kann entweder im Lexikon oder durch Zwei-Ebenen-Regeln behandelt werden: “In der hier beschriebenen Anwendung für Deutsch wird regelmäßige Allomorphie (wie z.B. Umlaut, ß/ss-Wechsel) durch 2-Ebenen-Regeln behandelt, unregelmäßige Stammvarianten (z.B. abgelautete Verbstämme) sind lexikalisiert.” (Schiller (1996), S. 40)

In Abbildung 3.10 sind die Einträge für vier verschiedene Verben im DMOR-Verblexikon dargestellt. *zähl* und *ruder* unterscheiden sich lediglich in ihrer Flexionsklasse. Auf *-el* und *-er* endende Verben werden in DMOR in einer eigenen Klasse behandelt, damit zum einen die Infinitivform und das Partizip Präsens richtig gebildet werden können, zum anderen ein Trigger für die optionale *e*-Elision (*ich rudere/rudre*) eingefügt werden kann. Beide Flexionsklassen verzweigen weiter und bilden jeweils das gesamte Verbparadigma ab. Für das starke Verb *treten*$^{\mathbb{P}}$ werden die veränderten Stämme für Teile der Flexion einzeln aufgelistet (*tret:trat*). Die Notation der Stämme entspricht der Oberflä-

[9] Mit Stand von 1999, als die Arbeiten an den DMOR-Lexika zugunsten einer relationalen Datenbank eingestellt wurden, vgl. Lezius et al. (2000).

Stamm	Flexionsklasse	Funktion
zähl	`VVReg`	reguläre Verbflexion
ruder	`VVReg-el/er`	reguläre Verbflexion
tret	`VVPres1`	Präsens + Imperfekt Konjunktiv + Imperativ Sg.
tret	`VVPP-en`	Partizip Perfekt
tret:trat	`VVPastIndStr`	Imperfekt Indikativ
tret:tritt	`VVPres2+Imp0`	2. Person Präsens Indikativ + Imperativ Pl.
tret:trät	`VVPastKonjStr`	Imperfekt Konjunktiv
frag	`VVReg`	reguläre Verbflexion
frag:fräg	`VVPres2`	2./3. Person Präsens Indikativ

Abbildung 3.10: Flexionsklassen und Allomorphie bei Verben in DMOR

chenebene (vor dem Doppelpunkt) und der lexikalischen Ebene (hinter dem Doppelpunkt). Es werden möglichst viele Generalisierungen wahrgenommen: `VVPres1` verzweigt weiter in die Präsensflexion, die Bildung der Formen für Imperfekt Konjunktiv und den Imperativ Singular.

Im letzten Beispiel in Abbildung 3.10 ist ein Sonderfall dargestellt: Für das Verb *fragen*$^{\mathbb{P}}$ ist zunächst das reguläre Verb-Flexionsparadigma angegeben. Da allerdings in dialektalen Varianten des Deutschen die zweite und dritte Person Singular Präsens Indikativ mit Umlaut gebildet werden kann, wird diese Form hier einfach im Lexikon mitnotiert. Hier zeigt sich die Flexibilität des Finite-State-Ansatzes: Anstatt eine eigene Klasse bilden zu müssen, können Sonderfälle einfach aufgelistet werden. Der Zusammenhang zwischen beiden (für die Angabe der Grundform *fragen* auf der Ausgabeebene) wird in DMOR durch den Stamm *frag* hergestellt.

Stamm	Flexionsklasse	Grundform	Pluralform
Solo	`NNeut_s_s`	*Solo*	*Solos*
Solo:Soli	`NNeut/Pl`	*Solo*	*Soli*
Serum	`NNeut-um/a`	*Serum*	*Sera*
Serum:Seren	`NNeut/Pl`	*Serum*	*Seren*
Komma	`NNeut_s_s`	*Komma*	*Kommas*
Komma:Kommata	`NNeut-a/ata`	*Komma*	*Kommata*

Abbildung 3.11: Allomorphie bei Pluralformen in DMOR

Sonderfälle der geschilderten Art gibt es auch bei anderen Wortarten. In Abbildung 3.11 sind Substantive dargestellt, die im Sprachgebrauch über eine alternative Pluralform verfügen. Eine zusätzliche Besonderheit bei *Komma*$^{\mathbb{P}}$ ist, dass beide dargestellten Flexionsklassen sowohl die Singular- wie die Plural-

flexion behandeln. Der Transducer erzeugt also für die Wortform *Komma* alle Singular-Analysen doppelt. Identische Analysen werden allerdings miteinander verschmolzen, so dass jeder Analysestring nur jeweils einmal ausgegeben wird.

Flexionsklasse	Beispiele	# Stämme
`NFem-Deriv`	*Abhärtung, Adoption, Entität*	4893
`NFem-a/en`	*Algebra, Tunika*	61
`NFem-in`	*Anwältin, Wanderin, Ärztin*	42
`NFem-is/en`	*Basis, Dosis, Synopsis*	16
`NFem-is/iden`	*Arthritis, Bronchitis*	16
`NFem-s/$sse`	*Bedrängnis, Nuß*	23
`NFem-s/ssen`	*Anakrusis, Hosteß, Stewardeß*	3
`NFem/Pl`	*City:Cities, Galaxis:Galaxien*	19
`NFem/Sg`	*Abkehr, Wucht, Ästhetik*	659
`NFem_0_$`	*Mutter, Tochter*	2
`NFem_0_$e`	*Auskunft, Faust, Herkunft*	65
`NFem_0_e`	*Drangsal, Supernova*	15
`NFem_0_en`	*Abart, Drängelei, Konferenz, Zäsur*	809
`NFem_0_n`	*Abrede, Achse, Ökologie*	2501
`NFem_0_s`	*Anaconda, Shell, Tramway, Troika*	94
`NFem_0_x`	*Anchovis, Iris, Jeans*	6
`N?/Pl_0`	*Gebrüder, Geschwister*	8
`N?/Pl_x`	*Annalen, Eltern*	19

Abbildung 3.12: DMOR-Flexionsklassen: Nomina femininum und Pluraliatantum

Weitere Beispiele für derartige 'unregelmäßige' Pluralbildungen finden sich in Abbildung 3.12 bei der Flexionsklasse `NFem/Pl`, für Adjektivbildungen in Abbildung 3.13 bei der Flexionsklasse `AdjComp`. Umlautung im Plural bei Substantiven und in Steigerungsformen bei Adjektiven dagegen wird im Flexionsklassenbezeichner durch ein Dollar-Zeichen markiert, das in einer Zwei-Ebenen-Regel als Trigger fungiert.

Spezielle Klassenbezeichner wie `NFem-Deriv, NFem-in` (vgl. Abbildung 3.12) und `NNeut-Dimin` (vgl. Abbildung 3.17) weisen darauf hin, dass es sich bei dem Substantiv um ein Derivativum handelt: `NFem-Deriv` umfasst Derivationen auf *-ung, -heit, -keit, -ion, -(i)tät* und *-schaft*. `NFem-in` beinhaltet Formen auf *-in*, die sich nicht durch eine =*in*-Klasse darstellen lassen, weil der Derivationsstamm umgelautet oder getilgt ist: *Köchin, Schurkin, Ruderin*. `NNeut-Dimin` enthält Verkleinerungsformen auf *-chen* und *-lein*.

In der Auflistung der Adjektiv-Flexionsklassen (vgl. Abbildung 3.13) zeigt sich an der Klasse `AdjPos` eine Besonderheit von DMOR: Für Adjektive, die aus semantischen oder morphologischen Gründen nicht steigerbar sind (**abendelänger*), ist nur die Flexion im Positiv möglich. Dadurch kann Übergenerierung eingeschränkt werden.[10]

[10] Aufgrund der Übereinstimmung der Flexionsendung *-er* und der Komparativendung *-er* gibt

Flexionsklasse	Beispiele	# Stämme
`Adj$`	*arg, stark, unklug*	18
`Adj$e`	*alt, gesund, ungesund*	9
`Adj+`	*bar, lose, entlegen, farbig, übersät*	5430
`Adj+(e)`	*abhold, antik, untreu, weh*	36
`Adj+Lang`	*afrikanisch, nepalesisch, westfälisch*	152
`Adj+e`	*gemäß, los, schwerverletzt, ungeahnt*	1231
`Adj-eler`	*adorabel, sauber, zappenduster, übel*	127
`Adj0`	*anderthalb, extra, wievielte*	11
`AdjComp`	*gut:bess, hoch:höh, nah:näh*	3
`AdjNN`	*recht, schuld*	2
`AdjPos`	*abendelang, allermeiste, ungekündigt, übrig*	48
`AdjPosAttr`	*alleinig, besonder, vorig, vorletzt*	28
`AdjPosPred`	*gang, barfuß, hoch, schade, zigfach*	31
`AdjPosSup`	*inner, mittler, ober, unter, vorder, äußer*	6
`AdjSup`	*gut:be, hoch:höch, nah:näch*	3
`Adj~+e`	*baß, gewiß, platschnaß*	11

Abbildung 3.13: DMOR-Flexionsklassen: Adjektive

Transposition

Durch die Flexibilität des Finite-State-Modells können regelmäßig ablaufende Prozesse auf verschiedene Weisen behandelt werden. Im Deutschen gibt es mit der Transposition, dem völlig regelmäßig ablaufenden Wortartwechsel ohne Formveränderung, einen solchen Prozess. In DMOR wird dieser Prozess durch Fortsetzungsklassen im Lexikon modelliert.

(3.6)
```
LEXICON   VInf   +V.Inf:0   V#;
                  ^VINF:0    NNeut/Sg_s;
```

In Beispiel 3.6 ist dargestellt, wie das Sublexikon `VInf` zur Behandlung der Transposition verwendet wird. Im ersten Fall gibt es die Information aus, dass es sich bei der Eingabe um ein Verb im Infinitiv handelt, und geht in einen Endzustand `V#` über. Im zweiten Fall wird die Information *^VINF* an die Ausgabe angehängt und es wird in das Sublexikon `NNeut/Sg_s` verzweigt. Das Resultat dieser Vorgehensweise ist, dass substantivierte Infinitive nicht in das Lexikon aufgenommen werden müssen, sondern für jedes Verb im Lexikon automatisch auch die Substantivierung der Infinitivform analysiert werden kann.[11]

Für die Eingabe *Spielen* ergeben sich die in Abbildung 3.14 dargestellten Analysestrings. Mit *^VINF* wird die Transposition markiert: Es handelt sich in

es bei adjektivischen Wortformen, die auf *-er* enden, eine Mehrdeutigkeit zwischen der Form im Positiv und der Form im Komparativ: *(ein) nagelneuer (Wagen)*. Die Mehrdeutigkeit fällt weg, wenn der Komparativ für ein Adjektiv nicht zugelassen ist.

[11]Dies spiegelt die Wirklichkeit wider. Jeder Sprecher des Deutschen kann Lehnverben oder Neubildungen wie *surfen* und *googeln* sofort substantivisch verwenden: *(das) Surfen, (das) Googeln*.

Wortform	Analysestring	Analysestring (Forts.)
Spielen	`Spiel+NN.Neut.Dat.Pl`	`*spielen+V.1.Pl.Pres.Ind`
	`spielen^VINF+NN.Neut.Akk.Sg`	`*spielen+V.1.Pl.Pres.Konj`
	`spielen^VINF+NN.Neut.Dat.Sg`	`*spielen+V.3.Pl.Pres.Ind`
	`spielen^VINF+NN.Neut.Nom.Sg`	`*spielen+V.3.Pl.Pres.Konj`
	`*spielen+V.Inf`	

Abbildung 3.14: Morphologische Analyse von *Spielen*

dieser Verwendung um ein Substantiv (`+NN`), das aber auf ein Verb im Infinitiv zurückgeht. Die Analysen auf der linken Seite sind diejenigen, die aus den in Beispiel 3.6 dargestellten Einträgen hervorgehen, während die vier auf der rechten Seite aus anderen Lexikonregeln in der Verbflexion erzeugt werden (mit Stern wegen der Großschreibung der Wortform). Die Substantivierung von flektierten Adjektiven (*das Gute, das Schöne*) wird analog gehandhabt.

Komposition

Flexionsklasse	Beispiele	# Stämme
`NCp#0`	*Abblend, Bagatell, Viel, Öko*	538
`NCp#en`	*Desiderat, Dokument, Zitat*	33
`NCp#es`	*Arm, Bund*	43
`NCp#s`	*Aushilf, Vorweihnacht*	24

Abbildung 3.15: DMOR-Flexionsklassen: Kompositionserstglieder

Komposition wird in DMOR bei substantivischem und adjektivischem Erstglied über die Fortsetzungsklassen, bei allen anderen über Auflistung bewerkstelligt. Bei Substantiven und Adjektiven ist für jede Flexionsklasse angegeben, ob und in welche der in Abbildung 3.15 dargestellten Fortsetzungsklassen sie verzweigt. Dies ist zwar mit Übergenerierung verbunden, da nicht immer alle Lexeme mit derselben Flexionsklasse auch dieselbe Fuge nehmen, aber immer noch wesentlich eingeschränkter, als wenn jeder Stamm mit jeder Fuge verwendet werden dürfte. Beispielsweise gäbe es für das Kompositum *Merkmal=Erkennung* eine falsche Lesart **Merkmaler=Kennung*, wenn *-er-* eine gültige Fuge für *Merkmal* wäre. Für häufig vorkommende Fugen wie *-s-, -e-, -er-* gibt es eine Vielzahl von Substantivpaaren, die diese Art von Mehrdeutigkeit hervorrufen: *Saal/Aal*, *Strumpf/Trumpf*, *Sturm/Turm*, *Sendung/Endung* für *-s-*, *Emission/Mission*, *Etat/Tat*, *Egel/Gel* für *-e-*, *Erfassung/Fassung*, *Erläuterung/Läuterung*, *Ersatz/Satz* für *-er-* u.v.a.m. (In CISLEX sind diese "Nomenpaare mit häufiger Endgliedambiguität" (Maier-Meyer (1995), S. 210) aufgelistet,

wobei gleichzeitig eine Präferenzierung vorgenommen wird, um keine Ambiguitäten aufkommen zu lassen.)

Für alle Wortarten außer Substantiven und Adjektiven, für Adjektive in Komparativ und Superlativ und für Sonderformen müssen die einzelnen Erstglieder aufgelistet und einer der in Abbildung 3.15 dargestellten Fortsetzungsklassen zugeordnet werden.[12] Verbstämme kommen zwar recht häufig als Erstglieder vor, sind in DMOR allerdings nicht generell als Erstglieder zugelassen, da die Vielzahl der Konversionen (*platz(en)/Platz, still(en)/still, ruf(en)/Ruf, fett(en)/fett/Fett* etc.) zu einer sehr großen Anzahl mehrdeutiger Analysen führen würde. Sie müssen daher mit Vertretern anderer Wortarten oder Sonderformen explizit aufgelistet werden (vgl. Abbildung 3.16).

Erstglied	Flexionsklasse	Beispiele	Typ
Senk	`NCp#0`	*Senk=Blei*	Verbstamm
Abbiege	`NCp#0`	*Abbiege=Spur*	Partikelverbstamm
Sofort	`NCp#0`	*Sofort=Hilfe*	Adverb
Allein	`NCp#0`	*Allein=Schuld*	
Mindest	`NCp#0`	*Mindest=Verzehr*	Adjektiv
Höchst	`NCp#0`	*Höchst=Form*	Adj. im Superlativ
Pseudo	`NCp#0`	*Pseudo=Lösung*	neoklassisch
Thermos	`NCp#0`	*Thermos=Kanne*	
Binnen	`NCp#0`	*Binnen=Hafen*	Sonderformen
Solidar	`NCp#0`	*Solidar=Gemeinsschaft*	
Vize	`NCp#0`	*Vize=Präsident*	

Abbildung 3.16: Separat aufgelistete Kompositionserstglieder in DMOR

Die Möglichkeit der Auflistung bietet den großen Vorteil, dass nicht geklärt werden muss, welchem Lexem Erstglieder wie *Binnen, Solidar* oder auch *Mindest* zugeordnet sind. Es geht ausschließlich darum, sie für die Analyse von Komposita zum Lexikon hinzuzufügen. Eine Phantasiewortform wie **Pseudohöchstthermossofortkanne* erhält in der Morphologiekomponente eine eindeutige Zerlegung *Pseudo=Höchst=Thermos=Sofort=Kanne*, da DMOR eine unbegrenzte Zahl von Erstgliedern in einer Wortform zulässt. Der Nachteil ist, dass bei Konversionen nicht klar ist, um welche Wortart es sich in der Zerlegung handelt: Die Zerlegungen *Platz=Konzert* und *Platz=Wunde* lassen nicht erkennen, ob sich das Erstglied auf das Lexem *platzen*$^{\mathbb{P}}{}_{V}$ oder das Lexem *Platz*$^{\mathbb{P}}{}_{NN}$ bezieht.

[12] Die Bildung zusammengesetzter Zahlwörter wie *einhundertunddrei* findet allerdings innerhalb der Klasse der Zahlwörter statt.

Flexionsklasse	Beispiele	# Stämme
NNeut-0/ien	*Adverb, Fossil*	11
NNeut-Dimin	*Bildchen, Körnchen, Örtchen*	397
NNeut-Herz	*Herz*	1
NNeut-a/ata	*Klima, Komma*	7
NNeut-a/en	*Aroma, Stigma*	17
NNeut-on/a	*Analogon, Enklitikon, Lexikon, Paradoxon*	4
NNeut-s/$sser	*Faß, Roß, Schloß, Vorhängeschloß*	4
NNeut-s/sse	*As, Bedürfnis, Roß*	41
NNeut-um/a	*Aktivum, Technikum*	56
NNeut-um/en	*Abstraktum, Ultimatum*	120
NNeut/Pl	*Agens:Agenzien, Sandwich:Sandwiches, Serum:Sera*	54
NNeut/Sg_0	*Ces, C, Tempus, Tennis*	95
NNeut/Sg_es	*All, Wild*	128
NNeut/Sg_s	*Ale, Badminton, Ticktack*	697
NNeut_0_x	*Avis, Rendezvous*	7
NNeut_es_$e	*Floß*	1
NNeut_es_$er	*Ei, Abbild, Vorland*	104
NNeut_es_e	*Mus, Gebäck, Mandat, Öl*	658
NNeut_es_en	*Bakelit, Hemd, Verb*	9
NNeut_s_$	*Kloster*	1
NNeut_s_0	*Abenteuer, Getriebe, Vehikel*	247
NNeut_s_e	*Ren, Portal*	218
NNeut_s_en	*Ion, Alkali, Requisit*	16
NNeut_s_n	*Auge, Interesse*	7
NNeut_s_s	*A, Abonnement, Email*	476
NNeut_s_x	*Abkommen, Volumen*	95

Abbildung 3.17: DMOR-Flexionsklassen: Nomina neutrum

Wortbestand

Da die Morphologiekomponente Komposition beherrscht, müssen Komposita im Lexikon nicht aufgelistet werden. Die Frage nach den lexikalisierten Formen wurde rein morphologisch beantwortet: *Bahnhof, Grundlage, Flugzeug* etc. befinden sich nicht im Lexikon, da sie morphologisch komplex sind und eine Zerlegung erhalten, deren Kopf sich morphologisch verhält wie die gesamte Wortform (*Bahn=Hof, Grund=Lage, Flug=Zeug*). Substantivierungen von Phrasen (*Zur·schau·stellung*) und Derivationen (*un·zu·rechn·ungs·fähig*) müssen im Lexikon aufgelistet werden. Für Komposita gilt dies nur, sofern das Erstglied weder als Stamm einer Wortart, noch als explizites Kompositionserstglied eingetragen ist: *Kommanditgesellschaft*. Partikelverben sind im DMOR-Lexikon nicht verzeichnet. Stattdessen sind Verbbasen, die nur oder fast ausschließlich in Partikelverben vorkommen, ohne eine weitere Markierung als normale Verben aufgelistet: *geistern*$^{\mathbb{P}}$, *gabeln*$^{\mathbb{P}}$, *kerkern*$^{\mathbb{P}}$, *quartieren*$^{\mathbb{P}}$ (für Bildungen wie *umher·geistern, auf·gabeln, ein·kerkern, ein·quartieren*). Die gebräuchlichsten Verbpartikeln und Verbzusätze sind in einer eigenen Lexikondatei aufge-

Flexionsklasse	Beispiele	# Stämme
`NMasc-Adj`	*Angehörige, Beamte*	8
`NMasc-ns`	*Buchstabe, Friede, Wille*	17
`NMasc-s/$sse`	*Abfluß, Kuß, Paß*	44
`NMasc-s/Sg`	*Beschuß, Haß*	6
`NMasc-s/sse`	*Abriß, Krokus, Regreß*	46
`NMasc-s0/sse`	*Albatros, Rebus*	11
`NMasc-us/en`	*Absolutismus, Logarithmus, Zyklus*	98
`NMasc-us/i`	*Ablativus, Nukleus*	18
`NMasc/Pl`	*Allroundman:Allroundmen, Anbau:Anbauten*	42
`NMasc/Sg_0`	*Nu, Moschus, Vokalismus*	308
`NMasc/Sg_es`	*Mars, Unterricht, Zustrom*	317
`NMasc/Sg_s`	*Adel, Pardon, Äther*	195

Abbildung 3.18: DMOR-Flexionsklassen: Nomina maskulinum (1/2)

listet. Über diese hinaus ist Komposition mit verbalem Kopf nicht zugelassen: *schlangestehen* und *gegensteuern* werden von DMOR nicht analysiert, da *schlange* und *gegen* nicht als Verbzusätze aufgelistet sind.[13] Dies gilt nicht, wenn die Wortform substantiviert ist: *(das) Schlange=stehen* und *(das) Gegen=steuern* werden über die Kompositionsregeln von DMOR erfasst.[14]

Das DMOR-Lexikon orientiert sich in seinem Wortbestand im Wesentlichen am HGC (vgl. Abschnitt 1.5). Daher sind Fremdwörter enthalten, sofern sie mit einer gewissen Häufigkeit im HGC vorkommen und damit als 'eingedeutscht' gelten können (Beispiele nach HGC-Vorkommenshäufigkeit absteigend sortiert): *live*$_{(3008)}$, *Clinch*$_{(488)}$, *Fayence*$_{(270)}$, *Dinner*$_{(260)}$, *Display*$_{(197)}$, *checken*$_{(93)}$, *Coupé*$_{(88)}$, *powern*$_{(39)}$, *Compiler*$_{(34)}$, *bleu*$_{(28)}$, *sprayen*$_{(21)}$, *groggy*$_{(12)}$.

Diskussion

DMOR ist ein sehr mächtiges Morphologiesystem, in dem Flexionsphänomene des Deutschen durch die Einbettung in das Zwei-Ebenen-Modell adäquat behandelt werden. Bei der Erkennung von Wortformen lag das System auf der *Morpholympics* bei verschiedenen Eingabewortlisten immer bei den besseren Systemen (vgl. Hausser (1996), S. 14, dort *PC-K* abgekürzt), was zum einen auf die Größe des Lexikons, zum anderen auf den sehr stark übergenerierenden Kompositionsmechanismus zurückzuführen ist. Reguläre systematische Phänomene wie die Substantivierung von Infinitiv- und Partizipformen und die Erkennung großgeschriebener Wortformen am Satzanfang werden systematisch

[13] Der Grund hierfür liegt vermutlich in der massiven Übergenerierung, wenn alle Substantive, Adjektive, Adverbien und Verben als Erstglieder für die Komposition mit Verben zugelassen werden: *(ich) *schlangestehe* würde analysiert.

[14] Die Berechtigung hierfür liegt natürlich in der Tatsache, dass flektierte Formen wie *(des) Schlangestehens* völlig akzeptierbare Wortformen sind.

Flexionsklasse	Beispiele	# Stämme
NMasc_0_x	*Felsen, Spekulatius, Tuareg*	19
NMasc_en_en	*Adressat, Kalif, Rezensent*	286
NMasc_en_en=in	*Abiturient, Korrespondent*	321
NMasc_es_$e	*Schaft, Korb, Vorwand*	426
NMasc_es_$er	*Geist, Mund*	20
NMasc_es_e	*Aal, Brief*	869
NMasc_es_en	*Dorn, Strahl, Zins*	18
NMasc_n_n	*Ahne, Rabe, Welpe*	126
NMasc_n_n=$in	*Affe, Spitzbube*	3
NMasc_n_n=in	*Archäologe, Schöffe*	63
NMasc_s_$	*Acker, Apfel, Bruder*	14
NMasc_s_$x	*Bindfaden, Ofen, Vorgarten*	20
NMasc_s_0	*Brösel, Kerker, Numismatiker*	947
NMasc_s_0=in	*Beifahrer, Kenner*	848
NMasc_s_e	*Ster, Abkömmling, Vokal*	302
NMasc_s_e=in	*Akteur, Gemahl, Sekretär*	31
NMasc_s_en	*Demonstrator, Lorbeer, Typ*	120
NMasc_s_en=in	*Administrator, Organisator*	62
NMasc_s_n	*Abbieger, Gevatter, Stachel*	14
NMasc_s_s	*Beatnik, Kognak, Transfer*	404
NMasc_s_x	*Balken, Rücken*	119

Abbildung 3.19: DMOR-Flexionsklassen: Nomina maskulinum (2/2)

behandelt. Die ebenso systematische Generierung von Kompositionsstammformen über die Fortsetzungsklasse und die Möglichkeit der Auflistung von Ausnahmeformen vermeidet Ambiguitäten bei der Erkennung von Komposita, wie sie bei Systemen auftreten, die uneingeschränkte Kompositionsfugen erlauben.

Dass dennoch eine starke Übergenerierung auftritt, liegt daran, dass in Flexionsklassen Lexeme nur aufgrund morphologischer Kriterien zusammengefasst werden, nicht aber, weil sie sich auf allen linguistischen Beschreibungsebenen gleich verhalten: Die Substantive *Blatt*$^{\mathbb{P}}$, *Buch*$^{\mathbb{P}}$, *Dach*$^{\mathbb{P}}$ und *Gehalt*$^{\mathbb{P}}$ haben dieselbe Flexionsklasse NNeut_es_$er (*Blatt, Blatt(e)s, Blätter*), treten aber in verschiedenen Kompositionsstammformen auf. Während es für *Gehalt*$^{\mathbb{P}}$ keinen Beleg im Korpus gibt, in dem die Grundform als Kompositionsstamm auftritt, finden sich einige hundert Belege für einen Kompositionsstamm mit *s*-Fuge (*Gehalts=Erhöhung, Gehalts=Liste, Gehalts=Zahlung* etc.), aber wiederum nur einige wenige für einen Kompositionsstamm mit Umlaut und *er*-Fuge (*Gehälter=Affäre, Gehälter=Kürzung*). Bei *Blatt*$^{\mathbb{P}}$, *Buch*$^{\mathbb{P}}$ und *Dach*$^{\mathbb{P}}$ ist es genau umgekehrt: Fugen-*s* kommt nur bei Falschzuordnungen vor (*Buchs=Baum* nicht zu *Buch*$^{\mathbb{P}}$, *Dachs=Berg* nicht zu *Dach*$^{\mathbb{P}}$). Umlaut und *er*-Fuge kommt relativ häufig vor (*Blätter=Wald, Bücher=Gilde, Dächer=Meer*), und die fugenlose Grundform ist der Normalfall: *Blatt=Laus, Buch=Messe, Dach=Boden* etc. Es ist eine offene Frage in der Forschung, welche Faktoren hier eine Rolle spielen.

(3.7) *bess<u>er</u>=Verdienende, höch<u>st</u>=Strafe, weit<u>est</u>=entfernt; krank<u>en</u>=Haus, ält<u>esten</u>=Rat*

Bei einigen Wortbildungsmustern, die nur in bestimmten Kontexten häufig auftreten, lässt DMOR Übergenerierung bewusst zu: Grundsätzlich sind die Komparativ- und Superlativform von Adjektiven als Erstglieder zugelassen. Dies erlaubt die Erkennung und Zerlegung der in 3.7 aufgelisteten Wortformen. Allerdings sind dafür im Korpus nur sehr wenige Belege zu finden.

(3.8) *möglich<u>er</u>weise, normal<u>er</u>weise, glücklich<u>er</u>weise, üblich<u>er</u>weise, notwendig<u>er</u>weise, fälschlich<u>er</u>weise*

Sehr häufig ist hingegen die Bildung von Adverbien auf *-weise* (vgl. 3.8), die jedoch keine Komposition, sondern Derivation darstellt. Da *weise* im Lexikon als Adjektiv eingetragen ist, wird für alle in 3.8 angegebenen Adverbien eine Analyse erzeugt, wenn auch nicht die richtige.

(3.9) *best=möglich, schnellst=möglich, frühest=möglich, weitest=gehend*

Für Superlativ beschränken sich die Funde auf ca. 30 Adjektive in recht eingeschränkten Fügungen (vgl. 3.9).

(3.10) *be<u>sten</u>=Liste, näch<u>sten</u>=Hilfe, jüng<u>sten</u>=Turnier*

be<u>sten</u>falls, schlimm<u>sten</u>falls, günstig<u>sten</u>falls, äußer<u>sten</u>falls, schlecht<u>esten</u>falls; schlecht<u>eren</u>falls

Für Superlativ mit *-en* als Fuge sind im Korpus vereinzelte Belege für Komposition mit substantivischem Kopf zu finden, aber zahlreiche Adverbbildungen auf *-falls* (vgl. 3.10). Da im DMOR-Lexikon jedoch *falls*$^{\mathbb{P}}$ nur als Konjunktion eingetragen ist und als solche nicht an Wortbildung teilhaben darf, werden die Bildungen auf *-falls*, die nicht im Lexikon eingetragen sind, nicht analysiert.

Dass Derivation in DMOR nicht behandelt wird, bedeutet, dass völlig regulär gebildete Ableitungen wie *brauch·bar, Brauch·bar·keit, Un·brauch·bar·keit* usw. im Lexikon eingetragen werden müssen und nicht über die in ihnen enthaltenen Morpheme analysiert werden können. Um bei besonders häufig vorkommenden Mustern dennoch die Analyse zu ermöglichen, ist im Substantivlexikon bspw. eine Form **Losigkeit* eingetragen, die aus der nicht erkannten Derivation eine erkannte Komposition macht und somit die Erkennung vieler Wortformen ermöglicht. Dass hier die Wortbildung nicht korrekt ist, wird für die korrekte Erkennung der Flexionsinformation in Kauf genommen. Diese Art des *Work-arounds* ist allerdings nur in sehr eingeschränkten Fällen möglich.

Neben der faktischen Gleichstellung von Simplizia und Derivativa gibt es einen zweiten Kritikpunkt an DMOR: Da das Lexikon allein der Kompilierung in

einen endlichen Automaten dient, ist für Erweiterungen allein die Möglichkeit der Hinzufügung von Stamm/Flexionsklasse-Paaren vorgesehen. Weitergehende Informationen lassen sich nur schwer integrieren. Dies führt z.B. dazu, dass eine vorhandene (semantische) Unterteilung der Personennamen in Vor- und Nachnamen im Flexionsklassenbezeichner kodiert wird. Diese Vermischung linguistischer Beschreibungsebenen in den zur Verfügung stehenden Mitteln (also den Sublexika) erschwert die Transparenz des Gesamtsystems und damit die Erweiterbarkeit erheblich.

3.2.2 Aspekte von Morphologiesystemen

Die Leistungsfähigkeit von Morphologiesystemen lässt sich anhand einiger Aspekte definieren. Diese sind *Effizienz, Korrektheit, Robustheit, Abdeckung* und *Spezifizität* (vgl. Abbildung 3.20, entnommen von Vortragsfolien zum Thema *Sprachtechnologie* von Hans Uszkoreit, vgl. Uszkoreit (2000), Folie 20).

efficiency	**geringer Zeit- und Speicherbedarf**
accuracy	**Fähigkeit, linguistisch korrekte Lösungen zu finden**
robustness	**Fähigkeit, mit allen möglichen Eingaben fertigzuwerden**
coverage	**größtmögliche Abdeckung der Sprache**
specifity	**Fähigkeit, die richtige Analyse zu selegieren**

Abbildung 3.20: Performanzkriterien nach Uszkoreit

Die **Effizienz** besagt zum einen, wie viele Wortformen in welcher Zeitspanne analysiert werden können, und zum anderen, wieviel Speicher dabei zur Laufzeit und für die Daten benötigt wird. Bei den auf der Morpholympics vorgestellten Systemen variierte der Zeitbedarf von einigen tausend Wortformen pro Sekunde bis hinunter zu weniger als 10 Wortformen pro Sekunde (vgl. Hausser (1996), S. 13), ein Unterschied, der sich bei der automatischen morphologischen Analyse eines Korpus mit mehreren Millionen Wortformen durchaus bemerkbar macht. Der Speicherbedarf ist heutzutage nicht mehr entscheidend, da Festplattengrößen im Gigabytebereich (ein Gigabyte entspricht 1024 Megabyte) und Hauptspeicher im Bereich mehrerer hundert Megabytes liegen: Der Speicherbedarf für Regeln und Lexika im DMOR-System liegt bei ungefähr **einem** Megabyte (vgl. Schiller (1996), S. 48).

Korrektheit ist ein relativer Begriff. Hundertprozentige Korrektheit kann ein Morphologiesystem nicht erreichen, wenn es keine allgemein anerkannte Theorie der Morphologie gibt. Solange umstritten ist, was genau eine Konversion ist und was nicht, kann kein System für sich reklamieren, Konversionen generell analysieren zu können. Neben der Theorie müssen insbesondere die Anforderungen der einer morphologischen Analyse nachfolgenden Komponenten berücksichtigt werden.

(3.11) [[zwei Farbe]$_{Phrase}$ ig]$_{Derivation}$+ADJ.Pos
[[zwei] [Farbe ig]$_{Derivation}$]$_{Komposition}$+ADJ.Pos
[[zwei] [farbig]]$_{Komposition}$+ADJ.Pos

Je nachdem, ob die Wortform *zweifarbig* als Komposition oder Derivation angesehen wird und wie tief die Zerlegung gehen soll, ist eine der in Beispiel 3.11 dargestellten Analysen die gewünschte. Als Eingabe für eine Syntaxkomponente sind alle drei Varianten als richtig anzusehen, da unabhängig von der inneren Struktur die morphosyntaktische Information bei allen dreien identisch ist. Ein Information-Retrieval-System hingegen kann aus der ersten und dritten Variante leichter die Information beziehen, dass ein Zusammenhang zum Substantiv *Farbe*$^{\mathbb{P}}$ besteht. Korrektheit lässt sich also immer nur in Bezug auf die vorher festgelegte Analysetiefe und Einordnung bestimmter morphologischer Phänomene messen.

Ein sehr wichtiges Kriterium für die automatische morphologische Analyse ist die **Robustheit**. Textkorpora enthalten sehr viel nur schwer analysierbares Material wie Fremdwörter, Tippfehler und Vermischungen von Buchstaben und Sonderzeichen jeder Art. In Morphologiesystemen sollte diese Art der Eingabe zumindest nicht zum Absturz des Systems führen. Ihre Verarbeitbarkeit hängt jedoch davon ab, ob solche Formen bereits im Lexikon berücksichtigt werden. Die einzige mir bekannte lexikalische Ressource, die Sonderformen aller Art berücksichtigt, ist *CISLEX* (vgl. Maier-Meyer (1995), S. 3).

Ein weiterer wichtiger Aspekt ist die erzielte **Abdeckung**, d.h. die Menge der Wortformen, für die das System eine richtige Analyse erzeugt. Die Abdeckung der Morphologiekomponente hängt nicht unbedingt direkt mit der Größe des Lexikons zusammen. Selbst mit einem kleinen Lexikon kann eine hohe Abdeckung erzielt werden, wenn die Morphologiekomponente stark übergeneriert. Ein Beispiel sind Fugenelemente bei der Komposition: In einem System, in dem zu jedem Substantiv die möglichen Kompositionserstglieder aufgelistet sind und keine Kompositionsfugen angegeben werden, hängt die Abdeckung der morphologischen Analyse direkt von der Anzahl und Qualität der eingetragenen Kompositionserstglieder ab. Werden hingegen Fugen im Lexikon abgelegt und die Wortbildungsregeln frei gestaltet, so wird die Anzahl der Analysen wesentlich höher sein, allerdings um den Preis einer größeren Anzahl von Ambiguitäten und Falschanalysen.

Das Kriterium der **Spezifizität** schließlich bezieht sich auf Ambiguitäten. Unter der Voraussetzung, dass unter den Ausgaben eines Morphologiesystems die gewünschte Analyse vorhanden ist, sollte eine **Disambiguierung** genau diese finden. Allerdings kann es auch durchaus gewünscht sein, mehrere Varianten als korrekt zuzulassen, z.B. bei der Zerlegung des Kompositums *Staubecken*. Sowohl die Lesart *Staub=Ecken* wie auch *Stau=Becken* sind morphologisch und semantisch möglich.

3.3 Von der Flexionsanalyse zur Wortbildungsanalyse

In diesem und dem vorangegangenen Kapitel wurden Theorie und Praxis der morphologischen Analyse behandelt. Die Verwendung von endlichen Automaten zur morphologischen Analyse ist aus Gründen der Geschwindigkeit und der allgemeinen Verfügbarkeit von Automatencompilern die zur Zeit vorherrschende Technik und wird in den großen bekannten Systemen – Gertwol, WordManager – sowie in zahlreichen kleineren Systemen mit Erfolg eingesetzt. Die zusätzliche Anwendung des Zwei-Ebenen-Modells, die ebenfalls über endliche Automaten verläuft, verspricht zusätzlich zu den genannten Vorteilen eine linguistisch adäquate Behandlung von morphologischen Prozessen in der Wortformenbildung. Die Behandlung morphologischer Phänomene erschöpft sich allerdings häufig in Flexion (die alle Systeme beherrschen) und Komposition (die die meisten Systeme beherrschen). Darüber hinausgehende Phänomene wie Derivation und neoklassische Wortbildung werden oft nur implizit behandelt, d.h., trotz der regelbasierten Vorgehensweise der Morphologiesysteme und des regelhaften Charakters der Phänomene werden sie als Simplizia behandelt. Auch zehn Jahre nach der ersten (und bislang einzigen) deutschen *Morpholympics* (vgl. Hausser (1996)) gilt die folgende Aussage:

> "Alle Systeme, die sich auf der MORPHOLYMPICS präsentierten, behandeln die Flexion des Deutschen. In der Regel verfügen die Systeme auch über Mechanismen der Kompositaanalyse, die allerdings häufig auf bestimmte Kompositionstypen beschränkt sind. Derivationssprozesse werden hingegen nur von den wenigsten Systemen behandelt, und auch hier ist [...] die Behandlung der suffixalen Derivation auf einige wenige Derivationssuffixe begrenzt." (Hausser (1996), S. 19f.)

Ten Hacken und Lüdeling stellen fest: "Word formation is usually not a separate issue. It is integrated with inflectional morphology or ignored altogether." (ten Hacken und Lüdeling (2002), S. 68) Dabei zeigt sich in der Praxis, dass der überwiegende Teil der Fehler, die eine Morphologiekomponente macht, in den morphologisch komplexen Formen begründet liegt. Zwei Hauptursachen lassen sich ausmachen:

1. Die Morphologiekomponente erzeugt keine Analyse, wenn ein Bestandteil einer Wortbildung nicht im Lexikon verzeichnet ist. Für ein produktives Wortbildungsmuster, wie die Komposition eines darstellt, kann ein fehlender Lexikoneintrag leicht in hunderten nicht analysierten Wortformen resultieren (vgl. z.B. die Auflistung der Wortbildungen mit *Polit-* in Abbildung 2.7 auf Seite 19).

2. Die Morphologiekomponente zerlegt die Wortform falsch, so dass die ausgegebene Flexionsinformation zwar für den gefundenen Kopf der Zerlegung richtig ist, aber für die Wortbildung nicht stimmt. Sehr häufig sind Eigennamen, die eine Zerlegung als Kompositum mit substantivischem Kopf erhalten, aber keine Analyse als Name: *Eisen=Berg, Fried=Berg, frei=Burg*. Zumindest die Städtenamen erhalten das falsche Genus und propagieren diesen Fehler in die nachfolgende Verarbeitungsstufe.

In keinem der beiden Fälle spielt die falsche Behandlung der Flexion eine Rolle. Der Status von Wortbildungen und den an ihr beteiligten Bestandteilen entscheidet über die Qualität und die Abdeckung einer Morphologiekomponente.

Für die regelbasierte Verarbeitung von Derivation, Konversionen und neoklassischer Wortbildung fehlt es an Konzepten. Der Grund dafür ist in der Tatsache zu suchen, dass derartige Phänomene mit dem für die Beschreibung von Flexion und Komposition entwickelten Inventar nicht zu erfassen sind. Es ist nötig, das Inventar zu entwickeln, das für die Beschreibung der über Flexion und Komposition hinausgehenden Phänomene benötigt wird. Der Nutzen ist zum einen die linguistisch adäquate Behandlung aller morphologischen Phänomene, zum anderen die damit einhergehende kontrollierte Erhöhung der Abdeckung, die Morphologiesysteme erzielen.

Im nachfolgenden Kapitel werden morphologische Einheiten und Prozesse ausführlich vorgestellt.

Kapitel 4

Morphologische Einheiten und Prozesse

Dieses Kapitel beschäftigt sich mit der Repräsentation und Strukturierung von Wortbildungsbestandteilen in einem Morphologiemodell. Zunächst wird eine in der Wortbildungsliteratur übliche Aufteilung in morphologische Prozessierungsmodelle erläutert (vgl. Abschnitt 4.1): Item and Arrangement (IA) und Item and Process (IP). Im Anschluss daran werden für das in dieser Arbeit gewählte Modell relevante Begriffe erklärt. Sie bilden die Grundlage für die Behandlung von Wortbildungsphänomenen. Die Prozessierungsmodelle geben eine Gliederung für die Beschreibung der Wortbildungsphänomene vor: Je nachdem, welche Arten von Prozessen in derartigen Phänomenen ablaufen, erfolgt die Beschreibung im Rahmen des IA-Modells (vgl. Abschnitt 4.2) oder im Rahmen des IP-Modells (vgl. Abschnitt 4.3). Die Beschreibung der Einheiten, ihrer Eigenschaften und der Wortbildungsmuster, in denen sie auftreten, bildet die Grundlage für die Konzeption eines Lexikons, das die morphologische Analyse optimal unterstützt.

4.1 Paradigmen der morphologischen Modellierung

In der Literatur werden häufig zwei gegensätzliche Modelle zur grundsätzlichen Beschreibung morphologischer Phänomene eingesetzt. Bei dem einen handelt es sich um **Item and Process (IP)**, bei dem anderen um **Item and Arrangement (IA)** (vgl. Hockett (1954)).

In IP werden morphologische Phänomene als Funktionen betrachtet, die eine Veränderung einer Einheit bewirken. Bei Umlautung wird der umzulautende Vokal einem Prozess unterzogen, der bewirkt, dass sich seine Eigenschaften (z.B. die Aussprache) ändern, d.h., aus einer bestehenden Form wird eine ver-

änderte Form erzeugt. Das Anhängen einer Flexionsendung kann als Prozess der 'Suffigierung' oder 'Konkatenation' angesehen werden.

Die Sicht von IA entspricht der von 'Dingen' und ihrer 'Anordnung'. In ihr lassen sich konkatenative Prozesse sehr leicht beschreiben, da lediglich die Einheiten benannt werden müssen, die sich zu größeren Strukturen zusammensetzen lassen, und Regeln angegeben werden müssen, die auf diesen Einheiten operieren. Die Beschreibung nicht-konkatenativer Prozesse führt hingegen zu Schwierigkeiten.

In dieser Arbeit wird die morphologische Analyse aus einer strikten IA-Sichtweise heraus behandelt. Die Hauptaufgabe ist dementsprechend, die 'Items' oder morphologischen Einheiten so zu wählen, dass durch ihre Anordnung alle linguistischen Phänomene, die eine morphologische Analyse zu berücksichtigen hat, abgedeckt werden können.

4.2 Einheiten und Prozesse in IA

Morphologische Einheiten sind schwer zu definieren, da sie über Ausprägungen auf allen linguistischen Beschreibungsebenen verfügen. Als Grapheme liegen sie als Zeichenkette vor, als Phoneme stellen sie (Sprech-)Laute dar, als Sememe stehen sie für eine bestimmte Bedeutung. Darüber hinaus treten sie in unterschiedlichen Funktionen auf, z.B. als Flexionselemente, Wortbildungselemente oder Stämme. Diese inhomogene Menge von Einheiten wird dennoch i.A. unter einem Oberbegriff zusammengefasst, dem Begriff des *Morphems*, der 'kleinsten bedeutungstragenden Einheit'. Allerdings fällt dessen Definition in einem Item-and-Arrangement-Umfeld anders aus als in einem Item-and-Process-Umfeld. In dieser Arbeit werden Morpheme ausschließlich aus der Item-and-Arrangement-Sichtweise betrachtet: Morphologische Einheiten werden als konkrete Bausteine angesehen, aus denen Wortbildungsregeln größere, definierte Einheiten zusammensetzen. Infolge dessen definiert sich ein Wortbildungsprodukt als eine disjunkte und vollständige Zusammensetzung von Bausteinen, deren Anordnung in einer Wortbildungsregel festgelegt ist. In den folgenden Abschnitten werden diese Bausteine nach ihren unterschiedlichen Funktionen spezifiziert. Zunächst wird jedoch eine Übersicht darüber gegeben, welche Arten von Morphemen traditionell unterschieden werden und welche Typen sich der traditionellen Sichtweise entziehen.

4.2.1 Übersicht: Das Morphem

Ein zentraler Begriff für die Beschreibung der Morphologie einer Sprache ist der Begriff des **Morphems**. An dieser Stelle wird nicht versucht, diesen Begriff zu definieren. Es wird lediglich eine terminologische Einordnung vorgenommen, die das Verständnis der folgenden Abschnitte erleichtern soll.

Traditionell werden drei Arten von Morphemen unterschieden:

1. **lexikalische Morpheme** (auch **Grundmorpheme** oder **Basismorpheme** genannt),
2. **Wortbildungsmorpheme** (auch **Affixe** genannt) und
3. **grammatische Morpheme** (auch **Flexionselemente** oder **Flexive** genannt).

Der Unterschied zwischen Basismorphemen und Wortbildungsmorphemen besteht darin, dass Basismorpheme frei vorkommen, Wortbildungsmorpheme nur in gebundener Form. Die traditionelle Unterscheidung zwischen Komposition und Derivation baut auf diesem Morphembegriff auf: Eine Komposition ist eine Verbindung aus mindestens zwei Basismorphemen, eine Derivation ist eine Verbindung aus einem Basismorphem und mindestens einem Wortbildungsmorphem.[1] Flexionselemente treten wie Wortbildungsmorpheme nur gebunden auf, sind aber durch ihre Einbindung in Flexionsparadigmen stärker restringiert.[2] In diesen drei Arten von Morphemen sind Unikale und Konfixe allerdings nicht erfasst. "Auf eine ganze Reihe von Morphemen der Hauptwortarten treffen die Merkmale für Grundmorpheme nicht uneingeschränkt zu. So sind eine Vielzahl entlehnter Elemente nicht wortfähig, sondern treten nur in Kombination mit anderen Morphemen auf: *therm, stat, bio*." (Fleischer und Barz (1995), S. 25; vgl. auch Erben (2000), S. 26, Fußnote 22)

Bei **Unikalen** handelt es sich um morphologische Einheiten, die keinem frei vorkommenden Basismorphem zugeordnet werden können: *Schorn* in *Schornstein*, *lier* und *zicht* in *verlieren* und *verzichten*, *Kinker* und *litz* in *Kinkerlitzchen*. Es handelt sich i.A. um früher frei vorkommende Einheiten, die im Gegenwartsdeutsch nur noch als Bestandteil einer oder weniger komplexen Formen erhalten sind. Für die Behandlung dieser Einheiten in einem Morphologiesystem oder in einem Lexikon gibt es keine klaren Regeln: Es muss entschieden und dokumentiert werden, ob sie als eigenständige Einheiten behandelt werden oder nicht.

