A Auszüge aus CISLEX-EF+

Einträge im Grundformenlexikon (ohne semantische Kodierung):

```
Buhlerei;fem;NS0;NP3
Buhlerin;fem;NS0;NP5
Buhler;mask;NS2;NP1

buhlerisch,.ADJ0

buhlen,.VSW1
```

Die entsprechenden Einträge im Vollformenlexikon:

```
%Buhlerei,Buhlerei.fem(NS0,NP3):snf:sgf:sdf:saf
%Buhlereien,Buhlerei.fem(NS0,NP3):pnf:pgf:pdf:paf
%Buhlerin,Buhlerin.fem(NS0,NP5):snf:sgf:sdf:saf
%Buhlerinnen,Buhlerin.fem(NS0,NP5):pnf:pgf:pdf:paf
%Buhler,Buhler.mask(NS2,NP1):snm:sdm:sam:pnm:pgm:pam
%Buhlern,Buhler.mask(NS2,NP1):pdm
%Buhlers,Buhler.mask(NS2,NP1):sgm

%buhlerisch,buhlerisch.ADJ0:pos
%buhlerische,buhlerisch.ADJ0:pos_e
%buhlerischem,buhlerisch.ADJ0:pos_m
%buhlerischen,buhlerisch.ADJ0:pos_n
%buhlerischer,buhlerisch.ADJ0:pos_r

%buhle,buhlen.VSW1:spi1:spk1:spk3
%buhlen,buhlen.VSW1:inf:ppi1:ppi3:ppk1:ppk3
%buhlend,buhlen.VSW1:part1
%buhlest,buhlen.VSW1:spk2
%buhlet,buhlen.VSW1:ppk2
%buhlst,buhlen.VSW1:spi2
%buhlt,buhlen.VSW1:spi3:ppi2
%buhlte,buhlen.VSW1:sii1:sii3:sik1:sik3
%buhlten,buhlen.VSW1:pii1:pii3:pik1:pik3
%buhltest,buhlen.VSW1:sii2:sik2
%buhltet,buhlen.VSW1:pii2:pik2
%gebuhlt,buhlen.VSW1:part2
```

5 Literatur

Bergenholtz, H. & B. Schaeder (1977): Die Wortarten des Deutschen. Stuttgart: Klett.

Courtois, B. (1989): Un système de dictionnaires électroniques pour les mots simples du français. Report No 9, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

DUDEN-Redaktion (Hrsg.) (1986): DUDEN "Rechtschreibung der deutschen Sprache und der Fremdwörter". Mannheim; Wien; Zürich: Bibliographisches Institut.

Gross, G. (1991): La forme d'un dictionnaire électronique. LADL-Report, Laboratoire d'Automatique Documentaire et Linguistique, Université Paris 7.

Gross, M. (1989): The Use of Finite Automata in the Lexical Representation of Natural Lnaguage. In: Gross, M. & D. Perrin (eds.), Electronic Dictionaries in Computational Linguistics, LITP Spring School on Theoretical Computer Science Saint-Pierre d'Oléron, France, May 1987 Proceedings. Berlin, Heidelberg, New York.

Mackensen, L. (1955): Deutsches Wörterbuch. Laupheim.

Maier, P. (1994): Die morphologische Kodierung des CISLEX. Interner Bericht, CIS, Universität München.

Sinclair, J.M., Fox, G. et al. (1988): Collins COBUILD English Language Dictionary. London: Collins.

Silberztein, M. (1989): Dictionnaires électroniques et reconnaissance lexicale automatique. Thèse de doctorat, Université Paris VII.

Wahrig, G. (1975): Deutsches Wörterbuch. Gütersloh.

ist zur Zeit am CIS in Bearbeitung.[22] Für die hier auftretenden Probleme werden lokale Grammatiken entwickelt, die in der Lage sind, solche Einheiten korrekt zu identifizieren.

4.2 Automatische Textindizierung

Die Indizierung der Dokumente ist die Grundvoraussetzung für Information Retrieval aus Volltexten. Da eine solche Indizierung bei großen Datenmengen nicht mehr manuell vorgenommen werden kann, müssen Methoden entwickelt werden, den Informationsgehalt eines Textes so exakt wie möglich durch die Indizierung abzubilden. Herkömmliche Verfahren zur automatischen Indizierung von Texten basieren im wesentlichen auf einer Trunkierung von vorgegebenen Schlagwörtern. Eine korrekte und vollständige Indizierung ist jedoch nur möglich, wenn zu jedem Stichwort alle Flexionsformen und alle relevanten Vorkommen in Komposita im Text identifiziert werden können. Das setzt aber eine linguistische Aufbereitung der Dokumente voraus. Diese linguistische Aufbereitung kann sich auf verschiedenen Ebenen abspielen. Die elementarste Ebene ist die morphologische Ebene, die Ebene der Textlemmatisierung also. Auf dieser Ebene werden alle Wortformen bzw. das Vorkommen von Wortformen in Komposita nach ihrer Grundform indiziert. Obwohl mit einer korrekten morphologischen Indizierung schon erheblicher Fortschritt erzielt werden kann, ist eigentlich eine Indizierung auf semantischer Ebene wünschenswert, die nicht nur einzelne Wörter, sondern größere Einheiten wie Phrasen, Kollokationen usw. auf eine kanonische *semantische Grundform*, den ausgedrückten Gedanken, zurückführt und nach diesem indiziert. Eine solche kanonische semantische Grundform wird beispielsweise durch die in Silberztein (1989) beschriebenen endlichen Automaten bereitgestellt. Mithilfe solcher Automaten ist es möglich nicht nur nach Stichworten, sondern beispielsweise nach allen möglichen Paraphrasen zu einer semantischen Grundform zu suchen.