Bei **Konfixen** handelt es sich ebenfalls um morphologische Einheiten, die als solche erkennbar sind, aber keinem frei vorkommenden Basismorphem zugeordnet werden können. Bei ihnen ist die Besonderheit, dass sie erkennbar aus einer anderen Sprache stammen: *Biologe, Biologie, biologisch* sind Beispiele für Wortformen, deren Endungen *-e, -ie, -isch* in Form und Bedeutung Derivationssuffixen entsprechen, deren erster Bestandteil BIOLOG aber keine frei

[1] Diese Darstellung ist stark vereinfacht, aber im Prinzip läuft es auf diese Unterscheidung hinaus.

[2] Da es hier nicht um die Abgrenzung von Flexion, Komposition und Derivation geht und Flexive für die Behandlung der Wortbildung in dieser Arbeit irrelevant sind, wird hier nicht weiter auf sie eingegangen.

vorkommende Form ist. Es handelt sich darüber hinaus um eine morphologisch komplexe Form, wie die Betrachtung der Bestandteile ergibt: *βίος 'das Leben'*, *λόγος 'das Wort'*. Eine Zuordnung innerhalb eines Morphologiesystems für das Deutsche muss – wie bei den Unikalen – gefunden und dokumentiert werden.

Neben den angesprochenen Einheiten gibt es noch den Begriff der **Fugenelemente**. Diese werden gemeinhin nicht als Morpheme angesehen, da sie nicht 'bedeutungstragend' sind, haben dann allerdings überhaupt keinen Status. In dieser Arbeit werden Fugenelemente als dem Bestandteil zugehörig angesehen, hinter dem sie in der Wortform vorkommen (vgl. den folgenden Abschnitt 4.2.2).

Abkürzungen oder Kurzwortbildungen wie *Hapag* und *Kripo* werden von keinem Morphembegriff erfasst (vgl. Erben (2000), S. 25, Fußnote 19). Dies gilt ebenfalls für sogenannte *Kontaminationen* (*Kurlaub* aus *Kur* und *Urlaub*).

Morpheme werden durch **Morphe** realisiert, das sind die **orthographischen Formen** von Morphemen. **Allomorphe** sind verschiedene Morphe desselben Morphems, z.B. *Haus* und *Häus*.[3]

Nach der Anzahl der Morpheme in einer nicht flektierten Wortform wird nach **morphologisch einfachen Formen** (Simplizia, sie bestehen nur aus einem Morphem) und **morphologisch komplexen Formen** (enthalten mindestens zwei Morpheme) unterschieden.

Das im Item-and-Arrangement-Ansatz vertretene Prinzip der disjunkten Zusammensetzung erlaubt die Angabe der **Morphemgrenzen**[4] für jede Wortform.

(4.1) a. *Un·be·denk·lich·keits·be·schein·ig·ung, Wirk·sam·keit, Kontra·zept·ion, Häus·chen, Blau·beere, Ein·heit, Bio·log·e*

b. *in·form·ier·en*

c. *ge·ruder·t, Häus·er*

d. *Apfel, grün, gegen*

In 4.1 sind einige Wortformen mit Morphemgrenzenmarkierungen abgebildet.[5] Basismorpheme sind unterstrichen dargestellt. Die Wortformen in a enthalten Basismorpheme und Wortbildungsmorpheme, die Wortform in b enthält

[3]Ein Morphem müsste eigentlich als die Menge all seiner Allomorphe dargestellt werden, aber aus Gründen der Einfachheit wird i.A. stellvertretend dafür das am wenigsten komplexe aus der Menge genommen, in diesem Fall also die nicht-umgelautete Form.

[4]Strenggenommen handelt es sich um die **Morph**grenzen, da sich die Wortform aus Morphen zusammensetzt, aber da die Morphe ja immer stellvertretend für ein Morphem stehen, wird, wie in der Literatur üblich, der Begriff **Morphemgrenze** verwendet.

[5]Hier wird keine universelle Gültigkeit beansprucht, die Zerlegungen beziehen sich auf das in dieser Arbeit dargelegte Modell einer Item-and-Arrangement-Morphologie. Generell gilt nach wie vor: "Noch weniger darf man glauben, dass die durch analyse gefundenen elemente die urelemente der sprache überhaupt sind. Unser unvermögen ein element zu analysieren beweist gar nichts für dessen primitive einheit." (Paul (1886), S. 297f.)

ein Basismorphem, zwei Wortbildungsmorpheme und ein Flexionsmorphem (*-en*), und die Wortformen in c enthalten Basismorpheme und Flexionsmorpheme. Simplizia bestehen per Definition aus genau einem Basismorphem (vgl. 4.1 d).

Basismorpheme sind die Träger lexikalischer Information. Als solche lassen sie sich Lexemen zuordnen: *denken*$^{\mathbb{P}}$, *scheinen*$^{\mathbb{P}}$, *Haus*$^{\mathbb{P}}$, *blau*$^{\mathbb{P}}$, ...[6]

4.2.2 Stammformen

Morphe bzw. zusammenhängende Gruppen von Morphen werden in der Sprache in wort- und wortformbildenden Funktionen realisiert, die in dieser Arbeit als **Stammformen** bezeichnet werden. Dieser Begriff wurde in Fuhrhop (1998) zur einheitlichen Darstellung von Flexion, Derivation und Komposition eingeführt. Fuhrhop unterscheidet drei Arten von Stammformen[7], die **Flexionsstammform** (kurz: Flexionsstamm) für die Wortformenbildung und die **Derivationsstammform** (kurz: Derivationsstamm) und **Kompositionsstammform** (kurz: Kompositionsstamm) für die Wortbildung. Mit diesen drei Stammformtypen lassen sich die **konkatenativen** Prozesse in der Morphologie modellieren:

Flexion stellt sich dar als Affigierung von Flexionsmorphemen an eine Flexionsstammform. In Beispiel 4.1 c auf Seite 50 treten die Basismorpheme *ruder* und *Häus* in ihrer Funktion als Flexionsstammformen auf.

Derivation stellt sich dar als Affigierung von Wortbildungsmorphemen an eine Derivationsstammform. In Beispiel 4.1 a auf Seite 50 treten die Basismorpheme *wirk* und *ein* (zu *einen*$^{\mathbb{P}}{}_{V}$) in ihrer Funktion als Derivationsstammformen auf.

Komposition stellt sich dar als Affigierung von Basismorphemen an eine Kompositionsstammform. In Beispiel 4.1 a auf Seite 50 treten die Basismorpheme *blau* und *Bio* in ihrer Funktion als Kompositionsstammformen auf.

Ein großer Vorteil des Konzepts hinsichtlich einer rein konkatenativen Sichtweise auf morphologische Prozesse ist die Tatsache, dass allomorphe Stammformen unabhängig davon, ob bei ihrer Erzeugung nicht-konkatenative Prozesse (**Tilgung** und **Umlautung**) stattgefunden haben oder nicht, denselben Status

[6] Im Falle der Lehnwörter führt die Zuordnung auf Lexeme einer anderen Sprache: *capere*$^{\mathbb{P}}$ (lateinisch *nehmen, fassen*) und entweder *forma*$^{\mathbb{P}}$ (lat. *Gestalt, Form*) oder *formare*$^{\mathbb{P}}$ (lat. *bilden, gestalten*). Dies ist sicherlich einer der Gründe dafür, dass die Behandlung neoklassischer Wortbildung in Morphologiesystemen für Deutsch nicht sehr weit verbreitet ist.

[7] Eine weitere von Fuhrhop eingeführte Form, die Vergleichssegmentform, die insbesondere für die Beschreibung neoklassischer Wortbildungsprozesse geeignet ist, wird in Abschnitt 4.2.4 beschrieben.

haben. *Haus* und *Häus* sind gleichberechtigte allomorphe Flexionsstammformen und gleichberechtigte allomorphe Derivationsstammformen: Der Umlaut in der einen Form ist weder für den Status noch für die Verarbeitung relevant. *Öf* und *Äug* lassen sich trotz Umlautung und Tilgung eindeutig den Paradigmen *Ofen*$^{\mathbb{P}}{}_{NN}$ und *Auge*$^{\mathbb{P}}{}_{NN}$ zuordnen, wie die Verkleinerungsformen *Öfchen 'kleiner Ofen'* und *Äuglein 'kleines Auge'* zeigen. Stammformen bieten ein sauberes Konzept für eine Behandlung von Wort(formen)bildung im Rahmen von Item and Arrangement.

Exkurs: Flexion und Wortbildung Kompositionsstammformen verbinden sich mit Flexionsstammformen zu Flexionsstammformen. Derivationsstammformen verbinden sich mit Wortbildungsaffixen ebenfalls zu Flexionsstammformen. Derivations- und Kompositionsstammformen können nicht am Ende einer Wortform auftreten, Flexionsstammformen hingegen schon. An sie können nur noch Flexionsaffixe angehängt werden. Fuhrhop (1998) bezeichnet Flexionsstammformen daher auch als **Grundstammformen**. Damit ist das funktionale Gegenstück zur Grundform in einem Paradigma (vgl. Abschnitt 2.1.2) benannt: Ein Lexem steht für ein Paradigma und wird durch eine Grundform repräsentiert. Einige der Wortformen aus dem Paradigma treten in der Funktion von Flexionsstammformen auf. Eine dieser Flexionsstammformen vertritt das Paradigma funktional, das ist die Grundstammform.

Die Bildung der **Wortform** kann im Deutschen auf die Flexionsstammform begrenzt werden, da in dieser Sprache die Flexion im Allgemeinen am Wortrand stattfindet, nicht im Wort. Eisenberg nennt zwei Gegenbeispiele, zum einen die Demonstrativpronomen *derjenige, diejenige, dasjenige*, zum anderen die Wortformen *Kindchen* und *Kinderchen* (vgl. Eisenberg (1994), S. 201), die man als Wortformen eines Paradigmas auffassen könnte, da die erste nur im Singular, die zweite nur im Plural verwendbar ist. Im vorliegenden Modell können sie allerdings auch als Wortformen zweier verschiedener (defektiver) Paradigmen *Kindchen*$^{\mathbb{P}}$ und *Kinderchen*$^{\mathbb{P}}$ angesehen werden, wobei sich die Wortform *Kinderchen* aus der Derivationsstammform *Kinder* und dem Derivationssuffix *-chen* zusammensetzt.[8] Die oben angesprochenen Demonstrativpronomen hingegen und Lehnwortformen wie *Singularetantum* mit Pluralform *Singulariatantum* (vgl. Duden (2001), S. 1455) und *Pluraletantum* mit Pluralform *Pluraliatantum* (vgl. Duden (2001), S. 1219) fasse ich als Ausnahmen auf. Dementsprechend sind im Deutschen Flexionssuffixe die einzigen Einheiten, die nicht in der Funktion einer Stammform auftreten, sondern eine Wortform 'abschließen'. Diese Tatsache rechtfertigt nach meiner Ansicht bereits

[8] An den Beispielen *Mütter·chen* und *Hühner·ei* ist leicht ersichtlich, dass eine Kompositionsstammform zwar aussehen kann wie eine Pluralform, aber semantisch nichts mit dieser zu tun haben muss: Weder die Verkleinerungsform einer Gruppe von Großmüttern noch das eine Ei von mehreren Hühnern erscheinen plausibel.

die Unterscheidung von Derivation und Flexion, die gelegentlich angezweifelt wird (vgl. Bauer (2003), S. 91ff). □

Fuhrhop sieht die Stammformen als in einem **Stammparadigma** eingebettet an. Demnach umfasst der Lexembegriff außer dem Flexionsparadigma auch das Stammparadigma (vgl. Fuhrhop (1998), S. 22). Diese Auffassung wird in dieser Arbeit geteilt.

Erstglied		Zweitglied	Produkt		
Lexem	Ksf	Fsf	Fsf	Prozess bei Sfb	Zeile
hoch$^{\mathbb{P}}{}_{ADJ}$	*hoch*	*Haus*	*Hochhaus*	–	1
Haus$^{\mathbb{P}}{}_{NN}$	*Haus*	*hoch*	*haushoch*		2
lesen$^{\mathbb{P}}{}_{V}$	*lese*	*Stunde*	*Lesestunde*	Fugung	3
Arbeit$^{\mathbb{P}}{}_{NN}$	*Arbeits*	*Amt*	*Arbeitsamt*		4
Licht$^{\mathbb{P}}{}_{NN}$	*Lichter*	*Kette*	*Lichterkette*		5
Westen$^{\mathbb{P}}{}_{NN}$	*West*	*Küste*	*Westküste*	Tilgung	6
Sprache$^{\mathbb{P}}{}_{NN}$	*Sprach*	*Kurs*	*Sprachkurs*		7
einzeln$^{\mathbb{P}}{}_{ADJ}$	*einzel*	*Fall*	*Einzelfall*		8
Buch$^{\mathbb{P}}{}_{NN}$	*Bücher*	*Wurm*	*Bücherwurm*	Fug. + Umlaut	9
hoch$^{\mathbb{P}}{}_{ADJ}$	*höchst*	*Strafe*	*Höchststrafe*		10

Abbildung 4.1: Kompositionsstammformen und Kompositabildung

In Abbildung 4.1 sind Beispiele für Kompositionsstammformen und ihre Verbindung mit Flexionsstammformen angegeben.[9] Bei den ersten beiden Beispielen entspricht die Kompositionsstammform der Grundform. In den Beispielzeilen 3-5 ist die Kompositionsstammform **gefugt**. In Zeile 3 tritt beispielsweise eine *e*-Fuge an den Verbstamm *les*. In Zeilen 6-8 sind bei der Kompositionsstammform Zeichen am Ende (bezogen auf die Grundform) **getilgt**. Aus der Grundform *Sprache* wird ein verkürzter Kompositionsstamm *Sprach*[10] (Zeile 7). Zeilen 9 und 10 schließlich zeigen Kompositionsstammformen, die gegenüber der Grundform **umgelautet und gefugt** sind. Die morphologischen Eigenschaften des Zweitgliedes bestimmen die morphologischen Eigenschaften des Wortbildungsprodukts (vgl. Olsen (1991), S. 336). Da die Flexionsstammformen in

[9] **Fsf** steht für 'Flexionsstammform', **Ksf** für 'Kompositionsstammform', **Sfb** für 'Stammformbildung'. Bei zweigliedrigen Komposita wird der linke Bestandteil allgemein als **Erstglied** bezeichnet, der rechte als *Zweitglied* oder **Kopf**. Die Angabe der Zeilennummer dient der Referenzierung.

[10] Dass es sich dabei nicht um das Imperfekt des Verbs *sprechen*$^{\mathbb{P}}$ handelt, liegt daran, dass flektierte Formen im Deutschen gewöhnlich nicht als Erstglieder in Komposita oder Basen von Derivativa auftreten können. Dies ist eine Grundannahme, die sich aus Gegenbeispielen, wie sie z.B. in Fußnote 8 dargestellt sind, ergibt.

den Beispielen mit den jeweiligen Grundformen übereinstimmen, sind die Lexeme nicht separat angegeben.

Basis		Affix	Produkt		
Lexem	Dsf	Fsf	Fsf	Prozess bei Sfb	Zeile
Stein$^{\mathbb{P}}{}_{NN}$	*stein*	*-ern*	*steinern*	–	1
schreiben$^{\mathbb{P}}{}_{V}$	*schreib*	*ver-*	*verschreiben*		2
lesen$^{\mathbb{P}}{}_{V}$	*leser*	*-lich*	*leserlich*	Fugung	3
Ehre$^{\mathbb{P}}{}_{NN}$	*ehren*	*-halber*	*ehrenhalber*		4
lachen$^{\mathbb{P}}{}_{V}$	*läch*	*-el*	*lächel*	Umlautung	5
krank$^{\mathbb{P}}{}_{ADJ}$	*kränk*	*-lich*	*kränklich*		6
Grube$^{\mathbb{P}}{}_{NN}$	*grüb*	*-chen*	*Grübchen*	Tilgung + Uml.	7
Blume$^{\mathbb{P}}{}_{NN}$	*blum*	*-ig*	*blumig*	Tilgung	8

Abbildung 4.2: Derivationsstammformen und Derivationsbildung

In Abbildung 4.2 sind Beispiele für Derivationsstammformen und ihre Verbindung mit Flexionsstammformen angegeben.[11] Bei den ersten beiden Beispielen entspricht die Derivationsstammform der Grundform. In den Beispielzeilen 3 und 4 ist die Derivationsstammform gefugt.[12] In den Zeilen 5 und 6 ist die Derivationsstammform umgelautet. In Zeilen 7 und 8 sind kombinierte Tilgung und Umlautung sowie Tilgung dargestellt.

Das Konzept der Derivations- und Kompositionsstammformen bewirkt eine **Vereinheitlichung** der Darstellung von Derivation und Komposition. Der Unterschied, der zwischen Abbildung 4.1 und Abbildung 4.2 besteht, bezieht sich allein auf den morphologischen Status des Zweitgliedes bzw. des Affixes, d.h., ob es frei oder gebunden vorkommt. Höhle kommt bei der Betrachtung der Struktur von Derivationen und Kompositionen zu einem ähnlichen Ergebnis und bezeichnet dies als die "Kompositionstheorie der Affigierung" (vgl. Höhle (1982), S. 82).

4.2.3 Affixe

Nachdem die Basismorpheme in ihren verschiedenen Auftretensweisen beschrieben sind, verbleiben noch die Affixe. In dem in dieser Arbeit vertretenen

[11] **Dsf** steht für 'Derivationsstammform', **Fsf** für 'Flexionsstammform', **Sfb** für 'Stammformbildung'. Bei Derivationen wird der Stamm allgemein als **Basis** bezeichnet, an die das **Affix** angehängt wird. Die Angabe der Zeilennummer dient der Referenzierung.

[12] Dass es sich im Beispiel *leserlich* um die Ableitung eines Verbs handelt, ergibt sich zum einen aus dem Affix (vgl. den folgenden Abschnitt 4.2.3), zum anderen aus der Bedeutung: *leserlich* heißt soviel wie *kann (gut) gelesen werden*.

Modell der deutschen Morphologie[13] sind Derivations**suffixe** Träger morphologischer Eigenschaften (vgl. Abschnitt 5.1.1). Dies erklärt, warum eine Derivation wie *blumig* (vgl. Zeile 8 in Abbildung 4.2) ein Adjektiv sein kann, obwohl die Basis substantivisch ist. Im Unterschied zu anderen morphologischen Einheiten **selegieren** Affixe Basen nach deren morphologischen Eigenschaften (vgl. Lüdeling und Fitschen (2002) und ten Hacken und Lüdeling (2002)). Dies bedeutet insbesondere, dass die Derivation im Deutschen mit wesentlich restriktiveren Regeln als die Komposition beschrieben werden kann. Affixen kommt somit eine besondere Rolle in der Beschreibung von Wortbildungsprozessen zu.

Außer als Suffixe treten Affixe im Deutschen als **Präfixe** und **Zirkumfixe** auf. Präfixe werden an die linke Seite einer Basis affigiert anstatt an die rechte. Sie beeinflussen im Gegensatz zu den Suffixen nicht die morphosyntaktischen Eigenschaften des Wortbildungsproduktes.[14] Dementsprechend gehören sie zu den morphologischen Einheiten, die keiner Wortart angehören.

Zirkumfixe bilden eine diskontinuierlich auftretende Kombination aus einem Präfix und einem Suffix. Das typische Muster ist die Nominalisierung von Verben mit dem Präfix *Ge-* und dem Suffix *-e* (*Ge·renn·e, Ge·heul·e, Ge·seufz·e*). Die Klassifizierung als Zirkumfix ergibt sich aus der Tatsache, dass keine der beiden möglichen Zerlegungen in unmittelbare Konstituenten (**Geseufz*, **Seufze*) belegt ist. Daher muss hier eine Wortbildung angenommen werden, bei der beide Affixe gleichzeitig an die Basis gehängt werden. Es handelt sich dabei um ein **Klammerparadox** (vgl. Spencer (1991), S. 397ff.).[15]

4.2.4 Zwischenkategorien

Unter der Bezeichnung *Zwischenkategorien* werden hier morphologische Einheiten beschrieben, die zwar durch ihr reihenbildendes Auftreten als eigenständige Einheiten identifiziert werden können, aber nicht frei vorkommen und daher nur schwer einem Morphem zuzuordnen sind. Es handelt sich einerseits um die **Affixoide**, andererseits um die in Abschnitt 4.2.1 angesprochenen **Unikale** und **Konfixe**.

[13] Dieses Modell ist im Rahmen des DeKo-Projekts konzipiert worden, vgl. Abschnitt 5.1.

[14] Dies ist für eine Untermenge der Präfix- und Partikelverben im Deutschen umstritten: Ich teile die in Olsen (1991) vertretene Meinung, dass in diesen Fällen der Präfigierung eine **Konversion** der Basis vorweggeht. Neben den in Olsen (1991), S. 342ff., gegebenen Gegenargumenten scheint mir insbesondere der Gedanke plausibel zu sein, dass eine Konversion allein z.B. im Falle von *feucht*$^{\mathbb{P}}{}_{ADJ}$ nach *feuchten*$^{\mathbb{P}}{}_{V}$ keine hinreichend gut unterscheidbare Form schafft: Das Resultat ist eine gebräuchliche Flexionsform des Adjektivs und damit von diesem nur schwer unterscheidbar. Erst das (nachfolgende) Hinzufügen eines Präfixes z.B. ermöglicht die eindeutige Unterscheidbarkeit.

[15] Ein Klammerparadox tritt auch bei anderen Phänomenen auf, beispielsweise bei der Kombination von Präfix- oder Partikelverben und Adjektivsuffixen (*be·schein·ig(en), un·aus·weich·lich*) etc.

Affixoide

Affixoide sind affix-ähnliche Gebilde, die die orthographische Form eines freien Morphems haben, deren Bedeutung aber nicht (mehr) mit der des freien Morphems übereinstimmt. Beispiele sind *-mäßig, -artig, -durstig; super-, Affen-*. Beispiele für Wortbildungsprodukte sind *kosten·mäßig, taten·durstig, flucht·artig; super·reich, Affen·schande*. An den Beispielen ist erkennbar, dass es Abstufungen gibt: *Affenschande* hat nichts mit *Affen* zu tun, sondern es handelt sich um eine Umschreibung für *große Schande*. Man könnte nun argumentieren, dass dies auch für Komposita gilt, deren Bedeutung sich nicht transparent aus der Bedeutung des Bestandteile erschließen lässt (*Augen=Schein, Fleisch=Wolf*). Der Unterschied ist jedoch, dass sich zahlreiche weitere Bildungen mit dem Erstglied *Affen-* finden lassen, die jeweils eine Verstärkung des Zweitgliedes ausdrücken: *Affen·hitze, Affen·tempo, Affen·theater*. Das Affixoid ist **reihenbildend**.

In einem Morphologiemodell, in dem Affixe sich von Grundmorphemen unter Anderem dadurch unterscheiden, dass sie ihre Basen nach morphologischen Eigenschaften selegieren, sind Affixoid eindeutig darstellbar: *Affen* (zu *Affe*$^{\mathbb{P}}{}_{NN}$) kann als Kompositionsstammform (z.B. in *Affenhaus*) auftreten. *Affen-* (zu *Affen*$^{\mathbb{P}}{}_{NNAff}$, *Aff* für 'Affix') kann als Derivationsstammform[16] (z.B. in *Affenschande*) auftreten. Für Adjektivsuffixoide kann man entweder ebenfalls ein eigenes Lexem vorsehen, oder man interpretiert die Wortbildungsprodukte ohnehin als Derivationen mit einer komplexen Basis (*Tatendurst·ig*, vgl. Abschnitt 4.2.5). In allen Fällen kann eine eindeutige Entscheidung innerhalb des vorgestellten Modells getroffen werden. Welche dies jeweils ist, wird bei der Realisierung des Lexikons entschieden.

Exkurs: Derivationsaffixe in der Sprachgeschichte Obwohl in dieser Arbeit eine rein synchrone Sichtweise beschrieben wird, ist an dieser Stelle der Blick auf die Sprachgeschichte nützlich: Sprachgeschichtlich gesehen ist die Derivation ein Ableger der Komposition, da die heute nur noch gebunden auftretenden Affixe ursprünglich eigenständige Morpheme mit festen Bedeutungen waren. Zum Beispiel *-bar* entstand demzufolge aus der althochdeutschen Form *beran* 'tragen', also ist z.B. *fruchtbar* der Herkunft nach ein Kompositum mit der Bedeutung 'Frucht tragend' (vgl. Erben (2000), S. 54; siehe weiterhin Fleischer und Barz (1995), S. 252, Wilmanns (1899), S. 496, Kluge (1995), S. 79). Die heutigen Derivationssuffixe entwickelten sich also von freien Morphemen über Affixoide erst zu Affixen, und heutige Affixoide entwickeln sich ebenso allmählich zu Affixen (vgl. Erben (2000), S. 136ff). Fleischer und Barz weisen darauf hin, dass die Zuordnung der Wortbildung mit Affixoiden zu Derivation oder Komposition in jüngerer Zeit wieder kontrovers diskutiert wird (vgl. Fleischer

[16] Da Affixoide in ihrer speziellen Bedeutung nicht frei vorkommen, muss man sie als Derivationen ansehen.

und Barz (1995), S. 26ff). In modernen Morphologie-Lehrbüchern wird die Zwischenstellung der Affixoide beschrieben: “With a word like *childlike* we have something which can be seen as being on the cusp between word-status and affix-status: we might not know whether to analyse this word as being a compound or a derivative.” (Bauer (2003), S. 270) □

Unikale und Konfixe

Unikale und Konfixe unterscheiden sich von Basismorphemen dadurch, dass sie – zumindest aus synchroner Sicht – über keine Wortart verfügen. Im Gegensatz zu Affixen selegieren sie keine Basen.[17] Wortbildungen wie *Stief·bruder, Schwieger·vater* genauso wie *pseudo·intellektuell, hyper·modern* lassen sich also nicht mit den Begriffen Derivation oder Komposition fassen. Die unterstrichen dargestellten Bestandteile müssen also – ähnlich wie Präfixoide – als gebundene Morpheme aufgefasst werden, für die eigene Lexeme geschaffen werden, bei denen sie wiederum als Derivations- oder Kompositionsstammformen fungieren können. In Fuhrhop (1998) wird für die Gruppe der Konfixe, deren Auftreten mit zwei verschiedenen Affixen belegt ist, der Begriff der **Vergleichssegmentform** verwendet (vgl. auch Lüdeling et al. (2002)). Für Paare wie *organisier(en)/Organisation, demonstrier(en)/Demonstration* etc. bilden demzufolge *organis* und *demonstr* Vergleichssegmentformen.

4.2.5 Komplexe Lexikoneinträge

Eine Besonderheit der Differenzierung morphologischer Einheiten nicht nur nach ihrem Status (frei oder gebunden vorkommend), sondern auch nach ihrer morphologischen Form (einfach oder komplex) ist es, dass morphologisch komplexe, aber gebundene Einheiten im Modell beschrieben werden können. Dies erlaubt eine adäquate Beschreibung des Phänomens der sogenannten **Zusammenbildungen**.

In der Literatur wurden Beispiele wie *Dickhäuter, viertürig, Appetithemmer* lange Zeit kontrovers behandelt (vgl. Leser (1990)). Mittlerweile scheinen sich die zwei Varianten durchzusetzen, dass es sich entweder um Komposition mit Argumentvererbung oder um Derivationen von Phrasen handelt (vgl. Donalies (2002), S. 93ff.). Im vorliegenden Modell können Phrasen in der Funktion einer komplexen gebundenen Derivationsstammform auftreten: *Dickhäut* zu einem Lexem *dicke Haut*$^{\mathbb{P}}{}_{Phrase}$. Der Vorteil dieser Vorgehensweise ist der, dass ähnliche Konstruktionen (*rechtskräftig, tatkräftig*) mit demselben Vorgehen model-

[17] Das Gegenteil ist der Fall: Neoklassische Affixe wie *-abel* wählen ihre Basen nach deren Herkunft, also *akzeptabel* vs. *annehmbar* (Derivationsstammformen unterstrichen dargestellt).

liert werden: *rechtskräft* als freie[18] komplexe Derivationsstammform zu einem Lexem *Rechtskraft*$^{\mathbb{P}}{}_{NN}$[19]. Auch die in Abschnitt 4.2.1 angeführten Beispiele *Biologe, Biologie, biologisch* passen in das Schema: *biolog* als komplexe gebundene Derivationsstammform zu einem Lexem *biolog*$^{\mathbb{P}}{}_{Konfix}$.

Während sich über die Zuordnung zu (hypothetischen) Lexemen sicherlich streiten lässt, bietet das Konzept der morphologisch komplexen freien oder gebundenen Stammformen einigen Spielraum für eine gleichartige Behandlung gleichartiger Phänomene, die in vorhandenen Morphologiekomponenten oft keine adäquate Behandlung erfahren.[20]

4.3 Nicht-konkatenativ ablaufende morphologische Prozesse (IP)

Prozesse, die sich nicht durch Aneinanderfügen von definierten Einheiten erklären lassen, sind in der Morphologie häufig anzutreffen: Es handelt sich um Veränderung von Stämmen durch z.B. Umlautung, Ablautung oder Tilgung. Sofern sich diese Prozesse innerhalb eines Flexionsparadigmas (vgl. Abschnitt 2.1.2) oder innerhalb eines Stammparadigmas (vgl. Abschnitt 4.2.2) abspielen, kann von ihnen abstrahiert werden, indem die veränderten Formen als Allomorphe angesehen werden. Anders verhält es sich hingegen, wenn ein Prozess lexemübergreifend stattfindet, wie dies beim **Wortartwechsel ohne Affigierung** der Fall ist. Hier wird eine Relation zwischen zwei Lexemen hergestellt, die nicht als Konkatenation zu beschreiben ist, sondern nur als Prozess.

Lexem	Dsf	Affix	Fsf	Lexem	Wortbildung
greifen$^{\mathbb{P}}{}_{V}$	*greif*	*-bar*	*greifbar*	*greifbar*$^{\mathbb{P}}{}_{ADJ}$	Derivation
greifen$^{\mathbb{P}}{}_{V}$/*Griff*$^{\mathbb{P}}{}_{NN}$	*griff*	*-ig*	*griffig*	*griffig*$^{\mathbb{P}}{}_{ADJ}$	Derivation
greifen$^{\mathbb{P}}{}_{V}$	*Griff*	$\emptyset$	*Griff*	*Griff*$^{\mathbb{P}}{}_{NN}$	abstr. Nom.
segeln$^{\mathbb{P}}{}_{V}$	*Segel*	$\emptyset$	*Segel*	*Segel*$^{\mathbb{P}}{}_{NN}$	Konversion

Abbildung 4.3: Derivation, Konversion und abstrakte Nominalisierung

In Abbildung 4.3 sind zwei Derivationen, eine abstrakte Nominalisierung

[18]Dass *rechtskräft* als 'frei vorkommend' und *Dickhäut* als 'gebunden vorkommend' klassifiziert werden, ist der Unterschied, der allerdings in der Unterscheidung zwischen Wortbildung und Phrase begründet liegt.

[19]Zur möglichen Segmentierung in die Bestandteile *rechts* und *kräftig* vgl. Schuch (1990), S. 136.

[20]Wenn eine Phrase als Lexem auftreten darf, kann man mit den hier beschriebenen Mitteln eine Wortbildung wie *Freiluftbühne* als Substantiv-Kompositum mit morphologisch komplexer, gebundener Kompositionsstammform, die auf ein frei vorkommendes Lexem zurückgeht, ansehen.

sowie eine Konversion dargestellt. Die Unterschiede liegen (in dieser Darstellung) darin, dass in den unteren beiden Zeilen kein Affix vorhanden ist (∅).[21] Obwohl in der Darstellung suggeriert wird, dass Konversion und abstrakte Nominalisierung sich sehr wohl konkatenativ darstellen lassen, gelingt dies nur unter Zuhilfenahme eines Null-Affixes (∅).

In den beiden folgenden Abschnitten werden Konversion und abstrakte Nominalisierung beschrieben.

4.3.1 Wortartwechsel ohne Stammveränderung

In dieser Arbeit wird unter **Konversion** der Wechsel der **Grundstammform** in eine andere Wortart verstanden.[22] Dieser Prozess ist im Deutschen sehr vielfältig: Eine Übersicht verschiedener Basen und ihrer Konversionsprodukte lässt sich Fleischer und Barz (1995), S. 50, entnehmen. Neben dem Auftreten einer leeren Einheit, dem Null-Affix, das in der maschinellen Verarbeitung sehr problematisch ist,[23] gibt es bei der IA-Darstellung von Konversion, wie sie in Abbildung 4.3 angedeutet ist, das Problem der **Ableitungsrichtung**: Ob *segeln*$^{\mathbb{P}}$ aus *Segel*$^{\mathbb{P}}$ abgeleitet ist oder *Segel*$^{\mathbb{P}}$ aus *segeln*$^{\mathbb{P}}$, kann nur unter Berücksichtigung der Sprachgeschichte eindeutig festgestellt werden. Aus diesen beiden Gründen wird Konversion in dieser Arbeit als ein Phänomen angesehen, das sich nur im IP-Modell adäquat beschreiben lässt: als (richtungsloser) Wortartwechsel ohne Stammveränderung.

4.3.2 Wortartwechsel mit Stammveränderung

Für die **abstrakte Nominalisierung** besteht das Problem der Ableitungsrichtung nicht, denn es handelt sich ausnahmslos um den Wechsel von einem starken Verb zu einem Substantiv.[24] Dafür muss die in Abbildung 4.3, Zeile 3, dargestellte Relation zwischen dem Lexem des zugrundeliegenden Verbs und der Derivationsstammform erklärt werden. Wenn diese Form der 'Ablautung' als nicht verschieden von den Prozessen von Umlautung und Fugung, wie sie in

[21] In der Literatur wird Konversion daher oft als *implizite Derivation* im Gegensatz zu *expliziter Derivation* bezeichnet.

[22] In der Literatur werden oft Phänomene der **Transposition** mit der Konversion vermischt. Dabei handelt es sich aber um den regelmäßig ablaufenden Wortartwechsel flektierter Wortformen, bei dem keine neuen Lexeme erzeugt werden, sondern rein syntaktisch das grammatische Verwendungspotential einer Einheit in einem Kontext erweitert wird (*laufen*$^{\mathbb{P}}$ → *(das) Laufen*).

[23] Es erhöht sehr stark das Auftreten von Mehrdeutigkeiten, da dann jedes Lexem potentiell in jeder Wortart auftreten kann.

[24] Dass es sich nicht um Transposition im Sinne von Fußnote 22 handelt, zeigen Formen wie *Gang* zu *gehen*$^{\mathbb{P}}$, die (im heutigen Deutsch) keine Flexionsformen des zugrundeliegenden Verbs (mehr) darstellen. Das Muster ist nicht mehr produktiv, aber Produktivität ist eine Voraussetzung für Transposition.

Derivationsstammformen auftreten, angesehen wird, bleibt bei abstrakten Nominalisierungen allein das Problem des Null-Affixes, das seine Darstellung im IP-Modell begründet: als (gerichteter) Wortartwechsel mit Stammveränderung.

4.4 Übersicht über Stammformtypen

In diesem Kapitel wurde gezeigt, dass mit dem Konzept von Stammformen und Affixen eine adäquate Beschreibung der morphologischen Phänomene der Wortbildung möglich ist. Dies wurde anhand einiger Problembereiche der morphologischen Beschreibung demonstriert. Weil Stammformen ein von morphologischer Komplexität unabhängiges Konzept darstellen, ist eine große Flexibilität in der Behandlung von Wortbildungsbestandteilen möglich.

Stammformtyp	Stammform	Lexem	Wort(form)bildung
Flexionsstamm	*Glas*	*Glas*$^{\mathbb{P}}{}_{NN}$	*(des) Glases*
	Äpfel	*Apfel*$^{\mathbb{P}}{}_{NN}$	*(den) Äpfeln*
	Äpfelchen	*Äpfelchen*$^{\mathbb{P}}{}_{NN}$	*(des) Äpfelchens*
	-ung	*-ung*$^{\mathbb{P}}{}_{NN}$	*(die Darstell)ungen*
Derivationsstamm	*Glas*	*Glas*$^{\mathbb{P}}{}_{NN}$	*glasig*
	bedeutsam	*bedeutsam*$^{\mathbb{P}}{}_{ADJ}$	*Bedeutsamkeit*
	gebetsmühlen	*Gebetsmühle*$^{\mathbb{P}}{}_{NN}$	*gebetsmühlenhaft*
	-bar	*-bar*$^{\mathbb{P}}{}_{ADJ}$	*(die Darstell)barkeit*
Kompositionsstamm	*Glas*	*Glas*$^{\mathbb{P}}{}_{NN}$	*Glastür*
	Hochhaus	*Hochhaus*$^{\mathbb{P}}{}_{NN}$	*Hochhaussiedlung*
	Wirksamkeits	*Wirksamkeit*$^{\mathbb{P}}{}_{NN}$	*Wirksamkeitsnachweis*
	-heits	*-heit*$^{\mathbb{P}}{}_{NN}$	*(das Einheits)beispiel*

Abbildung 4.4: Beispiele für Stammformen

In Abbildung 4.4 sind Beispiele für die einzelnen Stammform-Typen und Produkte ihrer Wort(formen)bildung angegeben. Einfache wie komplexe Formen können immer in allen drei Funktionen auftreten. Dabei erscheint das Auftreten einzelner Affixe in Stammformfunktion zunächst ungewöhnlich. Dies ist es auch, wenn man ein Modell von Wortbildungen in unmittelbaren Zerlegungen zugrundelegt, denn dann kann ein Affix nie allein Derivations- oder Kompositionsstammform sein, sondern immer nur in Verbindung mit einer Basis auftreten: *Bedeutsam·keit*. Aus der Sichtweise eines Lexikons hingegen, in dem auch gebundene Morpheme als Lexeme eingetragen werden, ist es sinnvoll, diesen auch die volle Stammformenfunktionalität zuzubilligen: Das Suffix *-heit* besitzt mit *-heits* eine Kompositionsstammform, die unabhängig von der Basis ist, an die *-heit* affigiert ist.

Kapitel 5

Vorhandene Lexikon-Systeme

In diesem Kapitel werden drei sehr heterogene Lexikonsysteme vorgestellt. Beim ersten handelt es sich lediglich um die Konzeption eines Lexikons, die im Rahmen eines Projektes zur Derivations- und Kompositionsmorphologie entstanden ist. Bei dieser handelt es sich allerdings um das Lexikonmodell, dessen Realisierung in dieser Arbeit beschrieben wird. Die Beschreibung des DeKo-Lexikonmodells erfolgt in Abschnitt 5.1. Das zweite Lexikon, CELEX, wurde bereits in Abschnitt 3.1.1 erwähnt, da es als Vollformensystem eine Art Zwischenstellung zwischen einer reinen Ressource und einem 'Morphologiesystem' einnimmt. An dieser Stelle erfolgt nun eine ausführliche Beschreibung von Inhalt und Struktur von CELEX (vgl. Abschnitt 5.2). Beim dritten Lexikonsystem, CISLEX, handelt es sich um ein Lexikon, bei dem der Aspekt der Abdeckung sehr wichtig ist. Es wird in Abschnitt 5.3 beschrieben.

5.1 DeKo

Bei *DeKo* handelte es sich um ein vom Land Baden-Württemberg gefördertes Projekt zur Derivations- und Kompositionsmorphologie mit eineinhalbjähriger Laufzeit, das in Schmid et al. (2001) vorgestellt wurde. Die Projektziele waren die "Beschreibung und Modellierung von Prozessen der deutschen Wortbildung", die "Erstellung eines robusten Systems zur Analyse und strukturellen Beschreibung komplexer Wörter" sowie die "Einbindung der Analyse komplexer Wörter in die deutsche Version des Text-to-Speech-Systems (TTS-Systems) FESTIVAL" (vgl. Heid (2001), S. 3). Aufgrund der bereits vorhandenen Flexionsmorphologie-Komponente DMOR (vgl. Abschnitt 3.2.1) war die Berücksichtigung von Flexion nicht nötig.

Zur Erreichung der Ziele wurde ein Lexikon konzipiert, das die Lemmata und Flexionsklassen aus DMOR inkorporierte, aber zusätzlich die erhebliche Anreicherung der Einträge durch phonetische, morphologische, syntaktische und semantische Informationen vorsah. Zur Behandlung der Derivation wur-

den Derivationsaffixe zum Lexikon hinzugefügt und mit ihren Eigenschaften umfassend beschrieben. Dazu wurden Tabellen angelegt, in denen Merkmale zu einigen hundert Affixen, zu den Wortbildungsprodukten sowie zu den von den Affixen selegierten Basen aufgelistet sind (vgl. z.B. die Tabellen in Schmid et al. (2001), S. 5f.). Besonderer Wert wurde dabei auf die Produktivität eines Affixes und auf die verschiedenen reihenbildenden Wortbildungsmuster gelegt. Das DMOR-Konzept der Koppelung von Kompositionsfugen an Flexionsklassen wurde zugunsten der Auflistung von Derivations- und Kompositionsstammformen (vgl. Fuhrhop (1998)) zu den Lexemen aufgegeben.

Mit der expliziten Auflistung von Derivationsaffixen und Stammformen war es erstmals möglich, eine Morphologiekomponente (den *DeKo-Automaten*) zu erzeugen, die die hierarchische Struktur von Derivationen und Kompositionen ermittelte. Es wurde eine kontextfreie Grammatik verwendet, um die Wortbildungsregeln zu kodieren und damit die Zuweisung einer Struktur an morphologisch komplexe Wortformen zu ermöglichen. Stämme und Regeln werden mit Hilfe der *AT&T Finite State Tools* (vgl. Sproat (2000)) in einen gewichteten endlichen Automaten kompiliert.[1] Eine Übersicht über die Architektur des Gesamtsystems wird in Schmid et al. (2001), S. 7, gegeben.

Die Erkennungsrate des DeKo-Automaten hängt entscheidend von der vollständigen Erfassung aller Derivations- und Kompositionsstammformen sowie der Merkmale, die die Selektion der Basen beeinflussen, ab. Die Methodik bei der Akquirierung von Stammformen ist in Heid et al. (2002) dargelegt. Die einzelnen Merkmale lexikalischer Einheiten werden im folgenden Abschnitt beschrieben. Bis zur vollständigen Erfassung der Stammformen müssen Zwei-Ebenen-Regeln angewendet werden, um Tilgungs- und Fugungsphänomene zu modellieren.

5.1.1 Eigenschaften lexikalischer Einheiten in DeKo

Im DeKo-Lexikonkonzept (vgl. Lüdeling et al. (2000)) werden sog. **lexikalische Einheiten** über mehrere Dimensionen spezifiziert.

form	morph_status	selegiert	Beispiele	Trad. Terminologie
`simplex`	`frei`	`nein`	*Haus*, *Baum*, *Auto*	**Stamm**
`simplex`	`gebunden`	`nein`	*elektr-*, *vibr-*, *ident-*	**gelehrter Stamm**
`simplex`	`frei`	`ja`	*·frei*, *·reich*	**Affixoid**
`simplex`	`gebunden`	`ja`	*-sam*, *-abel*	**Affix**

Abbildung 5.1: Eigenschaften der Simplizia im DeKo-Lexikonmodell

[1] Auf die Behandlung von Akzentuierung und Syllabifizierung, die ebenfalls stattfindet, wird hier nicht eingegangen.

Auf der Ebene der **morphologischen Form** werden morphologisch einfache und morphologisch komplexe Einheiten unterschieden, auf der Ebene des **morphologischen Status** werden frei und gebunden vorkommende Einheiten unterschieden, und hinsichtlich ihrer Fähigkeit zur Selektion schließlich werden **selegierende** und **nicht-selegierende** Einheiten unterschieden. Abbildung 5.1 veranschaulicht die Zusammenhänge für die Simplizia.[2] Neben der Tatsache, dass nun auch gebundene selegierende Elemente im Lexikon verzeichnet werden (*vibr-* im Beispiel für die Derivationen *vibr·ieren, Vibr·ation*), führt die feinere Differenzierung dazu, dass zwischen dem Adjektiv *frei*$^{\mathbb{P}}$ und dem Affixoid *·frei* im Lexikon unterschieden werden kann, da das Affixoid Basen selegiert, das Adjektiv hingegen nicht. Was den Status morphologisch komplexer Einheiten angeht, ist hingegen mit dieser Differenzierung noch nicht viel gewonnen: Derivationen und Kompositionen haben sicherlich die Belegungen `form:komplex`, `morph_status:frei` und `selegiert:nein`, aber ob *inform-* in der Wortbildung *inform·ieren* ein Simplex- oder ein Komplexstamm (*?in·form-*) ist, hängt immer noch von der zugrundegelegten Theorie ab. Dasselbe gilt für Phrasen in der Wortbildung: Dass der Derivationsstamm *Drittkläss* in *Drittkläss·ler* von komplexer Form ist, steht außer Frage, aber welchem Lemma oder welchen Lemmata er zugeordnet wird, muss dennoch unanbhängig von den Merkmalen entschieden werden. Dagegen gibt es für *abstrakte Nominalisierungen* ein Merkmal/Wert-Paar (`form:komplex_abstrakt`). Komplexe Formen, die morphologisch einem Derivationsmuster zu entsprechen scheinen, dies aber semantisch nicht tun, erhalten die Belegung `form:komplex_semi` (*Abstecher* im Sinne von *Umweg* ist keine Nominalisierung zu *abstechen*$^{\mathbb{P}}$). Es ist vorgesehen, zu jeder morphologisch komplexen Form die **Struktur** mit abzulegen, so dass im Lexikon gespeicherte Wortbildungen nicht mehr von einer Morphologiekomponente analysiert werden müssen.

Neben der morphologischen Form, dem morphologischen Status und der Fähigkeit zu Selektion werden einer lexikalischen Einheit im DeKo-Lexikonmodell weitere Merkmale zugeschrieben. Das wichtigste ist die **Kategorie**, die im Wesentlichen mit der Wortart übereinstimmt (da einige Einheiten keine Wortart haben, wurde die neutralere Bezeichnung gewählt). Für Präfixe und gebundene nicht-selegierende Einheiten, die nach traditioneller Sichtweise im Deutschen nicht über eine eigene Wortart verfügen[3], wird ein Platzhalter wie `Praefix` oder `Konfix` eingetragen. Daneben gibt es die **Herkunft**, die aus synchroner Sichtweise markiert, ob eine lexikalische Einheit eher als einheimisch (nativ), (neo)klassisch oder fremd empfunden wird:

> "Native Wörter sind (völlig unabhängig von ihrer Herkunft) solche Wörter, die den generellen grammatischen Regularitäten des Deut-

[2] Merkmal/Wert-Paare werden im Folgenden mit Doppelpunkt (`form:simplex`) notiert.

[3] Vgl. zur Wortart neoklassischer Elemente aber die Ausführungen in Lüdeling et al. (2002), S. 20ff.

> schen entsprechen; nichtnative Wörter sind (wiederum ungeachtet ihrer Herkunft) solche Wörter, die diesen Regularitäten nicht entsprechen." (Heidolph et al. (1981), S. 909, zitiert nach Fuhrhop (1998), S. 96)

Dieses Merkmal wird von Derivationsaffixen verwendet, die danach ihre Basen auswählen (Suffix *-abel* bspw. geht nur an neoklassische Einheiten wie *akzept-*). Ein weiteres Merkmal betrifft die **Lexikalisiertheit**. Dieses erlaubt die Differenzierung der rein morphologisch betrachteten Komplexheit lexikalischer Einheiten nach semantischen Kriterien: Während sich die Bedeutung von *Glas=Tür* rein kompositionell aus den Bedeutungen von *Glas* und *Tür* ergibt, gilt dies für *Bahn·hof* oder *Augen·blick* nicht ohne weiteres. *Bahnhof*$^{\mathbb{P}}$ kann nun als morphologisch komplexe, aber lexikalisierte Form angesehen werden. *Glastür*$^{\mathbb{P}}$ kann als morphologisch komplexe, aber **nicht** lexikalisierte Form angesehen werden.Im DMOR-Lexikon konnte nicht so stark differenziert werden: Jede eingetragene Form wurde automatisch als morphologisch einfach und damit lexikalisiert angesehen, morphologisch komplexe Formen waren nicht vorgesehen.