4.3 Mehrsprachige Wörterbücher

Die Kodierung des CISLEX-Kernlexikon erfolgte bewußt in Übereinstimmung mit den entsprechenden Wörterbüchern fürs Französische, Englische, Italienische, Spanische, Portugiesische usw., die entweder direkt am LADL in Paris oder in den entsprechenden Ländern unter Regie des LADL entwickelt wurden. Ein wichtiges Forschungsziel ist die wechselseitige Abbildung dieser Wörterbücher aufeinander. Dies wird mithilfe einer automatischen Auswertung paralleler mehrsprachiger Korpora geschehen. Ein derartiges System mehrsprachiger, nach Sprachpaaren verbundener Wörterbücher, zusammen mit der in Entwicklung befindlichen semantischen Klassifizierung der Einträge, eröffnet für die maschinelle Übersetzunug völlig neue Möglichkeiten. So ist die schrittweise Weiterentwicklung vom reinen Wörterbuch hin zum Phrasenwörterbuch, in dem sowohl Terminologie als auch Mehrwortlexeme mehrsprachig dargestellt werden, durch die Auswertung der Korpora auf der Basis der einfachen Wörterbücher möglich.

[22] Einzelne Konstruktionen, wie etwa die Analyse der Bindstrichkomposita oder der numerischen Ausdrücke sind bereits abgeschlossen.

Lemmatisierung verstehen wir die Rückführung von Wortformen auf eine kanonische Grundform und die Identifizierung der morpho-syntaktischen Merkmale, die durch diese Wortform repräsentiert werden. Der Begriff des Tagging ist wesentlich allgemeiner gefaßt und bezeichnet lediglich die Annotierung von Text mit bestimmten Etiketten (den sog. Tags). Diese Tags können sowohl lexikalische, semantische als auch Strukturinformation beinhalten. Wir verstehen unter Tagging hier im engeren Sinne eine Annotierung von Texten mit Kategorien, die sich aus dem Lexikon ableiten lassen. In diesem Sinne ist Lemmatisierung als Spezialfall des Tagging zu sehen.
Herkömmliche Lemmatisierungssysteme arbeiten in erster Linie regelbasiert und verwenden nur ein relativ kleines Stamm- oder Grundformenwörterbuch. In einem System wie dem CISLEX reduziert sich die Wortlemmatisierung[20] der einfachen Formen auf den 'lexical look-up' im bereits erzeugten Vollformenlexikon. Die eigentliche Hauptaufgabe besteht in der korrekten Zerlegung der Komposita und der Analyse von Sonderformen (Bindestrichwörter, Zahlen, Mischformen usw.). Auf diese Punkte werden wir im folgenden noch kurz eingehen.

4.1.1 Kompositabehandlung

Wie schon erwähnt, enthält der zentrale Teil des CISLEX zwei Teillexika: die einfachen Formen (EF+) und die komplexen Formen (KF). Obwohl bisher eine große Anzahl von Komposita erfaßt und kategorisiert worden ist, wird man bei der Lemmatisierung nicht davon ausgehen können, daß alle für eine Textanalyse relevanten Lemmata in KF vorhanden sein werden. Die Analyse neuer bzw. bisher noch nicht erfaßter Komposita geschieht anhand eines Kompositazerlegungsalgorithmus. Die Zerlegung beruht auf der Tatsache, daß nach der Definition jede komplexe Form zerlegbar ist in einen maximalen rechten Teil, dessen Lemmatisierung in EF+ (genauer gesagt im Vollformenlexikon, das auf EF+ basiert) erfolgen kann, und einen linken Teil, der entweder (i) als entsprechendes Präfix im Morphlexikon ist oder (ii) als mögliches Vorderglied in EF+ ist oder (iii) eine Folge von komplexen Formen ist, deren maximaler rechter EF+-Suffix ein mögliches Vorderglied ist.[21] Mit diesem Verfahren lassen sich die Komposita fast ausnahmslos korrekt segmentieren.

4.1.2 Erkennung von Sonderformen

Neben der Segmentierung von Komposita spielt die Erkennung und Analyse von Sonderformen in Texten eine sehr wichtige Rolle. An Sonderformen können die Abkürzungen und Akronyme direkt im entsprechenden CISLEX-Modul nachgeschlagen werden. Daneben gibt es aber noch eine Vielzahl von Formen, die speziell analysiert werden müssen, wie etwa Bindstrichkomposita, Bindestrich-Koordinationsglieder, Wörter, die aus einem numerischen Ausdruck oder Sonderzeichen und produktiven Prä- oder Suffixen bestehen, sowie eine Vielzahl weiterer Konstruktionen, die nicht-alphanumerischen Ketten enthalten (das sind unter anderem alle Zahlenangaben, Datumsangaben usw.). Eine umfassende Behandlung dieser Phänomene

[20] Für eine Disambiguierung der ambigen Formen ist natürlich weiterhin eine Lemmatisierung auf Phrasen- oder Satzebene notwendig.

[21] Mögliche EF+-Vorderglieder sind z.B. alle Fugenformen von Nomen, die Verbstämme usw.

3.3 Sonderfälle

Da die Wortklassen neben morphologischen Eigenschaften vor allem auf syntaktischen Funktionen beruhen, kommt es bei Wörtern, die verschiedene syntaktische Funktionen erfüllen können, zu einer Mehrfachklassifikation. Davon betroffen sind sowohl die Funktionswortarten als auch die Hauptwortarten, wie am Beispiel der adjektivischen Nomen bereits deutlich wurde. Hier muß, um eine unnötige Vervielfachung der Einträge zu verhindern, zwischen dem systematischen Vorkommen in anderen Funktionen und zufälligen Idiosynkrasien unterschieden werden. Gibt es eine Regularität, nach der alle (oder bis auf bestimmte Ausnahmen alle) Elemente einer Wortklasse auch in einer anderen syntaktischen Funktion auftreten, so ist dies als allgemeine Regel zu formulieren. Typisch sind hier die adjektivischen Nomen[17] und der adverbiale Gebrauch von Adjektiven[18]. Ein ähnlicher Fall wie bei den adjektivischen Nomen liegt bei Pronomen und Determinatoren vor. Im CISLEX wird davon ausgegangen, daß sämtliche Determinatoren auch pronominal auftreten können.[19] Im Gegensatz zu diesen Fällen handelt es sich bei der Mehrfachklassifizierung vieler Funktionswörter um zufällige Ambiguitäten, die alle kodiert werden mußten.
Während es sich bei den oben besprochenen Fällen um Wortartwechsel bei unveränderter Morphologie (höchstens bei einzelnen Formen eines Paradigmas konnten Unterschiede auftreten) handelte, sollen nun Fälle betrachtet werden, bei denen Lexeme ohne explizite Derivation in einer anderen Wortart und mit entsprechend anderer Morphologie auftreten. Beispiele hierfür sind unter anderem die partizipialen Adjektive und die nominalisierten Infinitive. Während im Fall der nominalisierten Infinitive die Morphologie prädiktabel ist (die Genitiv-Singular-Form lautet Infinitiv+s, Plural ist auch morphologisch nicht möglich), lassen sich die morphologischen und morphosyntaktischen Eigenschaften (Graduierbarkeit und attributiver, adverbialer oder prädikativer Gebrauch) nicht eindeutig vom Verb ableiten. Die partizipialen Adjektive wurden deshalb im CISLEX explizit als Adjektive kodiert.