Weitere Merkmale, die eine lexikalische Einheit haben kann und die die Selektion durch Derivationsaffixe beeinflusst, sind syntaktischer, phonologischer und semantischer Art. Die **Argumentstruktur** bei Verben ist oft ausschlaggebend dafür, ob ein Verb als Basis für eine deverbale Ableitung dienen kann (Suffix *-bar* gewöhnlich nur an transitive Verben, also nicht **schlafbar*). Einen auf Schwa endenden Derivationsstamm selegiert das Suffix *-ei*, während die Allomorphe *-erei* oder *-elei* dies nicht tun. Ein semantisches Muster, das das Suffix *-lich* darstellt, ist die mehrfache Wiederholung nach einer Zeitspanne (*täg·lich, stünd·lich, minüt·lich*)[4]. Hier muss die substantivische Basis eine Zeitspanne ausdrücken.

5.1.2 Das DeKo-Lexikonmodell

Die Konzeption des DeKo-Lexikons sieht die Speicherung von Informationen zur Flexion, Wortbildung, Phonetik, Syntax, Semantik und Korpusfrequenz zu einer lexikalischen Einheit vor. Eine lexikalische Einheit in DeKo wird durch eine **Zitierform** repräsentiert. Bei der Zitierform handelt es sich i.A. um dieselbe orthographische Form wie das Lemma oder das Affix. Die Zitierform dient allein der leichteren Identifikation eines Eintrags durch den Benutzer, sie wird nicht für die komputationelle Verarbeitung benötigt. Sie kann in verschiedenen Stammformen realisiert sein. Dies ermöglicht z.B. die Zusammenfassung von orthographischen Varianten wie *Veredlung* und *Veredelung* in einer lexikalischen Einheit. Jede Stammform erhält eine Flexionsklasse. Diese beiden Bestandteile sind quasi 'rückwärtskompatibel' zum DMOR-Lexikon. Bei den Wortarten,

[4]Alle Beispiele in diesem Absatz sind Lüdeling und Fitschen (2002), S. 2, entnommen.

die an Wortbildung teilnehmen, sind die Derivations- und Kompositionsstämme aufgelistet. Zu jedem Lexem ist die **phonetische Transkription** in *SAMPA*-Notation (phonetisches Alphabet aus ASCII-Zeichen, vgl. SAMPA (1989)) mitsamt der Silbenbetonung und der Anzahl der Sprechsilben angegeben.

An syntaktischen Informationen ist die Speicherung von Subkategorisierungsrahmen für Verben, Adjektive und Substantive vorgesehen. Verschiedene Arten semantischer Information sind vorgesehen, aber noch nicht weiter spezifiziert worden. Das Feld **Semantischer Typ** wird bislang als einziges verwendet. Hier werden Eigennamen unterschieden nach Vor-, Nach- und Städtenamen, bei Substantiven kann die Unterscheidung von Appellativa (*count nouns*) und Kontinuativa (*mass nouns*) sowie Titeln usw. vorgenommen werden.

Schließlich wird die Gesamtvorkommenshäufigkeit der Formen des Lexems im HGC verzeichnet. Es handelt sich allerdings nicht um die tatsächliche Lemmafrequenz, sondern um die Summe der **Tokenfrequenzen** aller in ihrer Oberflächenform verschiedenen Wortformen aus dem jeweiligen Paradigma.[5] Bei gebundenen lexikalischen Einheiten ist die Feststellung der Frequenz schwierig bis unmöglich, da hier zunächst die Art der Wortbildung feststehen müsste, bevor die Affixfrequenz gezählt werden kann (*Schwung* darf nicht als Affix *-ung* gezählt werden).

5.1.3 Diskussion

Mit der in DeKo vorgenommenen detaillierten Beschreibung von Derivationsaffixen und Kompositionsmustern wird die linguistisch adäquate maschinelle Behandlung von Derivation und Komposition ermöglicht. Bislang lagen derartige Beschreibungen für das Deutsche hauptsächlich in gedruckter Form und für einheimische Affixe vor (vgl. z.B. die Reihe zur Deutschen Wortbildung, Kühnhold und Wellmann (1973), Wellmann (1975), Kühnhold et al. (1978), Ortner et al. (1991), Pümpel-Mader et al. (1992), oder Fleischer und Barz (1995)). Die Kombination von fundierter linguistischer Beschreibung, Lexikonkonzeption und Realisierung in einem endlichen Automaten ist ein großer Schritt in Richtung maschinelle Analyse von Wortbildungsphänomenen, die in deutschen Textkorpora vorkommen.

Auf der anderen Seite erfordert die Umsetzung des Modells einen sehr großen Aufwand bei der Vergabe der Merkmalwerte für mehr als 50 000 Lexeme im (DMOR-)Lexikon. Obwohl das Projekt einige Jahre zurückliegt, konnten dennoch bislang längst nicht alle Informationen erhoben werden. Der Vorteil allerdings, ein wohldefiniertes Lexikonkonzept zu haben, so dass Phänomene

[5] Für *Schuster*$^{\mathbb{P}}{}_{NN}$ ist die Vorkommenshäufigkeit die Tokenfrequenz der Formen *Schuster, Schusters, Schustern* im HGC. Vorkommen des homonymen Eigennamens (*Schuster*$^{\mathbb{P}}{}_{NE}$) werden also mitgezählt, d.h., die angegebene Vorkommenshäufigkeit liegt häufig höher als der tatsächlich zu erwartende Wert.

linguistisch adäquat behandelt werden können, wiegt diesen Nachteil deutlich auf: Erst mit einem fundierten Lexikonmodell ist eine umfangreiche strukturierte und konsistente Erweiterung eines Lexikons zu erreichen. Ohne ein solches Modell wären weder die Wartbarkeit noch eine hohe Qualität der Ressource möglich.

5.2 CELEX

Die *CELEX Lexical Database* (vgl. Baayen et al. (1995); CELEX (1995a)) ist eine Sammlung von Dateien, die für etwa 365 000 Wortformen des Deutschen (das entspricht 51 000 Lemmata) Informationen zu Orthographie, Phonologie, Morphologie, Syntax sowie Vorkommenshäufigkeit in mehreren Korpora enthalten. Zu jedem der fünf Bereiche existiert eine ausführliche Dokumentation (vgl. Gulikers et al. (1995)). Die der Ressource zugrundeliegenden Korpora stammen vom Institut für Kommunikationsforschung und Phonetik (IKP) in Bonn und vom Institut für deutsche Sprache (IDS) in Mannheim.

5.2.1 Die Struktur der Ressource

Die Anordnung der Daten in den CELEX-Dateien geschieht zeilenweise: Einzelinformationen zu jeweils einem Lemma oder einer Wortform sind je Zeile durch einen Schrägstrich voneinander getrennt. Die Lemmata sind durchlaufend von 1 (*A*) bis 51 728 (*Zytostom*) numeriert. Diese Nummer verweist bei jedem Wortformeintrag auf das dazugehörige Lemma. Da die Ressource ursprünglich in Form einer relationalen Datenbank vorlag, sind die Dateien wie Abbildungen von Datenbanktabellen in Textdateien vorstellbar und die laufende Nummer wie eine eindeutige Identifikations-Nummer (ID).

In den folgenden Tabellen ist für die drei Lexeme *Haus*$^{\mathbb{P}}$, *Häuschen*$^{\mathbb{P}}$ und *Häuserblock*$^{\mathbb{P}}$ dargestellt, welche und wie die zugehörige linguistische Information in CELEX repräsentiert wird. Zur Illustration spezieller Merkmalbelegungen werden teilweise weitere Lexeme oder Wortformen hinzugenommen.

```
80\abdecken\19\ab-dek-ken\Y\abdeck\ab-deck\Y
7285\Bettuch\7\Bett-tuch\Y\Bettuch\Bett-tuch\Y
16406\Haus\2000\Haus\N\Haus\Haus\N
16412\H"auschen\29\H"aus-chen\N\H"auschen\H"aus-chen\N
16413\H"auserblock\2\H"au-ser-block\N\H"auserblock\H"au-ser-block\N
```

Abbildung 5.2: CELEX. Deutsche Orthographie, Lemma

Abbildung 5.2 zeigt fünf Orthographie-Einträge (aus der Datei `gol`, *German Orthography, Lemmas*). Für *Haus* ist `16406` die laufende Nummer, die in allen

weiteren Einträgen, die auf das Lexem *Haus*ℙ verweisen, verwendet wird. 2000 ist die Vorkommenshäufigkeit des Lemmas, also die aufaddierte Korpusfrequenz der Formen *Haus, Hauses, Hause, Häuser* und *Häusern*. In der nächsten Spalte sind die Silbengrenzen im Lemma in Form von Bindestrichen angegeben (*ä* wird in CELEX als "*a* bzw. teilweise als *ae* kodiert). Wie an den Beispielen `Bett-tuch` und `ab-dek-ken` zu erkennen ist, kodieren die Einträge zugleich die deutschen Worttrennungsregeln: *ck* wird *k-k* getrennt und ein in der Wortform an einer Morphemgrenze getilgter Konsonant tritt bei der Trennung wieder auf[6]. Die Änderung an der Wortform wird durch das `Y` (für 'yes') in der nächsten Spalte gekennzeichnet. Die letzten drei Spalten wiederholen die Information für den Stamm des Wortes. Im Falle des Verbs *abdecken*ℙ ist dies `abdeck`, wieder mit Trennung und Kennzeichnung der Formveränderung (bei der Belegung `Y` am Ende der ersten Zeile handelt es sich offenbar um einen Fehler).

```
296008\Haus\620\16406\Haus\N
296009\Hause\840\16406\Hau-se\N
296010\H"auser\213\16406\H"au-ser\N
296011\H"ausern\95\16406\H"au-sern\N
296012\Hauses\232\16406\Hau-ses\N
```

Abbildung 5.3: CELEX. Deutsche Orthographie, Wortform

Abbildung 5.3 zeigt die fünf möglichen Wortformen für das Lexem *Haus*ℙ (aus der Datei `gow`, *German Orthography, Wordforms*). Auch die Wortformen werden durchnumeriert (Spalte 1). Nach der Wortform folgt deren Vorkommenshäufigkeit im Korpus (Spalte 3). Die fünf Einzelhäufigkeiten zusammengezählt ergeben wieder die Gesamthäufigkeit 2000 für das Lexem *Haus*ℙ. In Spalte 4 steht die laufende Nummer oder ID des der Wortform zugrundeliegenden Lexems, 16406. Die letzten beiden Spalten geben wieder die Silbentrennung und etwaige dabei auftretenden Formveränderungen an.

```
16406\Haus\2000\619\336\2.5263\1829\341\2.5328\171\291\2.4639
```

Abbildung 5.4: CELEX. Korpusfrequenz, Lemma

Abbildung 5.4 zeigt verschiedene Korpusfrequenzen des Lexems *Haus*ℙ (aus der Datei `gfl`, *German Frequency, Lemmas*). Nach ID, Lemma und Lemmafrequenz (Spalten 1-3) folgt in Spalte 4 die berechnete mögliche Abweichung der Frequenz bei mehrfach vorkommenden identischen orthographischen Formen. Die CELEX-Dokumentation gibt als Beispiel die Wortform *nahe* an, die als Adjektiv, Präposition und Verb auftreten kann (vgl. Gulikers et al. (1995), S. 5-105).

[6]Nach der Rechtschreibreform, die am 1.8.1998 beschlossen wurde, gibt es die Konsonantenreduktionsregel nicht mehr, man schreibt *Betttuch*.

Die Korpusfrequenz von *nahe* ist 403, aber dieser Wert besagt nicht, in welcher Wortart die Wortform wie oft auftritt. Daher wird eine Stichprobe von Belegen durchgesehen, auf die Gesamtfrequenz hochgerechnet und mit einer Formel die mögliche Abweichung ermittelt. Im Falle von *nahe* werden die Frequenzen 250 für die Präposition-Lesart, 153 für die Adjektiv-Lesart und 0 für die Verb-Lesart errechnet, bei einer möglichen Abweichung von 33. Dies besagt für die Präposition: "This means that the true frequency for this form of *nahe* is almost certain–at least 95% certain–to lie between 120 and 186." (Gulikers et al. (1995), S. 5-105) Der nächste Wert, `336`, ist die normalisierte Frequenz auf eine Million Token. Wenn man die HGC-Frequenz für das Lexem *Haus*$^{\mathbb{P}}$ (90 207) auf diese Weise normalisiert, erhält man 442, also eine durchaus vergleichbare Anzahl. Die nächste Spalte enthält den (Zehner-)Logarithmus der normalisierten Frequenz. Die letzten sechs Spalten schließlich enthalten jeweils die Frequenz auf dem geschriebenen und dem gesprochenen Teil des Korpus mitsamt Normalisierung und Logarithmus. Im geschrieben Teil kommt das Lexem *Haus*$^{\mathbb{P}}$ 1 829 Mal vor, im gesprochenen 171 Mal, zusammen also wieder 2 000 Mal.

`296008\Haus\16406\620\61982\104`	46.453	227
`296009\Hause\16406\840\1494\141`	19.614	96
`296010\Haeuser\16406\213\135\36`	11.417	56
`296011\Haeusern\16406\95\7723\16`	4.412	22
`296012\Hauses\16406\232\10\39`	8.311	41

Abbildung 5.5: CELEX. Korpusfrequenz, Wortform (HGC zum Vergleich)

Abbildung 5.5 zeigt die verschiedenen Korpusfrequenzen der Wortformen des Lexems *Haus*$^{\mathbb{P}}$ (aus der Datei `gfw`, *German Frequency, Wordforms*). Aus Gründen der Übersichtlichkeit wurden die hinteren sieben Spalten, also die Differenzierung nach geschriebener oder gesprochener Sprache sowie der Logarithmus der normalisierten Frequenzen, weggelassen. Zum Vergleich sind die Einzelfrequenzen aus HGC (absolut und normalisiert auf eine Million Token) in je einer eigenen Spalte mit angegeben. Die Übereinstimmung bei den letzten drei angegebenen Wortformen ist sehr hoch (*Hauses* normalisiert 39 Mal in den CELEX-Korpora und 41 Mal im HGC).

Abbildung 5.6 zeigt die Kodierung der Morphologie der Lexeme in CELEX (aus der Datei `gml`, *German Morphology, Lemmas*). Nach der ID, dem Lemma und der Korpusfrequenz (Spalten 1-3) folgt der **morphologische Status**. CELEX unterscheidet sechs Varianten, drei für Lexeme mit morphologischer Analyse und drei für Lexeme ohne morphologische Analyse. Morphologisch analysierte Lexeme können **morphologisch komplex** (*Häuschen, Häuserblock*, Markierung `C`), **monomorphemisch** (*Aal, Haus*, Markierung `M`) oder **Konversionen** (*aalen, Absicht*, Markierung `Z` für "zero derivation") sein. Die anderen drei Möglichkeiten sind **morphologisch irrelevant** (*Photograph, Privileg*, Markierung `I`),

```
3\aalen\1\Z\1\Y\Y\Y\Aal\N\N\N\N\((Aal)[N])[V]\N\N\N\N\r1\N
4\Aal\80\M\1\Y\Y\Y\Aal\N\N\N\N\(Aal)[N]\N\N\N\N\S1/P1\N
684\Absicht\427\Z\1\Y\Y\Y\abseh\V\
    Y\N\N\(((ab)[V|.V],(seh)[V])[V])[N]\Y\N\N\N\S3/P3\N
6406\Haus\2000\M\1\Y\Y\Y\Haus\N\N\N\N\(Haus)[N]\N\N\N\N\S1/P4u\N
8767\Photograph\10\I\0\Y\Y\Y\\\\\\\\\\N\S2/P3\N
9649\Privileg\39\I\0\Y\Y\Y\\\\\\\\\\N\S1/P10\N
16412\Haeuschen\29\C\1\Y\Y\Y\Haus+chen\Nx\
      N\N\Y\((Haus)[N],(chen)[N|N.])[N]\N\N\Y\N\S1/P2\N
16413\Haeuserblock\2\C\1\Y\Y\Y\Haus+er+Block\NxN\
      N\N\Y\((Haus)[N],(er)[N|N.N],(Block)[N])[N]\N\N\Y\N\S1/P5\N
```

Abbildung 5.6: CELEX. Deutsche Morphologie, Lemma

morphologisch unbestimmt (*Adamit*, Markierung `U`) oder **lexikalisierte Flexion** (*anhaltend*, Markierung `F`; letzte zwei Beispiele aus Gulikers et al. (1995), S. 5-54). Derivation wird in CELEX also als morphologisch komplex markiert. Unter "morphologisch irrelevant" fallen neben den beiden neoklassischen Lexemen Kompositionen mit einem Eigennamenbestandteil (*Achensee*), Phrasen in der Wortbildung (*Aufundabgehen*) und Interjektionen (*ach*; vgl. Gulikers et al. (1995), S. 5-55). 'Morphologisch unbestimmt' ist alles, was nicht anderweitig eingeordnet werden kann (*Aerogramm, Rembours, Wirrwarr* sind Beispiele aus Gulikers et al. (1995), S. 5-55).

Die nächste Spalte enthält die Zahl der morphologischen Analysen. In den allermeisten Fällen ist dies genau eine Analyse, und in 764 Fällen gibt es zwei Analysen. Die mehrdeutigen Fälle betreffen Kompositionen, die auch als Derivationen mit komplexer Basis verstanden werden können: *Tellersammlung* als `Teller+Sammlung` oder `Teller+sammel+ung`. In Gulikers et al. (1995) wird die zweite Variante als *abgeleitetes Kompositum* bezeichnet. Die drei Spalten nach der Zahl der morphologischen Analysen geben an, ob es sich um ein solches abgeleitetes Kompositum handelt, ob es sich um ein nicht mehr abgeleitetes Kompositum handelt oder ob es sich um etwas Anderes handelt (bei eindeutiger Analyse sind alle drei auf `Y` ('yes') gesetzt). Eine weitere Mehrdeutigkeit tritt auf, wenn ein Bestandteil einer Komposition eine Konversion ist. Ein Beispiel ist das Kompositum *Zuchttier*, das in CELEX sowohl auf `Zucht+Tier` als auch auf `zuecht+Tier` zurückgeführt wird. Da andererseits die Komposita *Platzmangel* auf `Platz+Mangel` und *Platzwunde* auf `platz+Wunde` zurückgeführt werden, ist davon auszugehen, dass bei *Zucht* die Wahrung der Mehrdeutigkeit beabsichtigt ist.

In der Zeile folgt nun die Zerlegung des Lexems in unmittelbare Konstituenten mitsamt dem Zerlegungsmuster in der Spalte dahinter. Hier steht das `N` für *Nomen*, das `V` für *Verb* und das `x` für *Affix* (Kompositionsfugen zählen in

CELEX zu den Affixen). Wenn sich Stammveränderungen ergeben, ist dies in der nächsten Spalte mit einem `Y` markiert. Im Beispiel betrifft dies das Lexem *Absicht*$^{\mathbb{P}}$, da es sich um eine Konversion handelt, die auf *absehen*$^{\mathbb{P}}$ zurückgeführt wird. Die beiden folgenden Spalten markieren Opaqueheit und Umlautung. Opaqueheit wird hier im Sinne von Idiomatisierung verwendet, markiert werden "words whose analysis is *opaque* – that is, words made up of morphemes which are recognizable, but where the meaning of the head element isn't reflected in the meaning of the full word. An example of this is *Angsthase*" (Gulikers et al. (1995), S. 5-65). Das Auftreten von Umlauten in der morphologisch komplexen Form, die in der Grundform nicht auftreten, ist im Beispiel bei *Häuschen* und *Häuserblock* gekennzeichnet.

Es folgt eine weitere Zerlegung des Lexems, diesmal mit der Struktur der Wortbildung und der Kennzeichnung der einzelnen Bestandteile. Im Falle der Konversion bei *aalen*$^{\mathbb{P}}$ beispielsweise wird durch `((Aal)[N])[V]` der Zusammenhang zum Substantiv *Aal*$^{\mathbb{P}}$ angezeigt. Der Darstellung lässt sich sogar entnehmen, welche Wortart Affixe selegieren: `(chen)[N|N.]` bedeutet, dass ein Substantiv als Wortbildungsprodukt entsteht, wenn *-chen* an ein Substantiv angehängt wird. Es folgen wieder drei Spalten für die Anzeige von Stammveränderung, Opaqueheit und Umlautung. Die drei letzten Spalten schließlich markieren die Trennbarkeit des Lexems (geben also im Deutschen quasi an, dass es sich um ein Partikelverb handelt), geben das Flexionsparadigma an (`S` steht für *Singular*, `P` für *Plural*, `r` für *reguläres Verb*, `u` markiert Umlautung bei Plural etc.).

```
296008\Haus\620\16406\nS,dS,aS
296009\Hause\840\16406\dS
296010\Haeuser\213\16406\nP,gP,aP
296011\Haeusern\95\16406\dP
296012\Hauses\232\16406\gS
```

Abbildung 5.7: CELEX. Deutsche Morphologie, Wortform

Abbildung 5.7 zeigt die Kodierung der morphosyntaktischen Kategorien für die Wortformen (aus der Datei `gmw`, *German Morphology, Wordforms*). Neben der laufenden Nummer, der Wortform, der Frequenz und der ID des Lexems *Haus*$^{\mathbb{P}}$ ist noch eine weitere Spalte angegeben, in der Kürzel für die morphologischen Kategorien aufgelistet sind. Die Wortform *Haus* kann also Nominativ Singular, Dativ Singular oder Akkusativ Singular sein, *Hauses* nur Genitiv Singular, *Häusern* Dativ Plural usw.

Abbildung 5.8 zeigt die Kodierung von syntaktischen Eigenschaften der Lexeme (aus der Datei `gsl`, *German Syntax, Lemmas*). Unter den Begriff *Syntax* fallen in CELEX diverse Phänomene: die Wortart, das Genus bei Substantiven, der semantische Typ bei Eigennamen, Verbklassen, Subkategorisierung von Ver-

```
16406\Haus\2000\1\3\\N\N\\\\\\\\\\\\\\\\
24048\massiv\107\2\\\N\N\\\\\\\\\\\\\PCS\\\
28415\persiflieren\1\4\\\N\N\1\5\00N000000;\I\I\O\I\I\I\I\I\I\\\\
```

Abbildung 5.8: CELEX. Deutsche Syntax, Lemma

ben, die Steigerbarkeit von Adjektiven und der Numeraltyp bzw. die semantische Unterteilung der Pronomina. Hier wird nicht auf alle Spalten, sondern nur auf die relevanten Phänomene eingegangen (weitere Informationen befinden sich in Gulikers et al. (1995), S. 5-85ff.). In dieser Datei ist die Wortart der Lexeme kodiert, und zwar als eine Zahl zwischen 1 und 10. In der Tabelle sind ein Substantiv (Kodierung 1), ein Adjektiv (2) und ein Verb (4) angegeben. Beim Genus bei Substantiven gibt es die Besonderheit, dass es auch mehrere Genera pro Lexem geben kann, z.B. maskulin und feminin bei *Sellerie*$^{\mathbb{P}}$ (vgl. Gulikers et al. (1995), S. 5-88). Eigennamen werden nach geographischen Namen, Personennamen und Firmen- oder Produktnamen gegliedert. Singulariatantum und Pluraliatantum sind jeweils in einer eigenen Spalte gekennzeichnet. Bei Verben wird angegeben, ob sie das Perfekt mit *haben* oder *sein* bilden. Sie werden nach Hilfsverben, Modalverben, Kopulaverben, unpersönlichen, reflexiven oder anderen Verben unterteilt. Bei der Subkategorisierung der Verben wird unterschieden, ob diese möglich, unmöglich oder fakultativ ist. Neun verschiedene Komplement-Positionen werden unterschieden, und jede kann in sieben verschiedenen Arten realisiert sein. Bei Adverbialen wird noch nach lokativer, temporaler etc. Verwendung differenziert. Adjektive werden nach ihrer Steigerbarkeit unterschieden in solche, die nicht steigerbar sind, solche, die nur Komparativ oder nur Superlativ bilden können, und schließlich solche, die voll steigerbar sind (wie *massiv* im Beispiel, gekennzeichnet durch die Markierung PCS). Numerale können als Kardinale, Ordinal, Bruchzahlen und sog. *Multiplikative* (Beispiel: *achtfach*) auftreten. Pronomina werden in Personalpronomen, Demonstrativpronomen usw. unterteilt. Als letztes wird für Präpositionen angegeben, welche(n) Kasus sie regieren.

```
16406\Haus\2000\'hBs\[haus]\'hBs\[haus]\[CVVC]\[CVVC]\haus\haus
14508\gehen\7302\'ge-@n\[ge:][@n]\'ge\[ge:]\[CVV][VC]\[CVV]\ge:\ge:
296008\Haus\620\16406\'hBs\[haus]\[CVVC]
296009\Hause\840\16406\'hB-z@\[hau][z@]\[CVV][CV]
296010\Haeuser\213\16406\'hX-z@r\[hOy][z@r]\[CVV][CVC]
296011\Haeusern\95\16406\'hX-z@rn\[hOy][z@rn]\[CVV][CVCC]
296012\Hauses\232\16406\'hB-z@s\[hau][z@s]\[CVV][CVC]
```

Abbildung 5.9: CELEX. Deutsche Phonologie, Lemma und Wortform

Der Vollständigkeit halber seien hier auch noch Beispiele für Einträge aus der Phonologie-Lemma- und der Phonologie-Wortform-Datei angegeben (Dateien `gpl/gpw`, *German Phonology, Lemmas/Wordforms*; vgl. Abbildung 5.9). Neben den bekannten ersten Spalten für ID, Lemma/Wortform und Korpusfrequenz gibt es verschiedene Formate der phonetischen Transkription (DISC und SAMPA), das Konsonanten- und Vokalmuster und eine Unterscheidung der Aussprache jeweils für das Lemma und den Stamm bzw. für die Wortform.

5.2.2 Bewertung

Bei den deutschen Daten der *CELEX Lexical Database* handelt es sich bezüglich der Menge der kodierten Phänomene der deutschen Morphologie, Syntax, Semantik und Phonetik und die Art ihrer Repräsentation um eine beachtliche Ressource. Durch die Methode, für ein Lemma oder eine Wortform relevante Information auf jeweils einer Zeile zu halten, lässt sich die Ressource leicht in andere Sprachverarbeitungskomponenten einbinden. Die Berücksichtigung der Wortbildungsphänomene, die Angabe der Wortbildungsmuster bei morphologisch komplexen Wortformen und die Angabe von flacher und tiefer Analyse hebt die Ressource von Morphologiesystemen ab, die nur Flexion behandeln. Durch die Auflistung aller Wortformen zu einem Lexem umgeht CELEX ein Problem aller lexembasierten Ressourcen oder Komponenten: Es ist in einem Lexemlexikon schwierig, Frequenz- oder Ausspracheinformationen für alle Wortformen abzulegen, ohne diese Wortformen doch noch auflisten zu müssen. Schließlich sind sowohl der Bezug auf ein aus verschiedenen Textsorten zusammengesetztes Korpus als auch die dort erzielte Abdeckung als positiv hervorzuheben: “When compared with the 6 million word corpus of the Institute for German Language at Mannheim, the coverage of CELEX lemmata is 83% of the totalcorpus.” (CELEX (1995b)) Die Wahrscheinlichkeit, mit den 51 000 Lemmata einen gewissen *Grundwortschatz* auch in anderen Texten abzudecken, ist nicht gering.

Als Nachteil kann man nennen, dass die Gleichbehandlung aller Wortarten die Fehleranfälligkeit der Ressource erheblich erhöht. Wenn bspw. in der Syntaxtabelle immer Spalten für Substantiv, Adjektiv und Verb vorgesehen sind, dann können Informationen leicht falsch eingeordnet werden. Aus der Informatik (insbesondere der Datenbanktechnik) bekannte Prinzipien wie Redundanzfreiheit (vgl. Vossen (1994), S. 20) und Vermeidung von Abhängigkeiten (vgl. ebd., S. 191ff.) werden hier verletzt.

5.3 CISLEX

Das CISLEX ist am Centrum für Informations- und Sprachverarbeitung (CIS) der Universität München entstanden. “Das Ziel des CISLEX Projekts ist die Erstellung eines weitgehend vollständigen elektronischen Wörterbuchs des Deut-

schen mit morphologischer, syntaktischer und semantischer Information." (CISLEX (o.J.)).

Das CISLEX ist die von den vorgestellten Ressourcen am besten dokumentierte. Drei Dissertationen allein beschäftigen sich mit den Themen der *automatischen Lemmatisierung* (vgl. Maier-Meyer (1995); hier findet eine Dokumentation des Bestandes an Wortarten in CISLEX statt), der *semantischen Klassifikation* der Substantive (Langer (1996); hier wird eine vollständige Ontologie entwickelt, in die alle Substantive aus CISLEX eingefügt werden) und der syntaktischen und semantischen Beschreibung der Verbklassen (Schnorbusch (1998)).

5.3.1 Aufbau und Inhalt des CISLEX

Das CISLEX ist entsprechend der Unterscheidung von vier Typen von Wortformen modular aufgebaut. Die vier Typen sind "[e]infache und komplexe Wortformen", "Eigennamen aus den verschiedensten Bereichen", "Fremd- und Fachwörter" sowie "Kurz- und Sonderformen" (Maier-Meyer (1995), S. 26). Die vier korrespondierenden Lexika sind "**das deutsche Kernlexikon**", "**das Namenslexikon**", "**das Fremd- und Fachwörterbuch**" und "**das Lexikon der Sonderformen**" (ebd.). Der Lexikonaufbau und die Lexikonerweiterung erfolgten mit Hilfe von Wortlisten und Korpora: "Auf der Basis von verfügbaren Wortlisten wurde ein Grundstock von Lemmata angelegt, der zum einen durch den Vergleich mit gängigen Wörterbüchern und zum anderen durch Korpusuntersuchungen ständig aktualisiert und erweitert wird." (Maier-Meyer (1995), S. 30)

Die vier Hauptlexika sind jeweils wieder in Teillexika unterteilt. So gibt es im Kernlexikon ein Lexikon der einfachen Formen, eines der erweiterten einfachen Formen, eines der komplexen Formen und eines schließlich der flektierten Formen. "Bei der Aufteilung in ein Lexikon der einfachen Formen und ein Lexikon der komplexen Formen geht es lediglich um eine möglichst effiziente und möglichst redundanzfreie Darstellung des ausgewählten Wortschatzes." (Maier-Meyer (1995), S. 31) Diese dient dann einer möglichst effizienten kaskadierten morphologischen Verarbeitung von Wortformen: Zur morphologischen Analyse kann zunächst geschaut werden, ob die Wortform bei den flektierten einfachen Formen zu finden ist (das entspricht dem Nachschauen in der Vollformenliste). Dann wird von rechts nach links versucht, eine flektierte einfache Form abzutrennen, auf ihre Grundform zurückzuführen und wiederum zu schauen, ob die komplexe Grundform im Lexikon der komplexen Formen vorhanden ist. Erst danach wird bei Misserfolg ein Zerlegungsalgorithmus angewandt.

Die Unterscheidung von einfachen und komplexen Formen entspricht nicht ganz der Aufteilung in Simplizia und Wortbildungen: "[D]ie häufigsten Suffixe [werden] als spezielle Kategorien" in das Lexikon der einfachen Formen aufgenommen (vgl. Maier-Meyer (1995), S. 32). Suffixbildungen, die nicht mit einem dieser Suffixe stattfinden, gelten also als einfache Formen. Präfigierun-

gen zählen zu den komplexen Formen. Da jedoch "die Präfigierungen einfacher Basen sich häufig anders verhalten als die Basis" (ebd.), werden einfache Formen und Präfigierungen noch zum Lexikon der erweiterten einfachen Formen zusammengefasst.

Die Definition der *einfachen Form* in CISLEX lautet wie folgt: "Ein Wort W ist eine **einfache Form** genau dann, wenn es keine sinnvolle Zerlegung W = W_1 W_2 gibt, so daß W_1 eine Folge von Morphemen ist und W_2 ein Wort mit denselben morphologischen Eigenschaften wie W." (Maier-Meyer (1995), S. 31) Da vorher bereits festgestellt wird, dass das Kernlexikon "in erster Linie ein morphologisch-orientiertes Lexikon des Deutschen sein soll" (ebd.), gehe ich davon aus, dass idiomatisierte oder lexikalisierte Komposita wie *Augenschein, Brombeere* und *Bahnhof* in CISLEX als komplexe Formen behandelt werden.

Die flektierenden Wortarten werden in CISLEX in Flexionsklassen eingeteilt. Aufgrund des Anspruchs der Vollständigkeit, also der Abdeckung großer Textkorpora, gibt es bspw. für Substantive eine große Anzahl Flexionsklassen: Maier-Meyer listet allein 101 Klassen für die Plural-Deklination der Nomen auf (vgl. Maier-Meyer (1995), S. 46ff.).[7]

5.3.2 Bewertung

Bei CISLEX handelt es sich um ein Lexikonsystem, das sehr pragmatisch orientiert ist: Das Ziel der vollständigen Abdeckung von Wortformen in Textkorpora lässt sich derzeit nur durch ein breit angelegtes Lexikon erreichen, nicht durch eines, das bei bestimmten Phänomenen wie der Wortbildung in die Tiefe geht. Die konsequente Benennung und Behandlung von Problemklassen, die außerhalb der Kernbereiche der Morphologie angesiedelt sind, ist zur Erreichung des Ziels unerlässlich. Es ist zu vermuten, dass CISLEX von allen verfügbaren kombinierten Lexikon-/Morphologiesystemen dieser Aufgabe am besten gerecht wird.

Ein zweiter sehr beachtenswerter Aspekt bei CISLEX ist die Kaskade der Verarbeitungsschritte: Durch die Aufteilung der morphologischen Analyse vom Nachschauen in Lexika für einen schnellen Zugriff bis hin zur Anwendung von Zerlegungsalgorithmen, wenn vorher keine Analyse gefunden wurde, vermeidet das System ein Problem nicht-kaskadierter Systeme: dass mehrdeutige Zerlegungen gefunden werden, obwohl die richtige Lösung bereits bekannt ist.

[7] Darunter finden sich Beispiele wie die Grundform *Targi*$^{\mathbb{P}}$, deren Pluralform sich durch Abschneiden der letzten vier Buchstaben und Anhängen der Zeichenkette *uareg* ergibt.

Kapitel 6

Konzeption des IMSLEX

IMSLEX basiert in seiner Konzeption weitgehend auf dem DeKo-Lexikonmodell, das in Abschnitt 5.1.2 vorgestellt wurde. In diesem Kapitel wird das Repräsentationsformat mitsamt einigen praktischen Überlegungen zur Strukturierung der Daten und zu generellen Prinzipien, die die Ressource erfüllen soll, vorgestellt (vgl. Abschnitt 6.1). Das Resultat der angestellten Überlegungen ist eine Dokumenttyp-Beschreibung des IMSLEX, also eine Umsetzung der Lexikon-Konzeption in eine Datenstruktur. Die Dokumenttyp-Beschreibung erfolgt in Abschnitt 6.2.

6.1 Vorüberlegungen

Es hat sich gezeigt, dass das DMOR-Lexikon, das ausschließlich für das Einlesen in einen endlichen Automaten gedacht ist, unflexibel hinsichtlich Erweiterungen ist: Es sind keine Felder vorgesehen, in die weitere, über die Flexion hinausgehende Informationen eingetragen werden könnten. Darüber hinaus stehen Stämme, die ein ganzes Paradigma vertreten, und solche, die nur einen Teil eines irregulären Paradigmas vertreten (z.B. Suppletivstämme), gleichberechtigt nebeneinander. Es kann nicht ohne weiteres ermittelt werden, wie viele und welche Lexeme überhaupt im DMOR-Lexikon vorhanden sind. Das DeKo-Lexikonkonzept geht auf diese Anforderungen ein, indem es die **lexikalische Einheit** als ein Grundkonstrukt ansieht, für das einige weitere Informationen vorgesehen sind. Das Ziel ist die Umsetzung des DeKo-Lexikonkonzepts in eine Ressource und die Verschmelzung dieser Ressource mit den im DMOR-Lexikon enthaltenen Daten.

Die vier Hauptanforderungen an die Struktur der zu erstellenden Ressource lauten wie folgt:

Rückwärtskompatibilität Die verbesserte und erweiterte Lexikonressource muss sich leicht per Skript in das von der Morphologiekomponente geforderte Format abbilden lassen.

Erweiterbarkeit Die Ressource soll sowohl inhaltlich als auch strukturell erweiterbar sein.

Wartbarkeit Die Ressource soll gepflegt werden können, ohne die Konsistenz zu gefährden.

Flexibilität Die Struktur der neuen Ressource soll leicht an Veränderungen anpassbar sein. Auf diese Weise kann auf Neuerungen in der Behandlung von Morphologie und Wortbildung leichter eingegangen werden.

Aus inhaltlicher Sicht muss die Ressource das DeKo-Modell adäquat umsetzen, also die in DeKo definierten lexikalischen Einheiten mit all ihren Merkmalen enthalten, so dass die Ressource die Anwendung der Wortbildungsregeln auf Einheiten und Merkmalen optimal unterstützt.

6.1.1 Wahl des Repräsentationsformates

Die Wahl des Repräsentationsformates ist entscheidend für die Erweiterbarkeit und Flexibilität der Ressource. Das DMOR-Lexikon wurde bereits in eine relationale Datenbank überführt (vgl. Lezius et al. (2000)). In einer solchen sind allerdings Strukturanpassungen nur sehr umständlich durchzuführen: Eine Änderung des Datenschemas erfordert das Aus- und wieder Einlesen des gesamten Datenbestandes. Idealerweise sollten die Skripte, die zum Auslesen der für die Morphologiekomponente notwendigen Informationen dienen, von der Struktur der Ressource unabhängig sein, so dass sie nicht bei jeder Änderung angepasst werden müssen. Ein Formalismus, der die Erweiterbarkeit und Flexibilität der Ressource gewährleistet, ist XML.

Die Dokumentenbeschreibungssprache XML

Bei XML, der *eXtensible Markup Language* (vgl. Harold (2000)), handelt es sich um einen Formalismus, der die Definition von Klassen von Dokumenten[1] ermöglicht. Zwei Probleme werden durch XML gelöst: zum einen die Definition des **Zeichenvorrats** von Dokumenten, zum anderen die Definition der **Dokumentstruktur**. Der erste Punkt unterbindet eine Ad-hoc-Kodierung von Sonderzeichen, die bisher eines der größten Probleme beim Austausch von Ressourcen darstellte. Der zweite Punkt stellt Bausteine für ein standardisiertes und eindeutiges **Markup** (Auszeichnung) des Dokumentes zur Verfügung.

XML ist eine echte Teilmenge der Dokumentenbeschreibungssprache SGML (vgl. Goldfarb und Rubinsky (1990)), die Ende der 80er Jahre entwickelt

[1] In XML-Terminologie wird jede in XML repräsentierte Ressource als *Dokument* bezeichnet.

wurde.[2] Mit Hilfe von XML werden **Dokumenttypen** in einer Dokumenttyp-Definition (*Document Type Definition*, **DTD**) beschrieben. Jede **Instanz** eines solchen Dokumenttyps muss der vorgegebenen Dokumentstruktur entsprechen. Damit wird eine **automatische Validierung** der Dokumentstruktur gegen die Strukturdefinition ermöglicht.

Bei XML handelt es sich genaugenommen um eine Sprache zur Beschreibung von Beschreibungssprachen. D.h., XML stellt lediglich die Bausteine zur Verfügung, die benötigt werden, um eine Beschreibungssprache eindeutig zu definieren. Diese Bausteine sind auf der inhaltlichen Seite zwei Konstrukte **Elemente** und **Attribut/Wert-Paare**, auf der formalen Seite syntaktische Festlegungen auf die Notation dieser Konstrukte.

Vor- und Nachteile von XML

Mit den im vorangegangenen Abschnitt vorgestellten Mitteln lassen sich Klassen von Dokumenten definieren, aber auch Ressourcen, deren Struktur eindeutig definiert sein soll, so dass der Zugriff mit einem Computer leicht und ambiguitätsfrei möglich ist. Auf der einen Seite handelt es sich bei XML um einen Standard, der von der Forschungsgemeinde schnell angenommen wurde und sich seit einigen Jahren als Formalismus für die Repräsentation von Daten etabliert hat, auf der anderen Seite muss man aufgrund der Beschränkung auf gerade zwei Konstrukte zur Beschreibung von Daten gewisse Kompromisse bei der Modellierung der Ressource eingehen, die nicht der eigentlichen Komplexität gerecht werden.

Als standardisierter Formalismus profitiert XML von einer Fülle frei verfügbarer Software, mit der XML-Dokumente erstellt (Editoren), bearbeitet (Parser) und umformatiert bzw. in verschiedene Ausgabeformate umgewandelt werden können (Stylesheet-Prozessoren). Diese Software ist i.A. für die Standard-Programmiersprachen verfügbar, so dass nicht nur die Dokumente ausgetauscht werden können, sondern prinzipiell auch die Werkzeuge, die um sie herum entstehen. Dem Aufwand, mit XML ein neues Format erlernen zu müssen, steht der Nutzen gegenüber, damit eine Fülle neuer Anwendungen, die mit XML derzeit realisiert werden, erfassen zu können.

Die Beschränkung der Beschreibungsmittel bezieht sich auf direkte Abhängigkeit zwischen Entitäten. Es ist nicht möglich, Implikationen zu modellieren: Die Aussage *Wenn Attribut x den Wert y hat, dann muss Element z im Dokument vorkommen* kann in XML nicht dargestellt werden. Allerdings erlaubt der ebenfalls standardisierte Verarbeitungsmechanismus für XML-Dokumente, die *eXtensible Style Sheet Language for Transformations* (**XSLT**, vgl. Clark (1999)),

[2] Der wohl bekannteste Dokumenttyp, der mit Hilfe von SGML definiert wurde, ist HTML, die *Hypertext Markup Language*. Diese definiert eine **Klasse** von Dokumenten, die HTML-Dokumente, die den größten Teil aller Seiten ausmachen, die im Internet miteinander verbunden sind.

die Ausführung von Kontrollstrukturen, so dass Aussagen der dargestellten Art zumindest maschinell überprüfbar sind.

Der größte Vorteil von XML ist die Trennung von Ressource und Strukturbeschreibung: Jede Dokumentinstanz kann mit einem XML-Parser automatisch auf ihre Gültigkeit gemäß der Dokumenttyp-Definition geprüft werden, so dass bestimmte Arten von Fehlern von vornherein ausgeschlossen sind.

Modellierungsprinzipien

Trotz der wenigen Beschreibungskonstrukte, die im XML-Formalismus geboten werden, sind die Lösungsmöglichkeiten für Modellierungsaufgaben vielfältig. Die beiden Extreme sind der völlige Verzicht auf Dokumentinhalt, also die Kodierung sämtlicher Informationen als Elementhierarchie und in Form von Attributen, oder aber die Verwendung möglichst weniger Elemente bei einer sehr flachen Hierarchie. Im ersten Fall lässt sich sehr gezielt auf einzelne Informationseinheiten zugreifen, allerdings leidet die Übersichtlichkeit der Ressource an der Menge der Metadaten im Verhältnis zum Dokumentinhalt. Beim zweiten Extrem tritt das Markup auf Kosten der Granularität der gespeicherten Informationen in den Hintergrund. Die Entscheidung zwischen Dokumentinhalt, Attribut oder Element lässt sich nur in Abhängigkeit der zu modellierenden Ressource festlegen. Es gibt allerdings einige generelle Prinzipien, die die Verständlichkeit der gewählten Modellierung erhöhen.

- **Attribute** werden am besten dann verwendet, wenn ein Merkmal über eine vorgegebene, nicht zu große Menge von Werten verfügt. Wortarten sind ein Beispiel für einen solchen **Aufzählungstyp**, ebenso die Unterscheidung, ob eine Einheit morphologisch einfach oder komplex ist.

- Lässt sich eine Information in weitere Informationen untergliedern, dann empfiehlt sich die Modellierung als **Element**. Treten beispielsweise zwei Elemente stets gemeinsam auf, so können sie in ein übergeordnetes Element eingebettet werden.

- **Dokumentinhalt** schließlich ist den Informationen vorbehalten, die nicht weiter zerlegt werden müssen bzw. die nicht aufzählbar sind: Kommentare z.B. werden i.A. nicht für spezielle Anfragen benötigt. Die Zitierformen lassen sich nicht als Aufzählungstyp repräsentieren, etc.

Diese Empfehlungen lassen sich nicht immer einhalten, aber wo dies nicht geschieht, sollte dokumentiert werden, warum an dieser Stelle vom Standardvorgehen abgewichen wurde.

6.1.2 Prinzipien bei der Konzeption einer Ressource

Unabhängig vom Repräsentationsformat gibt es zu strukturellen und inhaltlichen Aspekten der Ressource einige Entscheidungen zu treffen. Unter inhaltlichen Gesichtspunkten ist dies die Frage nach der Abhängigkeit von einer bestimmten Theorie. Unter strukturellen Gesichtspunkten gibt es verschiedene Varianten, in denen ein Lexikonmodell realisiert werden kann. Die Frage und die Varianten werden im Folgenden beleuchtet.

Theorieunabhängigkeit

Es ist wünschenswert, sich nicht zu stark an Theorien zu binden, da mit einer Änderung an der Theorie immer auch Änderungen an einem darauf aufbauenden System verbunden sind. In IMSLEX besteht eine Abhängigkeit vom Modell der Zwei-Ebenen-Morphologie, was allomorphe Flexionsstämme angeht: Aus DMOR wird in IMSLEX die Kodierung umgelauteter Pluralformen von Substantiven über die Flexionsklasse übernommen. Im Eintrag *Apfel*$^{\mathbb{P}}$ verweist allein ein Dollar-Zeichen im Flexionsklassen-Bezeichner auf die Umlautung im Plural. Der umgelautete Pluralstamm selber wird nicht angegeben, sondern kann nur implizit über eine Analyse des Flexionsklassen-Bezeichners ermittelt werden.[3] Die Flexionsklassen-Bezeichner können nicht ohne weiteres in ein Paradigma übersetzt werden, welches unabhängig von einem Morphologiesystem ist, das die Zwei-Ebenen-Morphologie implementiert.

Eine Theorieunabhängigkeit ließe sich hier nur erreichen, wenn man alle Flexionsparadigmen ausmultiplizierte und nach Stämmen und Endungen neu gruppierte.[4] Da das Lexikon allerdings derzeit als Datenbasis für ein Zwei-Ebenen-Morphologiesystem (SMOR, vgl. Schmid et al. (2004)) dient, bleibt die Theorieabhängigkeit zunächst bestehen.

Redundanzvermeidung

Redundanz in einem System kann eine erhöhte Fehleranfälligkeit zur Folge haben. Daher wird gewöhnlich versucht, **Generalisierungen** wahrzunehmen, die die Fehleranfälligkeit eines Systems reduzieren. Ein Beispiel für eine solche Generalisierung sind die Flexionsparadigmen: Die Angabe einer Grundform und einer Flexionsklasse erspart die obligatorische Angabe aller Wortformen eines Paradigmas. Wird bei einer Wortform eine fehlerhafte Flexionsendung ent-

[3] Teilweise sind unregelmäßigen Stämme aber auch explizit kodiert, so bei Adjektiven (*höh, höch* für *hoch*$^{\mathbb{P}}$ und einige andere) und starken Verben (*bäck, buk, bük* etc. für *backen*$^{\mathbb{P}}$).

[4] Anstelle der e-Elisionsregel aus DMOR gäbe es dann z.B. beim Eintrag für das Verb *handeln*$^{\mathbb{P}}$ zwei Flexionsstammformen *handl* (*ich handl<u>e</u>*) und *handel* (*du handel<u>st</u>*). Dies würde bedeuten, dass Flexion und Wortbildung analog behandelt werden. Die jetzige Ungleichbehandlung erklärt sich aus der Kombination der bereits bestehenden Flexionskomponente mit einem neuen Wortbildungskonzept.

deckt, reicht es, diesen Fehler einmal in der Klasse zu korrigieren, anstatt ihn bei jedem betroffenen Lexem berichtigen zu müssen. Es ist eine globale Änderung vorgenommen worden statt einer lokalen.