4 Anwendungen

Zum Abschluß wollen wir nun noch näher auf einige typische Anwendungen elektronischer Wörterbücher eingehen. Wir werden anhand der Anwendungsbereiche Lemmatisierung, Indizierung und automatische Übersetzung die Notwendigkeit der Erstellung und Pflege elektronischer Wörterbücher im Sinne der eingangs gestellten Anforderungen aufzeigen.

4.1 Lemmatisierung und Tagging

Die korrekte Lemmatierung bzw. das korrekte Tagging ist für eine automatische Bearbeitung von Korpora jeglicher Art die wichtigste Grundvoraussetzung. Unter

[17] Nur adjektivische Nomen vom Typ *Beamter*, denen im Lexikon kein Adjektiv entspricht, müssen explizit als adjektivisches Nomen kodiert werden.

[18] Das adverbiale Auftreten der Adjektive wird durch ein Merkmal +adverbial beim Adjektiv kodiert. Nur in Fällen, in denen das Adverb lexikalisiert ist und die entsprechenden Konstruktionen ambig sind, wie in *Er sagte das bestimmt.* wird ein gleichlautendes Adverb explizit kodiert.

[19] Pronominale Formen, die im Determinatorparadigma nicht auftreten, müssen explizit kodiert werden.

werden. Die Konjunktionen werden unterschieden in subordinierende und koordinierende Konjunktionen. Typische Konjunktionen sind *aber, denn, und, weil*, usw.

Interjektion (INTJ) Die Interjektionen nehmen bei den Wortarten eine Sonderstellung ein, da sie selbst Satzfunktion übernehmen können. Wenn Interjektionen innerhalb eines Satzes vorkommen, so ist das Auftreten beschränkt auf Verbzweitsätze und dort auf die Position vor dem Vorfeld. Beispiele für Interjektionen sind: *ätsch, ahoi, hatschi, frischauf* usw.

3.2 Die morphologischen Kategorien

Die flektierenden Wortklassen Nomen, Verb, Adjektiv, Pronomen und Determinator müssen nach morphologischen Kriterien weitersubklassifiziert werden. Diese Subklassifizierung richtet sich danach, wie die Wortart-typischen morpho-syntaktischen Merkmale realisiert werden. Bei der Subklassifizierung stand die Operationalisierbarkeit der morphologischen Kategorien im Vordergrund. Die Frage, was im linguistischen Sinn ein korrekter Stamm und was die entsprechenden Flexionsendungen sind, wurde vernachlässigt zugunsten einer formal-orientierten Oberflächenbeschreibung des Verhältnisses zwischen den verschiedenen Flexionsformen. Die morphologischen Klassen sind einzig dadurch bestimmt, wie aus der Grundform (das ist in der Regel der verbale Infinitiv, die Prädikativform beim Adjektiv und die Nominativ-Singular-Form bei Nomen) die anderen Formen gebildet werden.
Die Nomen wurden hinsichtlich Singularmorphologie und Pluralmorphologie kodiert. Die Eigenschaft von Nomen nur pluralisch bzw. nur singularisch aufzutreten wird als spezielle morphologische Singular- bzw. Pluralklasse aufgefaßt.[16] Zusammen mit der Genusinformation führt diese Kodierung zu insgesamt ca. 250 morphologisch verschiedenen Nomentypen.
Bei den Adjektiven sind für die morphologische Subklassifikation die orthographischen Regularitäten bei der Bildung der Flexionsformen und der Komparationsformen relevant, die zu 18 verschiedenen Typen führen.
Die Verben sind unterteilt in starke Verben, deren Tempusmarkierung durch Ablaut geschieht, schwache Verben, bei denen dem gesamten Paradigma derselbe Stamm zugrundeliegt und unregelmäßige Verben, zu denen sowohl völlig unregelmäßige Bildungen, als auch Präteritopräsentia und die sogenannten Rückumlautverben gehören. Diese 3 Typen sind nach den orthographischen Besonderheiten bei der Bildung der Formen weiter unterteilt. Bei den schwachen Verben handelt es sich lediglich um phonologisch bzw. orthographisch bedingte Veränderungen, die bei der Konkatenation mit der Flexionsendung auftreten.
Die Determinatoren und Pronomen werden aufgrund ihrer geringen Anzahl direkt als Vollformen aufgenommen.

[16] Wir gehen hierbei von einem morphologischen Begriff Singularia bzw. Pluralia Tantum aus. Das heißt, wenn die entsprechenden Formen morphologisch möglich sind, so werden sie auch als entsprechende Singular- oder Pluralklasse kodiert, unabhängig von semantischen oder pragmatischen Einschränkungen.

Determinatoren (DET) Als Determinatoren werden im CISLEX Elemente bezeichnet, die an erster Stelle in einer Nominalphrase, also vor dem Adjektiv stehen (außer den pränominalen Genitivattributen). Typische Vertreter der Determinatoren sind der definite Artikel, Demonstrativa oder Possessiva. Auch ein Teil der traditionellerweise als Quantoren bezeichneten Elemente gehört zu den Determinatoren, falls sie bei einem folgenden Adjektiv ein bestimmtes Deklinationsmuster fordern.

Pronomen (PRON) Pronomen sind Wörter, die allein eine Nominalphrase bilden können und nicht zusammen mit Determinatoren und attributiven Adjektiven vorkommen. Zu den Pronomen gehören Personalpronomen, Relativpronomen, Reflexivpronomen, Fragepronomen usw. Neben diesen ausschließlich pronominal verwendbaren Wörtern gehen wir davon aus, daß jeder Determinator auch pronominal auftreten kann.

Adverb (ADV) Die Adverbien haben als einzige Klasse der nicht-flektierenden Wortarten Satzgliedcharakter, d.h. sie können allein das Vorfeld eines Verbzweit-Satzes bilden.

Partikel (PART) Der Terminus *Partikel* wird in der Literatur unterschiedlich gebraucht. Häufig werden als Partikel alle nicht-flektierenden Wortarten bezeichnet. Wir gehen hier von einem engeren Partikelbegriff aus, der eine Klasse von nicht-flektierenden Wörtern bezeichnet, die sich distributionell von Präpositionen, Adverbien und Konjunktionen unterscheiden. Von den Adverbien unterscheiden sich die Partikeln im wesentlichen dadurch, daß sie nicht vorfeldfähig sind. Im Gegensatz zu Konjunktionen verknüpfen sie keine Satzteile miteinander, und im Gegensatz zu Partikeln haben Präpositionen keine Kasusrektion. Die Klasse der Partikeln ist also in gewissem Sinn eine Restklasse im Bereich der nicht-flektierenden Wortarten. Typische Partikeln sind *also, bloß, eben, sehr, nur*, usw.

Verbpartikel (VPART) Der größte Teil der vorkommenden abtrennbaren Verbpartikeln sind Elemente anderer Wortklassen, die auch in anderer Funktion frei vorkommen und müssen deshalb nicht speziell klassifiziert werden. Die Klasse *Verbpartikel* ist nur für solche Lexeme gedacht, die nur in dieser Funktion vorkommen, wie beispielsweise *inne* in *innehalten* oder *abhanden* in *abhandenkommen*. Da diese Elemente in Verbzweitsätzen in abgetrennter Form vorkommen, müssen sie im Rahmen der einfachen Formen kodiert werden.

Präposition (PRAEP) Die Präpositionen sind dadurch gekennzeichnet, daß sie als Kopf einer Präpositionalphrase den Kasus der folgenden Nominalphrase (im Falle von Postpositionen, die hier nicht von den Präpositionen unterschieden werden, ist natürlich die vorangehende Nominalphrase zu betrachten) festlegen, in manchen Fällen auch die Präposition einer folgenden Präpositionalphrase. An syntaktischer Information wird bei den Präpositionen der Kasus bzw. die Präposition des Komplements kodiert.

Konjunktion (KONJ) Konjunktionen verknüpfen Konstituenten miteinander. Sie haben selbst keinen Satzgliedstatus, können also nicht allein im Vorfeld eines deutschen Satzes stehen. Sie können innerhalb des Satzes nicht verschoben

um eine konsistente Klassifizierung zu gewährleisten, und andererseits muß die Klassifikation so theorieneutral wie möglich und für die intendierten Anwendungen sinnvoll sein. Dies kann eine semantisch motivierte Klassifikation nicht leisten. Auch eine morphologisch motivierte Klassifikation (d.h. nach den realisierbaren morphologischen Merkmalen und der Art ihrer Realisierung) scheidet aus, da sie i) nicht in der Lage ist, den Bereich der Funktionswörter adäquat zu erfassen und ii) die Klassifizierung von nichtflektierenden Wörtern, die intuitiv flektierenden Wortklassen zuzurechnen sind, erhebliche Schwierigkeiten bereitet.[15] Wir gehen daher von einem Wortartensystem aus, das die Morphologie zwar nicht völlig ignoriert, bei dem aber die distributionellen Kriterien als primär gegenüber den morphologischen Kriterien angenommen werden. Dies erlaubt unter anderem auch eine zufriedenstellende Klassifikation von Nomen mit adjektivischer Deklination, wie *Verwandter*, *Angestellter* oder *Beamter*, die je nach Artikelwahl unterschiedlich dekliniert werden. Dieses Vorgehen hat allerdings auch zur Folge, daß insbesondere bei den Funktionswörtern zahlreiche Mehrfachklassifikationen zu verzeichnen sind, da gerade in diesem Bereich viele Wörter unterschiedliche syntaktische Funktionen erfüllen können. Ist für bestimmte Anwendungen eine so feine Unterscheidung nicht notwendig, oder durch die damit verbundene Mehrfach-Kategorisierung zu aufwendig, können die nichtflektierenden Wortarten als eine Klasse betrachtet werden. Im CISLEX werden folgende Wortarten unterschieden:

Nomen (N) Als Nomen werden im CISLEX alle lexikalischen Einheiten betrachtet, die Kopf einer Nominalphrase sein können und im Unterschied zu Pronomen zusammen mit einem Artikel auftreten können. Nomen sind bezüglich Genus und der Verwendung im Singular und Plural markiert. Nach morphologischen und syntaktischen Kriterien ergeben sich bei den Nomen folgende Subklassen: (i) N: reguläre Nomen mit nominaler Deklination, (ii) NA: Nomen mit adjektivischer Deklination und (iii) NQ: quantorenähnliche Nomen, die sich von regulären Nomen dadurch unterscheiden, daß sie als Maßnomen das Kopfnomen einer Nominalphrase modifizieren können.

Adjektiv (A) Adjektive sind dadurch bestimmt, daß sie pränominal (in der Nominalphrase zwischen Determinator und Kopfnomen) oder prädikativ vorkommen können. In der attributiven Verwendung kongruieren Adjektive, wenn sie flektierbar sind, hinsichtlich Kasus, Genus und Numerus mit dem Kopfnomen; die Art der Deklination ist vom vorangehenden Determinator abhängig. Die Adjektive sind hinsichtlich ihrer Distribution nach den Merkmalen attributiv, prädikativ und adverbial kodiert. Ebenso muß die Möglichkeit der Komparation, die nicht bei allen Adjektiven gegeben ist, explizit kodiert werden.