Ein Beispiel für Redundanz, die gewollt ist, ist die Angabe der Flexionsklasse bei jeder lexikalischen Einheit. Diese Angabe ist bei nicht flektierenden Einheiten eigentlich überflüssig, denn dort wird oft lediglich die Information der Wortart wiederholt, die sich im selben Eintrag an anderer Stelle noch einmal befindet. Bei morphologisch komplexen Einheiten richtet sich die Flexion nach der Flexion eines Bestandteils, auf den aus dem Eintrag heraus auch verwiesen wird: Auch hier wiederholt die Nennung der Flexionsklassen Information, die im selben Eintrag durch den Verweis implizit bereits vorhanden ist. Der Gewinn in beiden Fällen ist der der **Transparenz** oder Übersichtlichkeit: Dadurch, dass für alle Arten von Einheiten dieselben Konzepte verwendet werden, steht für jede lexikalische Einheiten stets fest, an welcher Stelle welche Art von Information vermerkt ist. Für eine maschinelle Verarbeitung bedeutet dies, dass von Unterschieden, die zwischen Einheiten bestehen, abstrahiert wird zugunsten einer klaren und einfachen Sicht auf die Daten.[5]

Modularisierung

Ein Prinzip, das zur Erhöhung der Transparenz beiträgt, ist das der **Modularisierung**. Dies betrifft die Aufteilung komplexer Strukturen in kleinere Teile, zwischen denen allerdings keine Abhängigkeiten bestehen dürfen. Im Falle des Lexikons wird die Modularisierung sowohl bei der Makrostruktur als auch bei der Mikrostruktur erreicht: Die Makrostruktur stellt sich als flache Organisation lexikalischer Einheiten dar. Dabei ist die für die morphologische Verarbeitung relevante Information jeweils in einem Eintrag gebündelt: Durch Wegnahme oder Hinzufügen lexikalischer Einheiten ändert sich nichts an der prinzipiellen Verarbeitbarkeit der vorhandenen Daten.[6] Die Mikrostruktur, also die Gliederung der einzelnen Einträge, fällt je nach Wortart leicht unterschiedlich aus, ist aber ebenfalls modular ausgerichtet. Zu jeder lexikalischen Einheit gibt es globale Merkmale und Angaben zur Flexionsmorphologie. Dazu können fakultativ Module zur Wortbildung, Syntax, Semantik, Phonetik und schließlich wortartspezifische Informationen hinzukommen.

Innerhalb der Module in der Mikrostruktur kommt es allerdings doch zu **Abhängigkeiten**. So hängt das Vorkommen eines Moduls für wortartspezifische Informationen von der Wortart ab, die im Modul für globale Merkmale abgelegt

[5]Da sich bei einer allgemein gehaltenen Struktur wiederum die Fehleranfälligkeit der Ressource erhöht, wird hier von der Möglichkeit der automatischen Konsistenzüberprüfung Gebrauch gemacht (vgl. Abschnitt 8.1.4).

[6]Dies wäre anders, wenn Information zwischen lexikalischen Einheiten redundanzarm gespeichert würde: Dann dürfte eine lexikalische Einheit nur entfernt werden, wenn sichergestellt wäre, dass dadurch die Integrität an anderer Stelle nicht gefährdet wäre.

ist. Derartige Abhängigkeiten versucht man in der Informatik normalerweise zu verhindern, da sie die Fehleranfälligkeit der Ressource erhöhen. In diesem Fall überwiegen jedoch die Vorteile des modularen Konzepts für die Übersichtlichkeit und die maschinelle Verarbeitung der Ressource die Nachteile, die durch die Abhängigkeit entstehen.[7]

6.2 Dokumenttyp-Definition (DTD)

In der Dokumenttyp-Definition werden zwei Entitäten unterschieden: Die **Elemente** dienen der Strukturierung des Dokuments. Sie werden ähnlich einer kontextfreien Grammatik miteinander in Beziehung gesetzt. Mittels regulärer Zeichen können Elemente miteinander kombiniert oder quantifiziert werden. Elemente können sequentiell angeordnet sein oder in Disjunktion auftreten. Die **Attribute** dienen der Spezifizierung der Eigenschaften von Elementen. Sie werden in Abschnitt 6.2.2 vorgestellt.

6.2.1 Elemente – Hierarchische Struktur

Reguläres Zeichen	Erklärung
?	0 oder ein Vorkommen (Optionalität)
*	0 oder beliebig viele Vorkommen
+	ein oder beliebig viele Vorkommen
,	Aufeinanderfolgen (Sequenz)
|	Ausschließendes Oder (Disjunktion)
( )	Gruppierung

Abbildung 6.1: Reguläre Zeichen in der DTD

In Tabelle 6.1 sind die in der DTD verwendeten Metazeichen mit Erklärung aufgeführt. In weiteren Verlauf dieses Kapitels werden Elementnamen (im Text sowie in den Abbildungen) stets kursiv gesetzt.

Die lexikalische Einheit

Jedes XML-Dokument verfügt über ein Wurzel-Element, hier *lexikon* (vgl. Abbildung 6.2). Gemäß der flachen Struktur des IMSLEX besteht ein Lexikon aus beliebig vielen lexikalischen Einheiten (*le*). Diese wiederum sind in Module gegliedert, von denen die ersten beiden, *Globale_Merkmale* und *Flexionsmorphologie*, obligatorisch sind, die anderen optional. Von den Modulen für die

[7] Auch hier sind automatische Konsistenzüberprüfungen möglich (vgl. Abschnitt 8.1.4).

```
<!ELEMENT   lexikon   ( le+ ) >

<!ELEMENT   le        (
                      Globale_Merkmale,
                      Flexionsmorphologie,
                      Wortbildung?,
                      Semantik?,
                      Syntax?,
                      (Substantiv_Merkmale | Adjektiv_Merkmale |
                       Adverb_Merkmale | Verb_Merkmale | ,
                       Abk_Merkmale | Verbpartikel_Merkmale)?,
                      Affix_Merkmale?,
                      Bearbeitungs_Merkmale?
                      ) >
```

Abbildung 6.2: IMSLEX-DTD. Lexikalische Einheit

wortartspezifischen Merkmale kann nur eines je lexikalischer Einheit auftreten. *Affix_Merkmale* können noch hinzukommen.[8] *Bearbeitungs_Merkmale* haben eine rein administrative Bedeutung.

Die Anordnung der Elemente entspricht dem Prinzip der größtmöglichen Übereinstimmung der Einträge unabhängig von den Eigenschaften einer lexikalischen Einheit: Die in Abbildung 6.2 skizzierte Struktur ist für alle in IMSLEX vertretenen Eintragstypen gültig, umfasst also die Beschreibung sämtlicher lexikalischer Einheiten.

Globale Merkmale

```
<!ELEMENT   Globale_Merkmale          (
                                      Zitierform,
                                      PhonetischeTranskription?,
                                      Vorkommenshaeufigkeit+
                                      ) >

<!ELEMENT   Zitierform                ( #PCDATA ) >
<!ELEMENT   PhonetischeTranskription  ( #PCDATA ) >
<!ELEMENT   Vorkommenshaeufigkeit     ( #PCDATA ) >
```

Abbildung 6.3: IMSLEX-DTD. Globale Merkmale

[8] Der Grund für die Aufteilung in wortartspezifische und affixspezifische Merkmale liegt darin, dass einige Einheiten über beide verfügen können, z.B. Substantivsuffixe.

Das Modul für *Globale_Merkmale* (vgl. Abbildung 6.3) setzt sich aus drei Elementen zusammen, die beliebige Zeichenketten als Inhalt haben können.[9] *Zitierform* und *Vorkommenshaeufigkeit* sind obligatorisch. Es darf nur genau eine Zitierform geben. Da die Vorkommenshäufigkeit immer relativ zu einem Korpus erhoben wird, kann es mehr als ein Element *Vorkommenshaeufigkeit* geben. Das Element *PhonetischeTranskription* tritt fakultativ auf. In diesem Modul sind die Merkmale einer lexikalischen Einheit versammelt, die weniger für die komputationelle Bearbeitung der Lexikoneinträge als vielmehr für den Benutzer interessant sind.

Flexionsmorphologie

<!ELEMENT	*Flexionsmorphologie*	(*Stammformen*) >
<!ELEMENT	*Stammformen*	(*DMORstamm*, *Stammform*+) >
<!ELEMENT	*Stammform*	(*Stamm*, *DMORklasse*) >
<!ELEMENT	*DMORstamm*	(#PCDATA) >
<!ELEMENT	*Stamm*	(#PCDATA) >
<!ELEMENT	*DMORklasse*	(#PCDATA) >

Abbildung 6.4: IMSLEX-DTD. Flexionsmorphologie

Das Modul für *Flexionsmorphologie* (vgl. Abbildung 6.4) enthält ein obligatorisches Element *Stammformen*. Dieses Element enthält beliebig viele *Stammform*-Elemente. Die Elemente, die zu einer *Stammform* zusammengefasst werden, sollen stets gemeinsam auftreten: *Stammform* ist immer ein Paar aus *Stamm* und *DMORklasse*. An dieser Stelle wird die Kompatibilität zu DMOR hergestellt: *DMORklasse* steht für die DMOR-Flexionsklasse.

Zusätzlich zu den Stamm/Flexionsklasse-Paaren muss ein Element *DMORstamm* angegeben werden. Es handelt sich dabei um die Grundstammform eines Flexionsparadigmas (vgl. Abschnitt 4.2.2). In DMOR wird sie benötigt, um bei irregulären Stämmen den Zusammenhang zu einem regulären Stammeintrag herzustellen (*back:buk*).

Neben der Auflistung von Suppletivstämmen bieten die *Stammform*-Elemente eine Möglichkeit, Schreibvarianten einer lexikalischen Einheit zu notieren oder nach alter und neuer Rechtschreibung zu differenzieren: Während für die Zitierform eine eindeutige Form gewählt werden muss, können beide Varianten als Stammformen angegeben werden: *Nuß*, *Nuss*.[10]

[9] `#PCDATA` steht für *parsable character data*.

[10] Ein Attribut markiert, ob es sich um alte oder neue Schreibung handelt (vgl. Abschnitt 6.2.2, Abbildung 6.15 auf Seite 90).

Wortbildung

<!ELEMENT	*Wortbildung*	(*Derivation*?, *Komposition*?, *Strukturen*?) >
<!ELEMENT	*Derivation*	(*Derivationsstaemme*?) >
<!ELEMENT	*Derivationsstaemme*	(*Derivationsstamm*+) >
<!ELEMENT	*Derivationsstamm*	(#PCDATA) >
<!ELEMENT	*Strukturen*	(*Struktur*+) >
<!ELEMENT	*Struktur*	(#PCDATA) >

Abbildung 6.5: IMSLEX-DTD. Wortbildung

Das Modul für das Element *Wortbildung* (vgl. Abbildung 6.5) kann die Elemente *Derivation*, *Komposition* und *Strukturen* enthalten. Die ersten beiden dienen der Auflistung von Derivations- und Kompositionsstammformen zu einer lexikalischen Einheit. Ihre Struktur ist identisch (in der Abbildung sind daher die Elemente *Komposition*, *Kompositionsstaemme* und *Kompositionsstamm* nicht mehr eigens aufgelistet). Der Zwischenschritt über das Element *Derivationsstaemme* (*Kompositionsstaemme*) erklärt sich dadurch, dass damit zukünftige Erweiterungen bei den Derivations- und Kompositionsstammformen leichter möglich sind. Ähnlich wie bei den Stamm/Flexionsklasse-Paaren aus der Flexionsmorphologie ist vorstellbar, dass zu den einzelnen Stammformen weitere Informationen anfallen.

Beim Element *Strukturen* handelt es sich um eine Auflistung von einzelnen *Struktur*-Elementen, die wiederum beliebige Zeichenketten enthalten. In diesen Elementen werden die Zerlegungen von morphologisch komplexen lexikalischen Einheiten in unmittelbare Konstituenten abgelegt.[11]

Syntax

Das Element *Syntax* (vgl. Abbildung 6.6) enthält Subkategorisierungsrahmen, repräsentiert als Zeichenketten (zum Format und zur Erstellung der Subkatrahmen vgl. Eckle-Kohler (1999)).

[11] Zur Zeit geschieht dies in einer Kurzform: Das Lexem *Darstellung*$^{\mathbb{P}}$ ist eine *ung*-Derivation des Verbs *darstellen*$^{\mathbb{P}}$ und erhält daher die Struktur `darstell(V)[++]ung(NNSuff)`. Eine Erweiterung auf eine eigene Hierarchie mit Anzahl der Bestandteile, direktem Verweis auf die ID dieser Bestandteile im Lexikon und Angabe der Art der Wortbildung ist der logische nächste Schritt.

```
<!ELEMENT   Syntax          ( Subkatrahmen* ) >

<!ELEMENT   Subkatrahmen    ( #PCDATA ) >
```

Abbildung 6.6: IMSLEX-DTD. Syntax

Semantik

```
<!ELEMENT   Semantik            (
                                SemantischerTyp?,
                                Kommentar?,
                                Lambdaausdruck?,
                                Praesupposition?,
                                Anwendungsbereich?
                                ) >

<!ELEMENT   SemantischerTyp     ( #PCDATA ) >
<!ELEMENT   Kommentar           ( #PCDATA ) >
<!ELEMENT   Lambdaausdruck      ( #PCDATA ) >
<!ELEMENT   Praesupposition     ( #PCDATA ) >
<!ELEMENT   Anwendungsbereich   ( #PCDATA ) >
```

Abbildung 6.7: IMSLEX-DTD. Semantik

Das Element *Semantik* (vgl. Abbildung 6.7) enthält als Elemente die im DeKo-Lexikonmodell (vgl. Abschnitt 5.1.2) spezifizierten Merkmale, jedoch zur Zeit alle ohne weitere Struktur. Sie sind alle optional.[12]

Wortartspezifische Merkmale

```
<!ELEMENT   Substantiv_Merkmale   ( Genus ) >
<!ELEMENT   Adjektiv_Merkmale     ( Verwendung ) >
<!ELEMENT   Adverb_Merkmale       ( Verwendung ) >

<!ELEMENT   Genus                 ( #PCDATA ) >
<!ELEMENT   Verwendung            ( #PCDATA ) >
```

Abbildung 6.8: IMSLEX-DTD. Wortartspezifische Merkmale (1/4)

[12] *SemantischerTyp* wird derzeit als einziges dieser Elemente (bei der Spezifizierung von Eigennamen) bereits verwendet.

Die wortartspezifischen Merkmale bei Substantiven, Adjektiven und Adverbien werden jeweils als Zeichenketten angegeben (vgl. Abbildung 6.8). Sie könnten ebenso durch Aufzählungstypen repräsentiert werden, können auf diese Weise jedoch auch leer gelassen werden, wenn entweder noch keine Verwendung ermittelt wurde oder aber (wie bei Pluraliatantum) kein Genus vorliegt.

```
<!ELEMENT   Verb_Merkmale              ( Aktionsart,
                                         VerbHatResultatszustand,
                                         IntensionalitaetLexikalisiert,
                                         SemantischeVerbklasse ) >

<!ELEMENT   Aktionsart                 ( #PCDATA ) >
<!ELEMENT   VerbHatResultatszustand    ( #PCDATA ) >
<!ELEMENT   IntensionalitaetLexikalisiert ( #PCDATA ) >
<!ELEMENT   SemantischeVerbklasse      ( #PCDATA ) >
```

Abbildung 6.9: IMSLEX-DTD. Wortartspezifische Merkmale (2/4)

Bei Verben sind, ähnlich wie beim Element *Semantik*, die im DeKo-Modell spezifizierten Informationen als Elemente aufgeführt (vgl. Abbildung 6.9), enthalten jedoch im Lexikon noch keinen Inhalt.

```
<!ELEMENT   Verbpartikel_Merkmale      (
                                         Basisverbzahl,
                                         Partikelverbklasse+
                                         ) >

<!ELEMENT   Basisverbzahl              ( #PCDATA ) >
<!ELEMENT   Partikelverbklasse         ( #PCDATA ) >
```

Abbildung 6.10: IMSLEX-DTD. Wortartspezifische Merkmale (3/4)

Die Verbpartikel bzw. Verbzusätze hingegen (vgl. Abbildung 6.10) verfügen über zwei Arten von Informationen, eine Klasse und die Anzahl der im HGC gefundenen Partikelverben mit dieser Partikel (vgl. Aldinger (2002)). Beide Informationen liegen wieder als beliebige Zeichenketten vor.

Die Merkmale für Abkürzungen bzw. für Affixe (vgl. Abbildung 6.11) beschließen die Strukturbeschreibung des Lexikons. Bei Abkürzungen können ausgeschriebene Formen angegeben werden. Auch hier ist das Element *Ausgeschr_Form* in ein anderes Element eingebettet, um ggf. Erweiterungen vorzunehmen: Es ist denkbar, dass noch weitere Erläuterungen zu einer ausgeschriebenen Form hinzukommen. Darüber hinaus kann es zu einer Abkürzung mehrere ausgeschriebene Formen geben.

```
<!ELEMENT   Abk_Merkmale      ( Ausgeschr_Formen? ) >
<!ELEMENT   Ausgeschr_Formen  ( Ausgeschr_Form+ ) >

<!ELEMENT   Ausgeschr_Form    ( #PCDATA ) >
<!ELEMENT   Affix_Merkmale    ( #PCDATA ) >
```

Abbildung 6.11: IMSLEX-DTD. Wortartspezifische Merkmale (4/4)

6.2.2 Attribute

Nachdem die hierarchische Struktur des Lexikons feststeht, werden nun die Merkmale der einzelnen Elemente beschrieben. Dazu dienen die Attribute, die für ein Element definiert werden können. Bei der Attributdeklaration werden neben dem Merkmalnamen die möglichen Merkmalwerte und ein Status angegeben. Die Aufzählung der möglichen Merkmalwerte bietet einen Schutz vor Fehlern in der Ressource: Ein XML-Parser gibt eine Fehlermeldung aus, wenn ein Merkmalwert im Dokument vorkommt, der nicht in der DTD deklariert wurde. Der 'Status' gibt an, ob ein Attribut verpflichtend gesetzt werden muss (`#REQUIRED`), fakultativ gesetzt werden kann (`#IMPLIED`) oder eine Default-Belegung erhält (Wert in doppelten Anführungsstrichen).

Lexikalische Einheit (le)

Die Attribute des Elements *le* (lexikalische Einheit) sind in Abbildung 6.12 dargestellt. Es handelt sich im Wesentlichen um die in DeKo definierten Merkmale (vgl. Abschnitt 5.1.1). Bei den obligatorischen Merkmalen ist außer bei **kategorie** immer ein Wert `undef` vorhanden, der als Platzhalter verwendet werden kann, wenn die genaue Belegung noch nicht klar ist.[13]

Die Merkmale **akzent** und **auslautverhaertung**[14] sind nur für Derivationsaffixe relevant und haben daher bei allen anderen Kategorien die Belegung `neutral`.

Die beiden Merkmale **erzeugt** und **geprueft** sind administrativer Natur: Um zu verhindern, dass bei der Lexikonpflege immer wieder dieselben Einträge durchgesehen werden, kann bei bereits vollständig bearbeiteten Einträgen das Merkmal **geprueft** auf `ja` gesetzt werden. Das Merkmal **erzeugt** dient der Unterscheidung zwischen maschinell und manuell erzeugten Lexikoneinträgen. Auf diese Weise kann die Qualität des Lexikons auf einem Stand gehalten werden, der bei unmarkiertem Hinzufügen von automatisch generierten Informationen nicht möglich wäre.

[13] Dies war vor allem beim Aufbau der Ressource hilfreich, da außer der Kategorie keine der Informationen im DMOR-Lexikon vorhanden ist.

[14] Der Bezeichner dieses Attributs wurde aus Platzgründen in der Abbildung abgekürzt.

```
<!ATTLIST   le
            id              ID                                          #REQUIRED
            kategorie
            ( Substantiv | Verb | Adjektiv | Name | Adverb |
            Numeral | Pronomen | Adposition | Verbpartikel |
            Konjunktion | Partikelverb | Konfix | Verbpraefix |
            Adjektivpraefix | Substantivpraefix | Interjektion |
            Artikel | Invar_Abk | Adjektivsuffix | Substantivsuffix |
            Verbsuffix | Adverbsuffix | Substantiv_Abk |
            Name_Abk | Adjektiv_Abk | Partikel )                       #REQUIRED
            m_status        ( Frei | Gebunden | undef )                  #REQUIRED
            m_form
            ( Simplex | Kurzwort | Nominalisierung | undef |
            Komplex | Komplex_semi | Komplex_abstrakt )                 #REQUIRED
            selegiert       ( ja | nein | undef )                       #REQUIRED
            lexikalisiert   ( ja | nein | undef )                       #REQUIRED
            herkunft        ( nativ | klassisch | englisch | unklar |
                            französisch | fremd | undef )               #REQUIRED
            akzent          ( neutral | beeinflusst | zieht_an)         "neutral"
            auslautverh.    (neutral | blockiert )                      "neutral"
            erzeugt         ( auto | manu )                             #IMPLIED
            geprueft        ( ja | nein )                               #IMPLIED
>
```

Abbildung 6.12: IMSLEX-DTD. Attribute der Lexikalischen Einheit

Globale Merkmale

Bei den globalen Merkmalen *PhonetischeTranskription* und *Vorkommenshaeufigkeit* gibt es jeweils zwei Attribute (vgl. Abbildung 6.13). Da es für die phonetische Transkription von Lexemen verschiedene Notationen gibt (vgl. z.B. Abschnitt 5.2), wird mit dem Merkmal **notation** angegeben, welche hier verwendet wird. Fakultativ kann noch ein Attribut **attr** hinzukommen, das beschreibt, ob die Erzeugung der phonetischen Transkription aufgrund von Systemwissen oder von Heuristiken geschah.

Bei der *Vorkommenshaeufigkeit* handelt es sich normalerweise um die addierten Tokenfrequenzen aller distinkten Wortformen aus dem Paradigma der lexikalischen Einheit im Korpus HGC. Alternativ kann auch ein anderes Korpus angegeben werden. Zur Zeit ist die einzige andere Belegung `Referenz`, ein hand-annotiertes deutsches Referenzkorpus. Will man bei einem Neueintrag nicht erst alle Frequenzen ermitteln, kann man als Wert entweder `-1` angeben oder aber zunächst die Tokenfrequenz der Grundform angeben und dazu das Merkmal **wert** auf `wortform` setzen.

```
<!ATTLIST   PhonetischeTranskription
            notation        ( SAMPA )       "SAMPA"
            attr            CDATA           #IMPLIED>

<!ATTLIST   Vorkommenshaeufigkeit
            korpus          ( HGC | Referenz )  "HGC"
            wert            ( wortform )        #IMPLIED>
```

Abbildung 6.13: IMSLEX-DTD. Attribute einiger globaler Merkmale

Flexionsmorphologie

```
<!ATTLIST   Flexionsmorphologie
            DMORlex
            ( VMod_Stems | VAux_Stems | V-0_Stems | V-ge_Stems |
            V-0_Stems_NoPref | V-ge_Stems_NoPref |
            NN_Stems_NoCp | NN_Stems_NoHead | NN_Stems |
            NE_Stems_NoCp | NE_Stems | NE_Stems_NoHead |
            ADJ_Stems_NoCp | ADJ_Abbr | NN_Abbr |
            NE_Abbr | INVAR_Abbr | VPrefSep )        #IMPLIED >
```

Abbildung 6.14: IMSLEX-DTD. Attribute der Flexionsmorphologie

Beim Element *Flexionsmorphologie* gibt es ein fakultatives Merkmal **DMORlex**. Dies ist neben den Elementen *Stamm* und *DMORklasse* die dritte Information, die benötigt wird, um die vollständige Kompatibilität zu DMOR herzustellen. Dass es als Attribut und nicht als Element repräsentiert wird, erklärt sich allein aus der Tatsache, dass die Merkmalwerte aufgezählt werden können.[15] Dass es nicht obligatorisch ist, liegt an der Tatsache, dass in IMSLEX auch Affixe und Konfixe eingetragen werden, die in DMOR nicht vorgesehen waren.

Flexionsmorphologie – Stammformen

Bei den verschiedenen Stämmen, die innerhalb des Elements *Flexionsmorphologie* auftreten können, kann nach alter und neuer Rechtschreibung differenziert werden. Da die meisten Stämme von der Rechtsschreibreform unberührt bleiben, gibt es die Defaultbelegung `beides`. Durch Setzen des Merkmalwertes `alt` oder `neu` wird das Auslesen spezifisch 'alter' oder 'neuer' Rechtschreibung ermöglicht.

[15] DMOR-Flexionsklassen könnten auch aufgezählt werden, enthalten aber teilweise Sonderzeichen, die in Attributwerten in einer DTD nicht erlaubt sind.

```
<!ATTLIST   Stammform
            id              ID                          #IMPLIED
            DMORtyp         ( reg | irreg | vollform )  #IMPLIED>

<!ATTLIST   DMORstamm
            orth            ( alt | neu | beides )      "beides">

<!ATTLIST   Stamm
            orth            ( alt | neu | beides )      "beides">
```

Abbildung 6.15: IMSLEX-DTD. Attribute von Stammformen

Beim Element *Stammform* kann ein Merkmal **DMORtyp** angegeben werden, der das Auslesen von *Stamm* und *DMORklasse* (vgl. Abbildung 6.4, S. 83) steuert: Ist das Merkmal nicht vorhanden oder lautet die Belegung `reg` (für 'regulär'), so wird das Paar aus Stamm und DMOR-Klasse ausgelesen. Lautet der Wert `irreg` (für 'irregulär'), so muss zusätzlich zum Stamm/DMOR-Klasse-Paar auch noch der DMOR-Stamm ausgelesen werden. Bei der Belegung `vollform` schließlich ist als Stamm bereits der Morphologiestring (vgl. Abschnitt 2.2) eingetragen, so dass keine DMOR-Klasse mehr ausgelesen werden muss. Ein *Stammform*-Element kann über eine ID direkt referenziert werden.

Derivation und Komposition

```
<!ATTLIST   Derivation
            typ             ( ja | nein )               #REQUIRED>

<!ATTLIST   Derivationsstamm
            id              ID                          #IMPLIED
            orth            ( alt | neu | beides )      "beides"
            typ
            ( umgelautet | kurz | lang | vorne_gefugt-getilgt |
            vorne_gefugt-hinten_gefugt | vorne_gefugt |
            hinten_gefugt | getilgt | umgelautet-getilgt |
            umgelautet-getilgt-hinten_gefugt | normal |
            umgelautet-hinten_gefugt | getilgt-hinten_gefugt )  "normal"
>
```

Abbildung 6.16: IMSLEX-DTD. Attribute von Derivation und Komposition

Die Elemente *Derivation/Komposition* und *Derivationsstamm/Kompositionsstamm* weisen dieselbe Attributstruktur auf, so dass hier stellvertretend

für beide nur die Attributdeklarationen für die Elemente *Derivation* und *Derivationsstamm* aufgelistet sind (vgl. Abbildung 6.16). Das **typ**-Attribut beim Element *Derivation* dient dazu, eine lexikalische Einheit explizit von Wortbildung auszuschließen. Sinnvoll ist dies bei Konversionen, die im Lexikon aufgelistet werden, wie *Pro*$^{\mathbb{P}}$ oder *Grün*$^{\mathbb{P}}$ als Substantive, da es dann nicht zu Falschzerlegungen wie **Pro=Gramm* oder mehrdeutigen Analysen wie *grün=Fläche, Grün=Fläche* durch die Morphologiekomponente kommen kann. Für Komposition wird dieses explizite Ausschließen bereits bei DMOR praktiziert, dort allerdings im Sublexikon kodiert (hier im Attribut **DMORlex**, vgl. Abbildung 6.14).

Beim Element *Derivationsstamm/Kompositionsstamm* ist eine ID angegeben, damit von Einträgen morphologisch komplexer lexikalischer Einheiten auf eine Stammform verwiesen werden kann. Da bei der Datenerhebung nicht immer auch sofort eine ID zugewiesen wird, ist der Status nur `#IMPLIED`. Es lassen sich im Nachhinein aus der ID der lexikalischen Einheit entsprechende eindeutige IDs automatisch erzeugen. Wie bei den anderen Stammformen ist hier durch das Attribut **orth** (für 'Orthographie') die Möglichkeit gegeben, nach alter und neuer Rechtschreibung zu unterscheiden. Das **typ**-Attribut bei *Derivationsstamm/Kompositionsstamm* beschreibt die morphologischen Prozesse, die zur Erzeugung der Stammform durchlaufen werden mussten.[16]

Affix_Merkmale

```
<!ATTLIST   Affix_Merkmale
            produktiv       ( ja | nein )   #REQUIRED >
```

Abbildung 6.17: IMSLEX-DTD. Attribute von Affix_Merkmalen

Zum Element *Affix_Merkmale* schließlich (vgl. Abbildung 6.17) gibt es ein Attribut **produktiv**. Auf diese Weise ist eine Unterscheidung zwischen produktiven und nicht-produktiven Affixen möglich. Das Attribut kann dazu verwendet werden, das Auslesen nicht mehr produktiver Affixe aus dem Lexikon zu verhindern, wenn z.B. Übergenerierung vermindert werden soll.

[16]Diese Information dient dem DeKo-Automaten zur Erkennung einer passenden gültigen Zerlegung, falls ein Derivationsaffix z.B. nur eine umgelautete Stammform selegiert.

Kapitel 7

Aufbau und Verwendung des IMSLEX

Nachdem die Konzeption des Lexikons vorgenommen wurde und eine Struktur in Form einer standardisierten Beschreibungssprache vorliegt, gilt es, die vorhandenen Daten in diese Struktur einzupassen und fehlende Informationen zu ergänzen (vgl. Abschnitt 7.1). Danach wird beschrieben, wie das Lexikon verwendet bzw. gepflegt werden kann (vgl. Abschnitt 7.2). Am Ende des Kapitels wird als Zusammenfassung aufgelistet, wie viele lexikalischen Einheiten je Kategorie aktuell ins IMSLEX eingetragen sind (Stand April 2004) und wie IMSLEX in ein Wörterbuchmodell eingeordnet werden kann (vgl. Abschnitt 7.3).

7.1 Anlegen des Lexikons

Das Anlegen der Lexikondaten erfolgt, nachdem die Struktur definiert ist, durch die Ausgestaltung der XML-Datei(en). In der DTD, der Strukturbeschreibung (vgl. Abschnitt 6.2), sind die Elementnamen, Attributnamen und Attributwerte vorgegeben, die verwendet werden dürfen oder müssen, aber die individuelle Ausgestaltung eines Dokuments kann von Dokumentinstanz zu Dokumentinstanz unterschiedlich ausfallen. Insbesondere der Dokument**inhalt**, also die Teile, die in der DTD als `#PCDATA` definiert sind, kann (im Rahmen der erlaubten Zeichen) beliebige Zeichenketten enthalten.[1]

7.1.1 Vorabentscheidungen

Zwei Fragen müssen geklärt werden, bevor eine vorhandene Lexikonressource in das neue Format überführt wird:

[1] Es gibt Bestrebungen, auch für die 'Semantik' eines Dokuments eine formale Beschreibungssprache zu definieren, analog zur DTD für die 'Syntax' des Dokuments, aber da bestehen bislang nur Ansätze.

1. Soll die Ressource aus einer oder aus mehreren Dateien bestehen?

2. Wie werden Attributwerte vorbelegt, die angegeben werden müssen, für die aber noch keine Daten vorhanden sind?

Aufteilung der IMSLEX-Daten in Dateien

Dadurch, dass die Struktur sämtlicher lexikalischer Einheiten in einer gemeinsamen DTD definiert wird, könnte die gesamte Ressource in einer einzigen Datei repräsentiert werden. Durch die Merkmalwerte kann jeder Eintrag jederzeit zweifelsfrei identifiziert werden.

Datei	Kategorie	Typ
IMSLEX_NN.xml	*Substantiv*	offene Klassen
IMSLEX_NE.xml	*Name*	
IMSLEX_ADJ.xml	*Adjektiv*	
IMSLEX_V.xml	*Verb*	
IMSLEX_PartV.xml	*Partikelverb*	
IMSLEX_ADV.xml	*Adposition, Adverb*	geschlossene Klassen
	Interjektion, Konjunktion, Partikel	
IMSLEX_PRON.xml	*Artikel, Pronomen*	
IMSLEX_NUM.xml	*Numeral*	
IMSLEX_Praefix.xml	(verschiedene Präfixe)	Affixe und Zusätze
IMSLEX_Suffix.xml	(verschiedene Suffixe)	
IMSLEX_Konfix.xml	*Konfix*	
IMSLEX_Erstglied.xml	*Erstglied*	
IMSLEX_VPartikel.xml	*Verbpartikel*	
IMSLEX_ABK.xml	(verschiedene Abkürzungen)	Sonderklassen

Abbildung 7.1: Einteilung der XML-Dateien in IMSLEX

Für das IMSLEX wird dennoch die in DMOR praktizierte Idee der Aufteilung in Dateien (grob) nach Wortarten übernommen, so dass für jeweils eine Wortart oder einige Wortarten eigene Dateien vorgesehen sind (vgl. Abbildung 7.1[2]). Das bietet den Vorteil, nicht mit einer einzigen sehr großen Textdatei arbeiten zu müssen[3], sondern mit mehreren kleinen Dateien, die auch separat bearbeitet werden können.

[2]Adverbien zählen zu den offenen Klassen. Da in der Adverb-Datei jedoch zumeist Vertreter geschlossener Klassen gesammelt sind, wird sie in dieser Abbildung bei den 'geschlossenen Klassen' dargestellt. 'Affixe und Zusätze' sind im DMOR-Modell teilweise nicht vorhanden.

[3]Momentan umfasst das Substantivlexikon in seiner XML-Repräsentation 20 Megabyte (MB) an Daten für 21 000 Einträge.

Vorbelegung von Attributwerten

Die Frage nach der Vorbelegung der Attributwerte ist schwieriger zu beantworten, denn bei Aufzählungstypen kann der Merkmalwert nicht leer gelassen werden. Dies war der Grund dafür, bei der Konzeption des Lexikons bei vielen Attributen einen Merkmalwert `undef` (für 'undefiniert') vorzusehen, der nun als Default verwendet werden kann. Einige Merkmalwerte können allerdings vorab bereits eine bestimmte Belegung erhalten, die später verfeinert werden kann. Bei den Merkmalen der lexiklaischen Einheit (Element *le*) ist die Vorbelegung wie nachfolgend dargestellt.

id Eine XML-ID muss mit einem Buchstaben beginnen und für das gesamte Dokument eindeutig sein. In IMSLEX besteht die ID aus der abgekürzten Kategorie und einer laufenden Nummer, also `n1` für das erste Substantiv, `n2` für das zweite, usw. Sie wird beim Erzeugen der Dateien eingesetzt.

kategorie Die Kategorie ergibt sich entweder aus der gerade bearbeiteten DMOR-Datei oder aus dem Flexionsklassenbezeichner. Eine Gegenüberstellung von IMSLEX-Kategoriebezeichner und traditioneller Wortart findet sich in Abbildung 7.11 auf S. 112. Für dieses Merkmal gibt es keinen Platzhalter, da bei jeder eingetragenen lexikalischen Einheit die Kategorie bekannt sein muss.

m_form, herkunft Morphologische Form und Herkunft einer lexikalischen Einheit werden in DMOR nicht erhoben. Diese Informationen müssen für jedes *le*-Element einzeln ermittelt werden. Daher werden beide zunächst mit dem Platzhalter `undef` belegt.

m_status Der morphologische Status wird wie folgt vorbelegt: Bei Affixen und Konfixen lautet die Belegung `gebunden`, bei allen anderen lexikalischen Einheiten `frei`. Dies ist insofern unproblematisch, als im DMOR-Lexikon außer einer kleinen Menge an Kompositionserstgliedern, die an ihrer Flexionklasse (vgl. z.B. Abbildung 3.15 auf Seite 37) eindeutig zu identifizieren sind, keine gebundenen Einheiten enthalten sind.

selegiert Bei der Selektion verhält es sich ähnlich wie beim morphologischen Status: Affixe selegieren (Belegung `ja`), alle anderen lexikalischen Einheiten nicht (Belegung `nein`). Diese Belegungen können von vornherein vergeben werden und müssen später nur dann verfeinert werden, wenn sich die Theorie ändert (also z.B. Affixoide nicht mehr als selegierend angesehen werden).

lexikalisiert Dieses Merkmal ist am schwierigsten zu behandeln, da hier Morphologie und Semantik vermischt werden. In der Annahme, dass in das DMOR-Lexikon nur lexikalisierte Einheiten aufgenommen wurden, wird

zunächst die Belegung `ja` vergeben. Partikelverben und Erstglieder erhalten die Belegung `undef`, da sie nicht für die Morphologiekomponente ausgelesen werden, sondern die Einträge nur für die Speicherung von Derivations- und Kompositionsstämmen sowie von Subkatrahmen[4] dienen.

Beim Element *Vorkommenshaeufigkeit* wird das Merkmal **korpus** mit dem Wert `HGC` vorbelegt. Beim Element *Flexionsmorphologie* kommt die Information für das Merkmal **DMORlex** aus den DMOR-Dateien. Das Merkmal **typ** bei Derivation und Komposition wird mit dem Wert `ja` vorbelegt, wenn die Wortart Substantiv, Adjektiv oder Verb ist. Eigennamen sind (anders als in DMOR) zunächst einmal von der Wortbildung ausgeschlossen (**typ**:`nein`).

Nach diesen Festlegungen können aus den DMOR-Dateien die XML-Einträge für die lexikalischen Einheiten erzeugt werden.

7.1.2 Die Übernahme der DMOR-Lexikondaten

Aus den DMOR-Dateien lassen sich für IMSLEX die Zitierform, die Kategorie sowie Flexionsstamm und Flexionsklasse herauslesen. Jedes Stamm/Flexionsklasse-Paar aus dem DMOR-Lexikon ist genau einem Sublexikon zugeordnet, so dass das Attribut **DMORlex** immer eine eindeutige Belegung hat. Teilweise ist eine Fallunterscheidung erforderlich, was die Wahl der Zitierform und die Art der Kodierung der Flexionsinformation angeht. Für das Attribut **DMORtyp** beim Element *Flexionsmorphologie* muss immer eine Fallunterscheidung getroffen werden. Ist im DMOR-Lexikon beim Stammeintrag zusätzlich zum Flexionsstamm eine allomorphe Form angegeben (*Atlas:Atlanten; back:bük*), so erfordert dies den Wert `irreg` beim Attribut **DMORtyp**.

Da es in DMOR keinen expliziten Lexembegriff gibt, muss überprüft werden, inwiefern die Zitierform oder der Stamm als Grundform in IMSLEX verwendet werden können. Im Allgemeinen stimmen diese überein; Verben sind die bekannte Ausnahme. Zunächst werden zwei Fälle regelbasierter morphologischer Verarbeitung in DMOR angesprochen, die ebenfalls einen Einfluss auf den Lexembestand in IMSLEX haben können.

Fehlende Lexeme

Für zwei Phänomene sind im DMOR-Lexikon keine direkten Lexikoneinträge vorgesehen, die Movierung mit dem Suffix *-in* und die Transposition. Während die Transposition in Morphologiekomponenten gewöhnlich regelbasiert behandelt wird, nicht durch Lexikoneinträge, muss die Verschmelzung von zwei Lexemen zu einem Lexem für die Übernahme in IMSLEX rückgängig gemacht werden.

[4] Letzteres gilt nur für Partikelverben, nicht für Erstglieder.

Movierung In einer DMOR-Flexionsklasse kann durch die Zeichenkette `=in` markiert werden, dass es zu einer Form auch ein auf *-in* endendes Pendant gibt: *Agent, Agentin*. Bei der Übernahme der Daten aus DMOR ins IMSLEX werden beide Formen also zunächst **einem** Lexem $Agent^{\mathbb{P}}$ zugerechnet. Da die Einträge im DMOR-Lexikon diesbezüglich auch inkonsistent gehandhabt wurden, gibt es für einige dieser Derivationen zwei Einträge ($Dieb^{\mathbb{P}}$, $Diebin^{\mathbb{P}}$), für viele aber nur einen ($Agent^{\mathbb{P}}$).[5]

Transposition Der substantivierte Infinitiv, das substantivierte Adjektiv sowie substantivierte Partizipien werden in DMOR regelbasiert über die Fortsetzungsklassen behandelt. Einige der Formen verfügen zusätzlich über einen Eintrag als lexikalisierte Substantive (z.B. $Verbrechen^{\mathbb{P}}{}_{NN}$, $Essen^{\mathbb{P}}{}_{NN}$). In IMSLEX wird Transposition ebenfalls als regelbasierter Prozess angesehen, so dass substantivierte Infinitive etc. – wie in DMOR – keine Lexikoneinträge erhalten.

Alle in diesem Kapitel erwähnten Umwandlungsschritte sind in der Programmiersprache *Perl* (*Practical Extraction and Report Language*, vgl. Wall et al. (2000)) programmiert. Die Sprache erlaubt die Verwendung regulärer Ausdrücke und wurde ursprünglich für die Verarbeitung von Textdateien entworfen.

Zitierform und Flexionsmorphologie

Im Folgenden wird für einzelne Wortarten beschrieben, wie die für IMSLEX relevanten Informationen aus den DMOR-Lexikondateien in der neuen Ressource repräsentiert werden.

Substantive Bei Substantiven, die über Singular- und Pluralflexion verfügen, wird die Nominativ-Singular-Form als Zitierform gewählt: *Apfel, Hündchen, Nuß*[6]. Das Merkmal **DMORtyp** beim Element *Stammform* wird mit dem Wert `reg` belegt. Kommt eine unregelmäßige oder zusätzliche Pluralform hinzu (*Komma* $\rightarrow$ *Kommas/Kommata*), so ändert dies nichts an der Zitierform, sondern es kommt eine Stammform mit dem Attribut/Wert-Paar **DMORtyp**:`irreg` hinzu.

Pluraliatantum haben keine Singularformen, daher wird bei ihnen die Nominativ-Plural-Form als Zitierform gewählt: *Kosten, Leute*. Bei lexikalisierten substantivischen Partizipien, die einen eigenen Lexikoneintrag erhalten,

[5] An dieser Stelle besteht noch Handlungsbedarf. Da in IMSLEX morphologisch komplexe Einheiten als solche markiert werden, können die in DMOR verschmolzenen Einträge ohne Informationsverlust wieder getrennt werden.

[6] Derzeit ist bei Lexemen, die von der Rechtschreibreform betroffen sind, noch die alte Schreibung als Zitierform gewählt.

wird die Nominativ-Singular-Form in schwacher Flexion als Zitierform gewählt: *(der/die) Angehörige, Beamte, Gefreite*. **DMORtyp** ist bei beiden `reg`.

Eigennamen Bei Eigennamen wird im Allgemeinen die Nominativ-Singular-Form als Zitierform gewählt: *Marisa, Berlin, Weizsäcker*. Bei geographischen Namen, die nur im Plural verwendet werden, wird die Nominativ-Plural-Form als Zitierform gewählt: *Malediven, Ardennen*. Da es bei Eigennamen nicht zu unregelmäßigen oder zusätzlichen Pluralformen kommt, gilt immer **DMORtyp**:`reg`. Die im Flexionsklassenbezeichner kodierte Information, dass es sich um einen Nachnamen, einen männlichen oder weiblichen Vornamen etc. handelt, wird zu diesem Zeitpunkt noch nicht verwertet.

Verben Bei Verben und Partikelverben wird die Infinitiv-Form als Zitierform gewählt: *gehen, rudern, überzeugen, abwandern*. Bei regulären Verben wird das Merkmal **DMORtyp** beim Element *Stammform* mit dem Wert `reg` belegt. Irreguläre oder starke Verben erhalten die Belegung `irreg`, damit beim Auslesen des Lexikons der *DMORStamm* mit ausgelesen wird: *back:buk, back:bük* etc. *DMORStamm* ist immer der Verbstamm ohne die Infinitivendung.[7]

Adjektive Bei Adjektiven wird die unflektierte Form als Zitierform gewählt: *blau, riesig, ideenreich*. Dies gilt auch dann, wenn ein Adjektiv überwiegend attributiv verwendet wird: *hellicht*. Das Merkmal **DMORtyp** beim Element *Stammform* erhält i.A. den Wert `reg`. Ausnahmen sind die Einträge der Suppletivstämme bei den Lexemen *gut*$^{\mathbb{P}}$, *hoch*$^{\mathbb{P}}$ und *nah*$^{\mathbb{P}}$ (**DMORtyp**:`irreg`).

Adverbien In der Adverb-Datei sind verschiedene nicht-flektierende Wortarten zusammengefasst. Die Frage nach der Zitierform stellt sich nur für Schreibvarianten einer Form: *andererseits/andrerseits*. Da diese in DMOR als zwei separate Einträge vorliegen, bilden sie in IMSLEX zunächst auch zwei separate Einträge. Beim Verschmelzen der beiden Einträge zu einem Eintrag mit zwei Stammformen beim Element *Flexionsmorphologie* wird diejenige der Formen, die eine größere Vorkommenshäufigkeit im HGC aufweist, zur Zitierform: *andererseits*$^{\mathbb{P}}$ wegen *andererseits*$_{(6152)}$ vs. *andrerseits*$_{(60)}$.

Pronomen Die Pronomen flektieren teilweise (*seine, seinen*), teilweise nicht (*allerlei, derlei*). Bei den flektierenden wird (abweichend von DMOR) nicht der längste gemeinsame Teilstring aller Formen aus dem Paradigma gewählt, sondern die Nominativ-Plural-Form (bzw. Femininum Singular, wenn keine Pluralform existiert) in starker Flexion: *diese, welche, alle*. Dies erlaubt die Un-

[7] Nach Höhle handelt es sich präzise um den Stamm der 2. Person Plural Präsens Indikativ, den 'unmarkierten' Stamm eines Verbs (vgl. Höhle (1982), S. 82, Fußnote 5).

terscheidung von Lexemen wie *welch*$^{\mathbb{P}}$ und *all*$^{\mathbb{P}}$, die nicht flektieren und nur vor Artikeln auftreten: *welch ein Tag, all die Kinder*. Bei den meisten Pronomen ist anstelle der Flexionsklasse (Element *DMORklasse*) zu jeder Vollform der Morphologiestring angegeben: `+DEM.Subst.MN.Gen.Sg.St` zur Form *dessen* beim Lexem *die*$^{\mathbb{P}}{}_{DEM}$. Zur Steuerung der Ausleseroutine erhält das Merkmal **DMORtyp** den Wert `vollform`.

Affixe und Zusätze Affixe und Konfixe sind nicht im DMOR-Lexikon eingetragen. Diese Einträge müssen neu erzeugt werden. Zitierform und damit Lexem ist das Affix mit einem Bindestrich an der Seite, an der die Basis affigiert wird: *-chen*$^{\mathbb{P}}$, *ent-*$^{\mathbb{P}}$. Bei Präfixen und Konfixen bleibt die Flexionsinformation leer, während bei Suffixen Stammform und Flexionsklasse eingetragen werden: Als *Stamm* wird die Form ohne Bindestrich (*chen*) eingetragen. Die Flexionsklasse ist identisch mit der Flexionsklasse von *chen*-Derivationen (*Hölzchen, Stöckchen*).

Bei Partikelverben und sogenannten 'Erstgliedern' bleibt ebenfalls die Flexionsinformation leer, da auch sie nur aufgrund ihrer Wortbildungsstämme eingetragen sind (*Ausgehanzug, Darstellung; Schrebergarten, Allroundtalent*), die einer Einheit zugeordnet werden müssen.