Verb (V) Die Verben lassen sich im Gegensatz zu den anderen Wortklassen auch rein morphologisch definieren: Ein Verb ist ein Wort, das ein verbales Paradigma hat. Wobei für ein verbales Paradigma in erster Linie die Tempusmarkierung ausschlaggebend ist. Nach der Distribution werden bei den Verben folgende Subklassen unterschieden: (i) Auxiliare und Modalverben, die nicht selbständig eine Verbalphrase bilden können, und (ii) Vollverben.

[15] *super* oder *lila* im adjektivischen Bereich sowie *Abakus* oder *Kasus* im nominalen Bereich wären Beispiele für solche Problemwörter, die, will man nicht die verschiedensten Nullmorpheme annehmen, nicht als Adjektiv bzw. Nomen klassifiziert werden könnten.

andererseits den Fachwortschatz[13] der nach verschiedenen Fachbereichen getrennt erfaßt wird.

2.4 Sonderformen

Einheiten, die nicht dem Wortbereich angehören, werden als Sonderformen erfaßt. Im Moment enthält das CISLEX an Sonderformen Abkürzungen, Akronyme und Morphe. Die Morphe sind nach ihrer Stellung weiter unterteilt in Präfixe, Suffixe, Fugen und Stämme. Die Morphlisten spielen insbesondere im Hinblick auf die Kompositazerlegung eine wichtige Rolle.

2.5 Lokale Grammatiken

Um auch Texteinheiten, die über den Wortbereich hinausgehen, identifizieren zu können, sind neben den Modulen des CISLEX, deren Aufgabe darin besteht, alle Einheiten, die zwischen Blanks (oder anderen Separatoren) stehen, korrekt zu identifizieren, sind sogenannte lokale Grammatiken notwendig. Lokale Grammatiken erlauben es, Information die über die reine Wortebene hinausgeht in Form von endlichen Übergangsautomaten darzustellen.[14] Lokale Grammatiken finden in verschiedensten Bereichen des Lexikons ihre Anwendung:

Darstellung orthographischer Varianten

Erkennung von Mehrwortlexemen

Repräsentation von Paraphrasenbeziehungen

3 Die morphologische Kodierung

Die bereits eingangs erwähnte Definition der *einfachen Formen* hat zur Folge, daß sich die morphologische Klassifikation weitgehend auf diesen Teil des Lexikons beschränken läßt, da sich nach dem Suffixkriterium die morphologischen Eigenschaften der komplexen Formen aus denen ihrer einfachen EF-Suffixe herleiten lassen. Die einfachen Formen sind nach minimalen syntaktischen Kriterien (im wesentlichen nur die Wortart) und nach ihrem Flexionsverhalten klassifiziert. Im folgenden soll auf diese beiden Bereiche näher eingegangen werden.

3.1 Das CISLEX-Wortartensystem

Im Gegensatz zu den Wortartklassifikationen herkömmlicher Lexika, die in der Regel eine inkonsistente Mischklassifizierung auf der Grundlage morphologischer, syntaktischer und semantischer Kriterien darstellen, muß eine Wortartklassifikation für ein elektronisches Wörterbuch einerseits auf operationalisierbaren Kriterien aufbauen,

[13] Unter Fachwortschatz verstehen wir Fachbegriffe, die nicht zum allgemeinen Wortschatz des Deutschen, wie er etwa durch die großen Lexika (DUDEN, Wahrig, Mackensen, usw.) beschrieben wird, gehören.

[14] Für eine detaillierte Beschreibung dieses Formalismus siehe Silberztein(1989).

Die restlichen Wortformen stammen überwiegend aus den Bereichen der Eigennamen und der Derivate von Eigennamen[10], des Fremd- und Fachwortschatzes und der Sonderformen inklusive der Abkürzungen und Akronyme.
Um diese Fälle alle erfassen zu können, ist das CISLEX im Zusammenhang mit konkreten Anwendungen in 4 Module gegliedert: das deutsche Kernlexikon (DKL), das Eigennamenlexikon (EN), Spezialvokabular und Sonderformen. Diese Module werden im folgenden näher erläutert.

2.1 Das deutsche Kernlexikon (DKL)

Das deutsche Kernlexikon besteht aus den bereits erwähnten einfachen Formen des Deutschen (EF), dem Abschluß dieser einfachen Formen unter Präfigierung (EF-PRF) und den Komposita (KF). EF und EF-PRF bilden zusammen das Basislexikon (EF+), mit dem sowohl die einfachen Formen des Deutschen als auch die Derivation auf einfachen Basen vollständig beschrieben sind. Aus diesem Basislexikon wird auf der Basis der morphologischen Kodierung mithilfe der Morphologiekomponente das Vollformenlexikon[11] (DKL-FLEX) generiert, das die eigentliche Grundlage für die verschiedenen Anwendungen des Lexikons darstellt. Unabhängig von den einfachen Formen bilden die komplexen Formen einen separaten Bestandteil des Kernlexikons. Alle Komponenten des DKL sind nach einem einheitlichen morphologischen System kodiert, das im nächsten Abschnitt ausführlich dargestellt wird. Beispiele für Einträge des EF- und des FLEX-Lexikons finden sich im Anhang.

2.2 Das Lexikon der Eigennamen (EN)

Der derzeitige Stand des CISLEX-Systems umfaßt ein strukturiertes Eigennamenlexikon (die EN-Formen); es ist eingeteilt in folgende Unterbereiche: i) Vornamen: dieses Lexikon umfaßt z.Zt. ca. 75 000 Einträge, wobei jeder Eintrag mit Angaben über das Geschlecht, die Nationalität und Varianten versehen ist; ii) Persönlichkeiten: dieses Lexikon umfaßt die Nachnamen, Vornamen, und Berufe von Personen; iii) Länder und Provinzen: dieses Lexikon umfaßt alle offiziellen Ländernamen und deren Derivate (Adjektive, Bewohnernamen); iv) natürliche geographische Objekte: dieses Lexikon umfaßt alle einschlägigen geographischen Objekte (Berge, Flüße, Wüsten, usw.); v) Städte und Straßennamen: dieses Lexikon umfaßt die Ortsnamen Deutschlands, für einige Städte alle Straßennamen[12]; vi) konstruierte geographische Objekte: dieses Lexikon umfaßt eine Anzahl von Namen von Bauten, Flughäfen, usw. und vii) Firmennamen: dieses Lexikon umfaßt Angaben zu Firmen (Sparte, Namensvarianten, Sitz, usw.).