Sonderklassen Abkürzungen, nicht ihre ausgeschriebene Form, werden als Zitierformen gewählt. Beim Element *Flexionsmorphologie* werden die Stamm/Flexionsklassen-Paare aus DMOR übernommen (**DMORtyp**:`irreg`). Wie bei den Adverbien können später Einträge miteinander verschmolzen werden, bei denen es verschiedene Abkürzungsvarianten für dieselbe ausgeschriebene Form gibt (z.B. *s*, *sek*, *sec* für *Sekunde*).

7.1.3 Auffüllen der DeKo-Merkmale

Nach dem Erzeugen des Grundlexikons aus den DMOR-Dateien müssen diejenigen Informationen aufgefüllt werden, die in DMOR nicht enthalten sind. Dies betrifft insbesondere die Merkmale der lexikalischen Einheit, die noch nicht mit einem sinnvollen Defaultwert belegt sind. Es handelt sich dabei um die Attribute **m_form** (morphologische Form) und **herkunft** sowie um die Derivations- und Kompositionsstämme.[8]

Während es sich bei der (halb)automatischen Umwandlung einer Ressource in eine andere um eine vergleichsweise einfache Operation handelt, ist die Auszeichnung von knapp 40 000 Substantiven, Adjektiven und Verben mit Informationen sehr zeitaufwendig. Da es sich bei den Einheiten aus dem DMOR-Lexikon um die Lexeme der in Texten häufig vorkommenden Wortformen handelt, sind

[8]Die ausführliche Beschreibung der Affixe fand bereits im Rahmen des DeKo-Projekts statt (vgl. 5.1) und konnte für das IMSLEX einfach übernommen werden.

sie häufig der Lexikalisierung oder Idiomatisierung unterworfen. Die semantischen Muster, die im Verlauf des DeKo-Projekts für Derivationsaffixe gesammelt wurden, können zwar hier Richtungen vorgeben, aber letztendlich ist es vom individuellen Sprachempfinden abhängig, ob Formen wie *sichtbar* und *offenbar* als komplex, semikomplex oder idiomatisch angesehen werden.

Für die Extraktion von Derivations- und Kompositionsstammformen aus Korpora wurden verschiedene Methoden angewendet, die in Heid et al. (2002) beschrieben sind.

Datei	FS	WS	# Ksf	Ksf-Beispiele	# Dsf	Dsf-Beispiel
IMSLEX_NN.xml	+	+	12.217	*Häusermeer*	1.115	*häuslich*
IMSLEX_NE.xml	+	+	214	*Elbtunnel*	0	*Kafkaesk*
IMSLEX_ADJ.xml	+	+	40	*Weitsprung*	23	*Klugheit*
IMSLEX_V.xml	+	+	422	*Gehversuch*	31	*lesbar*
IMSLEX_PartV.xml	–	+	102	*Abbiegespur*	1.159	*Darstellung*
IMSLEX_ADV.xml	+	+	9	*Soforthilfe*	0	*sofortig*
IMSLEX_PRON.xml	–	–	0	–	0	–
IMSLEX_NUM.xml	–	–	51	–	0	–
IMSLEX_Praefix.xml	–	–	0	–	47	*Monokultur*
IMSLEX_Suffix.xml	+	+	147	*Übungsflug*	139	*Sparsamkeit*
IMSLEX_Konfix.xml	–	–	0	–	446	*identifizieren*
IMSLEX_Erstglied.xml	–	+	0	–	87	*Benefizkonzert*
IMSLEX_VPartikel.xml	+	–	8	(*Zwischenruf*)	372	(*abgeben*)
IMSLEX_ABK.xml	+	+	0	*AIDS-Hilfe*	0	*FPÖler*

Abbildung 7.2: IMSLEX-Dateien und Stammformen

In Abbildung 7.2 sind für die einzelnen IMSLEX-Dateien[9] die Anzahlen der in ihnen enthalten Derivationsstammformen (Dsf) und Kompositionsstammformen (Ksf) aufgelistet (mit Stand Mai 2004). Die mit 'FS' (Flexionsstämme) und 'WS' (Wortbildungsstämme) überschriebenen Spalten geben an, ob aus der Datei diese Art von Stämmen für die Morphologiekomponente ausgelesen wird (+) oder nicht (–, vgl. Abschnitt 8.1).

7.1.4 Zwischenstand: Ein IMSLEX-Eintrag

Mit den bis hier beschriebenen Aktionen ist ein Lexikon entstanden, das für eine morphologische Analyse, die auch die Derivation berücksichtigt, einsetzbar ist. Bevor beschrieben wird, welche weiteren Merkmale noch hinzugekommen

[9]Der Einheitlichkeit halber entspricht die Auflistungsreihenfolge der in Abbildung 7.1 auf Seite 94.

sind oder in Zukunft hinzukommen sollen, soll hier ein Beispiel für einen Lexikoneintrag gegeben werden.

```
<le form="Simplex" herkunft="nativ" id="n25854" kategorie="Substantiv"
    lexikalisiert="ja" m_status="Frei" selegiert="nein">
<Globale_Merkmale>
    <Zitierform>Haus</Zitierform>
    <PhonetischeTranskription attr="0">h"aUs</PhonetischeTranskription>
    <Vorkommenshaeufigkeit korpus="HGC">90214</Vorkommenshaeufigkeit>
</Globale_Merkmale>
<Flexionsmorphologie DMORlex="NN_Stems">
    <Stammformen>
        <DMORStamm>Haus</DMORStamm>
        <Stammform DMORtyp="reg">
            <Stamm>Haus</Stamm>
            <DMORklasse>NNeut_es_$er</DMORklasse>
        </Stammform>
    </Stammformen>
</Flexionsmorphologie>
</le>
```

Abbildung 7.3: Die lexikalische Einheit $Haus^{\mathbb{P}}{}_{NN}$ in XML

In Abbildung 7.3 sind die globalen Merkmale und die Flexionsinformation dargestellt, wie sie in XML repräsentiert werden. Bei $Haus^{\mathbb{P}}{}_{NN}$ handelt es sich um eine morphologisch einfache, native Form, die frei vorkommt und dementsprechend auch keine Basen selegiert. Die Aussprache ist im SAMPA-Format (vgl. SAMPA (1989)) angegeben. Aus den im *Flexionsmorphologie*-Teil abgelegten Informationen lässt sich wieder die DMOR-Information extrahieren: Flexionsstamm, Flexionsklasse und Sublexikon (`NN_Stems`). Der Eintrag ist gemäß der DTD *valide*, d.h., er entspricht der in der DTD definierten Struktur.

Das Modul für das Element *Wortbildung* stellt sich wie in 7.4 abgebildet dar. Im Stammparadigma (vgl. Abschnitt 4.2.2) des Lexems $Haus^{\mathbb{P}}$ sind sowohl Derivationsstammformen wie Kompositionsstammformen enthalten. Mit den Derivationsstammformen lassen sich Formen wie *Häuschen* und *Häuserchen*, mit den Kompositionsstammformen Formen wie *Haushund* und *Häusermeer* bilden. Weitere Beispiele für XML-Einträge finden sich am Ende von Anhang D.

Beim Struktureintrag steht der Platzhalter für Simplex-Formen: `(ohne)`.

7.1.5 Auffüllen weiterer Merkmale

Es gab in der Ressource zwei Arten von Informationen, bei denen die DMOR-Flexionsklasse teilweise Rückschlüsse auf eine Belegung zuließ. Dies sind zum einen die Struktur einer morphologisch komplexen Einheit, zum anderen der 'semantische Typ'.

```
<Wortbildung>
    <Derivation typ="ja">
        <Derivationsstaemme>
            <Derivationsstamm id="ndsf25854_1"
                typ="umgelautet">Häus</Derivationsstamm>
            <Derivationsstamm id="ndsf25854_2"
                typ="umgelautet-hinten_gefugt">Häuser</Derivationsstamm>
        </Derivationsstaemme>
    </Derivation>
    <Komposition typ="ja">
        <Kompositionsstaemme>
            <Kompositionsstamm id="nksf25854_1">Haus</Kompositionsstamm>
            <Kompositionsstamm id="nksf25854_2"
                typ="umgelautet-hinten_gefugt">Häuser</Kompositionsstamm>
        </Kompositionsstaemme>
    </Komposition>
    <Strukturen>
        <Struktur>(ohne)</Struktur>
    </Strukturen>
</Wortbildung>
```

Abbildung 7.4: Derivation- und Kompositionsstämme von *Haus*$^{\mathbb{P}}{}_{NN}$ in XML

Strukturen

Die DMOR-Flexionsklasse `NFem-Deriv` umfasst Derivationen auf *-ung, -heit, -keit, -ion, -(i)tät* und *-schaft*. In diesem Fall konnte halbautomatisch überprüft werden, ob die potentielle Basis den Selektionskriterien des jeweiligen Affixes entsprach. Zusammen mit der vorher vergebenen Belegung der morphologischen Form ließen sich auf diese Weise die eindeutigen Fälle automatisch eintragen. In den Fällen, in denen ein Muster nicht mit dem Attributwert für die morphologische Form übereinstimmte oder in denen die Basis noch nicht als Derivationsstamm eingetragen war, musste intellektuell entschieden werden.

In Abbildung 7.5 sind einige Wortbildungsstruktur-Einträge aus IMSLEX aufgeführt. In runden Klammern wird die Wortart markiert.[10] `[++]` markiert die Grenze zum Derivationsaffix. Es fällt auf, dass einige Präfigierungen mit *un-* nicht markiert sind (Zeilen 6 und 8; Zeile 4). Dies liegt an Inkonsistenzen im DMOR-Lexikon, die ins IMSLEX übernommen wurden.[11]

(7.1) *unzivilisiert*$^{\mathbb{P}}{}_{ADJ}$, *zivilisatorisch*$^{\mathbb{P}}{}_{ADJ}$, *zivilistisch*$^{\mathbb{P}}{}_{ADJ}$, *Entzivilisierung*$^{\mathbb{P}}{}_{NN}$, *Unzivilisiertheit*$^{\mathbb{P}}{}_{NN}$, *zivilisieren*$^{\mathbb{P}}{}_{V}$

[10] Mit `PREF` werden Präfixe gekennzeichnet, mit `PART2` die häufig als Dervationsstamm auftretende Partizip-II-Form.

[11] Die Möglichkeit der systematischen Beseitigung von Inkonsistenzen im Lexikon ist durch die bewusste Unterscheidung von Simplex- und Komplex-Formen erst jetzt gegeben.

Wortbildungsstruktur	Zeile
un(PREF)vollkommen(ADJ)[++]heit(NNSuff)	1
un(PREF)voreingenommen(ADJ)[++]heit(NNSuff)	2
un(PREF)wahr(ADJ)[++]heit(NNSuff)	3
unwissend:unwissen(ADJ)[++]heit(NNSuff)	4
un(PREF)zeitgemäß(ADJ)[++]heit(NNSuff)	5
unzivilisiert(ADJ)[++]heit(NNSuff)	6
un(PREF)zufrieden(ADJ)[++]heit(NNSuff)	7
unüberlegt(ADJ)[++]heit(NNSuff)	8
vage:vag(ADJ)[++]heit(NNSuff)	9
verbissen:verbeißen(PART2)[++]heit(NNSuff)	10
verbockt:verbocken(PART2)[++]heit(NNSuff)	11
verbohrt:verbohren(PART2)[++]heit(NNSuff)	12
verborgen:verbergen(PART2)[++]heit(NNSuff)	13

Abbildung 7.5: Struktureinträge in IMSLEX, *-heit*-Derivationen

Im DMOR-Lexikon finden sich die in 7.1 aufgelisteten Einträge, die die Zeichenkette *zivili* enthalten. Dies erklärt, warum in Zeile 6 in Abbildung 7.5 keine weitere Unterteilung der Basis vorgenommen wurde. Dass der Bestandteil *zeitgemäß* in Zeile 5 nicht weiter zerlegt wurde, liegt daran, dass den Struktureinträgen das Prinzip der Zerlegung in unmittelbare Konstituenten zugrundeliegt, also die Zerlegung in nächst kleinere Einheiten, die ebenfalls im Lexikon verzeichnet sind.

Die Zeilen in Abbildung 7.5, die einen Doppelpunkt enthalten (Zeile 4, Zeilen 9-13), markieren eine Alternative, was die Interpretation der Basis angeht. In Zeile 9 handelt es sich bei *vag* schlicht um eine getilgte Derivationsstammform, die bei $vage^{\mathbb{P}}{}_{ADJ}$ noch nicht eingetragen war.[12] Die Beispiele in den Zeilen 10-13 hingegen weisen auf eine bekannte Abgrenzungsproblematik zwischen Adjektiven und Verbpartizipien hin.

(7.2) `Verbissenheit+NN.Fem.NGDA.Sg`

`verbissen(ADJ)heit(NNSuff)+NN.Fem.NGDA.Sg`

`verbeißen(V)heit(NNSuff)+NN.Fem.NGDA.Sg`

`verbissen(PART2):verbeißen(V)heit(NNSuff)+NN.Fem.NGDA.Sg`

Die Frage, welche der Analyse-Varianten in 7.2[13] eine Morphologiekomponente bei der Analyse der Wortform *Verbissenheit* ausgeben soll, richtet sich

[12] Ein positiver Nebeneffekt der Eintragung der Wortbildungsstruktur ist, dass die für die Erklärung der Wortbildungsprodukte, die bereits im Lexikon verzeichnet sind, notwendigen Stammformen gefunden werden, sofern sie nicht ohnehin schon eingetragen waren.

[13] `NGDA` steht in den Beispielen stellvertretend für 'Nominativ, Genitiv, Dativ oder Akkusativ'.

allein nach den Bedürfnissen der Anwendung, die das Analyseresultat entgegennimmt.[14]

Semantik

Die DMOR-Flexionsklassen der Eigennamen enthalten teilweise die Angabe, ob es sich um einen Personennamen (`Name`) oder um einen geographischen Namen (`Geo`) handelt. Diese Vorgabe wurde verwendet, um bei der halbautomatischen Vergabe der Information zum Element *SemantischerTyp*[15] bereits die plausibelste Information vorzugeben. Mit Hilfe eines Perl-Programms wurden alle in IMSLEX eingetragenen Eigennamen angezeigt, eine aufgrund der Flexionsklasse getroffene Hypothese über den 'Typ' angezeigt und mittels der Eingabe einer Nummer die gewählte Information gespeichert.[16] Dabei konnte, wenn zu einem angezeigten Eigennamen aus dem Lexikon noch kein passender Typ benannt war, zur Laufzeit des Programmes ein Typ mitsamt einer Nummer hinzugefügt werden. Auf diese Weise genügte ein einziger Durchlauf durch die Eigennamenliste, um alle Namen zu klassifizieren.

Abbildung 7.6 zeigt das Resultat der Annotation (Stand Juni 2004): Dargestellt sind die im Lexikon vorkommenden semantischen Typen von Eigennamen mitsamt der Anzahl lexikalischer Einheiten, für die sie vergeben wurden.[17]

7.1.6 Informationen aus anderen Ressourcen

Drei weitere Arten von Informationen, die in das IMSLEX aufgenommen werden konnten, sollen hier noch erwähnt werden: Daten zur **Subkategorisierung**, Daten zur **phonetischen Transkription** sowie Daten zum **syntaktischen Verhalten** von Adjektiven.

Die Subkategorisierungsrahmen für Verben, Adjektive und Substantive entstammen den Arbeiten von Judith Eckle-Kohler (vgl. Eckle-Kohler (1999)) und sind Ende der neunziger Jahre am IMS entstanden. Die Listen liegen im ASCII-Format vor und konnten automatisch ins IMSLEX eingefügt werden. Die In-

[14] Meiner Meinung nach sollte die 'maximale' Information ausgegeben werden, also die letzte der dargestellten Varianten, so dass eine nachfolgende Anwendung die Information herausfiltern kann, die sie benötigt.

[15] Die Bezeichnung *EigennamenTyp* wäre evtl. angemessener gewesen, passt aber nicht in das generelle Schema der Unabhängigkeit der Module von Eigenschaften wie der Wortart.

[16] Es ist denkbar, dass sich die Art von Informationen auch im Internet finden lässt. Zum einen jedoch erfüllen automatisch generierte Listen nicht den Qualitätsanspruch des Lexikons, zum anderen stellt die unbesehene Übernahme von Informationen aus dem Internet in den meisten Fällen eine Urheberrechtsverletzung dar. Dies begründet die Notwendigkeit für ein eigenständig durchgeführtes Verfahren.

[17] Dass nur eine Währung im Lexikon vorkommt, liegt daran, dass Währungsbezeichnungen im DMOR-Lexikon zu den Substantiven zählen, also nicht im Eigennamenlexikon zu finden sind. Die Eintragung eines weiteren Vulkans ist in Anhang D dargestellt.

# in IMSLEX	Semantischer Typ	Beispiel
2	Geo: Berg	*Belchen, Kaiserstuhl*
29	Geo: Bewohner einer Stadt	*Tübinger*
33	Geo: Bewohner eines Landes	*Afrikaner*
241	Geo: Fluß, See, Gebirge, Region	*Walachei*
38	Geo: Insel	*Java*
6	Geo: Kontinent	*Australien*
251	Geo: Land	*Abchasien*
4358	Geo: Stadt	*Hamburg*
1	Geo: Stamm	*Issak*
2	Geo: Vulkan	*Pinatubo, Ätna*
128	NE: Firma	*Schwabenbräu*
2	NE: Gestirn	*Andromeda, Uranus*
10	NE: Märchenfigur	*Rumpelstilzchen*
1087	NE: Nachname	*Röntgen*
14	NE: Name männlich	*Caesar, Rembrandt*
18	NE: Name unbestimmt	*Prater, Walhalla*
13	NE: Namenszusatz	*Tel, Sri, San*
1172	NE: Vorname männlich	*Wolfgang*
1074	NE: Vorname weiblich	*Katharina*
1	NE: Währung	*Sterling*

Abbildung 7.6: 'Semantischer Typ' von Eigennamen in IMSLEX

formationen zur phonetischen Transkription der Stämme wurden vom Lehrstuhl für Experimentelle Phonetik der Universität Stuttgart zur Verfügung gestellt. Die Daten zur syntaktischen Verwendung von Adjektiven wurden Mitte der neunziger Jahre am IMS für ein EU-Projekt im Rahmen des 'Language Engineering'-Programmes erstellt. Es handelt sich um das Projekt PAROLE (*Preparatory Action for Linguistic Resources Organization for Language Engineering*, LE2-4017). Sie wurden als Inhalt in das Element *Verwendung* bei den wortartspezifischen Eigenschaften der Adjektive übernommen.

7.2 Lexikonverwendung und Pflege

Nachdem beschrieben wurde, welche verschiedenen Arten von Informationen auf welche Weise ins IMSLEX gelangt sind, muss die Frage beantwortet werden, wie auf diese Daten zugegriffen werden kann. In diesem Abschnitt geht es dabei allein um die 'menschlichen' Aspekte, also wie sich das System einem Benutzer darstellt und wie Änderungen am Datenbestand vorgenommen werden können. Der Aspekt des Auslesens für eine Verarbeitungskomponente wird im nachfolgenden Kapitel 8 ausführlich dargestellt.

Zweierlei Methoden des Zugriffs auf eine Ressource können unterschieden werden: der **lesende** Zugriff und der **schreibende** Zugriff. Für den lesenden Zugriff wurde ein Programm entwickelt, das die in den einzelnen Einträgen vorhandenen Informationen anzeigt. Dieses Programm, der **IMSLEX-Browser**, wird in Abschnitt 7.2.1 vorgestellt. Für den schreibenden Zugriff wurde ebenfalls ein Programm entwickelt, das die Bearbeitung einzelner Einträge erlaubt: **IMSLexEdit**.[18] Daneben gibt es kleine Perl-Programme, die in Form eines Benutzer/Programm-Dialogs die schnelle Erstellung eines kompletten Neueintrags ermöglichen. Die Perl-Programme werden in Abschnitt 7.2.2 vorgestellt.

7.2.1 Der IMSLEX-Browser

Auf eine elektronisch gespeicherte Ressource kann gewöhnlich über eine **Schnittstelle** zugegriffen werden. Diese versteckt implementatorische Details, ermöglicht die Überprüfung von Zugriffsrechten und kann vordefinierte Sichten auf die Daten anbieten. Diese drei Aspekte werden i.A. als Vorteile angesehen. Allerdings ist es schwierig, im Vorhinein die gewünschten Sichten auf die Daten festzulegen, zumal sich mit steigender Komplexität der Ressource immer mehr Möglichkeiten ergeben. Es besteht die Gefahr, dass Informationen, die eigentlich verfügbar sind, nicht eingesehen werden können.

Der Lexikonbrowser[19] sollte zwei Ziele erfüllen:

1. eine vollständige Sicht auf den Datenbestand,
2. die Verwendbarkeit ohne die Kenntnis der Datenstruktur des Lexikons.

Darüber hinaus sollte eine maximale Unabhängigkeit von der Struktur der Ressource eingehalten werden, damit eine Strukturveränderung keine Anpassung des Programms erfordert. Diese Ziele waren mit der Wahl von XML als Repräsentationsformat relativ leicht zu erzielen. Anhand zweier 'Screenshots' soll die Funktionsweise im Folgenden erläutert werden.

Das Hauptenster – die Makrostruktur

In Abbildung 7.7 ist das Hauptfenster des Zugriffsprogramms dargestellt.[20] Dieses besteht aus drei Teilen. Diese Information sowie die, was im jeweiligen Teilfenster gezeigt wird, sind in einer kleinen Konfigurationsdatei enthalten, die ebenfalls in XML repräsentiert ist.

[18] Die Verschmelzung der beiden Programme in den IMSLEX-Browser wird derzeit durchgeführt.

[19] Der Begriff des 'Browsers' wird in Ermangelung eines verständlichen deutschen Pendants (*Blätterer* wäre die direkte Übersetzung) beibehalten.

[20] Die Programmierung erfolgte durch André Blessing in der Programmiersprache Java.

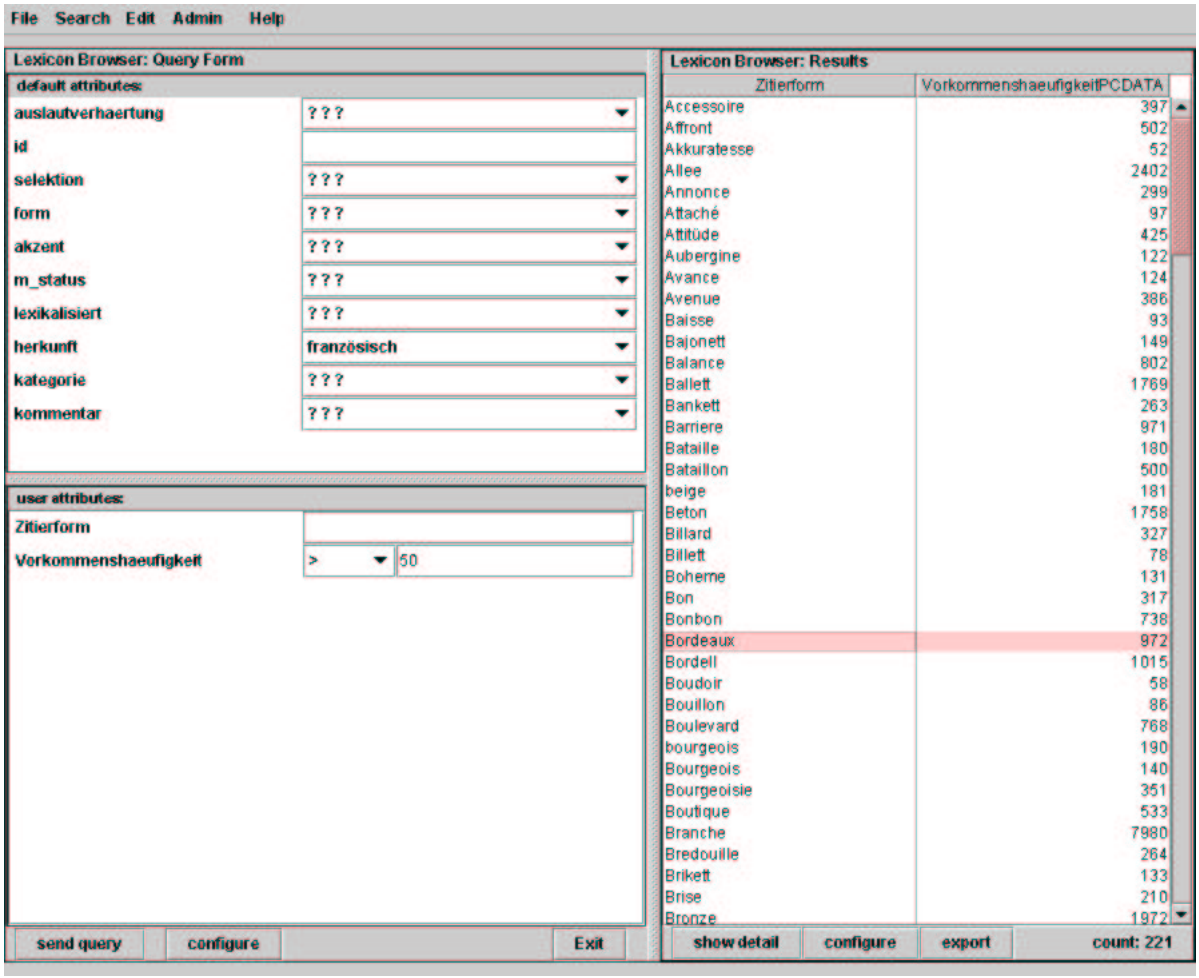

Abbildung 7.7: IMSLexApp – Ein Lexikonbrowser, Hauptfenster

Abbildung 7.8 zeigt den Teil der Konfigurationsdatei, der den Inhalt des linken unteren Teilfensters beschreibt, das hier mit dem Elementnamen *Search* bezeichnet wird (Suchfenster). Für jede Zeile in diesem Fenster gibt es ein Element *Searchitem*, dessen Attribut **name** jeweils vorgibt, wonach gesucht werden kann. In den *Pathobject*-Elementen wird der absolute Pfad im XML-Dokument zu dem Element hergestellt, dessen Inhalt durchsucht werden kann. Beide Elemente befinden sich innerhalb der Struktur von *Globale_Merkmale* (vgl. Abbildung 6.3 auf Seite 82). Wird die DTD geändert, müssen allein in dieser Konfigurationsdatei die beiden Pfade angepasst werden, der Programmcode kann vollständig unverändert bleiben.

Abbildung 7.9 zeigt den Teil der Konfigurationsdatei, der den Inhalt des rechten Teilfensters beschreibt, das hier mit dem Elementnamen *Content* bezeichnet wird (Ergebnisfenster). Es werden wiederum die beiden Elemente vor-

```
<Search>
    <Searchitem name="Zitierform" type="STRING">
        <Pathobject>lexikon</Pathobject>
        <Pathobject>le</Pathobject>
        <Pathobject>Globale_Merkmale</Pathobject>
    </Searchitem>
    <Searchitem name="Vorkommenshaeufigkeit" type="NUMBER">
        <Pathobject>lexikon</Pathobject>
        <Pathobject>le</Pathobject>
        <Pathobject>Globale_Merkmale</Pathobject>
    </Searchitem>
</Search>
```

Abbildung 7.8: Die XML-Konfigurationsdatei für das Suchfenster

```
<Listbox>
    <Content name="Zitierform" type="STRING" sort="YES">
        <Pathobject>lexikon</Pathobject>
        <Pathobject>le</Pathobject>
        <Pathobject>Globale_Merkmale</Pathobject>
    </Content>
    <Content name="Vorkommenshaeufigkeit" type="NUMBER" sort="NO">
        <Pathobject>lexikon</Pathobject>
        <Pathobject>le</Pathobject>
        <Pathobject>Globale_Merkmale</Pathobject>
    </Content>
</Listbox>
```

Abbildung 7.9: Die XML-Konfigurationsdatei für das Ergebnisfenster

gegeben, deren Inhalte in diesem Fall angezeigt werden. Das Attribut **sort** in der Belegung `YES` gibt an, dass initial nach der Zitierform alphabetisch (wegen **type**:`STRING`) sortiert wird.[21]

Das linke obere Teilfenster in Abbildung 7.7 ist das einzige, dessen Inhalt fest vorgegeben ist. Allerdings sind nicht etwa die einzelnen Attributnamen im Programmcode oder in einer Konfigurationsdatei aufgelistet, sondern es ist vorgegeben, dass in diesem Fenster alle Attribute des Elements *le* aufgelistet werden.[22] Die drei Fragezeichen rechts von den Attributnamen deuten an, dass sich hier weitere Informationen verbergen. Bei einem Mausklick auf eine der Flächen werden alle Werte angezeigt, die das Attribut laut DTD annehmen kann.

[21] Durch Mausklick auf das graue Feld im Hauptfenster mit der Bezeichnung `Vorkommenshaeufigkeit` wird nach dieser aufsteigend sortiert. Durch Mausklick in Verbindung mit der `Shift`-Taste wird absteigend sortiert, bei `Zitierform` analog.

[22] Ändern sich die Attribute zwischen zwei Programmaufrufen, so wird beim zweiten Aufruf der neue Stand angezeigt, da die Attribute und ihre möglichen Werte jedes Mal bei Programmstart aus der DTD ausgelesen werden.

Im Beispiel wurde die Belegung **herkunft**:`französisch` ausgewählt. Wird nun auf `send query` links unten im Fenster geklickt, erscheinen im Ergebnisfenster rechts alle Einträge aus dem IMSLEX, bei denen das Attribut **herkunft** mit dem Wert `französisch` versehen ist, alphabetisch sortiert und mit Vorkommenshäufigkeit im HGC.[23] Im Suchfenster kann nach beliebigen Zeichenketten gesucht werden, also z.B. auch nach allen Einträgen, die mit *be* beginnen, mit *ung* enden oder die Sequenz *eau* enthalten.[24]

Wird in den beiden links angeordneten Fenstern keine Aktion durchgeführt, so werden bei einem Mausklick auf `send query` im Ergebnisfenster **alle Zitierformen** aus dem IMSLEX mitsamt ihrer Vorkommenshäufigkeit angezeigt. Auf diese Weise ist das Ziel, die gesamte Makrostruktur transparent zu machen, erreicht. Sämtliche Attribute und Attributwerte einer lexikalischen Einheit sowie sämtliche lexikalischen Einheiten sind ohne jede Kenntnis der Struktur des Lexikons einsehbar.

Das Detailfenster – die Mikrostruktur

Im Ergebnisfenster in Abbildung 7.7 ist eine Zeile (durch Mausklick) farbig unterlegt worden. Wird auf `show detail` am unteren Fensterrand geklickt, so wird von der Ansicht der Makrostruktur in die Ansicht der Mikrostruktur gewechselt.

In Abbildung 7.10 ist ein Beispiel für ein Detailfenster des IMSLEX-Browsers dargestellt. In diesem ist im linken Teil die komplette *Eintragsstruktur* dargestellt, wobei verschiedene Entitäten farblich verschieden dargestellt sind. Die Strukturansicht ist 'ausklappbar' gehalten, es kann also je nach Bedarf eine Hierarchiestufe sichtbar gemacht oder wieder 'eingeklappt' werden. Im Beispiel[25] wurde das Element *Flexionsmorphologie* bis zum Element *Stammform* hinunter 'aufgeklappt'. Im rechten Teilfenster wird ein im linken Teilfenster markiertes Element automatisch farbig unterlegt. Damit ist die Verbindung zwischen beiden Teilfenstern kenntlich gemacht.

Alle im rechten Teilfenster dargestellten Informationen können geändert werden. Bei der Fläche mit dem kleinen schwarzen auf dem Kopf stehenden Dreieck handelt es sich wieder um ein sogenanntes 'Pull-down'-Menü, das bei Mausklick alle bei diesem Attribut möglichen Attributwerte anzeigt. Bei den anderen Flächen handelt es sich um Textfenster, in denen beliebige Änderungen vollzogen werden können. Allerdings werden diese Änderungen nur dann in die

[23] Im Suchfenster unten links wurde die Vorkommenshäufigkeit noch auf alle Zitierformen eingeschränkt, deren Wortformen häufiger als 50 Mal im Korpus vorkommen.

[24] Es werden reguläre Ausdrücke verwendet, daher sind die Suchmöglichkeiten nahezu unbegrenzt.

[25] Diese Darstellung spiegelt leider noch eine Vorversion des aktuellen Lexikons wider, in der *Globale_Merkmale* alle anderen Elemente umschließen. Dies ändert aber nichts daran, dass hier die Mikrostruktur in zwei verschiedenen Sichten erfasst wird.

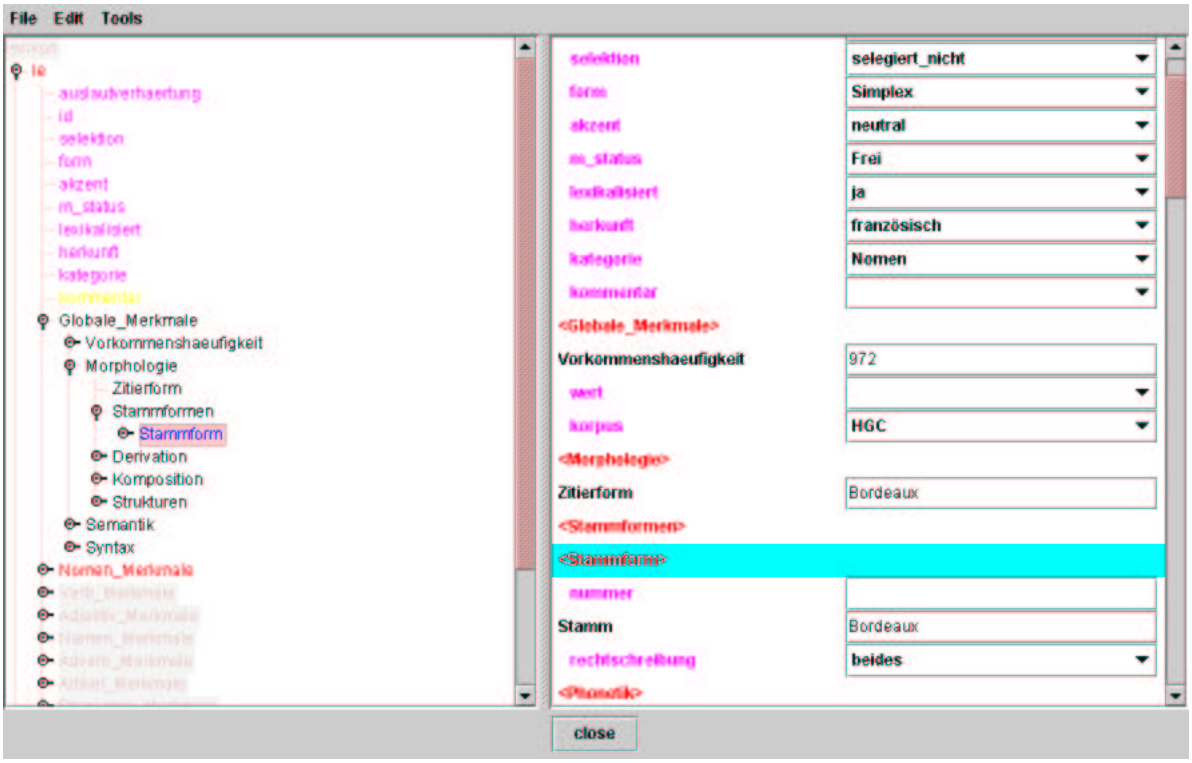

Abbildung 7.10: IMSLexApp – Ein Lexikonbrowser, Detailfenster

Datenbank[26] zurückgeschrieben, wenn das Programm im Administrator-Modus aufgerufen wurde.

7.2.2 Lexikonerweiterung

Da es sich bei IMSLEX-Edit um eine Vorgängerversion des IMSLEX-Browsers handelt, der im vorangegangenen Abschnitt vorgestellt wurde, wird hier auf eine Darstellung verzichtet. IMSLEX-Edit wurde im DeKo-Projekt (vgl. Abschnitt 5.1) eingesetzt, um die Einträge für Derivationsaffixe vorzunehmen. Derzeit finden Lexikonerweiterungen und -änderungen direkt auf den XML-Dateien statt. Im Folgenden werden die typischen Vorgänge beschrieben.

Zwei Arten von Lexikonerweiterung sind zu unterscheiden: das Hinzufügen von Informationen in der Mikrostruktur und die Erweiterung der Makrostruktur. Das Hinzufügen von Informationen in der Mikrostruktur verhält sich wie

[26]Der IMSLEX-Browser setzt nicht direkt auf dem XML-Dateien auf, sondern auf einer relationalen Datenbank, die ebenfalls unabhängig von der Datenstruktur aus den XML-Dateien erzeugt wurde. Dieser Aspekt ist aber lediglich wichtig für die Geschwindigkeit des Zugriffs auf die Daten, nicht für die Prinzipien bei der Programmierung, und wurde daher hier nicht gesondert erwähnt.

die in Abschnitt 7.1.5 beschriebenen Aktionen, also das nachträgliche Hinzufügen von Informationen in bereits bestehende Einträge. Dieses nachträgliche Hinzufügen gestaltet sich oft nach Listen, also beispielsweise durch eine Suche nach potentiellen Kompositionsstammformen speziell für lexikalische Einheiten, bei denen diese noch nicht eingetragen sind. Daher ist es relativ schwierig, ein Programm zu konzipieren, das alle potentiellen Vorgehensweisen zum systematischen Erweitern der vorhandenen Einträge umfasst. Aus diesem Grund werden bislang kleine Ad-hoc-Programme geschrieben, die eine bestimmte Aufgabe erledigen und dann ggf. als Vorlage für ein anderes Programm dienen.

Bei der Erweiterung der Makrostruktur müssen vollständige Einträge neu erzeugt werden. Dies stellt sich so dar wie der Zustand beim Erzeugen der Ressource, mit dem Unterschied, dass nicht mehr Teile der für einen Eintrag benötigten Information aus einer anderen Ressource entnommen werden können. Aus diesem Grund sind hier – wie auch bei Erweiterung bestehender Einträge – oft interaktive Programme am besten geeignet. Ein Beispiel für ein solches interaktives Perl-Programm ist in Anhang E (vgl. S. 157) angegeben. Es realisiert einen Dialog mit einem Benutzer, der ein neues Lexem in das Lexikon eintragen möchte. Die relevanten Attribute werden abgefragt, wenn möglich, werden Daten automatisch erzeugt, und wenn am Ende der Neueintrag bestätigt wird, wird eine vollständige XML-Struktur in eine Datei geschrieben. Nach Ablauf aller Eintragevorgänge kann diese Datei dann in die entsprechende XML-Datei kopiert werden.[27]

Zur Veranschaulichung der Funktionsweise des Programms sind in Anhang D zwei Pflegedialoge für je einen Substanitv- und einen Eigennameneintrag wiedergegeben (vgl. S. 151).

7.3 IMSLEX: Zusammenfassung

In den vorangehenden Abschnitten wurde ausführlich dargestellt, wie aus der Strukturdefinition des Lexikons eine Lexikoninstanz generiert und wie diese Instanz mit Inhalten aufgefüllt wurde. Es wurde weiterhin gezeigt, wie die Lexikondaten angesehen und verändert werden können.

In den beiden folgenden Abschnitten wird eine zusammenfassende Übersicht über einige Aspekte des IMSLEX gegeben: die Anzahl der je Wortart in IMSLEX aktuell gespeicherten lexikalischen Einheiten, die Zusammenhänge zwischen Modulen und Wortarten in der Mikrostruktur des Lexikons sowie abschließend eine Einordnung in ein Wörterbuchmodell, das den Vergleich des IMSLEX mit anderen maschinenlesbaren Wörterbüchern erleichtern soll. Das dafür verwendete Wörterbuchmodell wurde Heid (1997) entnommen.

[27] Dies ist unkomplizierter, als es klingt. Dadurch, dass feste Attributwerte vorgegeben werden, wird die Möglichkeit von Eingabefehlern reduziert.

7.3.1 Statistik und Übersicht der Module

Trotz der flachen Struktur des IMSLEX sind durch die verschiedenen Wortarten, die Unterscheidung nach offenen und geschlossenen Klassen oder die Trennung in flektierende und nicht-flektierende Klassen (vgl. Abschnitt 2.1.1) gewisse Gliederungsmöglichkeiten vorgegeben, die sich auf die Auswahl der Module in einem Eintrag auswirken. Ausgehend von den Wortarten, die im **STTS** unterschieden werden, werden im Folgenden in einer Übersicht die Beziehungen zwischen Wortarten und Modulen im IMSLEX dargestellt.

Kategorie	# Lexeme	STTS-Wortart	G	F	W	Syn	Sem	Spez
Substantiv	22.717	Nomina	+	+	+	+	–	+
Name	8.491	Nomina	+	+	+	–	+	–
Adjektiv	11.051	Adjektive	+	+	+	+	–	+
Verb	5.813	Verben	+	+	+	+	–	–
Partikelverb	6.394	(Verben)	+	+	+	+	–	–
Pronomen	103	Pronomina	+	+	+	–	–	–
Artikel	2	Artikel	+	+	–	–	–	–
Numeral	32	Kardinalzahlen	+	+	(+)	–	–	–
Adverb	1.095	Adverbien	+	+	+	–	–	–
Adposition	162	Adpositionen	+	+	–	+	–	–
Konjunktion	67	Konjunktionen	+	+	–	–	–	–
Interjektion	27	Interjektionen	+	+	–	–	–	–
Partikel	9	Partikeln	+	+	+	–	–	–
Substantivpraefix	43	–	+	+	+	–	–	+
Adjektivpraefix	30	–	+	+	+	–	–	+
Verbpraefix	21	–	+	+	+	–	–	+
Verbpartikel	387	–	+	+	+	–	–	+
Substantivsuffix	49	–	+	+	+	–	–	+
Adjektivsuffix	100	–	+	+	+	–	–	+
Verbsuffix	5	–	+	+	+	–	–	+
Adverbsuffix	13	–	+	+	+	–	–	+
Konfix	223	–	+	+	+	–	–	–
Substantiv_Abk	364	(Nomina)	+	+	+	–	–	+
Name_Abk	1.857	(Nomina)	+	+	+	–	–	+
Adjektiv_Abk	381	(Adjektive)	+	+	–	–	–	+
Invar_Abk	93	(diverse)	+	+	–	–	–	+

Abbildung 7.11: Kategorien, Wortarten und Module in IMSLEX

Der Zusammenhang zwischen den Kategorien der lexikalischen Einheiten, den Wortarten im STTS und den Modulen ist in Tabelle 7.11 dargestellt. Die Module sind *Globale Merkmale* (G), *Flexionsmorphologie* (F), *Wortbildung* (W), *Syntax* (Syn), *Semantik* (Sem) und *wortartspezifische Merkmale (Spez)*. Die Zah-

len in der zweiten Spalte geben die Anzahl der Lexeme je Kategorie im IMSLEX an (Stand April 2004). Das STTS nennt explizit elf Wortarten: Nomina, Verben, Artikel, Adjektive, Pronomina, Kardinalzahlen, Adverbien, Konjunktionen, Adpositionen, Interjektionen, Partikeln (vgl. Schiller et al. (1999), S. 4).[28]

Bei den Kategorien frei vorkommender lexikalischer Einheiten stimmen die Bezeichner im Wesentlichen mit denen der STTS-Wortart überein. Anstelle der allgemeineren Bezeichnung *Nomina* werden in IMSLEX *Substantiv* und *Name* verwendet. Partikelverben werden in IMSLEX im Gegensatz zum DMOR-Lexikon aufgelistet. Dies hat zwei Gründe: Zum einen verfügen Partikelverben über Subkategorisierungsrahmen, zum anderen bilden sie häufig die Basis bei Wortbildungen, so dass für sie Derivations- und Kompositionsstammformen eingetragen werden müssen.

Die globalen Merkmale dürfen als gewissermaßen konstituierende Merkmale einer lexikalischen Einheit bei keiner Kategorie fehlen. Auch das Modul der Flexionsmorphologie[29] ist obligatorisch. Bei den im DMOR-Lexikon nicht enthaltenen Präfixen und Konfixen dient dieses Modul wie bei den nichtflektierenden Klassen nur dem Auslesen von Stammform und Wortart[30]. Derivationssuffixe hingegen erhalten eine Flexionsklasse, da sie die morphologischen Eigenschaften des Derivats bestimmen. Bei Partikelverben werden nur die Wortbildungsstammformen ausgelesen, da die Bildung der Partikelverben über die separat gespeicherten Verbpartikel kombiniert mit den Verben abläuft.

Das Modul zur Wortbildung enthält die Derivations- und Kompositionsstammformen einer lexikalischen Einheit und bei morphologisch komplexen Einheiten die morphologische Struktur (oder Zerlegung). Einige Vertreter der geschlossenen Wortklassen nehmen nicht an Wortbildung teil. Abkürzungen können als Basis für das Derivationssuffix *-ler* herangezogen werden (*DDRler, ABMler*). Bei den Partikeln sind es die Antwortpartikel, die in Kompositionen angetroffen werden können (*Jasager, Neinstimme*). Pronomina treten häufig als mit Bindestrich abgetrennte Erstglieder auf, aber vereinzelt sind auch zusammengeschriebene Formen im Korpus belegt (*Ichfunktion*$_{(8)}$, *Ichform*$_{(8)}$, *Wirgefühl*$_{(1)}$). Numeralia bilden bezüglich der Wortbildung eine eigene Klasse mit eigenen Bildungsregeln, auf die hier nicht näher eingegangen wird.

Das Modul für die Syntax sieht derzeit ausschließlich Informationen zur Subkategorisierung vor.

Das Modul für Semantik enthält zur Zeit ausschließlich Informationen zu Eigennamen. Diese beziehen sich auf geographische Namen, bei Substantiven auch auf die Bewohner von Städten, Ländern und Regionen. Eine Ontologie wie

[28] Abkürzungen tragen die Wortart ihrer ausgeschriebenen Form bzw., wenn es sich um zusammengeschriebene Abkürzungen von Mehrwortlexemen handelt, ihrer "syntaktischen Funktion" (Schiller et al. (1999), S. 9).

[29] Der Begriff umfasst hier auch nicht-flektierende Klassen, vgl. Abschnitt 3.2.1.

[30] Da Präfixe nicht über eine Wortart verfügen, wird hier eine Phantasieform `PRAEF` verwendet.

in CISLEX oder eine Auszeichnung mit semantischen Verbklassen sind derzeit nicht geplant.

Das letzte Modul schließlich enthält Informationen, die spezifisch für eine Kategorie sind. Bei Substantiven (und Substantivsuffixen) ist dies die Paradigmenkategorisierung *Genus*. Bei Adjektiven (und Adjektivsuffixen) ist es die Tatsache, ob ein Adjektiv nur attributiv oder prädikativ verwendet werden kann oder beides. Bei Derivationssuffixen sind hier die Einschränkungen für die Wahl der Basen in Form von Merkmalen verzeichnet. Bei Präfixen gilt dies in eingeschränkter Form ebenso: Hier werden typische Verbpräfixe von solchen unterschieden, die sich mit Substantiven und Adjektiven verbinden. Bei den Verbpartikeln ist die Klasse nach Aldinger (2002) angegeben, und bei den Abkürzungen kann die ausgeschriebene Form verzeichnet werden.

7.3.2 Einordnung in ein Wörterbuchmodell

Heid führt fünf allgemeine Beschreibungskriterien für elektronische Wörterbücher auf (vgl. Heid (1997), S. 9ff.), die im Folgenden für die Einordnung von IMSLEX genutzt werden sollen: “Anwendungsorientierung”, “inhaltliche Beschreibung”, “formale Organisation”, “technische Eigenschaften” und “Zusammenhang [. . .] mit anderen [. . .] Ressourcen” (Heid (1997), S. 9).

“[D]ie Anwendungsorientierung eines Wörterbuchs bezeichnet die angestrebte hauptsächliche Benutzung, die der Wörterbuchentwickler für das Wörterbuch vorsieht.” (ebd., S. 10) In Falle des IMSLEX handelt es sich um eine Ressource, die den Datenbestand für die Erzeugung eines Systems zur automatischen morphologischen Analyse des Deutschen zur Verfügung stellt. Sie ist allerdings so flexibel und modular gestaltet, dass der in ihr gespeicherte Datenbestand auch von anderen computerlinguistischen Anwendungen genutzt werden kann.