2.3 Spezialvokabular

Unter Spezialvokabular verstehen wir einerseits fremdsprachliche Wörter, die sich nicht in das morphologische Klassifikationsschema des DKL einordnen lassen und

[10] Irreguläre Derivate, wie etwa *Israeli* müssen im Lexikon im Gegensatz zu den regulären Ableitungen explizit kodiert sein.

[11] Die Vollformen umfassen sowohl das vollständige Flexionsparadigma der Einträge in EF+ als auch die jeweiligen Fugenformen.

[12] Dies ist vor allem für die Auswertung der Zeitungen relevant.

wollen wir sie hier nicht weiter betrachten. Während es für die syntaktische Klassifikation von Lexikoneinträgen wenigstens konkrete Kodierungsvorschläge in der Praxis gibt, kann man bis heute auf keine systematische semantische Klassifikation *aller* Einträge eines größeren Wörterbuchs hinweisen[7]. Im CISLEX geht es z.Zt. darum, jedem Eintrag der einfachen Formen, insbesondere jedem Nomen, Verb und Adjektiv, eine semantische Kategorie zuzuweisen. Für die Beschreibung der Nomen wurden bisher über 300 semantische Klassen aufgestellt, anhand derer linguistische Verhaltensweisen festgehalten werden. In Zusammenarbeit mit Gaston Gross (Paris) wird die Liste dieser Kategorien vervollständigt und auf ihre Anwendung in zwei großen Bereichen überprüft[8]. Obwohl diese Arbeiten noch nicht abgeschlossen sind, zeigt sich, daß die deskriptiven Möglichkeiten weitreichende Konsequenzen haben. Das Beispiel der Menschenbezeichnungen möge hier genügen: Die Menschenbezeichnungen sind bisher in 25 Unterklassen aufgeteilt. Einige der Klassen, wie die Kategorie [BERUFSBEZEICHNUNGEN] (Tischler, Professor ...) sind wiederum in Subklassen unterteilt. Diese Unterklassen der Berufsbezeichnungen erlauben eine feinere Unterscheidung verschiedener Berufstypen wie akademische Berufe, Handwerksberufe, usw. Im Zusammenhang mit typischen syntaktischen Mustern, in denen diese Bezeichnungen vorkommen, ermöglichen es die semantischen Klassen, bestimmte Angaben, wie zum Beispiel den Beruf einer Person X, automatisch aus Texten zu extrahieren. Beispiele für solche Muster sind:

X ist [BERUFSBEZEICHNUNG]

X lehrt als [AKAD. BERUF] in X

Zudem lassen die semantischen Klassen eine korrekte Übersetzung syntaktischer Muster der Ausgangssprache in syntaktische Muster einer Zielsprache zu:

X travaille comme menuisier [BERUF] — X arbeitet als Schreiner

X travaille comme un fou [KRANKER] — X arbeitet wie ein Verrückter

Wie schaut nun ein Eintrag im Kernlexikon aus? Zu den Angaben, die wir bisher besprochen haben, kommt noch die Bestimmung des *Bereichs*, dem das jeweilige Wort zuzuordnen ist. Hier orientieren wir uns – so weit es geht – an der gängigen Einteilung der Disziplinen (etwa Medizin, Jura, usw.), obwohl wahrscheinlich kein Vademecum hier einschlägig sein wird. Umsomehr stellt sich auch hier die Frage nach der intendierten Anwendung, für die das Lexikonsystem eingesetzt werden soll. Eine Klassifikation der Bereiche hinsichtlich einer Anwendung im Bereich Information Retrieval wird nicht notwendigerweise gleich ausschauen müßen wie eine Klassifikation der gleichen Einträge hinsichtlich einer Anwendung in der automatischen Übersetzung. Inwiefern hier minimale Entscheidungen getroffen werden, die für ein breites Spektrum an Anwendungen in Frage kommen, ist noch ungeklärt.
In einem gewöhnlichen Text stellen jedoch die Wortformen, die auf einen Eintrag im Kernlexikon des Deutschen zurückgeführt werden können, nur einen Bruchteil dar[9].

[7] Einen Versuch in dieser Richtung stellt der Ansatz des COBUILD-Wörterbuchs dar, der in Sinclair, Fox et al. (1988) beschrieben ist.

[8] Diese sind die automatische Übersetzung und die automatische Indizierung.

[9] So enthält ein größeres Korpus einer typischen Tageszeitung (bezogen auf die Types) nur ca ein Fünftel einfache Formen.