Bei der inhaltlichen Beschreibung “sind makrostrukturelle und mikrostrukturelle Aspekte zu unterscheiden” (ebd., S. 10). In der Makrostruktur enthält das Lexikon ca. 60 000 Lexeme zu Wortformen, die in einem großen Korpus deutschsprachiger Zeitungstexte vorkommen. Die Zeitungen stammen aus den Jahren 1988 bis 1994, und das Korpus umfasst 200 Millionen Token. Neben den Lexemen umfasst die Makrostruktur ca. 260 Derivationsaffixe. Mehrwortlexeme sind erst in Ansätzen vorhanden. Die Gruppierung der Einträge geschieht nach Wortarten. In der Mikrostruktur wird differenziert nach Wortart und Flexionsklasse, weiterhin nach morphologischer Form (einfach oder komplex), Fähigkeit zur Selektion (Affix vs. Stamm) und Herkunft. Neben der orthographischen Form wird eine phonetische Transkription angegeben, weiterhin Flexions-, Derivations- und Kompositionsstammformen und schließlich Subkategorisierungsinformationen für Verben, Substantive und Adjektive. Eigennamen sind nach semantischen Kriterien gegliedert. Durch die Verwendung von

Stammformen wird eine rein konkatenativ ablaufende morphologische Analyse unterstützt.

Das Lexikon ist explizit organisiert, d.h., jeder Angabetyp ist eindeutig identifizierbar. Als Repräsentationsformat wird XML verwendet, so dass die Übereinstimmung der syntaktischen Struktur der Ressource mit der Strukturdefinition automatisch überprüft werden kann. Die strukturelle Konsistenz der Ressource ist damit gewährleistet.

Das Lexikon liegt in Form von Dateien vor, die grob nach Wortart, Typ des Lexems oder Affixes bzw. einer Mischung aus beidem untergliedert sind. Die größte Datei, die der Substantive, umfasst 25 Megabyte (MB) an Daten, insgesamt umfasst die Ressource etwas über 60 MB. Die Größe erklärt sich hauptsächlich durch den Verzicht auf platzsparende Abkürzungen bei der im XML-Format reichlich vorhandenen Markup-Information. Komprimiert (`gzip`) umfasst das gesamte Lexikon etwas über zwei MB Daten. Die Daten können vollautomatisch in eine relationale Datenbank (`mysql`) übertragen und wieder ausgelesen werden. Neben der Morphologiekomponente bietet die Datenbank eine zweite (und direkte) Zugriffsschnittstelle.

Das IMSLEX ist aus dem Datenbestand der Lexikondateien für das Lexikon- und Regelsystem DMOR entstanden, das am IMS für das Morphologiesystem PC-Kimmo entworfen wurde (vgl. Schiller (1996, 1995)). Als Quelle für eine derivations- und kompositionsmorphologische Analysekomponente wurde es mit einem neuen und erweiterten Datenmodell versehen. Seither wird die Mikrostruktur beständig aufgefüllt, während die Makrostruktur in Schüben nach Wortbildungsphänomenen systematisch erweitert wird.

Kapitel 8

Zusammenspiel von IMSLEX und Morphologiekomponente

In diesem Kapitel schließt sich der Kreis des Wartungszyklus. Morphologische Einheiten und Prozesse wurden erklärt, ein Lexikon wurde konzipiert und realisiert, nun muss noch beschrieben werden, wie die im Lexikon gespeicherten Einheiten der Morphologiekomponente wieder zugute kommen. Zunächst werden 'Stylesheets' vorgestellt, die eine Daten-Schnittstelle zwischen dem Lexikon und nachfolgenden (automatischen) Verarbeitungsstufen erzeugen (vgl. Abschnitt 8.1). Im Anschluss daran wird der Zusammenhang von Lexikon und Morphologiekomponente im Hinblick auf über die Schnittstelle hinausgehende Abhängigkeiten zwischen den beiden Komponenten diskutiert (vgl. Abschnitt 8.2). Zum Abschluss des Kapitels wird gezeigt, wie durch eine saubere konzeptionelle Trennung zwischen konkatenativ und nicht-konkatenativ beschreibbaren Phänomenen im Lexikon die Qualität der morphologischen Analyse erhöht werden kann (vgl. Abschnitt 8.3).

8.1 Auslesen des Lexikons

Der ursprüngliche Zweck des DMOR-Lexikons, die Zurverfügungstellung von lexikalischen Einheiten, die die Morphologiekomponente zur Durchführung der morphologischen Analyse von Wortformen benötigt, ist auch der Hauptzweck des IMSLEX.[1] Es gibt i.A. zwei Möglichkeiten der Realisierung einer Schnittstelle zwischen zwei Komponenten, zum einen den Zugriff über Funktions- oder Prozeduraufrufe, zum anderen die Definition eines Daten-Austauschformats und die Erzeugung dieses Formats durch eine Komponente. Da es im diesem

[1] Dass es daneben mit seinem Lexemkonzept mittlerweile die Möglichkeit bietet, weitere über die Morphologie hinausgehende linguistische Informationen in einer gemeinsamen Ressource zu speichern, wurde im vorangegangenen Kapitel gezeigt. In diesem Kapitel geht es ausschließlich um die morphologische Information.

Fall um die Weitergabe statischer Daten geht, also Einheiten ausgelesen werden, die im Lexikon vorhanden sind und nicht erst explizit errechnet werden müssen, wurde die zweite Variante gewählt. Die für die Morphologiekomponente SMOR (vgl. Schmid et al. (2004)) relevanten Daten werden mit Hilfe von *XSLT-Stylesheets* aus dem Lexikon ausgelesen und in eine Datei geschrieben, die die Morphologiekomponente wiederum einliest. Die in dieser Datei enthaltenen Informationen und das verwendete Format definieren die Schnittstelle zwischen den beiden Komponenten.

8.1.1 XSLT-Stylesheets

Bei der *eXtensible Style Sheet Language for Transformations* (**XSLT**, vgl. Clark (1999)) handelt es sich um einen standardisierten Verarbeitungsmechanismus für XML-Dokumente, der den deklarativen[2] Zugriff auf die Einheiten eines XML-Dokuments erlaubt. Dies hat zur Folge, dass auch bei den Stylesheets, wie schon beim IMSLEX-Browser (vgl. Abschnitt 7.2.1), eine maximal mögliche Unabhängigkeit von der Struktur der Ressource besteht: Informationen, auf die nicht zugegriffen werden muss, können in der Ressource neu angeordnet werden, ohne dass dies einen Einfluss auf die Funktionsfähigkeit der Stylesheets oder die Daten-Schnittstelle hat.

Für das Auslesen der verschiedenen Stammformentypen gibt es zwei Stylesheets, die sehr ähnlich arbeiten, aber ihre Informationen aus zwei unterschiedlichen Modulen eines XML-Eintrags beziehen: Das Stylesheet für die Flexionsinformation greift auf die Merkmale einer lexikalischen Einheit sowie auf Inhalte im Element *Flexionsmorphologie* zu, das Stylesheet für die Wortbildungsinformation greift auf die Merkmale einer lexikalischen Einheit sowie auf Inhalte im Element *Wortbildung* zu. Die Stylesheets werden in den beiden folgenden Abschnitten beschrieben und sind in Anhang F komplett abgedruckt (vgl. S. 163).

8.1.2 Stylesheet für die Flexionsinformation

Anhand dieses Stylesheets soll die Funktionsweise des Auslesens erläutert werden.

In Abbildung 8.1 ist der Anfang des Stylesheets, die Verarbeitung des *le*-Elements, dargestellt. Ein Stylesheet-Prozessor durchläuft das XML-Dokument und sucht nach passenden Elementnamen (`match` in der Abbildung). Daher wird zunächst das Wurzelelement angegeben, also *lexikon* (vgl. Abbildung 6.2

[2] 'Deklarativ' im Sinne eines Programmierprinzips: In einem Stylesheet wird nicht angegeben, **wie** die Information extrahiert werden soll, sondern **welche** Information in welches Format überführt werden soll. Einheiten, auf die nicht zugegriffen werden muss, werden auch nicht berücksichtigt.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
<xsl:output method="text" encoding="ISO-8859-1"/>

<xsl:template match="lexikon">
    <xsl:apply-templates select="le"/>
</xsl:template>

</xsl:stylesheet>
```

Abbildung 8.1: XSLT-Stylesheet für Flexion – *lexikon*-Element

auf Seite 82). Von diesem aus sollen alle lexikalischen Einheiten bearbeitet werden, also wird das 'Template' für das Element *le* aufgerufen.

```
<xsl:template match="le">
    <xsl:variable name="katsymbol">
        <xsl:call-template name="ersetze">
            <xsl:with-param name="quelle" select="kategorie"/>
        </xsl:call-template>
    </xsl:variable>
    <xsl:apply-templates select="Flexionsmorphologie/Stammformen/Stammform">
        <xsl:with-param name="kat" select="$katsymbol" />
        <xsl:with-param name="herk" select="herkunft" />
        <xsl:with-param name="form" select="form" />
        <xsl:with-param name="stamm"
            select="Flexionsmorphologie/Stammformen/DMORStamm" />
    </xsl:apply-templates>
</xsl:template>
```

Abbildung 8.2: XSLT-Stylesheet für Flexion – *le*-Element

Im *le*-Element werden die Attributwerte aufgesammelt, die später ausgegeben werden sollen (vgl. Abbildung 8.2). Zunächst wird ein weiteres Template aufgerufen, das das **kategorie**-Attribut der lexikalischen Einheiten durch ein Kürzel ersetzt (vgl. Anhang F). Es ist fortan als Variable `$katsymbol` im Stylesheet verfügbar. Neben dem Attribut **kategorie** werden noch die Attribute **herkunft** und (morphologische) **form** ausgelesen. Zusätzlich dazu wird der Inhalt des **DMORStamm**-Elements ausgelesen.[3] Schließlich wird ein Template für **alle**[4] *Stammform*-Elemente, die sich in der Hierarchie unterhalb des *Flexionsmorphologie*-Elements und des *Stammformen*-Elements befinden, aufgerufen. Die Parameter werden beim Templateaufruf übergeben.

[3] Dass an dieser Stelle nicht getestet wird, ob die Elemente oder Attribute vorhanden sind, liegt daran, dass sie allesamt in der DTD als obligatorisch definiert sind. Dies stellt allerdings eine Abhängigkeit dar.

[4] Dieser Aufruf ist **rekursiv** und erfasst daher alle *Stammform*-Elemente.

```
<xsl:template match="Stammform">
    <xsl:param name="kat" select="'FEHLER'" />
    <xsl:param name="herk" select="'FEHLER'" />
    <xsl:param name="form" select="'FEHLER'" />
    <xsl:param name="stamm" select="'FEHLER'" />

    <xsl:apply-templates select="../../../Flexionsmorphologie">
        <xsl:with-param name="kat" select="$kat" />
    </xsl:apply-templates>
    <xsl:choose>
        <xsl:when test="DMORtyp='irreg'">
            <xsl:value-of select="$stamm"/><xsl:text>:</xsl:text>
        </xsl:when>
    </xsl:choose>
    <xsl:value-of select="./Stamm"/>
    <xsl:text>&#60;</xsl:text><xsl:value-of select="$kat"/><xsl:text>&#62;</xsl:text>
    <xsl:text>&#60;base&#62;</xsl:text>
    <xsl:text>&#60;</xsl:text><xsl:value-of select="$herk"/><xsl:text>&#62;</xsl:text>
    <xsl:text>&#60;</xsl:text><xsl:value-of select="$form"/><xsl:text>&#62;</xsl:text>
    <xsl:text>&#60;</xsl:text><xsl:value-of select="./DMORklasse"/>
        <xsl:text>&#62;</xsl:text>
    <xsl:text>&#10;</xsl:text>
</xsl:template>
```

Abbildung 8.3: XSLT-Stylesheet für Flexion – *Stammform*-Element

Die Verarbeitung des *stammform*-Elements ist in Abbildung 8.3 dargestellt. Am Anfang werden die Parameter aufgelistet, die beim Aufruf übergeben wurden. Ist ein Parameter nicht definiert, wird in die Ausgabe die Zeichenkette hinter `select` übernommen.[5] Dann wird das Template für *Flexionsmorphologie* aufgerufen.[6] Danach findet ein Test statt (`xsl:choose`): Hat ein Eintrag beim Attribut **DMORtyp** den Wert `irreg`, so wird hier der DMOR-Stamm, gefolgt von einem Doppelpunkt, ausgegeben, weil danach eine unregelmäßige Form folgt (z.B. *back:buk*). Nach Verlassen der `xsl:choose`-Anweisung wird der Stamm ausgegeben, gefolgt von den Angaben für die Kategorie, den Typ des Stammes (hier einfach `base` zur Unterscheidung von den Wortbildungsstämmen), die Herkunft, die morphologische Form und schließlich die Flexionsklasse. Vor und hinter jeder Information wird eine öffnende und schließende spitze Klammer gesetzt.[7] Als letztes wird noch ein Zeilenumbruch-Zeichen ausgegeben (`&#10`), dann ist das Template abgearbeitet.

[5] Da dieses Stylesheet recht klein ist, reicht in diesem Fall der sehr unspezifische String *FEHLER* aus.

[6] Dieser Aufruf setzt für die Ausgabe das Kategoriekürzel mit der Zeichenkette `_Stems` zusammen, vgl. Anhang F, S. 163.

[7] Diese verbergen sich hinter ihrer Nummer im Zeichensatz, 60 für '<' und 62 für '>'. Der Zeichensatz wird am Anfang des Dokuments angegeben, vgl. Abbildung 8.1.

Ausgabe des Stylesheets			Zeile
<NN_Stems>	*Achtbarkeit*	<NN><base><nativ><Komplex><NFem-Deriv>	1
<NE_Stems>	*Siegen*	<NE><base><nativ><Simplex><NGeo-Neut+Loc>	2
<ADJ_Stems>	*evokativ*	<ADJ><base><undef><undef><Adj+>	3
<ADJ_Stems>	*exakt*	<ADJ><base><klassisch><Simplex><Adj+e>	4
<V-ge_Stems>	*beiz*	<V><base><nativ><Simplex><VVReg>	5
<V-ge_Stems>	*beiß:biss*	<V><base><nativ><Simplex><VVPP-en>	6
<ADP_Stems>	*abzüglich*	<ADP><base><undef><undef><Prep-Gen>	7
<ADP_Stems>	*an*	<ADP><base><nativ><Simplex><Prep-Dat>	8
<VPrefSep>	*gegen*	<VPART><base><nativ><Simplex><Pref/Sep>	9
<NN_Abbr>	*Mwst.*	<ABK><base><undef><undef><Abk_NN>	10

Abbildung 8.4: Stylesheet-Ausgabe für die Flexionsmorphologie

Die Ausgabe des Stylesheets sieht aus wie in Abbildung 8.4 dargestellt.[8] In Zeilen 3, 7 und 10 ist erkennbar, dass die Arbeiten am Lexikon noch nicht abgeschlossen sind: Hier sind Herkunft und morphologische Form noch undefiniert. In Zeile 6 ist ein Beispiel für einen unregelmäßigen Flexionsstamm aufgeführt. An der Flexionsklasse lässt sich erkennen, dass es sich um den Partizip-Stamm handelt (*ge<u>biss</u>en*). Die Ausgaben in der ersten Spalte geben das DMOR-Sublexikon wieder, das auch in SMOR dazu verwendet wird, die Wortbildung zu steuern (vgl. Abschnitt 3.2.1). In Zeile 9 ist ein Beispiel für eine trennbare Verbpartikel gegeben (*gegen* für Partikelverben wie *gegensteuern*$^{\mathbb{P}}$, *gegenzeichnen*$^{\mathbb{P}}$ etc.). Ebenso wie beim Beispiel für eine Abkürzung in Zeile 10 folgt hier der Name des Sublexikons nicht dem Namensschema bei den anderen Beispielen.

8.1.3 Stylesheet für die Wortbildungsinformation

Die Verarbeitung des Elements für *Derivationsstammform* und *Kompositionsstammform* im Stylesheet für die Ausgabe der Wortbildungsinformationen ist in Abbildung 8.5 dargestellt.[9] In der ersten Zeile ist zu erkennen, dass das dargestellte Template beide Wortbildungselemente bearbeitet. Während bei den Flexionsstämmen als Typ `base` ausgegeben wurde, findet hier eine Fallunterscheidung statt: Je nachdem, welches der beiden genannten Elemente gerade bearbeitet wird, wird dem Parameter `stammtyp` entweder die Zeichenkette `kompos` oder `deriv` zugewiesen. Ein weiterer Unterschied zum Flexionsstamm-Stylesheet ist, dass hier ein Test durchgeführt werden muss, ob eine Stammform überhaupt vorhanden ist. Während dies bei den Flexionsstammformen

[8] In der Originaldatei kommen die Leerzeichen nicht vor, die hier der Lesbarkeit halber vor und hinter der Stammform eingefügt wurden.

[9] Aus Platzgründen wurde die Ausgabe der schließenden spitzen Klammer '>' (`<xsl:text>&#62;</xsl:text>`) in drei Zeilen durch '...' ersetzt.

```
<xsl:template match="Kompositionsstamm|Derivationsstamm">
    <xsl:param name="kat" select="'FEHLER'" />
    <xsl:param name="herk" select="'FEHLER'" />
    <xsl:param name="stamm" select="'FEHLER'" />
    <xsl:param name="stammtyp">
        <xsl:choose>
            <xsl:when test="local-name(.)='Kompositionsstamm'">kompos</xsl:when>
            <xsl:otherwise>deriv</xsl:otherwise>
        </xsl:choose>
    </xsl:param>
    <xsl:choose>
        <xsl:when test="string-length(.)>0">
            <xsl:text>&#60;</xsl:text>
            <xsl:text>DK_Stems&#62;</xsl:text>
            <xsl:value-of select="$stamm"/><xsl:text>:</xsl:text>
            <xsl:value-of select="."/>
            <xsl:text>&#60;</xsl:text><xsl:value-of select="$kat"/>...
            <xsl:text>&#60;</xsl:text><xsl:value-of select="$stammtyp"/>...
            <xsl:text>&#60;</xsl:text><xsl:value-of select="$herk"/>...
            <xsl:text>&#10;</xsl:text>
        </xsl:when>
    </xsl:choose>
</xsl:template>
```

Abbildung 8.5: XSLT-Stylesheet für Wortbildung

vorausgesetzt wird, können im Falle der Wortbildung die Elemente *Derivationsstammform* und *Kompositionsstammform* auch einfach leer sein. Der Test über die Länge des Elementinhalts (`string-length(.)>0`) verhindert hier unsinnige Ausgaben. Anstelle der Sublexikon-Information wird bei allen Wortbildungsstammformen die Zeichenkette `DK_Stems` ausgegeben. Im Unterschied zu den Flexionsstammformen wird weiterhin der Inhalt des Elements *DMOR-Stamm* grundsätzlich mit ausgegeben, so dass auch formveränderte Wortbildungsstämme immer auf die Grundstammform zurückgeführt werden können (*Haus:Häuser*). Die eigentliche Ausgabe der Stammform geschieht in der Anweisung `select="."`. Innerhalb des Stylesheets kann jede Position des Dokuments angesprochen werden, z.B. durch die Angabe des direkten Pfades oder durch die Angabe eines relativen Pfades, von einem Element im Dokument aus gesehen.

Die morphologische Form wird derzeit noch nicht mit ausgegeben, da SMOR davon noch keinen Gebrauch macht.

In Abbildung 8.6 sind Beispiele für die Ausgabe des Wortbildungs-Stylesheets angegeben. Da sie eindeutig von den Ausgaben des Flexionsformen-Stylesheets unterscheidbar sind, können die Ausgaben miteinander vermischt werden. Die Morphologiekomponente nutzt die Informationen als Bausteine

Ausgabe des Stylesheets			Zeile
<DK_Stems>	*Amt:Ämt*	<NN><deriv><nativ>	1
<DK_Stems>	*Amt:Amts*	<NN><kompos><nativ>	2
<DK_Stems>	*Maria:Marien*	<NE><kompos><nativ>	3
<DK_Stems>	*niedrig:Niedrigst*	<ADJ><kompos><nativ>	4
<DK_Stems>	*entwickeln:Entwickl*	<V><deriv><nativ>	5
<DK_Stems>	*außen:Außen*	<INVAR><kompos><nativ>	6
<DK_Stems>	*vorne:Vorder*	<INVAR><kompos><undef>	7
<DK_Stems>	*abreißen:Abreiß*	<PV><kompos><nativ>	8
<DK_Stems>	*ler:ler*	<NNSUFF><kompos><nativ>	9
<DK_Stems>	*ling:lings*	<NNSUFF><kompos><nativ>	10
<DK_Stems>	*Miß:Miss*	<DSF><deriv><unklar>	11
<DK_Stems>	*Miß:Miß*	<DSF><deriv><unklar>	12

Abbildung 8.6: Stylesheet-Ausgabe für die Wortbildung

zur Durchführung der morphologischen Analyse.

8.1.4 Automatische Konsistenzüberprüfung mit Stylesheets

Durch die Wahl des Repräsentationsformates XML sind bestimmte Möglichkeiten gegeben, die Konsistenz der Ressource zu überprüfen. In jedem komplexen System, das Zusammenhänge zwischen den in ihm enthaltenen Einheiten enthält, können Änderungen an einer Stelle ungewolltes Verhalten oder Inkonsistenzen an einer anderen Stelle bewirken. Ein Beispiel für das in dieser Arbeit vorgestellte Lexikonsystem ist das Löschen oder Ändern eines Eintrags, auf den andere Einträge verweisen. Definiert sich beispielsweise das Lexem *Darstellung*$^{\mathbb{P}}$ durch den Verweis auf die Einträge für *darstellen*$^{\mathbb{P}}{}_{PV}$ und *-ung*$^{\mathbb{P}}$, so führt das Entfernen eines dieser beiden Einträge dazu, dass Verweise ins Leere gehen oder Analysen, die vorher erzielt werden konnten, nun nicht mehr nachvollzogen werden können. Im folgenden werden beispielhaft einige Möglichkeiten der Konsistenzüberprüfung mit Stylesheets aufgeführt.

Überprüfung von redundant gespeicherten Informationen Ein wesentlicher Aspekt in dem in dieser Arbeit vorgestellten Lexikon ist die Möglichkeit, neben Simplexformen auch morphologisch komplexe Formen zu speichern, die sich aus Simplexformen zusammensetzen und teilweise deren Eigenschaften übernehmen. Das Argument, das i.A. gegen eine solche Vorgehensweise spricht, ist das Prinzip der Redundanzvermeidung: Jeder doppelt eingetragene Sachverhalt erfordert grundsätzlich den doppelten Pflegeaufwand. Haben etwa die Lexeme *Tür*$^{\mathbb{P}}$ und *Haustür*$^{\mathbb{P}}$ eine unterschiedliche Flexionsklasse, obwohl der Eintrag *Haustür*$^{\mathbb{P}}$ auf den Eintrag *Tür*$^{\mathbb{P}}$ als seinen morphologischen Kopf verweist, so muss einer der beiden Flexionsklasseneinträge fehlerhaft sein. Da das Lexi-

kon in seiner derzeitigen Repräsentation keine Vererbung von Merkmalen vorsieht, ist eine regelmäßig durchgeführte automatische Überprüfung der redundanten Informationen zu empfehlen: Ein Stylesheet verfolgt die Links in den Struktureinträgen und vergleicht die Flexionsklassen im Eintrag und beim morphologischen Kopf. Bei Nicht-Übereinstimmung ist ein Fehler gefunden worden, und einer der beiden Einträge muss korrigiert werden.

Überprüfung von Abhängigkeiten Ein Nachteil, der sich aus der Einheitlichkeit der Struktur der Lexikonressource ergibt[10], ist die Möglichkeit von Einträgen an Stellen, wo ein Eintrag keinen Sinn ergibt. Ein Beispiel ist das Strukturfeld im Lexikoneintrag, das nur ausgefüllt werden soll, wenn es sich um eine morphologisch komplexe Form handelt. Als Konvention wurde für Simplizia festgelegt, dass das Strukturfeld den Eintrag `(ohne)` erhält. Auf diese Weise kann durch ein Stylesheet sowohl überprüft werden, ob alle als Simplizia markierten Elemente über diesen Eintrag verfügen, als auch, ob als morphologisch komplex markierte Elemente im Strukturfeld auf andere Einträge verweisen.

Überprüfung der Plausibilität von Merkmalwerten Ein spannender Fall, der sich wiederum aus der Kennzeichnung morphologisch komplexer Einträge und dem Wissen um Wortbildungsbestandteile ergibt, ist die Überprüfung der Plausibilität von Partizipbildungsweisen bei morphologisch komplexen Verben. Ausgehend von der Annahme, dass Simplexverben ihr Partizip mit *ge-* bilden, Präfixverben dies aber nicht tun, kann leicht überprüft werden, ob die Kodierung der Partizipbildung im Lexikon mit der Markierung der morphologischen Form übereinstimmt. Ein Stylesheet kann alle die Einträge ausgeben, in denen die morphologische Form als `Komplex` eingetragen ist, das Sublexikon aber nicht als `V-ge_Stems`, sondern als `V-0_Stems` eingetragen ist.

Die Resultate sind zunächst vorhersehbar: Morphologisch komplexe Verben, die im Lexikon wegen eines möglicherweise vorhandenen veralteten Bestandteils (*bezichtigen, erbarmen*) als Simplizia markiert sind, bilden ihr Partizip ohne *ge-*. Dasselbe gilt für die zumeist nicht heimischen *-ieren*-Verben. Übrig bleiben schließlich Verben, die Konversionen von oder zu Substantiven darstellen (*containern*$^{\mathbb{P}}$, *orakeln*$^{\mathbb{P}}$, *posaunen*$^{\mathbb{P}}$) sowie die in der folgenden Auflistung dargestellten:

- Lexikonfehler: *bowlen, tränken* in der falschen Partizipklasse: (*ich habe gebowlt*),
- komplexe Verben, die weder Präfix- noch Partikelverben sind: *frohlocken, prophezeihen*,

[10] Der Vorteil ist die größere Übersichtlichkeit der Ressource.

- fremde Verben, bei denen das *ge-* (vermutlich aufgrund der Betonung auf der zweiten Silbe) blockiert wird: *performen, kasteien, krakeelen, kredenzen, rumoren, schmarotzen, stibitzen*.

Es darf dabei nicht vergessen werden, dass die Lexikoneinträge an sich bereits eine Bearbeitung darstellen, so dass auf diese Weise evtl. nur die Intuition der Person überprüft wird, die die gefundenen Einträge bearbeitet hat. In jedem Fall ist ein derartiges systematisches Vorgehen sehr hilfreich, um schnell auf interessante Fälle, Problemfälle oder auch fehlerhafte Einträge zu stoßen.

8.2 Vorschläge zur Durchführung der morphologischen Analyse

In diesem Abschnitt wird zunächst ein mehrstufiges Verarbeitungsmodell für eine Morphologiekomponente entwickelt, und die einzelnen Stufen werden in Relation zum IMSLEX gesetzt. Im Anschluss daran wird die morphologische Analyse einiger der in dieser Arbeit als problematisch bezeichneten Phänomene diskutiert. Dieser Abschnitt bezieht sich dabei allein auf konkatenativ beschreibbare Phänomene.[11]

8.2.1 Ein Verarbeitungsmodell für eine Morphologiekomponente

Ein mehrstufiges Verarbeitungssystem, wie es z.B. für das CISLEX verwendet wird (vgl. Abschnitt 5.3), scheint unumgänglich zu sein, wenn das Auftreten von Mehrdeutigkeiten bei Zerlegungen von Wortformen verringert werden soll. Das Prinzip dahinter ist, dass eine Stufe nur durchlaufen werden muss, wenn in der Stufe vorher keine Anlayse gefunden wurde. Jede Folgestufe nimmt Restriktionen weg, die in der Vorstufe die Chance auf eine Analyse eingeschränkt haben. Der Nutzen ist allerdings gewaltig, denn im Gegensatz zu einem einstufigen Morphologiesystem, bei dem alle erzielten Analysen gleichberechtigt nebeneinanderstehen, wird hier die wahrscheinlichste Analyse zuerst ausgegeben. Die einzelnen Stufen werden in der folgenden Aufzählung benannt:

1. Ein Lexikon oder besser eine Liste für den direkten Zugriff, die eine Auflistung häufig vorkommender oder bekanntermaßen mehrdeutig analysierter Formen enthält. Dies können Lexeme mit Flexionsklassen sein: Dann ist allerdings ein Analyseschritt notwendig, der die Gefahr birgt, gerade

[11] Nicht konkatenativ beschreibbare Phänomene folgen in Abschnitt 8.3.

wieder die ungewollten Analysen zu erzeugen. Alternativ können es Wortformen sein, die mitsamt ihrer Analyse eingetragen sind: Damit kann sicher eine gewünschte Analyse erzielt werden, allerdings handelt es sich nun bereits um ein Vollformenlexikon, keine Analysekomponente mehr.

Nur die Vollformenlösung ergibt Sinn, da nur so sicher ungewollte Analysen ausgeschlossen werden können. Das Auftreten von Homonymie und Synkretismus lässt sich zwar dadurch auch nicht verhindern, aber zumindest kommt die regelgesteuerte Analyse nicht als weitere Quelle für Ambiguitäten hinzu.

2. Eine morphologische Analyse, die auf den morphologischen Einheiten operiert, die im Lexikon gespeichert sind, gemischt mit einer festen Menge an Wortbildungsregeln, die in der Morphologiekomponente abgelegt sind.

3. Eine morphologische Analyse, die neben den morphologischen Einheiten, die sie aus dem Lexikon bezieht, noch bestimmte Generalisierungen auf die Einheiten anwendet.[12] Die Generalisierungen auf dieser Stufe betreffen ausschließlich die Derivations- und Kompositionsstammformen:

   - Substantive: Alle Nominativ-Singular-, Genitiv-Singular- und Nominativ-Plural-Formen werden als Derivations- und Kompositionsstammformen zugelassen, zusätzlich die Grundform mit angehängtem Fugen-s von Simplizia mit Genus Femininum (*Arbeits*) und Derivationen auf *-ung, -heit, -keit* etc. (*Sicherheits, Sicherungs*).
   - Eigennamen: Alle Nominativ-Singular-Formen werden als Derivations- und Kompositionsstammformen zugelassen.
   - Adjektive: Alle Formen im Positiv werden als Derivations- und Kompositionsstammformen zugelassen.
   - Verben: Alle Verbstämme werden als Derivations- und Kompositionsstammformen zugelassen.

4. Eine morphologische Analyse, die neben den Eigenschaften aus 3. noch morphologische Prozesse zur Ad-hoc-Bildung von Derivations- und Kompositionsstammformen zulässt:

   - Substantive: Tilgung, Umlautung und Fugung
   - Adjektive: Komparativ- und Superlativ-Formen werden als Derivations- und Kompositionsstammformen zugelassen.

[12] Dies ist die Variante, der DMOR und SMOR am ehesten entsprechen.

5. Der 'Guesser' (engl.). Falls Wortformen nach Durchlaufen der Stufen 1 bis 4 noch immer nicht erkannt wurden, wird nach Suffix entschieden (dann kann zumindest die Flexion richtig erkannt werden), sonst handelt es sich bei großgeschriebenen Formen um Eigennamen.

Die Übergänge zwischen den Varianten 3 bis 5 sind fließend. Wenn eine Morphologiekomponente die Wortbildungsregeln nicht einschränkt, können einige der in Stufe 3 vorgenommenen Generalisierungen bereits eine Stufe vorher durchgeführt werden.

Für das IMSLEX ist das Modell deswegen attraktiv, weil anstelle einer großen, uniformen Makrostruktur, wie sie aus dem DMOR-Lexikon übernommen wurde, eine Menge von fein unterschiedenen Lexemen vorliegt, die nach verschiedenen Attributen gruppiert werden können. Es ist ein Leichtes, die als morphologisch komplex markierten Einheiten in die Liste für den direkten Zugriff auszulesen. Da ein Großteil von ihnen über einen Struktureintrag verfügt, sind die unmittelbaren Konstituenten bekannt und können mit ausgegeben werden. Allerdings handelt es sich nicht um Vollformen, so dass zusätzlich zum Auslesen noch die Generierung der Vollformen erfolgen muss.[13]

Für die folgenden Stufen wird dann das gesamte Lexikon ausgelesen. In Stufe 2 werden alle Phänomene erfasst, die völlig regulär ablaufen, also sich aus den im Lexikon gespeicherten Einheiten bilden lassen. Diese Stufe bietet eine sehr gute Möglichkeit, das Lexikon zu testen: Da an dieser Stelle noch keine Heuristiken eingesetzt werden, können die in dieser Stufe nicht analysierten Einheiten direkt auf fehlende morphologische Einheiten zurückgeführt werden.[14] Allerdings kann es hier auch zum umgekehrten Fehlerfall kommen: Es gibt eine Analyse, aber sie ist falsch. Ein Beispiel dafür ist die Zerlegung der im Korpus belegten französischen Wortform *Beaucoup* (an einem Satzanfang) in die Substantive *(der) Beau* und *(der) Coup*.

Stufe 3 bildet den Kompromiss zwischen Freiheit und Restriktion der Generalisierungen. Da die erwähnten Formen aus den Paradigmen der Wortarten oft formgleich mit Fugenelementen sind, wird ein geringeres Risiko eingegangen als bei der Freigabe von bestimmten typischen Fugenelementen für alle Erstglieder unabhängig vom Paradigma.

Stufe 4 bildet alle Prozesse ab, denen die Bildung von Stammformen unterliegen kann. Dies trägt der Tatsache Rechnung, dass z.B. eine umgelautete Derivationsstammform oft nur in ein oder zwei verschiedenen Wortbildungsmustern belegt ist (*Öfchen, Öflein*) und daher auch nur entsprechend schwer zu finden und im Lexikon zu verzeichnen ist.[15] Diese Stufe kann dafür verwendet

[13] Da der morphologische Kopf und das Paradigma bekannt sind, ist dies nicht schwer.

[14] Natürlich können auch Wortbildungsregeln fehlen, aber da diese gewöhnlich sehr generell gehalten werden, ist das Fehlen von Einheiten wahrscheinlicher.

[15] Für einen umgelauteten Derivations- und Kompositionsstamm für das Lexem *Klang*$^{\mathbb{P}}{}_{NN}$ sind genau zwei Wortbildungen im HGC belegt, die darüber hinaus auch nur je einmal vorkom-

werden, Kandidatenlisten für das Auffüllen der Mikrostruktur des Lexikons zu generieren.

In Stufe 5 schließlich kann nur noch über die Wortart der unbekannten Wortform spekuliert werden. Bei der weitaus größten Menge bis in diese Stufe nicht erkannter Formen handelt es sich um Eigennamen und Tippfehler. Ist beim Tippfehler das Derivationssuffix nicht betroffen, erfolgt zumindest eine Teilanalyse.

Die 'longest match'-Heuristik

Eine andere Heuristik soll nicht unerwähnt bleiben. Wird das Stufenmodell nicht angewendet, umfasst das Lexikon aber einfache wie komplexe Einheiten, so kommt es oft zu mehrdeutigen Zerlegungen mit einer unterschiedlichen Anzahl an Zerlegungsgliedern. Hier gilt fast immer der Grundsatz, dass eine Analyse mit weniger Zerlegungen besser ist als eine mit vielen. Diese Heuristik kann man als eine Art 'Stufenmodell im Kleinen' ansehen: Ist eine morphologisch komplexe Einheit wie *Bahnhof* neben den morphologisch einfachen Einheiten *Bahn* und *Hof* im Lexikon gespeichert, so ist sicherlich die nicht zerlegte Form die gewünschte Analyse. Dasselbe gilt für alle weiteren Wortbildungen mit dieser Einheit (*Bahnhofs=Kneipe* mit zwei Zerlegungsgliedern ist *Bahn=Hofs=Kneipe* mit drei Gliedern vorzuziehen, etc.).[16]

8.2.2 Verbesserung der morphologischen Analyse

Im Folgenden werden für einige der im Verlauf dieser Arbeit angesprochenen Phänomene, die einer Morphologiekomponente Schwierigkeiten bereiten können, die Alternativen besprochen, die sich in Kombination von IMSLEX und SMOR für ihre Behandlung ergeben. Es wird jeweils eine Empfehlung ausgesprochen, die jedoch lediglich als Diskussionsbasis dient.

Der Typ *blauäugig*

1. Eintrag von *äugig* als gebundenes Lexem. Das erlaubt eine Gleichbehandlung von *blau·äugig* und *rechts·kräftig*, allerdings wird eine falsche Wortbildungsstruktur suggeriert (vgl. Abschnitt 4.2.5).

men: *Schönklängler, Klängespektrum*.

[16]Mir sind bislang nur sehr wenige Gegenbeispiele für das Funktionieren dieser Heuristik begegnet, und die entstammen alle demselben Muster 'adjektivischer Kopf': *amtsdeutsch*$^{\mathbb{P}}$ führt zu **Kataster=amtsdeutsch*, *waffentechnisch*$^{\mathbb{P}}$ führt zu **Atom=waffentechnisch*, *sandfarben*$^{\mathbb{P}}$ führt zu **Wüsten=sandfarben*, *rechtswidrig*$^{\mathbb{P}}$ führt zu **Völker=rechtswidrig*, und *ostafrikanisch*$^{\mathbb{P}}$ führt zu **Nord=ostafrikanisch*. Ein * markiert die ungewollte Struktur.

2. Eintrag von *blauäug* als gebundener komplexer Derivationsstamm. Dieser muss einem Lexem zugewiesen werden, also der Phrase '*blaue Augen*$^{\mathbb{P}}{}_{Phrase}$'[17]. Diese Variante halte ich für gut, allerdings berücksichtigt sie die Produktivität nicht (vgl. Variante 4).

3. Eintrag von *äug* als Derivationsstamm zu *Auge*$^{\mathbb{P}}{}_{NN}$. In der Morphologiekomponente kann dann die flache Struktur *blau + äug + ig* erkannt werden, allerdings nicht die Wortbildungsstruktur. Da jedoch die einzelnen Bestandteile identifiziert werden (*blau*$^{\mathbb{P}}{}_{ADJ}$ + *Auge*$^{\mathbb{P}}{}_{NN}$ + *-ig*$^{\mathbb{P}}{}_{ADJSuff}$), ist eine Untersuchungen des Musters `ADJ + NN + ADJSuff` hinsichtlich der Häufigkeit des Auftretens der Struktur einer Derivation mit morphologisch komplexer Basis vielversprechend.

4. Eintrag der komplexen Form *blauäugig* mit einer der möglichen Zerlegungen als Struktureintrag. Diese Möglichkeit hat zwar zur Folge, dass eine Analyse in Stufe 1 im Stufenmodell in Abschnitt 8.2.1 erzielt werden kann, wird allerdings angesichts der Produktivität des Musters verworfen (vgl. die Auflistung in Abbildung 2.6 auf Seite 19).

Für die Kombination aus IMSLEX und SMOR schlage ich vor, zunächst Variante 3 zu untersuchen. Da die Derivations- und Kompositionsstammformen ohnehin kontinuierlich erhoben werden, ist hiermit sicherlich die höchste Neuerkennungsquote zu erzielen, ohne Gefahr zu laufen, die linguistischen Zusammenhänge aus den Augen zu verlieren.

Das eben Gesagte betrifft neben weiteren Beispielen desselben Typs (*viertürig, dreiachsig, breitschultrig* mit Suffix *-ig*, *Dickhäuter* mit Suffix *-er*) auch Wortformen wie *zweifarbig*[18], die von einer Morphologiekomponente als Komposition aus *zwei*$^{\mathbb{P}}{}_{CARD}$ und *farbig*$^{\mathbb{P}}{}_{ADJ}$ analysiert werden kann, da *farbig* im Gegensatz zu *äugig* frei vorkommt. Wenn in diesem Fall ein Derivationsstamm *farb* zu *Farbe*$^{\mathbb{P}}{}_{NN}$ hinzukommt und Variante 3 in der Morphologiekomponente eingesetzt wird, kommt zwar eine Analyse hinzu, aber eine, die mir linguistisch adäquater erscheint.

Das 'Produktivitäts'-Argument aus Varianten 2 und 4 ist nur dann zu berücksichtigen, wenn es keine anderen Bildungen mit dem gebundenen komplexen Stamm gibt. Im Falle von *dickhäut* sind im Korpus zwei direkte Ableitungen zu finden (*dickhäutig*$_{(3)}$, *Dickhäuter*$_{(102)}$), die allerdings beide selber wieder Basen für Folgeableitungen (*dickhäuterischer*$_{(2)}$, *Dickhäutigkeit*$_{(5)}$, *Dickhäuterin*$_{(1)}$) bzw. Komposita (*Dickhäuterhaus*$_{(2)}$, *Dickhäuterjunge*$_{(1)}$) bilden.

[17] Angesichts des offenkundigen Zusammenhangs von *blauäugig* und *blaue Augen* wäre die Schaffung eines gebundenen Lexems **blauaug*$^{\mathbb{P}}$ oder **Blauauge*$^{\mathbb{P}}$ unverständlich.

[18] Ebenso *kleinstädtisch, Hauptstädter* etc.

Der Typ *Arbeitnehmer*

Unter der Annahme, dass es tatsächlich einen häufig auftretenden Typ 'Komposition mit Argumentvererbung' gibt (vgl. Abschnitt 4.2.5), könnte hier zunächst (analog zu Variante 3 aus der Aufzählung im vorangehenden Abschnitt) eine 'flache' Zerlegung nach dem Muster `NN + V` + *-er* zugelassen werden. Alternativ kann der Kopf *Nehmer* als gebunden auftretende Einheit ins Lexikon aufgenommen werden.

Die Behandlung von Derivationen mit neoklassischen Basen

Das Auftreten neoklassischer Einheiten im Deutschen ist nicht auf wenige Fälle beschränkt, sondern gang und gäbe. Lüdeling und Schmid (2001) widerlegen die gängige Auffassung, dass die Wortbildungsmöglichkeiten neoklassischer Elemente durch ihre Herkunft eingeschränkt seien. Eine Morphologiekomponente kann dieser Tatsache Rechnung tragen, indem sie neoklassische Einheiten nicht grundsätzlich anders behandelt als native Einheiten. An dieser Stelle soll nur auf zwei Teilbereiche neoklassischer Wortbildung hingewiesen werden, die im Lexikon repräsentiert werden sollten.

Wie in Abschnitt 4.2.1 erwähnt wurde, bereitet die Tatsache, dass bei Derivationen mit neoklassischen Suffixen oft nicht nur die Suffixe, sondern auch die Basen gebundene Einheiten sind, gewöhnlich Schwierigkeiten bei deren morphologischer Analyse. Da lexikalische Einheiten im IMSLEX problemlos mit morphologischem Status `gebunden` eingetragen werden können, ist das im vorliegenden Lexikon nicht der Fall. Für die Eintragung ins Lexikon qualifiziert sich eine neoklassische Basis genau dann, wenn sie mit mindestens zwei verschiedenen Suffixen belegt ist. Ein Beispiel hierfür sind *illustrieren* und *Illustration*. Die beiden jeweils unterstrichenen Zeichenketten werden als Derivationsstammformen eingetragen. Die Frage, die bestehen bleibt, ist die nach der Wahl der Zitierform. Bislang wird hierfür im IMSLEX die Vergleichssegmentform (vgl. 4.2.4) gewählt. Diese ist allerdings – gerade bei recht kurzen Formen – meistens nicht besonders aussagekräftig, weswegen hier alternativ eine Auswahl aus einer der derivierten Formen hergenommen werden könnte, beispielsweise *illustr(ieren)* $^{\mathbb{P}}{}_{NEO}$ für die genannten Beispiele.[19]

Der zweite Teilbereich sind die neoklassischen 'Erstglieder'[20] Diese sind für einen großen Teil der nicht analysierten Formen zuständig und treten oft mit vielen verschiedenen Köpfen auf: In der HGC-Wortliste finden sich allein 253 Wortformen, die mit der Zeichenkette *pseudo* beginnen.

[19] In Lüdeling et al. (2002) sind zahlreiche Beispiele neoklassischer Wortbildung und neoklassischer Stämme aufgelistet.

[20] Da diese Einheiten nur gebunden auftreten, handelt es sich strenggenommen um Derivation. Wegen der aufgrund des frei vorkommenden Kopfes großen Ähnlichkeit zur Komposition werde ich sie dennoch 'Erstglieder' nennen.

8.3 Darstellung von IA und IP: Lexikon als komplexes System

Die Schwierigkeiten der in Abschnitt 4.3 angesprochenen nicht konkatenativ beschreibbaren Prozesse sind sowohl für das Lexikon als auch für die Morphologiekomponente offensichtlich. Im Lexikon sind sie schwer zu modellieren, in der morphologischen Analyse sind sie für viele Ambiguitäten verantwortlich. Im Folgenden schlage ich ein Vorgehen der **Vernetzung** von Lexikoneinträgen vor, das spezielle Relationen zwischen diesen Einheiten vorsieht. Diese Relationen können von einer **Filterkomponente** oder **Disambiguierungskomponente** ausgelesen werden, so dass einige der Ambiguitäten, die durch die derart ausgezeichneten Einheiten zustandekommen, aufgelöst werden können.

Mit der Realisierung dieses Vorschlags liegt ein computerlinguistisches Lexikon als komplexes System vor.

8.3.1 Vernetzung im Lexikon

Eine Vernetzung zwischen Lexikoneinträgen ist im *Struktur*-Element (vgl. Abbildung 7.5 auf Seite 103) bereits angelegt: Die Bestandteile, aus denen sich eine morphologisch komplexe Einheit zusammensetzt, sind dort zunächst mit ihrer Kategorie aufgelistet. Der logische nächste Schritt ist die Ersetzung dieser expliziten Nennungen durch XML-Links auf die Einträge der Bestandteile im Lexikon.

(8.1) <*Struktur*>`drehen(V)[++]ung(NNSuff)`</*Struktur*>

```
<Struktur typ="Derivation"  bestandteile="2">
    <Bestandteil nr="b1"  kategorie="V"  idref="v3976"/>
    <Bestandteil nr="b2"  kategorie="NNSuff"  idref="aff42"/>
</Struktur>
```

Abbildung 8.7: IMSLEX-Struktureintrag für *Drehung*$^{\mathbb{P}}{}_{NN}$

Anstelle des in 8.1 dargestellten Elementinhalts nähme dies die Form einer Hierarchie wie der in Abbildung 8.7 dargestellten an.[21] Ein XSLT-Stylesheet kann die in 8.1 gezeigte Darstellung leicht wieder erzeugen, indem die einzelnen *Bestandteil*-Elemente der Reihe nach durchlaufen werden und die Zitierform der über den `idref` referenzierten lexikalischen Einheit ausgegeben wird.[22]

[21] Dokumente ohne Dokumentinhalt, also nur mit Attributen, können in XML durch die Angabe eines Slash (/) vor der schließenden spitzen Klammer geschlossen werden.

[22] Beim Parsen eines XML-Dokuments mit einem XML-Parser wird automatisch überprüft, ob die Links eindeutig sind und ob sie auf einen bestehenden Eintrag verweisen. Nur wenn beides

Relationen zur Modellierung von IP-Phänomenen

Nach demselben Schema können nun Relationen zwischen Einträgen hergestellt werden, die dafür geeignet sind, nicht-konkatenativ ablaufende Prozesse zu modellieren. Dies soll an den Beispielen *Flug, Abflug* und *Platz* demonstriert werden.

Abstrakte Nominalisierung Lexeme wie *Flug*$^{\mathbb{P}}$, *Griff*$^{\mathbb{P}}$, *Tritt*$^{\mathbb{P}}$ etc. sind dafür verantwortlich, dass es in der DTD des IMSLEX beim Attribut **m_form** (morphologische Form) neben den beiden zu erwartenden Belegungen `Simplex` und `Komplex` eine weitere Belegung gibt: `Nominalisierung` (vgl. Abbildung 6.12 auf Seite 88). Dies stellte bislang die einzige Möglichkeit dar, die Tatsache auszudrücken, dass ein Lexem wie *Flug*$^{\mathbb{P}}{}_{NN}$ zwar wie ein Simplex wirkt, aber in Wirklichkeit einen morphologisch begründeten Zusammenhang zum Lexem *fliegen*$^{\mathbb{P}}{}_{V}$ aufweist.

```
<Relation id="r1"  typ="abstr_Nominalisierung"  bestandteile="1">
    <Bestandteil nr="r1b1"  kategorie="V"  idref="v6718"/>
</Relation>
```

Abbildung 8.8: IMSLEX-Struktureintrag für *Flug*$^{\mathbb{P}}{}_{NN}$

In Abbildung 8.8 ist dargestellt, wie die Beziehung zwischen *Flug*$^{\mathbb{P}}$ und *fliegen*$^{\mathbb{P}}$ (ID `v6718`) in einem Element *Relation*[23] ausgedrückt werden kann. Auf diese Weise kann beim Attribut für die morphologische Form nun auf die Vermischung von reinem Merkmal (`Simplex/Komplex`) und der Erklärung eines Merkmals (`Nominalisierung`) verzichtet werden. Der Unterschied zwischen IA und IP wird jetzt in der Unterscheidung von *Struktur* und *Relation* ausgedrückt.

Sekundär komplexe abstrakte Nominalisierung Lexeme wie *Abflug*$^{\mathbb{P}}$, *Angriff*$^{\mathbb{P}}$, *Aufstand*$^{\mathbb{P}}$ etc. sind im Attribut **m_form** einer lexikalischen Einheit noch schwieriger zu modellieren als 'einfache' abstrakte Nominalisierungen. Dies hat den Grund, dass hier scheinbar beide Modelle zusammenkommen: IP für die Behandlung von *Flug, Griff* etc., IA für die Zerlegung in ein 'Erstglied' *ab, an*. Das Lexem *Abflug*$^{\mathbb{P}}$ erhält im IMSLEX bislang die Belegung `Komplex_abstrakt`, was ähnlich wie `Nominalisierung` oben das Merkmal der morphologischen Form und die Erklärung dafür miteinander vermischt. In Wirklichkeit sind die

zutrifft, ist das Dokument valide.

[23] Die Darstellung ist insofern übertrieben, als es vermutlich immer nur einen Bestandteil gibt. Über den `idref` kann die Information, um welche Kategorie es sich handelt, ebenfalls bezogen werden, so dass das Attribut **kategorie** hier redundante Information darstellt. `id, typ` und `idref` würden ausreichen. Ich belasse es hier bei dieser Darstellung, da sie meiner Meinung nach für den Leser anschaulicher ist.

genannten Lexeme **sekundär komplex**, d.h., es handelt sich um Nominalisierungen einer morphologisch komplexen Form.

```
<Relation id="r2" typ="sek_kompl_abstr_Nominalisierung" bestandteile="1">
    <Bestandteil nr="r2b1" kategorie="PV" idref="pv86"/>
</Relation>
```

Abbildung 8.9: IMSLEX-Struktureintrag für *Abflug*$^{\mathbb{P}}{}_{NN}$

In einem *Relation*-Eintrag (hier: für das Lexem *Abflug*$^{\mathbb{P}}$) können sekundär komplexe abstrakte Nominalisierungen nun modelliert werden (vgl. Abbildung 8.9). Die ID `pv86` verweist auf den Eintrag für das Partikelverb *abfliegen*$^{\mathbb{P}}{}_{PV}$.

Konversion Lexeme wie *Platz*$^{\mathbb{P}}{}_{NN}$/*platzen*$^{\mathbb{P}}{}_{V}$, *Spiel*$^{\mathbb{P}}{}_{NN}$/*spielen*$^{\mathbb{P}}{}_{V}$, *Feuer*$^{\mathbb{P}}{}_{NN}$/*feuern*$^{\mathbb{P}}{}_{V}$, *Licht*$^{\mathbb{P}}{}_{NN}$/*lichten*$^{\mathbb{P}}{}_{V}$/*licht*$^{\mathbb{P}}{}_{ADJ}$ sind im Attribut **m_form** einer lexikalischen Einheit einfach zu modellieren, sofern der morphologische Zusammenhang zwischen ihnen nicht explizit gemacht werden soll. Soll die Ableitungsbeziehung doch expliziert werden, so stellt sich die Frage nach der Ableitungsrichtung (vgl. Abschnitt 4.3.1), die darüber entscheidet, welche der beiden (oder im letzten Fall sogar: drei) Einheiten als `komplex` ausgezeichnet wird.[24]

```
<Relation id="r3" typ="Konversion" bestandteile="1">
    <Bestandteil nr="r3b1" kategorie="V" idref="v11515"
            styp="ohne_Zusammenhang"/>
</Relation>
```

Abbildung 8.10: IMSLEX-Struktureintrag für *Platz*$^{\mathbb{P}}{}_{NN}$

Der Eintrag des Elements *Relation* für das Lexem *Platz*$^{\mathbb{P}}{}_{NN}$ ist in Abbildung 8.10 dargestellt. Die ID `v11515` verweist auf den Eintrag für *platzen*$^{\mathbb{P}}{}_{V}$. In dieser Darstellung ist ein weiteres Attribut vorhanden (**styp**). Dieses Attribut erlaubt die Herstellung 'semantischer' Zusammenhänge, die einen Einfluss auf die Behandlung von Ambiguitäten in der Disambiguierungskomponente haben. Es wird unten in 8.3.2 erläutert.

In Abbildung 8.11 ist die 'Gegenrichtung' abgebildet, also das Element *Relation*, wie es sich beim Verb *platzen*$^{\mathbb{P}}$ darstellt. Die ID `n47408` verweist auf den Eintrag für *Platz*$^{\mathbb{P}}{}_{NN}$. Die Redundanz zum in Abbildung 8.10 dargestellten Element erklärt sich dadurch, dass in dieser Arbeit Konversion immer als **bidirektionaler** Prozess verstanden wird. Dies wird hier durch zwei gerichtete Links modelliert.

[24] Es kann natürlich ein neuer Attributwert `Konversion` eingeführt werden, der richtungsneutral ist, also bei allen beteiligten Einheiten eingetragen wird. Dies vermischt aber wiederum Merkmal und Erklärung.

```
<Relation id="r4" typ="Konversion" bestandteile="1">
    <Bestandteil nr="r4b1" kategorie="NN" idref="n47408"/>
            styp="ohne_Zusammenhang"/>
</Relation>
```

Abbildung 8.11: IMSLEX-Struktureintrag für $platzen^{\mathbb{P}}{}_{V}$

8.3.2 Der Nutzen der Vernetzung für die Disambiguierung

Die im vorangegangenen Abschnitt erwähnten Konversionen wie $Platz^{\mathbb{P}}{}_{NN}/platzen^{\mathbb{P}}{}_{V}$, $Spiel^{\mathbb{P}}{}_{NN}/spielen^{\mathbb{P}}{}_{V}$, $Feuer^{\mathbb{P}}{}_{NN}/feuern^{\mathbb{P}}{}_{V}$, $Licht^{\mathbb{P}}{}_{NN}/lichten^{\mathbb{P}}{}_{V}/licht^{\mathbb{P}}{}_{ADJ}$ sind für sehr viele Ambiguitäten bei der morphologischen Analyse verantwortlich.

Wortbildungsstruktur	Wortbildungsmuster	Zeile
`Spiel=Platz+NN`	NN + NN	1
`spiel=Platz+NN`	V + NN	2
`Platz=Geräusch+NN`	NN + NN	3
`platz=Geräusch+NN`	V + NN	4
`Platz=Konzert+NN`	NN + NN	5
`platz=Konzert+NN`	V + NN	6

Abbildung 8.12: Mehrdeutige Zerlegungen aufgrund von Konversionen

Alle in Abbildung 8.12 dargestellten Wortbildungen erhalten zwei Analysen, jeweils eine Analyse mit verbalem Erstglied und eine Analyse mit nominalem Erstglied.[25] Wird nun das in Abbildung 8.10 dargestellte Attribut **styp** dazu verwendet, zwischen Konversionen, die einen semantischen Zusammenhang aufweisen, und solchen, die keinen semantischen Zusammenhang aufweisen, zu unterscheiden, so können die Analysen in Zeilen 1 und 2 zu einer Analyse verschmolzen werden.

```
<Relation id="r5" typ="Konversion" bestandteile="1">
    <Bestandteil nr="r5b1" kategorie="V" idref="v14224"
            styp="mit_Zusammenhang"/>
</Relation>
```

Abbildung 8.13: IMSLEX-Struktureintrag für $Spiel^{\mathbb{P}}{}_{NN}$

Die Begründung für die Verschmelzung ist in Abbildung 8.13 zu erkennen: Im Attribut **styp** wird hier ein semantischer Zusammenhang zwischen $Spiel^{\mathbb{P}}{}_{NN}$

[25] DMOR lässt Verbstämme als Erstglieder nicht generell zu, daher gäbe es hier keine Mehrdeutigkeiten, allerdings auch nicht die (semantisch) richtige Analyse in Zeile 4.

und *spielen*$^{\mathbb{P}}{}_{V}$ (ID v14224) hergestellt. Dieser kann von der Morphologiekomponente so interpretiert werden, dass die Kompositionsstammformen beider Lexeme miteinander verschmolzen werden. Somit fällt zumindest für die semantisch eng zusammenhängenden Konversionen[26] eine häufig auftretende Mehrdeutigkeit weg. Bei den Analysen in Zeilen 1 bis 4 hingegen ist der semantische Zusammenhang gerade nicht gegeben, so dass hier die Ausgabe beider Analysen garantiert, dass die gewünschte in der Ergebnismenge ist.

Auch gefugte oder getilgte Kompositionsstammformen können miteinander verschmolzen werden: Bei *Bade=Meister/bade=Meister* und *Such=Maschine/such=Maschine* lassen sich die Kompositionsstammformen auf *baden*$^{\mathbb{P}}{}_{V}$ oder *Bad*$^{\mathbb{P}}{}_{NN}$ bzw. *suchen*$^{\mathbb{P}}{}_{V}$ oder *Suche*$^{\mathbb{P}}{}_{NN}$ zurückführen, aber durch eine Verschmelzung ergeben sich wieder eindeutige Analysen.

Analysetiefe bei Konversionen

Ein Problem bei der morphologischen Analyse, das nur im Hinblick auf die Vorstellungen eines Anwenders geklärt werden kann, ist die **Tiefe** der ausgegebenen Analyse. Bei konkatenativ ablaufenden Prozessen wird gewöhnlich die Zerlegung in unmittelbare Konstituenten gewählt, da diese rekursiv weiter zerlegt werden können. Bei Konversionen hingegen stellt sich die Frage, ob eine solche Relation in der Ausgabe der Analyseergebnisse angezeigt werden soll.

(8.2) a. *Sturzflug*: *stürzen*$^{\mathbb{P}}{}_{V}$ *fliegen*$^{\mathbb{P}}{}_{V}$

b. *Gleitflug*: *gleiten*$^{\mathbb{P}}{}_{V}$ *fliegen*$^{\mathbb{P}}{}_{V}$

c. *Gleitflugmodell*: *gleiten*$^{\mathbb{P}}{}_{V}$ *fliegen*$^{\mathbb{P}}{}_{V}$ *Modell*$^{\mathbb{P}}{}_{NN}$

d. *Spielplatz*: *spielen*$^{\mathbb{P}}{}_{V}$ *platzen*$^{\mathbb{P}}{}_{V}$

In 8.2 sind einige Ausgaben einer (fiktiven) morphologischen Analyse dargestellt, die Konversion 'zurückverfolgen'.[27] Nun kann die wahrscheinlich nicht gewollte Analyse in 8.2 d durch die Relation zwischen *Platz*$^{\mathbb{P}}$ und *platzen*$^{\mathbb{P}}$ verhindert werden, wenn das Attribut **styp** den nicht bestehenden semantischen Zusammenhang zwischen beiden markiert (wie in Abbildung 8.10 dargestellt wurde). Die Analyse in 8.2 a hingegen kann ausgegeben werden, da das Attribut **styp** explizit einen semantischen Zusammenhang zwischen *Sturz*$^{\mathbb{P}}$ und *stürzen*$^{\mathbb{P}}$ herzustellen erlaubt.[28]

[26] Streng genommen handelt es sich bei Paaren wie *platzen/Platz* vermutlich nicht um Konversionen, weil gerade dieser Zusammenhang fehlt. Da dies allerdings der Form nicht angesehen werden kann, sind für die Morphologiekomponente zunächst alle Formübereinstimmungen mit Konversionen gleichzusetzen.

[27] Für diese Darstellung wird an dieser Stelle davon ausgegangen, dass eine Ableitungsrichtung bekannt ist und vom Verb zum Substantiv führt.

[28] Nicht alle semantischen Zusammenhänge stellen sich vermutlich so deutlich dar wie in diesen beiden Beispielen geschildert, aber Abgrenzungsproblematiken sind ohnehin in nahezu jedem Bereich des Lexikons anzutreffen.

Die Verwendung von Relationen ermöglicht also die Disambiguierung von Analysen, die einen oder mehrere Bestandteile enthalten, die Konversionen darstellen. Weitere Beispiele vom Typ `ohne Zusammenhang` sind Paare wie *Gehör*$^{\mathbb{P}}{}_{NN}$*/gehören*$^{\mathbb{P}}{}_{V}$, *Rausch*$^{\mathbb{P}}{}_{NN}$*/rauschen*$^{\mathbb{P}}{}_{V}$, *Rat*$^{\mathbb{P}}{}_{NN}$*/raten*$^{\mathbb{P}}{}_{V}$. Beispiele vom Typ `mit Zusammenhang` sind Paare wie *Segel*$^{\mathbb{P}}{}_{NN}$*/segeln*$^{\mathbb{P}}{}_{V}$, *Rauch*$^{\mathbb{P}}{}_{NN}$*/rauchen*$^{\mathbb{P}}{}_{V}$, *Krümel*$^{\mathbb{P}}{}_{NN}$*/krümeln*$^{\mathbb{P}}{}_{V}$, *Start*$^{\mathbb{P}}{}_{NN}$*/starten*$^{\mathbb{P}}{}_{V}$, *Schlaf*$^{\mathbb{P}}{}_{NN}$*/schlafen*$^{\mathbb{P}}{}_{V}$.

Es soll hier nicht der Eindruck erweckt werden, eine Behandlung von 'Semantik' wäre mit so einfachen Mitteln wie Links zwischen lexikalischen Einheiten möglich. Da das IMSLEX hauptsächlich morphologisch motiviert ist, umfassen die Einträge oft verschiedene semantische Varianten, die sich nicht in morphologischer Unterscheidung ausdrücken. Die vorgestellte Methode des Einfügens bidirektionaler, annotierter Verweise zwischen Lexikoneinträgen kann jedoch dazu führen, die Ambiguitäten, die sich bei der Strukturierung von Wortformen ergeben, einzuschränken.

Kapitel 9

Zusammenfassung

In dieser Arbeit wurde ein computerlinguistisches Lexikon vorgestellt, das eine übersichtliche, leicht nachvollziehbare Struktur aufweist, die jedoch auch funktional ist. Innerhalb der Ressource können Relationen morphologischer und semantischer Art repräsentiert werden. Es wurde gezeigt, wie Informationen in das Lexikon gelangen und wie die für eine nachfolgende Verarbeitungsstufe relevanten Daten mithilfe einer standardisierten Sprache ausgelesen werden können. Eine exemplarisch ausgewählte Anwendungskomponente besteht aus einem Morphologiewerkzeug, das ein Zwei-Ebenen-Modell implementiert. Es zeigt sich, dass in der Ressource mehr Informationen vorhanden sind, als in der Anwendungskomponente gebraucht werden. Durch die leicht zu erstellenden Ausleseroutinen können schnell Schnittstellen zu anderen Anwendungskomponenten geschaffen werden. Eine Besonderheit des gewählten Repräsentationsformates, der Dokumentenbeschreibungssprache XML, ist es, dass Werkzeuge bei Erweiterungen des Strukturmodells nicht notwendigerweise angepasst werden müssen: Kommen neue Strukturelemente zur Ressource hinzu, so funktionieren alle vorhandenen Ausleseroutinen unverändert, es sei denn, sie wollen auf die neuen Informationen zugreifen. Auf diese Weise ist eine besondere Flexibilität gewährleistet, die auch Strukturveränderungen in der Ressource ohne Folgekosten in der Infrastruktur erlaubt.

Ein computerlinguistisches Lexikon ist kein Selbstzweck, sondern dient i.A. der Aufbereitung bzw. Anreicherung von Daten für eine weitere Verarbeitung. Der Hauptabnehmer des Lexikons am IMS ist eine Morphologiekomponente, die als Schnittstelle zu weiteren computerlinguistischen Verarbeitungsstufen wie der Syntax oder dem Tagging fungiert. In der Arbeit wurde eine Übersicht über das Verfahren der morphologischen Analyse gegeben, und es wurden Morphologiekomponenten vorgestellt. Da ohne eine zugrundeliegende morphologische Theorie keine evaluierbare morphologische Analyse möglich ist, wurden die für eine Behandlung der Morphologie des Deutschen relevanten Phänomen benannt und in zwei Standardmodelle für die Darstellung morphologischer Prozesse eingeteilt: *Item and Arrangement* und *Item and Process*. Zusätzlich wurde

ein Lexikonmodell implementiert, das im Rahmen des DeKo-Projektes konzipiert worden war. Nun konnten die einzelnen morphologischen Phänomene in das Lexikonmodell eingeordnet werden. Es zeigte sich, dass die Phänomene des Gegenwartsdeutschen im Modell berücksichtigt sind und im Lexikon angemessen repräsentiert werden können. Dies gilt auch für Phänomene, die in anderen Lexikon- und Morphologiesystemen nur halbherzig oder gar nicht behandelt werden: neoklassische Wortbildung, Konversion, Phrasen in der Wortbildung. Für die Behandlung von Derivation und Konversion wurde im Lexikonmodell auf die Theorie der Derivations- und Kompositionsstammformen aus Fuhrhop (1998) zurückgegriffen. Diese ermöglicht eine wesentlich feinere Behandlung, als dies mit den traditionellen Konzepten der Tilgung und Fugung möglich ist: Übergenerierung und Falschanalysen können damit verhindert werden.

Schließlich wurde der Zusammenhang zwischen Abdeckung und Korrektheit einer Morphologiekomponente bzw. einer morphologischen Analyse hergestellt. Abdeckung wird oft in der Literatur bei der Beschreibung von Morphologiekomponenten verwendet, um dem Leser einen Eindruck zu vermitteln, wie 'gut' das System arbeitet. Allerdings lässt dieser Wert nicht den geringsten Aufschluss darüber zu, wie viele der Analysen korrekt sind. Korrektheit lässt sich nur durch Identifikation der relevanten Phänomene der Morphologie herstellen, denn anders lässt sich gar nicht sagen, was eine morphologische Analyse überhaupt ergeben soll. Oftmals ist mehr als eine korrekte Lösung möglich, so dass allein die Angabe von Zahlen zur Korrektheit und Vollständigkeit von Analysen ein System bewertbar macht. Eine Evaluierung eines Morphologiesystems und damit implizit auch des zugrundeliegenden Lexikons ist am besten im direkten Vergleich mit anderen Systemen zu erreichen, die dieselben morphologischen Phänomene behandeln und dies auch transparent machen.

Ich hoffe, mit dieser Arbeit einen Beitrag dazu zu leisten, dass in nicht allzu ferner Zukunft Systeme für die morphologische Analyse deutschsprachiger Texte vergleichbar werden und somit zielgerichtet weiterentwickelt und verbessert werden können.

Anhang A

EBNF für Analysestrings

Verwendete Abkürzungen:

Abkürzung	Erklärung
`AdjPos`	Adjektiv mit Steigerung
`Analyses`	Analysestring
`Großb`	Großbuchstabe
`Großkl`	Groß-/Kleinschreibung
`Grundf`	Grundform
`Kleinb`	Kleinbuchstabe
`Morphg`	Morphemgrenzenmarkierer
`Morphm`	Morphologiemerkmale
`Morphs`	Morphologiestring
`Movieru`	Movierung mit *-in* + Umlaut
`Partiz`	Partizip-Ableitung
`Zeichenk`	Zeichenkette
`EBNF`	Erweiterte Backus-Naur-Form

Abbildung A.1: Abkürzungen in der EBNF

Formal beschreibt die morphologische Analyse eine Abbildung von einer Wortform auf eine Menge von Analysestrings:

MA : Wortf → { "Analyses_1" , "Analyses_2" , . . . , "Analyses_n" }

Die Operanden können in einer EBNF-Notation definiert werden wie in Abbildung A.2 dargestellt. Aus Platzgründen wurden die Morphologiemerkmale in Abbildung A.3 ausgelagert. Zusammen beschreibt diese Grammatik alle Analysestrings, die DMOR ausgeben kann.

Großb	=	("A"|"B"|"C"|...|"Z"|"Ä"|"Ö"|"Ü")
Kleinb	=	("a"|"b"|"c"|...|"z"|"ä"|"ö"|"ü"|"ß"|"é")
Zeichenk	=	(Großb Kleinb*)+ | Kleinb+
AdjPos	=	("ADJ.Pos" | "ADJ.Comp" | "ADJ.Sup")
Partiz	=	("^VPAST" | "^VPRES")
Großkl	=	"*"
Movieru	=	"=$in"
Morphg	=	("=" | "·" | "# ")
Morphm	=	("1" | "2" | . . . | "subst" | "zu")
Grundf	=	Zeichenk (Morphg Zeichenk)* Movieru?
Morphs	=	"+" Großb+ Partiz? ("^" (Großb+ | AdjPos))? ("." Morphm)*
Analyses	=	Großklein? Grundf Morphs

Abbildung A.2: EBNF für Analysestrings und Morphologiestrings

Morphm = ("1" | "2" | "3" | "Adj" | "Adv" | "Akk" | "Ant" | "Comp" | "Dat" | "Def" | "Fem" | "Gen" | "Imp" | "Ind" | "Indef" | "Inf" | "Invar" | "Kon" | "Konj" | "Masc" | "NN" | "Neg" | "Neut" | "NoGend" | "Nom" | "PPast" | "PPres" | "Past" | "Pl" | "Pos" | "Pred" | "Pres" | "Sg" | "St" | "St/Mix" | "Sub" | "Sup" | "Sw" | "Sw/Mix" | "Vgl" | "attr" | "mD" | "oD" | "pers" | "prfl" | "pro" | "refl" | "rez" | "subst" | "zu")

Abbildung A.3: Vollständige Auflistung der Morphologiemerkmale

Anhang B

Abkürzungen morphologischer Kategorien im STTS

Wortarten im STTS

Diese Übersicht in Abbildung B.1 ist Schiller et al. (1999), S. 4, entnommen.

1.	Nomina (N)	7.	Adverbien (ADV)
2.	Verben (V)	8.	Konjunktionen (KO)
3.	Artikel (ART)	9.	Adpositionen (AP)
4.	Adjektive (ADJ)	10.	Interjektionen (ITJ)
5.	Pronomina (P)	11.	Partikeln (PTK)
6.	Kardinalzahlen (CARD)		

Abbildung B.1: Morphologische Kategorien und ihre Werte

Morphosyntaktische Kategorien

Kategorisierung	Kategoriekürzel	Kategorie
Kasus	*Akk*	Akkusativ
	Dat	Dativ
	Gen	Genitiv
	Nom	Nominativ
Numerus	*Pl*	Plural
	Sg	Singular
Genus	*Neut*	Neutrum
	Fem	Femininum
	Masc	Maskulinum
	NoGend	ohne Genus
Flexion	*St*	stark
	Sw	schwach
	Mix	gemischt
Grad	*Comp*	Komparativ
	Pos	Positiv
	Sup	Superlativ
Person	*1*	erste
	2	zweite
	3	dritte
Tempus	*Past*	Imperfekt
	Pres	Präsens
Modus	*Ind*	Indikativ
	Konj	Konjunktiv
Definitheit	*Def*	definit
	Indef	indefinit

Abbildung B.2: Morphosyntaktische Kategorien und ihre Werte (1/2)

Kategorisierung	Kategoriekürzel	Kategorie
Verwendung	*Adj*	adjektivisch
	Adv	adverbial
Pronomentyp	*pers*	Personalpronomen
	prfl	reflexives Personalpronomen
	refl	reflexiv
	rez	reziprok
Partikeltyp	*Ant*	Antwortpartikel
	Neg	Negationspartikel
	zu	*zu*
Konjunktionstyp	*Kon*	koordinierend
	Sub	subordinierend
	Vgl	vergleichend
Verbformen	*Inf*	Infinitiv
	Imp	Imperativ
	PPast	Partizip Imperfekt
	PPres	Partizip Präsens
Pronomenflexion	*attr*	attribuierend
	pro	pronominal
	subst	substituierend
Sonderformen	*Invar*	invariant
	Pred	Prädikativ
	mD	mit Determiner
	oD	ohne Determiner
	NN	Substantiv als Verbpartikel (*danksagen*)

Abbildung B.3: Morphosyntaktische Kategorien und ihre Werte (2/2)

Anhang C

Die IMSLEX-DTD

In diesem Abschnitt ist die vollständige DTD für das IMSLEX wiedergegeben, wie sie in Abschnitt 6.2 beschrieben wird (Stand Mai 2004).

```
  <?xml version="1.0" encoding="ISO-8859-1"?>

<!-- IMS, Universität Stuttgart -->
<!-- Arne Fitschen, fitschen@ims.uni-stuttgart.de -->
<!-- letzte Änderung: 14.5.2004 -->

<!-- IMSLex besteht aus beliebig vielen Einträgen von -->
<!-- lexikalischen Einheiten, mindestens aber einer.  -->
<!-- Zusätzlich darf es Schreibvarianten geben, die  -->
<!-- auf eine linguistische Einheit zeigen.  -->

<!ELEMENT lexikon  ( le+ ) >

<!-- Allen lexikalischen Einheit gemein sind gewisse Merkmale -->
<!-- wie Zitierform und Vorkommenshäufigkeit. Weiterhin -->
<!-- unterscheiden wir (Flexions-)Morphologie, Wortbildung, -->
<!-- Semantik und Syntax (Phonetik wurde ausgelagert).  -->
<!-- Die wortartspezifischen -->
<!-- Merkmale werden in Untergruppen zusammengefaßt. -->
<!-- Affix_Merkmale sind spezielle Features für den DeKo-Automaten. -->


<!ELEMENT le     (Globale_Merkmale,
                  Flexionsmorphologie,
                  Wortbildung?,
                  Semantik?,
                  Syntax?,

                  (Substantiv_Merkmale|Verb_Merkmale|Adjektiv_Merkmale|
                   Adverb_Merkmale|Abk_Merkmale|Verbpartikel_Merkmale)?,

                  Affix_Merkmale?,
                  Bearbeitungs_Merkmale?)>
```

```
<!-- Attribute zu lexikalischen Einheiten -->
<!-- Attribute können festgelegt Werte besitzen oder einfach -->
<!-- Text (CDATA). Sie können obligatorisch oder optional sein; -->
<!-- bei festgelegten Werten kann ein Defaultwert vorgegeben -->
<!-- werden, so daß man dieses Attribut im Normalfall nicht -->
<!-- extra anführen muß. -->
<!-- Die id dient der eindeutigen Kennzeichnung der lexikalischen -->
<!-- Einheit. "undef" ist die Default-Belegung, bevor ein Eintrag -->
<!-- bearbeitet wurde. -->

<!ATTLIST le  id        ID                                    #REQUIRED
              kategorie (Substantiv|Verb|Adjektiv|Name|Partikel|
                         Adverb|Numeral|Pronomen|Adposition|Verbpartikel|
                         Konjunktion|Partikelverb|Konfix|Verbpraefix|
                         Adjektivpraefix|Substantivpraefix|Erstglied|
                         Interjektion|Artikel|Invar_Abk|Verbsuffix|
                         Adjektivsuffix|Substantivsuffix|Adverbsuffix|
                         Substantiv_Abk|Name_Abk|Adjektiv_Abk ) #REQUIRED
              m_status  (Frei|Gebunden|Verbbasis|undef)     #REQUIRED
              form      (Simplex|Komplex|Komplex_abstrakt|Kurzwort|
                        Komplex_semi|Nominalisierung|undef) #REQUIRED
              selegiert     (ja|nein|undef)                 #REQUIRED
              lexikalisiert (ja|nein|undef)                 #REQUIRED
              herkunft  (nativ|klassisch|englisch|unklar|
                         französisch|fremd|undef)           #REQUIRED
              akzent    (neutral|beeinflusst|zieht_an) "neutral"
              auslautverhaertung (neutral|blockiert)   "neutral"
              kommentar (nicht_benutzen|
                         in_Phrasen_und_Zusammensetzungen)  #IMPLIED
              erzeugt   (auto|manu)                         #IMPLIED
              geprueft  (ja|nein)                           #IMPLIED
>

<!-- Zitierform ist Infinitiv bei Verben, Nominativ Singular -->
<!-- bei Substantiven, Positiv bei Adjektiven und Nominativ Plural  -->
<!-- bei Pronomen und Artikelwörtern. -->

<!ELEMENT Globale_Merkmale  (
                             Zitierform,
                             PhonetischeTranskription?,
                             Vorkommenshaeufigkeit
                            ) >

<!ELEMENT Vorkommenshaeufigkeit  (#PCDATA) >
<!ATTLIST Vorkommenshaeufigkeit  korpus  (HGC|Referenz)  "HGC"
                                 wert    (wortform)      #IMPLIED>

<!-- Der Teil "Flexionsmorphologie" bildet die Rückwärts- -->
<!-- kompatibilität zum alten Morphologiesystem DMOR. Hier werden  -->
<!-- die Stammformen mit ihren Flexionsklassen aufgelistet. Da -->
```

```
<!-- das alte (und auch das neue) System eine 2-Ebenen-Morphologie -->
<!-- implementieren, müssen z.B. selbst umgelautete Pluralstämme -->
<!-- nicht eigens aufgelistet werden, sondern können aus dem -->
<!-- Singularstamm erschlossen werden. -->

<!ELEMENT Flexionsmorphologie  (
                        Stammformen
                       ) >
<!ATTLIST Flexionsmorphologie
              DMORlex  ( VMod_Stems|VAux_Stems|
                         V-0_Stems|V-ge_Stems|
                         V-0_Stems_NoPref|V-ge_Stems_NoPref|
                         NN_Stems_NoCp|NN_Stems_NoHead|
                         NN_Stems|NE_Stems_NoCp|
                         NE_Stems|NE_Stems_NoHead|
                         ADJ_Stems_NoCp|ADJ_Abbr|NN_Abbr|
                         NE_Abbr|INVAR_Abbr|VPrefSep )    #IMPLIED >

<!-- Unter "Wortbildung" werden zum einen die Phänomene -->
<!-- Derivation und Komposition behandelt, wo sie auftreten. -->
<!-- Bei lexikalischen Einheiten mit morphologischer Form -->
<!-- "komplex" gibt es eine innere Struktur, die unter -->
<!-- "Strukturen" abgelegt werden kann.  -->

<!ELEMENT Wortbildung  (
                        Derivation?,
                        Komposition?,
                        Strukturen?
                       ) >

<!-- "Semantik" ist noch nicht ausgefüllt; wünschenswert wäre -->
<!--  zumindest eine Ontologie für "Anwendungsbereich"... -->
<!-- Zur Zeit wird das Feld benutzt, um bei seltenen Wörtern -->
<!-- eine Erklärung mitzuliefern, so daß nicht immer wieder -->
<!-- nachgeschlagen werden muß, und um andererseits bei -->
<!-- Homographen die Unterscheidung zu ermöglichen: -->
<!-- "Vogelbauer" vs. "Landwirt" bei "Bauer" -->

<!ELEMENT Semantik     (
                        SemantischerTyp?,
                        Kommentar?,
                        Lambdaausdruck?,
                        Praesupposition?,
                        Anwendungsbereich?
                       ) >

<!-- Subkatrahmen existieren bislang für Verben (TSNLP-Format) -->
<!-- und für Adpositionen: "ab" regiert Dativ und Akkusativ, -->
<!-- also Rahmen "obj(NP_dat)" und "obj(NP_acc)" -->

<!ELEMENT Syntax       (
```

```
                        Subkatrahmen*
                       ) >

<!-- Es folgen die einzelnen Wortarten-Merkmale. Es gibt zwar -->
<!-- bereits bei der Flexion und Wortbildung Unterschiede, aber -->
<!-- aber hier sind explizit die grammatischen Eigenschaften -->
<!-- aufgelistet, die z.B. in jedem Wörterbuch bei einem Eintrag -->
<!-- vermerkt sind oder sonstwie interessant (bei Adjektiven und  -->
<!-- Adverbien eher pragmatischer, bei Verben semantischer Art). -->

<!ELEMENT  Substantiv_Merkmale  (
                             Genus
                            ) >

<!-- Verwendung der Adjektive umfaßt "att(ributiv)", "pr(ä)d(ikativ)", -->
<!-- "attprd" (beides) -->

<!ELEMENT  Adjektiv_Merkmale  (
                             Verwendung
                            ) >

<!ELEMENT  Adverb_Merkmale  (
                             Verwendung
                            ) >

<!ELEMENT  Verb_Merkmale  (  Aktionsart,
                             VerbHatResultatszustand,
                             IntensionalitaetLexikalisiert,
                             SemantischeVerbklasse
                            ) >

<!-- Häufiger vorkommende Partikeln für die Bildung -->
<!-- von trennbaren Verben (aufnehmen/ich nehme auf) -->
<!-- werden gelistet. Theoretisch können das Vertreter  -->
<!-- der Klassen Adj, Adv, NN, V, Präp etc. sein.  -->
<!-- Die Partikelverbklasse kommt aus der Klassifikation -->
<!-- von Nadine Aldinger, die Basisverbzahl aus einer -->
<!-- Untersuchung von Partikelverben im HGC. Sie gibt -->
<!-- die Anzahl verschiedener Verben an, die mit dieser -->
<!-- Partikel gebildet wurden. -->

<!ELEMENT  Verbpartikel_Merkmale  (
                               Basisverbzahl,
                               Partikelverbklasse+
                             ) >

<!ELEMENT  Basisverbzahl          (#PCDATA) >
<!ELEMENT  Partikelverbklasse     (#PCDATA) >

<!ELEMENT  Abk_Merkmale  (
                               Ausgeschr_Formen
```

```
                          ) >
<!ELEMENT  Ausgeschr_Formen    ( Ausgeschr_Form+ ) >
<!ELEMENT  Ausgeschr_Form      (#PCDATA) >

<!ELEMENT  Affix_Merkmale   (#PCDATA) >
<!ATTLIST  Affix_Merkmale   produktiv (ja|nein)  #REQUIRED >


<!ELEMENT  Bearbeitungs_Merkmale  (
                           Bearbeitungs_Benutzer,
                           Bearbeitungs_Datum
                          ) >

<!ELEMENT  Bearbeitungs_Benutzer      (#PCDATA) >
<!ELEMENT  Bearbeitungs_Datum         (#PCDATA) >


<!-- Bei jeder Stammform ist die Aussprache angegeben (Sampa) -->
<!-- und die DMORklasse (Rückwärtskompatibilität zur  -->
<!-- alten IMS-Morphologie). Ein Stamm kann in alter und/oder -->
<!-- neuer Rechtschreibung vorliegen. -->

<!ELEMENT   Zitierform                (#PCDATA) >

<!ELEMENT   PhonetischeTranskription  (#PCDATA) >
<!ATTLIST   PhonetischeTranskription
                           notation  ( SAMPA )    "SAMPA"
                           attr        CDATA      #IMPLIED>

<!ELEMENT   Stammformen               ( DMORStamm, Stammform+ ) >

<!ELEMENT   DMORStamm                 ( #PCDATA ) >
<!ATTLIST   DMORStamm  orth     (alt|neu|beides)   "beides">

<!ELEMENT   Stammform                 (Stamm, DMORklasse) >
<!ATTLIST   Stammform id               ID          #IMPLIED
                DMORtyp   (reg|irreg|vollform)    #IMPLIED>

<!ELEMENT   DMORklasse                (#PCDATA) >
<!ELEMENT   Stamm                     (#PCDATA) >
<!ATTLIST   Stamm  orth        (alt|neu|beides)   "beides">

<!ELEMENT   Lambdaausdruck            (#PCDATA) >
<!ELEMENT   Praesupposition           (#PCDATA) >
<!ELEMENT   Anwendungsbereich         (#PCDATA) >


<!-- Generell: Elemente, die mehrfache Vorkommen haben können, -->
<!-- sollten in Container gesteckt werden: Derivationsstaemme  -->
<!-- enthält eine bis viele Derivationsstamm -->
```

```
<!ELEMENT   Derivation           ( Derivationsstaemme? )        >
<!ATTLIST   Derivation  typ      ( ja|nein )      #REQUIRED>

<!ELEMENT   Derivationsstaemme   ( Derivationsstamm+ )         >
<!ELEMENT   Derivationsstamm     ( #PCDATA )                     >
<!ATTLIST   Derivationsstamm       id            ID     #IMPLIED
              typ (normal|umgelautet|kurz|lang|
                vorne_gefugt-getilgt|vorne_gefugt-hinten_gefugt|
                vorne_gefugt|hinten_gefugt|getilgt|
                umgelautet-getilgt|umgelautet-getilgt-hinten_gefugt|
                umgelautet-hinten_gefugt|getilgt-hinten_gefugt)  "normal"
              orth   ( alt|neu|beides )     "beides"
              >

<!ELEMENT   Komposition          ( Kompositionsstaemme? ) >
<!ATTLIST   Komposition  typ     ( ja|nein )  #REQUIRED >

<!ELEMENT   Kompositionsstaemme  ( Kompositionsstamm+ ) >
<!ELEMENT   Kompositionsstamm    ( #PCDATA ) >
<!ATTLIST   Kompositionsstamm    id     ID               #IMPLIED
              typ (normal|umgelautet|kurz|lang|
                vorne_gefugt-getilgt|vorne_gefugt-hinten_gefugt|
                vorne_gefugt|hinten_gefugt|getilgt|verkürzt|
                umgelautet-getilgt|umgelautet-getilgt-hinten_gefugt|
                umgelautet-hinten_gefugt|getilgt-hinten_gefugt)  "normal"
              idiom  ( ja )                #IMPLIED
              orth   ( alt|neu|beides )    "beides"
              >

<!ELEMENT   Strukturen                   (Struktur+) >
<!ELEMENT   Struktur                     (#PCDATA)   >

<!ELEMENT   Genus                        (#PCDATA) >
<!ELEMENT   Subkatrahmen                 (#PCDATA) >
<!ATTLIST   Subkatrahmen     hv    ( haben|sein|
                             haben-variant|sein-variant )  #IMPLIED >

<!ELEMENT   SemantischerTyp              (#PCDATA) >
<!ELEMENT   Kommentar                    (#PCDATA) >

<!ELEMENT   Verwendung                   (#PCDATA) >

<!ELEMENT   Aktionsart                   (#PCDATA) >

<!ELEMENT   VerbHatResultatszustand      (#PCDATA) >
<!ELEMENT   IntensionalitaetLexikalisiert (#PCDATA) >
<!ELEMENT   SemantischeVerbklasse        (#PCDATA) >
```

Anhang D

Beispiele für einen Pflegedialog

In diesem Abschnitt ist der Bildschirmabzug zweier Pflegedialoge wiedergegeben, wie er sich beim Neueintrag der Formen *(die) Wehr* (für Wortbildungen wie *Feuerwehr, Bürgerwehr* etc.) und *Vesuv* darstellt. Durch den Aufruf mit der Option *-f* wird vom Programm zunächst die HGC-Wortliste eingelesen, um später die Vorkommenshäufigkeit im Korpus angeben zu können. Dann wird eine 'Schablone' eingelesen, das ist ein minimaler XML-Eintrag für ein Substantiv. Danach werden nacheinander die *Zitierform, Morphologische Form, Herkunft* und *Flexionsklasse* abgefragt (an dieser Stelle könnten alle möglichen Flexionsklassen ausgegeben werden, was hier aus Gründen der Übersichtlichkeit ausgelassen wurde). Die angegebenen Defaults beziehen sich auf den *undefiniert*-Fall bzw., bei der Flexionsklasse, auf die häufigste Klasse bei Neueinträgen. Als nächstes wird ein Perl-Programm aufgerufen, das alle Formen des Paradigmas generiert (hier nur *Wehr* und *Wehren*) und deren HGC-Vorkommenshäufigkeiten addiert (das Ergebnis ist hier allerdings völlig falsch, da vermutlich die meisten Vorkommen im Korpus auf **(das) Wehr** zurückgehen). Im Anschluss an die Frequenzermittlung wird ein einfaches `grep` (Unix-Kommando) auf allen IMSLEX-Dateien durchgeführt, um zu analysieren, ob es anderswo bereits Formen gibt, die die neue Zitierform als Zeichenkette enthalten. Das Resultat sind einige Eigennamen, der bereits vorhandene Eintrag in der Substantivdatei für *(das) Wehr*, eine Derivation (*Wehrhaftigkeit*) sowie jeweils ein Derivations- und Kompositionsstamm beim Verb *wehren*$^{\mathbb{P}}$. Der Eintrag kann also bestätigt werden, das Resultat ist der komplette XML-Eintrag, der zugleich in das Fenster und in eine Datei geschrieben wird.

Struktureinträge und Derivations- und Kompositionsstammformen werden in diesem Skript noch nicht abgefragt, um den Prozess des Eintragens nicht zu langwierig werden zu lassen. Die `id` beim *le*-Element ergibt sich aus der Systemzeit des Unix-Rechners, also der Anzahl aller Sekunden seit dem 1.1.1970. Dies gewährleistet, dass in den allermeisten Fällen eine eindeutige ID gefunden wird.

Interaktives Erzeugen eines NN-Neueintrags

```
  ./IMSLEX-NN-Neueintrag.perl -f

Lese Frequenzen aus HGC-Wortliste ein (HGC-wortliste-mit-frequenz)...
Done.
Lese Schablone ein (IMSLEX-EINTRAG-SCHABLONE-NN)


   Zitierform? [q] > Wehr


   Morphologische Form?

        0  undef
        1  Simplex
        2  Komplex
        3  Kurzwort


   [0] > 1


   Herkunft?

        0  undef
        1  klassisch
        2  nativ
        3  fremd
        4  unklar
        5  englisch
        6  französisch


   [0] > 2


   Flexionsklasse? [NMasc_s_s] > NFem_0_en


  ==>  Wehr (2165)  Wehren (275)        == macht 2440

IMSLEX_NE.xml:  <Zitierform>Wehr</Zitierform>
IMSLEX_NE.xml:  <DMORStamm>Wehr</DMORStamm>
IMSLEX_NE.xml:      <Stamm>Wehr</Stamm>
IMSLEX_NE.xml:  <Zitierform>Wehrbleck</Zitierform>
IMSLEX_NE.xml:  <DMORStamm>Wehrbleck</DMORStamm>
IMSLEX_NE.xml:      <Stamm>Wehrbleck</Stamm>
IMSLEX_NE.xml:  <Zitierform>Wehrheim</Zitierform>
IMSLEX_NE.xml:  <DMORStamm>Wehrheim</DMORStamm>
IMSLEX_NE.xml:      <Stamm>Wehrheim</Stamm>
IMSLEX_NN.xml:  <Zitierform>Wehr</Zitierform>
```

```
IMSLEX_NN.xml:  <DMORStamm>Wehr</DMORStamm>
IMSLEX_NN.xml:      <Stamm>Wehr</Stamm>
IMSLEX_NN.xml:      <Kompositionsstamm id="nksf74780_1">Wehr</Kompositionsstamm>
IMSLEX_NN.xml:  <Zitierform>Wehrhaftigkeit</Zitierform>
IMSLEX_NN.xml:  <DMORStamm>Wehrhaftigkeit</DMORStamm>
IMSLEX_NN.xml:      <Stamm>Wehrhaftigkeit</Stamm>
IMSLEX_V.xml:      <Derivationsstamm typ="normal">Wehr</Derivationsstamm>
IMSLEX_V.xml:      <Kompositionsstamm id="vksf18871">Wehr</Kompositionsstamm>


Eintragen jetzt ja/nein? [j] > j
<le form="Simplex"  herkunft="nativ"  id="i1087047340"  kategorie="Substantiv"
    lexikalisiert="ja"  m_status="Frei"  selegiert="nein">
<Globale_Merkmale>
  <Zitierform>Wehr</Zitierform>
  <Vorkommenshaeufigkeit korpus="HGC">2440</Vorkommenshaeufigkeit>
</Globale_Merkmale>
<Flexionsmorphologie DMORlex="NN_Stems">
  <Stammformen>
  <DMORStamm>Wehr</DMORStamm>
    <Stammform DMORtyp="reg">
      <Stamm>Wehr</Stamm>
      <DMORklasse>NFem_0_en</DMORklasse>
    </Stammform>
  </Stammformen>
</Flexionsmorphologie>
<Wortbildung>
  <Derivation typ="ja">
    <Derivationsstaemme>
      <Derivationsstamm typ="normal"></Derivationsstamm>
    </Derivationsstaemme>
  </Derivation>
  <Komposition typ="ja">
    <Kompositionsstaemme>
      <Kompositionsstamm id="ksfi1087047340" typ="normal"></Kompositionsstamm>
    </Kompositionsstaemme>
  </Komposition>
  <Strukturen>
    <Struktur></Struktur>
  </Strukturen>
</Wortbildung>
<Semantik>
</Semantik>
<Syntax>
  <Subkatrahmen></Subkatrahmen>
</Syntax>

<Substantiv_Merkmale>
  <Genus>Fem</Genus>
</Substantiv_Merkmale>
</le>
```

```
   Zitierform? [q] > q
```

Ein zweites Beispiel für einen Pflegedialog zeigt den Eintrag eines dritten Vulkans in das Eigennamenlexikon. Hier ist die Auflistung aller in Frage kommenden Flexionsklassen mit dargestellt (es sind bei Eigennamen wesentlich weniger als bei Substantiven), außerdem wird die Auflistung der semantischen Typen gezeigt. Wie an den drei Zeilen zwischen den beiden Auflistungen erkennbar ist, ist der Vesuv deswegen noch nicht als Eigenname im Lexikon eingetragen, weil er bereits über einen Eintrag als Substantiv verfügt.

In diesem Fall dürfte die Angabe der Vorkommenshäufigkeit durchaus realistisch sein, da die beiden verschiedenen Formen im Paradigma vermutlich keine Übereinstimmung mit anderen Wortformen aufweisen.

Interaktives Erzeugen eines NE-Neueintrags

```
  ./IMSLEX-NE-Neueintrag.perl -f
Lese Frequenzen aus HGC-Wortliste ein (HGC-wortliste-mit-frequenz)...
Done.
Lese Schablone ein (IMSLEX-EINTRAG-SCHABLONE-NE)


   Zitierform? [q] > Vesuv


   Morphologische Form?

        0  undef
        1  Simplex
        2  Komplex
        3  Kurzwort


   [0] > 1


   Herkunft?

        0  undef
        1  klassisch
        2  nativ
        3  fremd
        4  unklar
        5  englisch
        6  französisch


   [0] > 1
```

```
   Flexionsklasse?

        0  FamName_0
        1  FamName_s
        2  NGeo+er/in
        3  NGeo-Fem_0
        4  NGeo-Invar
        5  NGeo-Masc_0
        6  NGeo-Masc_s
        7  NGeo-Neut+Loc
        8  NGeo-Neut_0
        9  NGeo-Neut_s
        10  NGeo-Pl_0
        11  NGeo-Pl_x
        12  Name-Fem_0
        13  Name-Fem_s
        14  Name-Masc_0
        15  Name-Masc_s
        16  Name-Neut_s


   [1] > 6


  ==>  Vesuv (63)  Vesuvs (17)  == macht 80

IMSLEX_NN.xml:  <Zitierform>Vesuv</Zitierform>
IMSLEX_NN.xml:  <DMORStamm>Vesuv</DMORStamm>
IMSLEX_NN.xml:      <Stamm>Vesuv</Stamm>


   Semantischer Typ?