i) *einfachen* Formen (EF), ii) *komplexen* Formen (KF), iii) *Eigennamen* oder EN-Formen, iv) *Fremd- und Fachwörtern* oder F-Formen, *Kurzformen* oder *Sonderformen*. Was ist unter dieser Einteilung der in Texten vorkommenden Wortformen zu verstehen? Die Grundidee dieser Klassifikation ist wie folgt: in einem Text findet man in der Regel Vorkommen von elementaren Wörtern (im CISLEX-System, von einfachen oder komplexen Formen); dies sind in gewisser Weise die interessanten Bausteine der Sprache und zwar aus zwei Gründen: erstens, weil wir hier ein Vollständigkeitskriterium ansetzen können und zweitens, weil die meisten anderen Vorkommen von Wortformen aus diesen konstruiert sind. Wir definieren eine *einfache Form* W des Deutschen als ein Wort, für welches es keine Zerlegung W1W2 gibt, so daß W1 eine Folge von Morphemen ist und W2 ein Wort, aus dessen morphologischen Eigenschaften sich die morphologischen Eigenschaften von W direkt herleiten lassen. In anderen Worten, eine einfache Form ist ein Wort, welches keine einfache Form als (sinnvolles) Suffix hat (aus diesem Grund nennen wir dieses Kriterium auch *Suffixkriterium*). Nach diesem Kriterium sind z.B. *Bahnhof* oder *Frühjahr* und natürlich die meisten Ausdrücke, die umgangssprachlich als Komposita beschrieben sind, keine einfachen Formen. Eine Wortform W heißt *komplexe* Form, wenn es eine Zerlegung W1W2 gibt, so daß W1 eine Folge von Morphemen ist und W2 ein Wort, aus dessen morphologischen Eigenschaften sich die morphologischen Eigenschaften von W direkt herleiten lassen. Die einfachen und die komplexen Formen machen zusammen im CISLEX-System das *Kernlexikon* des Deutschen aus. Hinsichtlich der *einfachen Formen* sind folgende Aspekte zu klären: i) welche Klassen gibt es hinsichtlich des (groben) syntaktischen Verhaltens der Einträge? (Syntaktisches Adäquatheitskriterium); ii) welche Klassen gibt es hinsichtlich des *morphologischen* Verhaltens der Einträge? (Morphologisches Adäquatheitskriterium); und iii) welche Klassen gibt es hinsichtlich des *semantischen* Verhaltens der Einträge? (Semantisches Adäquatheitskriterium). Die Aspekte i) und ii) sind im CISLEX flächendeckend abgehandelt indem ca. 150 000 Einträge in 16 verschiedene syntaktische Klassen und insgesamt über 300 verschiedene morphologische Klassen eingeteilt worden sind (eine ausführliche Beschreibung dieser Klassifikation erfolgt in Abschnitt 3). Diese Klassifikation beschreibt in impliziter Form[5] ca. 700 000 verschiedene flektierte Wortformen. Die morphologische Beschreibung der einfachen Formen ist daher als adäquat zu bezeichnen, wenn sie alle und nur die richtig flektierten Formen der Einträge im Lexikon der einfachen Formen umfaßt.

Die morphologische Beschreibung läßt sich (mit sehr wenigen Ausnahmen) direkt auf die Menge der komplexen Formen übertragen, da diese ihre morphologischen Eigenschaften von ihren einfachen Formen als Suffixen übernehmen. Bisher wurden im CISLEX-System über 1 000 000 komplexe Formen erfaßt.

Die Kriterien der syntaktischen und semantischen Adäquatheit sind verhältnismäßig viel schwieriger zu gewährleisten. Über die minimale Anforderung hinaus, daß jeder Lexikoneintrag mit einer syntaktischen Kategorie versehen werden muß, gibt es kaum Konsensus darüber, wann ein Klassifikationssystem als adäquat aufzufassen ist[6]. Vielleicht die konkreteste Antwort auf diese Frage wurde von Maurice Gross im Rahmen der *lexique-grammaires* hinsichtlich vor allem der syntaktischen Beschreibung von Verben geliefert. Da diese Problematik weit über das Lexikon hinausführt,

[5] Dies geschieht durch eine Morphologiekomponente in Form eines Prologprogramms.

[6] Einen immer noch aktuellen Überblick zu diesem Thema findet man in Bergenholtz und Schaeder (1977).

1.4.1 Portabilität

Unter Portabilität verstehen wir die Möglichkeit das gleiche Wörterbuch in verschiedenen Typen von Anwendungen, wie auch in verschiedenen Implementierungsformen einzusetzen. Dies bedingt zum Beispiel, daß das elektronische Wörterbuch zunächst in implementierungsunabhängiger Form existiert und nicht "festverdrahtet" in einer bestimmten Datenstruktur.

1.4.2 Wiederverwertbarkeit ("Reusability")

Die hochgepriesene Reusability ist eigentlich nichts anderes als eine Form der Portabilität, die dafür verantwortlich ist, daß die im Wörterbuch enthaltenen Informationen nicht eindeutig nur mit einem Anwendungstyp zusammenhängen.

1.4.3 "Standards"

Die äußere Form sowie die im Wörterbuch kodierten Informationen sollten weder von der Sprache, für die das Wörterbuch entworfen wurde, noch von den Anwendungen abhängen. Anders formuliert sollten elektronische Wörterbücher für beliebige Sprachen den gleichen formalen Charakter haben.

1.4.4 Das LADL-Format

Ein wichtiger Ansatz, die oben genannten Anforderungen zu erfüllen, wurde vor einigen Jahren am LADL in Paris begonnen. Inzwischen liegen für fünf europäische Sprachen sehr umfangreiche Wörterbücher im gleichen Format vor. Sie sind dadurch gekennzeichnet, daß sie nach den gleichen Kriterien der Vollständigkeit, Korrektheit und Abgeschlossenheit entwickelt wurden, in ihrer externen Form (zumindest was die Lexika der Grundformen und der Vollformen betrifft) identisch konzipiert worden sind und von der Wahl der verwendeten Kategorien soweit wie möglich kompatibel gehalten wurden. Die vielen Anwendungen dieses Lexikons haben gezeigt, daß mit dem gewählten Format und der Auswahlmethodologie ganz verschiedene Anwendungen (Schreibkorrektur, Lemmatisierung, Indizierung, usw.) und ganz verschiedene Implementierungen (relationale Datenbank, DBM-files, Buchstabenautomaten, usw.) realisiert werden können.

2 Die Komponenten des CISLEX

In diesem System gehen wir von geschriebenen Sprachvorkommen[3] aus, wie wir sie aus Korpora verschiedenster Art belegen können. Eine *Wortform* ist eine Buchstabenfolge zwischen *Separatoren*, wobei wir der Einfachheit halber hier davon ausgehen können, daß man unter Separatoren *Leerzeichen*[4] verstehen kann. Im CISLEX-System entsprechen beobachtbare Wortformen im allgemeinen entweder

[3] Eine der geplanten Erweiterungen des CISLEX-Systems besteht in der Phonetisierung aller Teillexika. Phonetisierung bedeutet die Bildung von Lexika der kanonischen phonetischen Umschriften der orthographischen Wörter.