        0  Geo: Berg
        1  Geo: Bewohner einer Stadt
        2  Geo: Bewohner eines Landes
        3  Geo: Fluß, See, Gebirge, Region
        4  Geo: Insel
        5  Geo: Kontinent
        6  Geo: Land
        7  Geo: Stadt
        8  Geo: Stamm
        9  Geo: Vulkan
        10  NE: Firma
        11  NE: Gestirn
        12  NE: Märchenfigur
        13  NE: Nachname
        14  NE: Name männlich
        15  NE: Name unbestimmt
```

```
        16  NE: Namenszusatz
        17  NE: Vorname männlich
        18  NE: Vorname weiblich
        19  NE: Währung


   [13] > 9



Eintragen jetzt ja/nein? [j] > j
<le form="Simplex"  herkunft="klassisch"  id="i1087049566"  kategorie="Name"
    lexikalisiert="ja"  m_status="Frei"  selegiert="nein">
<Globale_Merkmale>
  <Zitierform>Vesuv</Zitierform>
  <Vorkommenshaeufigkeit korpus="HGC">80</Vorkommenshaeufigkeit>
</Globale_Merkmale>
<Flexionsmorphologie DMORlex="NE_Stems">
  <Stammformen>
  <DMORStamm>Vesuv</DMORStamm>
    <Stammform DMORtyp="reg">
      <Stamm>Vesuv</Stamm>
      <DMORklasse>NGeo-Masc_s</DMORklasse>
    </Stammform>
  </Stammformen>
</Flexionsmorphologie>
<Wortbildung>
  <Derivation typ="ja">
    <Derivationsstaemme>
      <Derivationsstamm typ="normal"></Derivationsstamm>
    </Derivationsstaemme>
  </Derivation>
  <Komposition typ="ja">
    <Kompositionsstaemme>
      <Kompositionsstamm id="ksfi1087049566" typ="normal"></Kompositionsstamm>
    </Kompositionsstaemme>
  </Komposition>
  <Strukturen>
    <Struktur></Struktur>
  </Strukturen>
</Wortbildung>
<Semantik>
  <SemantischerTyp>Geo: Vulkan</SemantischerTyp>
</Semantik>
<Syntax>
</Syntax>
</le>

   Zitierform? [q] > q
```

Anhang E

Perl-Programm zur Erzeugung des Pflegedialogs

In diesem Abschnitt ist der Quelltext des Perl-Programmes zur Erzeugung des in Anhang D dargestellten Pflegedialoges abgedruckt. Das Programm macht Gebrauch von einem externen Programm, in dem zu einer gegebenen Grundform das Flexionsparadigma ausgegeben wird.

Aufruf und Ausgabe eines externen Programmes

```
  amor.perl -bc Vesuv NGeo-Masc_s

Singular Nominativ (Masc): Vesuv
Singular Genitiv   (Masc): Vesuvs
Singular Dativ     (Masc): Vesuv
Singular Akkusativ (Masc): Vesuv
Plural   Nominativ (Masc): -
Plural   Genitiv   (Masc): -
Plural   Dativ     (Masc): -
Plural   Akkusativ (Masc): -
```

Perl-Programm zum Erzeugen eines NN-Neueintrags

```
  #!/usr/bin/perl -w

# Was macht das Programm?
#
# erstellt IMSLEX-NN-Neueintrag in Datei IMSLEX-kopiere-neu.txt

# Aufruf: ./IMSLEX-NN-Neueintrag.perl -f

# Arne Fitschen, IMS,
# letzte Änderung: Fri May 14 17:57:26 MEST 2004
```

```
######################################################################
## 1. Optionen einlesen
######################################################################

use Getopt::Std;
getopts('hHvdfn:'); # help, verbose, debug, freq
use vars qw( $opt_H $opt_h $opt_v $opt_d $opt_f );

######################################################################
## init
######################################################################

$pfad = "/IMSLEX/";
$outdat = "IMSLEX-Neueintrag.txt";

## Sicherungskopie der alten Ausgabedatei anlegen
if (-f $outdat){system("cp $outdat ${outdat}~");};
open(O, ">$outdat") || die "$!";
%hgc = ();

if ( $opt_f ){

  $hgcdat = "HGC-wortliste-mit-frequenz";
  print STDERR "Lese Frequenzen aus HGC-Wortliste ein ($hgcdat)...\n";
  open(A, "${pfad}$hgcdat") || die "$!";
  while(<A>){
    $hgc{$2} = "$1" if /(\d+)\s+([^\s]+)/;
  }
  close A;
  print STDERR "Done.\n";
}

## Hier Variablen definieren
$h{"XXFORM"} = "undef";     # Default-Belegung
$h{"XXHERK"} = "undef";     #
$h{"XXID"} = "FEHLER";      #
$h{"XXZIT"} = "FEHLER";     #
$h{"XXHGC"} = 0;            #
$h{"XXFLEX"} = "FEHLER";    #
$h{"XXSTRUK"} = "";         #
$h{"XXGEN"} = "FEHLER";     #

$XXFORM{0} = "undef";       # mögliche Werte
$XXFORM{1} = "Simplex";     #
$XXFORM{2} = "Komplex";     #
$XXFORM{3} = "Kurzwort";    #

$XXHERK{0} = "undef";       # mögliche Werte
$XXHERK{1} = "klassisch";   #
```

```
$XXHERK{2} = "nativ";       #
$XXHERK{3} = "fremd";       #
$XXHERK{4} = "unklar";      #
$XXHERK{5} = "englisch";    #
$XXHERK{6} = "französisch"; #


######################################################################
## main
######################################################################

$dat = "IMSLEX-EINTRAG-SCHABLONE-NN";
print STDERR "Lese Schablone ein ($dat)\n";
open(A, "${pfad}$dat") || die "$!";

%file = ();
$cnt = 0;
while ( <A> ){
  $cnt++;
  chomp;
  $file{$cnt} = $_;  ## Schablone zeilenweise in einen Hash lesen
}

############################
## ab hier Benutzer-Dialog
############################

$in = "";
while ( $in ne "q" ){  ## Schleife bis Eingabe 'q' (quit) kommt

  print "\n\n   Zitierform? [q] > ";
  $in = <>; chomp $in; exit if $in eq "q";
  die "Bis dann\n" unless $in;
  $h{"XXZIT"} = $in;  ## Vorsicht:
                      ##  nimmt alles an, was eingegeben wird!!

  print "\n\n   Morphologische Form?\n\n";
  foreach $xy (sort {$a<=>$b} keys %XXFORM){
    print "\t$xy  $XXFORM{$xy}\n"
  }
  print "\n\n   [0] > ";
  $in = <>; chomp $in; exit if $in eq "q";
  $in = 0 unless $in;  ## default auf 1 setzen bei leerer Eingabe
  $h{"XXFORM"} =  $XXFORM{$in} if exists $XXFORM{$in};

  print "\n\n   Herkunft?\n\n";
  foreach $xy (sort {$a<=>$b} keys %XXHERK){
    print "\t$xy  $XXHERK{$xy}\n"
  }
  print "\n\n   [0] > ";
  $in = <>; chomp $in; exit if $in eq "q";
```

```
  $in = 0 unless $in;  ## default auf 1 setzen bei leerer Eingabe
  $h{"XXHERK"} =  $XXHERK{$in} if exists $XXHERK{$in};

  $zeit = time;           ## gibt die Systemzeit zurück
  $h{"XXID"} = "i$zeit";  ## IDs in XML müssen mit Buchstabe anfangen

  print "\n\n   Flexionsklasse? [NMasc_s_s] > ";
  $in = <>; chomp $in; exit if $in eq "q";
  $in = "NMasc_s_s" if $in eq "";
  $h{"XXFLEX"} = $in;   ## Vorsicht:
                        ##  nimmt alles an, was eingegeben wird!!

  if ( $in =~ /(Masc|Neut|Fem)/ ){   ## Genus aus Flexionsklasse
    $h{"XXGEN"} = "$1";
  }

  if ( $opt_f && $h{"XXFLEX"} ){  ## Vorkommenshäufigkeiten ermitteln

    my $w = $h{"XXZIT"};
    my $f = quotemeta( $h{"XXFLEX"} );
    open(B,"amor.perl -bc $w $f|");   ## externes Programm: Generierung
    my %neu = ();                     ## erzeugt Wortformen
    my $out = "";
    while ( <B> ){
      next if /\-\s*$/;
      if ( /.*:\s([^\s]+)\/([^\s]+)$/ ){
$neu{$1} = 1;
$neu{$2} = 1;
      }
      elsif ( /.*:\s([^\s]+)$/ ){
$neu{$1} = 1;
      }
    }
    my @f = sort keys %neu;
    print "\n\n  ==> ";
    grep{
      my $fh = 0;
      $fh = $hgc{$_} if exists $hgc{$_};
      print " $_ ($fh) ";
      $ges += $fh
    }@f;
    print "\t== macht $ges\n\n";
    $h{"XXHGC"} = $ges;
    @f = ();
    $ges = 0;
    close B;
  }

  system("grep $h{'XXZIT'} IMSLEX*xml");

  print "\n\n\nEintragen jetzt ja/nein? [j] > ";
```

```
  $in = <>; chomp $in; $in = "j" unless $in;
  next unless $in eq "j";


  ########################################
  ## Ende: Ausgabe
  ########################################
  foreach $k (sort {$a<=>$b} keys %file){

    $zeile = $file{$k};
    while ($zeile =~
   /(.+)(XXFORM|XXHERK|XXID|XXZIT|XXHGC|XXFLEX|XXSTRUK|XXGEN)(.+)/){

      $zeile = "$1".$h{$2}."$3";
    };

    print "$zeile\n";    ## Ausgabe anzeigen
    print O "$zeile\n";  ## Ausgabe in Datei schreiben
  }

  print STDERR "\n\n";
}

close O;
print STDERR "\n\nDatei $outdat geschrieben.\n\n";
```

Anhang F

XSLT-Stylesheets zum Auslesen des Lexikons

In diesem Abschnitt werden die beiden Stylesheets abgedruckt, mit denen die Flexions- und die Wortbildungsinformation aus dem IMSLEX für die Morphologiekomponente SMOR ausgelesen werden. Beide Stylesheets lesen eine Datei *ersetzungen.xml* ein, in der die im IMSLEX verwendeten ausgeschriebenen (Wortart-)Kategorien zu Abkürzungen in Relation gesetzt werden, wie sie die Morphologiekomponente verwendet. (Für den Hinweis auf diese Möglichkeit danke ich Dr. Wolfgang Lezius.)

Datei *ersetzungen.xml*

```
  <?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>

<ersetzung>

<ersetze quelle="Substantiv"        ziel="NN"/>
<ersetze quelle="Adjektiv"          ziel="ADJ"/>
<ersetze quelle="Verb"              ziel="V"/>
<ersetze quelle="Partikelverb"      ziel="PV"/>
<ersetze quelle="Name"              ziel="NE"/>
<ersetze quelle="Adverb"            ziel="INVAR"/>
<ersetze quelle="Verbpartikel"      ziel="VPART"/>
<ersetze quelle="Name_Abk"          ziel="ABK"/>
<ersetze quelle="Adjektiv_Abk"      ziel="ABK"/>
<ersetze quelle="Substantiv_Abk"    ziel="ABK"/>
<ersetze quelle="Invar_Abk"         ziel="ABK"/>
<ersetze quelle="Adposition"        ziel="ADP"/>
<ersetze quelle="Pronomialadverb"   ziel="PRONADV"/>
<ersetze quelle="Konjunktion"       ziel="KONJ"/>
<ersetze quelle="Numeral"           ziel="CARD"/>
<ersetze quelle="Interjektion"      ziel="INTJ"/>
<ersetze quelle="Pronomen"          ziel="PRON"/>
```

```
<ersetze quelle="Partikel"          ziel="PTKL"/>
<ersetze quelle="Artikel"           ziel="DET"/>
<ersetze quelle="Substantivsuffix"  ziel="NNSUFF"/>
<ersetze quelle="Adjektivsuffix"    ziel="ADJSUFF"/>
<ersetze quelle="Verbsuffix"        ziel="VSUFF"/>
<ersetze quelle="Adverbsuffix"      ziel="ADVSUFF"/>
<ersetze quelle="Substantivpraefix" ziel="NNPRAEF"/>
<ersetze quelle="Adjektivpraefix"   ziel="ADJPRAEF"/>
<ersetze quelle="Verbpraefix"       ziel="VPRAEF"/>
<ersetze quelle="Konfix"            ziel="KONFIX"/>
<ersetze quelle="Erstglied"         ziel="DSF"/>

</ersetzung>
```

Stylesheet für Flexionsinformation

```
  <?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
<xsl:output method="text" encoding="ISO-8859-1"/>


<!-- ******************************************** -->
<!-- Stylesheet-NVA-auslesen.xsl                  -->
<!-- Arne Fitschen, IMS, 2004                     -->
<!--                                              -->
<!-- Dateien Luedelex_*.xsl auslesen:             -->
<!-- Stämme, DMOR-Klassen, DMOR-Unterlexikon      -->
<!-- (mit Kategorie, Herkunft, Stammtyp)          -->
<!--                                              -->
<!-- ******************************************** -->


<!-- ********************************** -->
<!-- root: weitergehen zu jedem le-Element -->
<!-- ********************************** -->

<xsl:template match="lexikon">
  <xsl:apply-templates select="le"/>
</xsl:template>


<!-- ********************************** -->
<!-- le: ab hier Stammformen untersuchen   -->
<!-- ********************************** -->

<xsl:template match="le">

  <!-- zuerst Kategorie durch Kategoriesymbol ersetzen mit Hilfe -->
  <!-- einer Ersetzungsdatei, in der zu 'Nomen' 'NN' steht etc. -->
```

```
  <xsl:variable name="katsymbol">

    <xsl:call-template name="ersetze">

      <xsl:with-param name="quelle" select="@kategorie"/>

    </xsl:call-template>

  </xsl:variable>

  <!-- dann die Stammform-Elemente auslesen, falls vorhanden -->
  <!-- Parameter sind Form, Herkunft und das Kategoriesymbol -->

  <xsl:apply-templates select="Flexionsmorphologie/Stammformen/Stammform">

    <xsl:with-param name="kat"   select="$katsymbol" />
    <xsl:with-param name="herk"  select="@herkunft" />
    <xsl:with-param name="form"  select="@form" />
    <xsl:with-param name="stamm"
                    select="Flexionsmorphologie/Stammformen/DMORStamm" />

  </xsl:apply-templates>

</xsl:template>


<!-- ******************************************************** -->
<!-- Stammformen: je nach DMORtyp Stamm und Klasse ausgeben -->
<!-- ******************************************************** -->

<!-- ******************************************************** -->
<!-- Dies ist das einzige Template, in dem wortart-         -->
<!-- spezifische Information verwendet wird. Alle anderen   -->
<!-- nehmen nur das Kategoriesymbol, das als Variable       -->
<!-- übergeben wird.                                        -->
<!--                                                        -->
<!-- Zuerst Unterlexikon auslesen aus <Komposition>         -->
<!--                                                        -->
<!-- Dann Fallunterscheidung DMORtyp:                       -->
<!-- (<default> ist Aktion "Stamm/DMORklasse auslesen")     -->
<!--                                                        -->
<!--   generell                                             -->
<!--   ========                                             -->
<!--   reg:               <default>                         -->
<!--                                                        -->
<!--   sonst                                                -->
<!--   =====                                                -->
<!--   irreg:  DMORStamm:<default>                          -->
<!--                                                        -->
<!-- ******************************************************** -->
```

```
<xsl:template match="Stammform">

  <xsl:param name="kat"    select="'FEHLER'" />
  <xsl:param name="herk"   select="'FEHLER'" />
  <xsl:param name="form"   select="'FEHLER'" />
  <xsl:param name="stamm"  select="'FEHLER'" />

  <!-- für jede Stammform zunächst Unterlexikon ausgeben -->
  <xsl:apply-templates select="../../../Flexionsmorphologie">
    <xsl:with-param name="kat" select="$kat" />
  </xsl:apply-templates>


  <!-- bei bestimmten DMOR-Typen zusätzlich zum Stamm noch -->
  <!-- den DMORStamm vorher ausgeben -->

  <xsl:choose>

    <!-- Sonderfälle NN/ADJ/V: Komma:Kommata, hoch:höh, back:buk -->
    <xsl:when test="@DMORtyp='irreg'">

      <!-- DMORStamm und : vor dem Stamm ausgeben...  -->
      <xsl:value-of select="$stamm"/><xsl:text>:</xsl:text>

    </xsl:when>

    <xsl:otherwise>
      <!-- default: nur Stamm ausgeben, also hier nichts machen -->
    </xsl:otherwise>

  </xsl:choose>

  <!-- AB HIER wieder für alle Fälle:                 -->
  <!-- Stamm, Kategoriesymbol, Stammtyp, Herkunft und -->
  <!-- DMOR-Klasse ausgeben je Stamm, dann Zeilenende -->

  <!-- zuerst den Stamm ausgeben -->
  <xsl:value-of select="./Stamm"/>

  <!-- Kategorie-Kuerzel -->
  <xsl:text>&#60;</xsl:text>
  <xsl:value-of select="$kat"/>
  <xsl:text>&#62;</xsl:text>

  <!-- Stamm-Typ (hier immer Default 'base') -->
  <xsl:text>&#60;base&#62;</xsl:text>

  <!-- Herkunft -->
  <xsl:text>&#60;</xsl:text>
  <xsl:value-of select="$herk"/>
  <xsl:text>&#62;</xsl:text>
```

```
  <!-- Morphologische Form -->
  <xsl:text>&#60;</xsl:text>
  <xsl:value-of select="$form"/>
  <xsl:text>&#62;</xsl:text>

  <!-- dann die DMOR-Klasse in spitzen Klammern -->
  <xsl:text>&#60;</xsl:text>
  <xsl:value-of select="./DMORklasse"/>
  <xsl:text>&#62;</xsl:text>

  <xsl:text>&#10;</xsl:text>

</xsl:template>


<!-- ******************************************************** -->
<!-- je nach DMORlex DMOR-Unterlexikon ausgeben               -->
<!-- ******************************************************** -->

<!-- ******************************************************** -->
<!-- Fallunterscheidung DMORlex:                              -->
<!--                                                          -->
<!--   nicht vorhanden: Normalfall <Kategoriesymbol>_Stems    -->
<!--   vorhanden: Normalfall oder Sonderfall                  -->
<!--                                                          -->
<!-- ******************************************************** -->

<xsl:template match="Flexionsmorphologie">

 <xsl:param name="kat" select="'FEHLER'" />

 <xsl:choose>

  <!-- wenn Unterlexikonname vorhanden, ausgeben (Vorsicht: -->
  <!-- leerer Wert wird als solcher ausgegeben: <>, darf -->
  <!-- aber nicht vorkommen nach DTD: Aufzählungstyp! -->

  <xsl:when test="@DMORlex">

    <!-- z.B. Sonderfälle NN_Stems/NoHead, NN_Stems/NoCp,  -->
    <!-- z.B. Sonderfälle ADJ_Stems/NoHead, NE_Stems/NoCp etc.  -->
    <xsl:text>&#60;</xsl:text>
    <xsl:value-of select="@DMORlex"/>
    <xsl:text>&#62;</xsl:text>

  </xsl:when>

  <!-- Default-Fall: Lexikonname setzt sich zusammen aus -->
  <!-- dem Kategoriesymbol und "_Stems" -->
  <xsl:otherwise>
```

```
    <xsl:text>&#60;</xsl:text>
    <xsl:value-of select="$kat"/>
    <xsl:text>_Stems</xsl:text>
    <xsl:text>&#62;</xsl:text>

  </xsl:otherwise>

 </xsl:choose>

</xsl:template>


<!-- ****************************************** -->
<!-- "matchall" am Ende: keine weiteren Ausgaben! -->
<!-- ****************************************** -->

<xsl:template match="*"/>


<!-- ****************************************** -->
<!-- Funktion 'ersetze'                         -->
<!-- ****************************************** -->

<xsl:template name="ersetze">

  <xsl:param name="quelle" select="'FEHLER'" />

  <xsl:variable name="myersetze"
    select="document('ersetzungen.xml')/ersetzung/ersetze[@quelle=$quelle]" />

  <xsl:choose>

    <xsl:when test="$myersetze">

      <xsl:value-of select="$myersetze/@ziel"/>

    </xsl:when>

    <xsl:otherwise>
      <xsl:text>FEHLER (Kategoriesymbol nicht in ersetzungen.xml)</xsl:text>
    </xsl:otherwise>

  </xsl:choose>

</xsl:template>


</xsl:stylesheet>
```

Stylesheet für Wortbildungsinformation

```
  <?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
<xsl:output method="text" encoding="ISO-8859-1"/>


<!-- ********************************************** -->
<!-- Stylesheet-DeKo-auslesen.xsl                   -->
<!-- Arne Fitschen, IMS, 2004                       -->
<!--                                                -->
<!-- Dateien Luedelex_*.xsl auslesen:               -->
<!-- Derivations- und Kompositionsstämme mit        -->
<!-- Kategorie, Herkunft und Stammtyp               -->
<!--                                                -->
<!-- ********************************************** -->


<!-- ********************************** -->
<!-- root: weitergehen zu jedem le-Element -->
<!-- ********************************** -->

<xsl:template match="lexikon">
  <xsl:apply-templates select="le"/>
</xsl:template>


<!-- ********************************** -->
<!-- le: ab hier Stammformen untersuchen   -->
<!-- ********************************** -->

<xsl:template match="le">

  <!-- zuerst Kategorie durch Kategoriesymbol ersetzen mit Hilfe -->
  <!-- einer Ersetzungsdatei, in der zu 'Nomen' 'NN' steht etc. -->

  <xsl:variable name="katsymbol">

    <xsl:call-template name="ersetze">

      <xsl:with-param name="quelle" select="@kategorie"/>

    </xsl:call-template>

  </xsl:variable>
```

```
  <!-- dann die Komposition/Derivation-Elemente auslesen -->
  <!-- Parameter sind Herkunft und Kategoriesymbol -->

  <xsl:apply-templates select="Wortbildung/Derivation">

    <xsl:with-param name="kat"  select="$katsymbol" />
    <xsl:with-param name="herk" select="@herkunft" />
    <xsl:with-param name="stamm"
                    select="Flexionsmorphologie/Stammformen/DMORStamm" />

  </xsl:apply-templates>

  <xsl:apply-templates select="Wortbildung/Komposition">

    <xsl:with-param name="kat"  select="$katsymbol" />
    <xsl:with-param name="herk" select="@herkunft" />
    <xsl:with-param name="stamm"
                    select="Flexionsmorphologie/Stammformen/DMORStamm" />

  </xsl:apply-templates>

</xsl:template>


<!-- ************************************************** -->
<!--  Derivation/Komposition: gibt es die für diese Form?   -->
<!-- ************************************************** -->

<xsl:template match="Komposition|Derivation">

  <xsl:param name="kat"    select="'FEHLER'" />
  <xsl:param name="herk"   select="'FEHLER'" />
  <xsl:param name="stamm"  select="'FEHLER'" />

  <xsl:param name="wobiart">

<xsl:choose>
<xsl:when test="local-name(.)='Komposition'">K</xsl:when>
<xsl:otherwise>D</xsl:otherwise>
</xsl:choose>

  </xsl:param>

  <!-- ********************************************************** -->
  <!--  Je nach Wortbildungsart in Derivation/Komposition verzweigen -->
  <!-- ********************************************************** -->

  <xsl:choose>

    <xsl:when test="@typ='ja' and
              $wobiart='D'">
```

```
      <xsl:apply-templates select="Derivationsstaemme/Derivationsstamm">

        <xsl:with-param name="kat"   select="$kat" />
        <xsl:with-param name="herk"  select="$herk" />
        <xsl:with-param name="stamm" select="$stamm" />

      </xsl:apply-templates>

    </xsl:when>

    <xsl:when test="@typ='ja' and
              $wobiart='K'">

      <xsl:apply-templates select="Kompositionsstaemme/Kompositionsstamm">

        <xsl:with-param name="kat"   select="$kat" />
        <xsl:with-param name="herk"  select="$herk" />
        <xsl:with-param name="stamm" select="$stamm" />

      </xsl:apply-templates>

    </xsl:when>

  </xsl:choose>

</xsl:template>



<!-- ********************************************************** -->
<!--  Derivation/Komposition: Stamm und Herkunft etc. ausgeben  -->
<!-- ********************************************************** -->

<xsl:template match="Kompositionsstamm|Derivationsstamm">

  <xsl:param name="kat"    select="'FEHLER'" />
  <xsl:param name="herk"   select="'FEHLER'" />
  <xsl:param name="stamm"  select="'FEHLER'" />

  <xsl:param name="stammtyp">
     <xsl:choose>
<xsl:when test="local-name(.)='Kompositionsstamm'">kompos</xsl:when>
<xsl:otherwise>deriv</xsl:otherwise>
     </xsl:choose>
  </xsl:param>

  <!-- Stamm ausgeben, falls überhaupt vorhanden! -->
  <xsl:choose>

    <!-- ******************************************** -->
```

```
    <!-- ******************  WHEN  ****************** -->
    <!-- ********************************************* -->

    <xsl:when test="string-length(.)>0">

      <!-- für jede Stammform zunächst Unterlexikon ausgeben -->
      <!-- Lexikonname ist entweder D_Stems oder K_Stems,  -->
      <!-- Derivation oder Komposition -->

      <xsl:text>&#60;</xsl:text>
      <xsl:text>DK_Stems&#62;</xsl:text>

      <!-- dann DMORStamm ausgeben und : und Stammform -->
      <xsl:value-of select="$stamm"/><xsl:text>:</xsl:text>

      <!-- Stamm, Kategoriesymbol und Herkunft ausgeben -->
      <!-- je Kompositionsstamm, dann Zeilenende -->

      <!-- zuerst den Stamm ausgeben -->
      <xsl:value-of select="."/>

      <!-- Kategorie-Kuerzel -->
      <xsl:text>&#60;</xsl:text>
      <xsl:value-of select="$kat"/>
      <xsl:text>&#62;</xsl:text>

      <!-- Stamm-Typ ('kompos' bei Komposition, 'deriv' sonst) -->
      <xsl:text>&#60;</xsl:text>
      <xsl:value-of select="$stammtyp"/>
      <xsl:text>&#62;</xsl:text>

      <!-- Herkunft -->
      <xsl:text>&#60;</xsl:text>
      <xsl:value-of select="$herk"/>
      <xsl:text>&#62;</xsl:text>

      <!-- und schließlich Zeilenumbruch -->
      <xsl:text>&#10;</xsl:text>

    </xsl:when>

  </xsl:choose>

</xsl:template>



<!-- ****************************************** -->
<!-- "matchall" am Ende: keine weiteren Ausgaben! -->
<!-- ****************************************** -->
```

```
<xsl:template match="*"/>


<!-- ****************************************** -->
<!-- Funktion 'ersetze'                         -->
<!-- ****************************************** -->

<xsl:template name="ersetze">

  <xsl:param name="quelle" select="'FEHLER'" />

  <xsl:variable name="myersetze"
    select="document('ersetzungen.xml')/ersetzung/ersetze[@quelle=$quelle]" />

  <xsl:choose>

    <xsl:when test="$myersetze">

      <xsl:value-of select="$myersetze/@ziel"/>

    </xsl:when>

    <xsl:otherwise>
      <xsl:text>FEHLER(Kategoriesymbol nicht in ersetze.xml)</xsl:text>
    </xsl:otherwise>

  </xsl:choose>

</xsl:template>


</xsl:stylesheet>
```

Englischsprachige Zusammenfassung

1 Introduction

The morphological analysis of wordforms is impossible without a lexicon. The lexicon, however, has long been neglected in systems for natural language processing. It has usually been seen as a combination of a list of idiosyncratic forms, and a listing of morphemes that word formation rules and inflection rules operate. In this view, there is a direct link between the coverage of a morphological analysis component, and lexicon size. Furthermore, the lexicon, in this model, does not need to have an internal structure at all.

For German, this view does not hold. The main reason for this are morphological processes that change the surface form of morphological units. A morphological analysis component for German has, for example, to deal with the treatment of forms that undergo the process of 'umlautung'. For example, the wordform *Häuschen 'small house'* in German is the diminutive form of the noun *Haus 'house'*. In a strictly concatenative model of morphology, the wordform *Häuschen* can be divided into two parts, *Häus*, and *-chen*.[1] It is generally agreed upon that *Häus* is some kind of realisation of the lexeme *Haus*$^{\mathbb{P}}{}_{NN}$. *-chen*, in contrast, occurs as a suffix in literally thousands of wordforms in the function of a diminutive marker.

Besides these, there are phenomena like neoclassical word formation (wordforms such as *demonstrieren 'to demonstrate', Demonstration 'demonstration'; Biologie 'biology', Biologe 'biologist'*), and conversion (*Segel 'sail', segel(n) 'to sail'*) that are hard to arrive at in morphological analysis. This is largely due to the fact that current morphological analysis systems such as finite-state transducers are based on a model of concatenative processing of defined units. There are, however, no indisputable definitions of neoclassical units, and, in conversion, there is no extra element except for the 'zero morpheme', which is difficult to handle in automatic processing. These, and some other phenomena, pose difficulties for the treatment of word formation in German.

[1] Affixes are marked with a hyphen at the side they attach to a base.

This thesis describes a lexicon that allows for the detailed specification of lexemes, thereby using diverse features to enable an adequate treatment of the phenomena mentioned above. The question asked in this thesis is:

> How must a computational-linguistic lexicon be constructed to optimally support the fully automatical morphological analysis of German?

The thesis deals specifically with word formation phenomena, but the treatment of inflection is included in the description of existing systems.

2 The Idea and Realisation

The idea pursued in this thesis is to identify the relevant morphological processes occurring in the German language, and to assign them to one of two models of linguistic description (see below). Thus, an adequate treatment for any of these can be found, and furthermore, the representation of the corresponding lexical units in the lexicon can be stated. The benefit for the reader is the documentation of a large German morphological lexicon, particularly with regard to its **internal structure**, the **description of the morphological units** contained, and the **knowledge about the relations** between the structure, the units, and morphological analysis.

Two models of linguistic description

Morphological phenomena can be classified into two types (cf. Hockett (1954)):

1. phenomena, that can be explained in a strictly concatenative fashion (the so-called **item and arrangement (IA)** model), and
2. phenomena, that cannot be explained in a strictly concatenative fashion (the so-called **item and process (IP)** model).

A great deal of confusion in word formation analysis stems from the fact that these two models are not sufficiently distinct from one another in the description of phenomena.

The concept of the stem form

Fuhrhop (1998) introduces the concept of the **stem forms**. In a **stem paradigm** – comparable to the notion of the inflectional paradigm – the **compounding stem forms** and the **derivation stem forms** of a lexeme are listed. These are

the surface form a lexeme can take in word formation processes like compounding and derivation, respectively. For example, the string *Häus* in the wordform *Häuschen* can be seen as a derivation stem form of the lexeme *Haus*$^{\mathbb{P}}{}_{NN}$. Similarly, in compounding, in a wordform *Brückenkopf 'bridgehead'*, *Brücken* can be seen as a compounding stem forms of the lexeme *Brücke*$^{\mathbb{P}}{}_{NN}$ *'bridge*'. There may be more than one stem form of each type for a lexeme. All the stem forms for a lexeme put together form the **stem paradigm** of this lexeme.

As a consequence, the morphological analysis component does not have to account for forms that occur in surface forms different from the base form, but they can just pick the units from the lexicon and use them in the word formation rules. With regard to the two models of linguistic description mentioned above it can be said that stem allomorphy allows for the classification of word formation elements as belonging to the realm of IA.

The treatment of IP phenomena

As for the treatment of phenomena which are not as easily transferable into the domain of IA, as stem variations are, in the thesis the concept of **relations** between lexicon entries has been proposed. Here, an XML link from one entry to another can state a relation between the two.

```
<Relation id="r5"  type="Conversion"  no_of_parts="1">
    <part no="r5b1"  category="V"  idref="v14224"
          stype="sem_connection"/>
</Relation>
```

Figure 1: Linking *Spiel*$^{\mathbb{P}}{}_{NN}$, and *spielen*$^{\mathbb{P}}{}_{V}$, in IMSLEX

In the example (see figure 1), this is illustrated for conversion. An XML element named *Relation* is added to the entry for the lexeme *Spiel*$^{\mathbb{P}}{}_{NN}$ *'game'*. Here, the lexeme is linked to another lexeme, uniquely identified by its ID, `v14224` (*spielen*$^{\mathbb{P}}{}_{V}$ *'to play'*). An attribute **stype** ('semantic type') states the type of the relation. In the example, there is a semantic relation between the two. The consequence for morphological analysis is that, in a wordform such as *Spielplatz 'playground'*, the ambiguity between the morphological category of the first part of the compound, *Spiel*, can be resolved because they express the same statement. In contrast to this, there is no semantic connection between the two forms, *Platz*$^{\mathbb{P}}{}_{NN}$ *'place, site'*, and *platzen*$^{\mathbb{P}}{}_{V}$ *'to burst'*, which seem to be related judging by their surface form. Thus, in compounds such as *Platzkarte 'reservation card'* and *Platzregen 'cloudburst'*, both morphological categories of the first part have to be considered for morphological analysis.

In IMSLEX, different kinds of morphological phenomena are described with

different means. This assures transparency, and adequacy of the resources' linguistic description.

The lexicon 'IMSLEX'

The realisation of the lexicon follows the conception developed in the *DeKo* project (cf. Schmid et al. (2001)). In the thesis, the internal structure of the resource is described in great detail. Besides this, the range of subjects mentioned concerning the lexicon comprises the questions of how to fill the lexicon with information, how the human interaction with the lexicon works, how the interface to a subsequent processing component is constructed, and, last but not least, how the lexicon can be maintained.

The lexicon has been realised using the XML standard, thereby ensuring interchangeability of the data. Besides, there is a vast range of tools supporting the automatic processing of XML documents, thereby reducing the need of reinventing the wheel for standard applications like parsing the data, checking for the validity of the resources' structure, and so on.

3 The Contribution of this thesis

The contribution of this thesis is the detailed description of a comprehensive morphological lexicon for German that is based on a sound model of morphological units and processes. Not only have some controversial phenomena been described, but furthermore their treatment in a morphological analysis component has been placed in context of their representation in the lexicon.

As a result, the comparison, and the purposeful refinement of morphological lexicons and morphological analysis components for German is greatly improved.

Literaturverzeichnis

[Aldinger 2002] ALDINGER, Nadine: *Die Argumentstruktur trennbarer Verben im Deutschen*, University of Stuttgart, Master's thesis, 2002

[Baayen 2001] BAAYEN, R. H.: *Word Frequency Distributions*. Dordrecht : Kluwer Academic Publishers, 2001

[Baayen et al. 1995] BAAYEN, R. H. ; PIEPENBROCK, Richard ; GULIKERS, Léon: *The CELEX lexical database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA : Linguistic Data Consortium, 1995

[Bauer 2003] BAUER, Laurie: *Introducing Linguistic Morphology*. 2. Edinburgh : Edinburgh University Press, 2003

[CANOO o.J.] CANOO: o.J.. – URL: http://www.canoo.net/index.html

[CELEX 1995a] CELEX: 1995. – URL: http://www.kun.nl/celex/

[CELEX 1995b] CELEX: 1995. – URL: http://www.kun.nl/celex/subsecs/section_source.html

[CISLEX o.J.] CISLEX: o.J.. – URL: http://www.cis.uni-muenchen.de/projects/CISLEX.html

[Clark 1999] CLARK, James: XSL Transformations (XSLT) 1.0 / W3C. URL `http://www.w3.org/TR/xslt`, 1999. – W3C Recommendation

[Domenig und ten Hacken 1992] DOMENIG, Marc ; HACKEN, Pius ten: *Word Manager: A system for Morphological Dictionaries*. Hildesheim : Olms, 1992

[Donalies 2002] DONALIES, Elke: *Die Wortbildung des Deutschen. Ein Überblick*. Tübingen : Gunter Narr Verlag, 2002 (Studien zur Deutschen Sprache; Bd. 27)

[Duden 2001] DUDEN: *Duden. Deutsches Universalwörterbuch*. 4. Mannheim, Leipzig, Wien, Zürich : Dudenverlag, 2001

[Eckle-Kohler 1999] ECKLE-KOHLER, Judith: *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*. Berlin : Logos Verlag, 1999

[Eisenberg 1994] EISENBERG, Peter: *Grundriß der deutschen Grammatik*. 3. Stuttgart : J.B. Metzler, 1994

[Erben 2000] ERBEN, Johannes: *Einführung in die deutsche Wortbildungslehre*. 4. Berlin : Erich Schmidt Verlag, 2000

[Finkler und Lutzy 1996] FINKLER, Wolfgang ; LUTZY, Ottmar: Standardisierte Selbstdarstellung des Systems MORPHIX. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 67–88

[Fleischer und Barz 1995] FLEISCHER, Wolfgang ; BARZ, Irmhild: *Wortbildung der deutschen Gegenwartssprache*. 2. Tübingen : Max Niemeyer Verlag, 1995

[Fuhrhop 1998] FUHRHOP, Nanna: *Grenzfälle morphologischer Einheiten*. Tübingen : Stauffenburg-Verlag, 1998

[Goldfarb und Rubinsky 1990] GOLDFARB, Charles F. ; RUBINSKY, Yuri: *The SGML handbook*. Oxford, UK : Clarendon Press, 1990

[Gulikers et al. 1995] GULIKERS, Léon ; RATTINK, Gilbert ; PIEPENBROCK, Richard: German Linguistic Guide / Max Planck Institue, Nijmwegen. 1995. – Forschungsbericht

[ten Hacken und Lüdeling 2002] HACKEN, Pius ten ; LÜDELING, Anke: Word Formation in Computational Linguistics. In: *Proceedings of Traitement Automatique de Langue Naturelle* Bd. 2. Nancy, Frankreich, 2002, S. 61–87

[Hanrieder 1996] HANRIEDER, Gerhard: MORPH - Ein modulares und robustes Morphologieprogramm für das Deutsche in Common Lisp. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 53–66

[Harold 2000] HAROLD, Elliotte R.: *Die XML-Bibel*. Bonn : MITP-Verlag, 2000

[Hausser 1996] HAUSSER, Roland: *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics*. Niemeyer, 1996

[Heid 1997] HEID, Ulrich: *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern*. Tübingen : Niemeyer, 1997 (Lexicographica. Series maior, 77)

[Heid 2000] HEID, Ulrich: Morphologie und Lexikon. In: GÖRZ, Günther (Hrsg.): *Handbuch der Künstlichen Intelligenz*. München : Oldenbourg, 2000, S. 665–709

[Heid 2001] HEID, Ulrich: DeKo: Derivations- und Kompositionsmorphologie, Zwischenbericht / IMS, University of Stuttgart. 2001. – Forschungsbericht

[Heid et al. 2002] HEID, Ulrich ; SÄUBERLICH, Bettina ; FITSCHEN, Arne: Using Descriptive Generalisations in the Acquisition of Lexical Data for Word Formation. In: *Proceedings of the 3rd Conference on Language Resources and Evaluation* Bd. IV. Las Palmas de Gran Canaria, Spain : LREC, 2002, S. 86–92

[Heidolph et al. 1981] HEIDOLPH, K. E. ; W., Flämig. ; MOTSCH, W. (Hrsg.): *Grundzüge einer deutschen Grammatik*. Berlin : Akademie Verlag, 1981

[Hockett 1954] HOCKETT, C. F.: Two models of grammatical description. In: *Word* 10 (1954), S. 210 – 231

[Höhle 1982] HÖHLE, Tilman: Über Komposition und Derivation: Zur Konstituentenstruktur von Wortbildungsprodukten im Deutschen. In: *Zeitschrift für Sprachwissenschaft* 1 (1982), S. 76–112

[Kluge 1995] KLUGE: *Etymologisches Wörterbuch*. 23. Berlin, New York : Walter de Gruyter, 1995

[Koskeniemmi und Haapalainen 1996] KOSKENIEMMI, Kimmo ; HAAPALAINEN, Mariikka: GERTWOL – Lingsoft Oy. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 121–140

[Koskenniemi 1983] KOSKENNIEMI, Kimmo: *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Helsinki, University of Helsinki, Dept. General Linguistics, Dissertation, 1983

[Kühnhold et al. 1978] KÜHNHOLD, Ingeburg ; PUTZER, Oskar ; WELLMANN, Hans: *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 3: Das Adjektiv*. Düsseldorf : Pädagogischer Verlag Schwann, 1978

[Kühnhold und Wellmann 1973] KÜHNHOLD, Ingeburg ; WELLMANN, Hans: *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 1: Das Verb*. Düsseldorf : Schwann, 1973

[Langer 1996] LANGER, Stefan: *Selektionsklassen und Hyponymie im Lexikon*, Universität München, Dissertation, 1996. – CIS-Bericht-96-94

[Langer et al. 1996] LANGER, Stefan ; MAIER, Petra ; OESTERLE, Jürgen: CIS-LEX – An Electronic Dictionary for German: Its Structure and a Lexicographic Application. In: *Papers in Computational Lexicography*. Budapest : COMPLEX '96, 1996, S. 155–163

[Leser 1990] LESER, Martin: *Das Problem der 'Zusammenbildungen'. Eine lexikalistische Studie*. Trier : Wissenschaftlicher Verlag, 1990

[Lezius 1996] LEZIUS, Wolfgang: Morphologiesystem MORPHY. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 25–35

[Lezius et al. 2000] LEZIUS, Wolfgang ; DIPPER, Stefanie ; FITSCHEN, Arne: IMSLex – Representing Morphological and Syntactical Information in a Relational Database. In: HEID, Ulrich ; EVERT, Stefan ; LEHMANN, Egbert ; ROHRER, Christian (Hrsg.): *Proceedings of the 9th EURALEX International Congress, Stuttgart, Germany*, 2000, S. 133–139

[Lüdeling und Fitschen 2002] LÜDELING, Anke ; FITSCHEN, Arne: An integrated lexicon for the analysis of complex words. In: *Proceedings of EURALEX 2002* Bd. I. Copenhagen, Denmark : CST Center for Sprogteknologi, 2002, S. 145–152

[Lüdeling und Schmid 2001] LÜDELING, Anke ; SCHMID, Tanja: Does origin determine the combinatory properties of morphological elements in German? In: DECESARIS, Janet (Hrsg.): *Proceedings of the third Mediterranean Meeting on Morphology*. Barcelona, 2001

[Lüdeling et al. 2000] LÜDELING, Anke ; SCHMID, Tanja ; HEID, Ulrich ; SÄUBERLICH, Bettina ; FITSCHEN, Arne ; MÖBIUS, Bernd: Ein integriertes Lexikon / IMS, Universität Stuttgart. 2000. – Forschungsbericht. Manuskript

[Lüdeling et al. 2002] LÜDELING, Anke ; SCHMID, Tanja ; KIOKPASOGLOU, Sawwas: Neoclassical word formation in German. In: *Yearbook of Morphology 2001* (2002)

[Maas 1996] MAAS, Heinz D.: MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 141–166

[Maier-Meyer 1995] MAIER-MEYER, Petra: *Lexikon und automatische Lemmatisierung*, Universität München, Dissertation, 1995. – CIS-Bericht-95-84

[Olsen 1991] OLSEN, Susan: Ge-Präfigierungen im heutigen Deutsch. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur* 113 (1991), S. 332 – 366

[Ortner et al. 1991] ORTNER, Lorelies ; BOLLHAGEN-MÜLLER, Elgin ; ORTNER, Hanspeter ; WELLMANN, Hans ; PÜMPEL-MADER, Maria ; GÄRTNER, Hildegard: *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 4: Substantivkomposita*. Berlin, New York : Walter de Gruyter, 1991

[Paul 1886] PAUL, Hermann: *Principien der Sprachgeschichte*. 2. Halle : Max Niemeyer, 1886

[Pümpel-Mader et al. 1992] PÜMPEL-MADER, Maria ; GASSNER-KOCH, Elsbeth ; WELLMANN, Hans: *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 5: Adjektivkomposita und Partizipialbildungen*. Berlin : Walter de Gruyter, 1992

[SAMPA 1989] SAMPA: 1989. – URL: http://www.phon.ucl.ac.uk/home/sampa/home.htm

[Schiller 1995] SCHILLER, Anne: DMOR: Entwicklerhandbuch. Interner Report. / Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. 1995. – Forschungsbericht

[Schiller 1996] SCHILLER, Anne: Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 37–52

[Schiller et al. 1999] SCHILLER, Anne ; TEUFEL, Simone ; STÖCKERT, Christine ; THIELEN, Christine: Guidelines für das Tagging deutscher Textcorpora mit STTS. Kleines und großes Tagset / Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart and Seminar für Sprachwissenschaft, Universität Tübingen. 1999. – Forschungsbericht

[Schüller und Lorenz 1996] SCHÜLLER, Gerald ; LORENZ, Oliver: LA-Morph – ein linksassoziatives Morphologiesystem. In: HAUSSER, Roland (Hrsg.): *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. 1996, S. 103–119

[Schmid et al. 2004] SCHMID, Helmut ; FITSCHEN, Arne ; HEID, Ulrich: SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In: *Proceedings of the 4th Conference on Language Resources and Evaluation* Bd. ?? Lissabon, Portugal : LREC, 2004, S. ??

[Schmid et al. 2001] SCHMID, Tanja ; LÜDELING, Anke ; SÄUBERLICH, Bettina ; HEID, Ulrich ; MÖBIUS, Bernd: DeKo: Ein System zur Analyse komplexer Wörter. In: LOBIN, Henning (Hrsg.): *Proceedings der GLDV-Frühjahrstagung 2001*, 2001, S. 49 – 57

[Schnorbusch 1998] SCHNORBUSCH, Dieter: *Einfache deutsche Verben. Eine syntaktische und semantische Beschreibung der verbalen Simplizia für das elektronische Lexikonsystem CISLEX*, Universität München, Dissertation, 1998. – to appear?

[Schuch 1990] SCHUCH, Gerhild v.: *Einführung in die Sprachwissenschaft*. München : Ars Una, 1990

[Spencer 1991] SPENCER, Andrew: *Morphological Theory. An Introduction to Word Structure in Generative Grammar*. Oxford : Blackwell, 1991

[Sproat 2000] SPROAT, Richard: Lextools: a toolkit for finite-state linguistic analysis. 2000. – Forschungsbericht. `http://www.research.att.com/sw/tools/lextools/`

[Trommer 2001] TROMMER, Jochen: Morphologie. In: CARSTENSEN, Kai-Uwe ; EBERT, Christian ; ENDRISS, Cornelia ; JEKAT, Susanne ; KLABUNDE, Ralf ; LANGER, Hagen (Hrsg.): *Computerlinguistik und Sprachtechnologie - Eine Einführung*. Heidelberg, Berlin : Spektrum Akademischer Verlag, 2001, S. 175–202

[Trost 2003] TROST, Harald: Morphology. In: MITKOV, Ruslan (Hrsg.): *The Oxford Handbook of Computational Linguistics*. Oxford, New York : Oxford University Press, 2003, S. 25–47

[Uszkoreit 2000] USZKOREIT, Hans: 2000. – URL: http://www.coli.uni-sb.de/ hansu/VLCL_Sprachtechnologie.PDF

[Vossen 1994] VOSSEN, Gottfried: *Datenmodelle, Datenbanksprachen und Datenbank-Management-Systeme*. 2. Addison-Wesley, 1994

[Wall et al. 2000] WALL, Larry ; CHRISTIANSEN, Tom ; ORWANT, Jon: *Programming Perl*. 3rd edition. O'Reilly, 2000

[Wellmann 1975] WELLMANN, Hans: *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 2: Das Substantiv*. Düsseldorf : Pädagogischer Verlag Schwann, 1975

[Wilmanns 1899] WILMANNS, W.: *Deutsche Grammatik. Gotisch, Alt-, Mittel- und Neuhochdeutsch. Zweite Abteilung: Wortbildung*. 2. Berlin : Walter de Gruyter & Co., 1899

[Zipf 1949] ZIPF, George K.: *Human Behavior and the Principle of Least Effort*. Cambridge, MA : Addison-Wesley, 1949