[4] Blanks, neue Zeilen, Tabs, usw. In reellen Anwendungen muß man davon ausgehen, daß man nicht nur ein Satzende-Erkennungsprogramm, sondern auch einen Wortformenpräprozessor braucht, der die im Lexikon nachzuschlagenden Wörter in eine normalisierte Form übersetzt.

eine konkrete Korpusmenge ist denkbar[2]. Die nächste Intuition die Vollständigkeit betreffend hat damit zu tun, daß alle Wortformen in Texten systematisch in Verbindung gesetzt werden müßen mit den Einträgen in der ersten Basismenge. Dies betrifft sowohl die üblichen flektierten Vollformen als auch die zusammengesetzten Wörter. Jede sinnvolle Analyse bzw. jede Zerlegung von komplexen Einheiten muß auf weniger komplexe - und bekannte - Entitäten im Lexikonsystem beruhen. Nur somit ist ein Anspruch auf Vollständigkeit zu gewährleisten. Solche Anforderungen sind rein formeller Art und betreffen zunächst nur das reine Erkennen von lexikalischen Einheiten.

1.2 Korrektheit

Die Kriterien der Korrektheit betreffen das Herleiten von flektierten Formen aus Grundformen und das Herleiten der Grundformen aus flektierten Formen. Die Minimalanforderung betrifft auf der einen Seite die Flexionsparadigmen der Sprache und auf der anderen die weitaus produktiveren und unregelmäßigeren Bereiche der Derivation.

1.3 Abgeschlossenheit

Das Kriterium der Abgeschlossenheit hängt eng mit dem der Korrektheit zusammen. Hierunter verstehen wir in erster Linie die Kriterien mit denen die morphologischen Regularitäten im Lexikon abgedeckt werden. So ist zum Beispiel das weiter unten angeführte Suffixkriterium ein typisches Beispiel dafür, welche Typen von Einträgen und welche konkreten Einträge man im Lexikon erwarten kann. Kriterien der Abgeschlossenheit betreffen andererseits auch semantische Regularitäten, insbesondere Regularitäten der Lexikalisierung. Hier ist zum Beispiel zu erwarten, daß im Lexikon der komplexen Formen systematisch unterschieden wird zwischen kompositionell interpretierbaren und nicht kompositionell interpretierbaren Formen. Darüberhinaus sind Kriterien anzugeben, mit denen über die Lexikalisierung von Komposita systematisch entschieden wird.

1.4 Weniger ist mehr ...

In den letzten Jahren wird immer mehr auf pragmatisch relevante Eigenschaften von Wörterbüchern hingewiesen, die mit ihrer Beziehung sowohl zu anderen Wörterbüchern als auch zu Anwendungen zu tun haben. Auch wenn diese Kriterien (Portabilität, Wiederverwendbarkeit und Standards) heute kaum von existierenden elektronischen Wörterbüchern erfüllt werden, ist man sich im klaren darüber, daß sie für industrielle Anwendungen eine wichtige Rolle spielen werden. Hier sollte vor allem nach dem Prinzip "weniger ist mehr" vorgegangen werden. Damit meinen wir, daß es sinnvoller ist, im Lexikon elementare Informationen in korrekter und vollständiger Weise zu repräsentieren, als eine Vielzahl theoretisch interessanter Eigenschaften der Lexeme, für die man i) keine exakten Kriterien definieren kann, und die sich ii) nicht auf breiter Basis kodieren lassen.

[2] "Minikorpora" sind sicherlich für den Aufbau eines brauchbaren elektronischen Wörterbuchs nicht als Datenbasis geeignet

Das CISLEX-Wörterbuchsystem

Franz Guenthner, Petra Maier

Ein vollständiges, theorieneutrales elektronisches Wörterbuch des Deutschen ist seit langer Zeit ein Desiderat der deutschen Computerlinguistik. Mit dem CISLEX-System wurde der Versuch unternommen, ein solches Wörterbuch zu erstellen. Der Aufbau und die Anforderungen, die an ein solches Wörterbuch gestellt werden, sollen hier vorgestellt werden. Eine detaillierte Beschreibung der Komponenten und der Kodierung des CISLEX wird dann zeigen, wie die beschriebenen Anforderungen realisiert wurden. Im letzten Abschnitt skizzieren wir einige systematische Anwendungen, die mit solchen Wörterbüchern unternommen werden können.

1 Elektronische Wörterbücher für das Deutsche: Anforderungen und Aufgaben

Elektronische Wörterbücher unterscheiden sich in mehreren Punkten von traditionellen gedruckten Wörterbüchern[1] einerseits und den Lexika, die in natürlichsprachlichen Anwendungen als lexikalische Datenbasis eingesetzt werden, andererseits. Die Hauptunterschiede sind in den Anwendungen elektronischer Wörterbücher als Datenbasis für sprachverarbeitende Programme jeglicher Art begründet und beziehen sich im wesentlichen auf die Vollständigkeit, die Abgeschlossenheit und die Korrektheit.

1.1 Vollständigkeit

Ein elektronisches Wörterbuch ist mehr als eine mit Merkmalen versehene Wortliste; es impliziert eine theoretische Einteilung der Objektklassen, denen die verschiedenen beobachtbaren lexikalischen Entitäten zuzuordnen sind. Durch die explizite Auflistung von Grundentitäten und durch die Bestimmung von Operationen auf diesen, werden Mengen von Wortformen identifiziert, bezüglich welcher verschiedene Adäquatheitsbedingungen gestellt werden können. In anderen Worten, man wird für diverse Anwendungen ziemlich genau sagen können, welche Inputbedingungen durch welche Algorithmen zu welchem Output führen bzw. erklären können, warum bestimmte Input-Output-Transformationen nicht zustandekommen. Auch wenn es sich prinizipiell um *offene* Mengen handelt, ist die Bestimmung einer Grundmenge von lexikalischen Entitäten für alle späteren Vergleiche und Evaluationen von besonderer Bedeutung. Über welche Entitäten wird gesprochen und wie werden sie charakterisiert? Im CISLEX ist eine wichtige Grundmenge des Systems die Menge der *einfachen Formen*; in einem anderen System könnte es eine Morphmenge oder eine anders bestimmte Menge sein. Solange jedoch keine solche Bestimmung vorliegt kann keine Evaluation des Lexikons beginnen. Diese Grundmenge sollte in der Regel so vollständig wie möglich sein, in dem alle möglichen zugänglichen Ressourcen und Intuitionen herangezogen werden; auch eine explizite Einschränkung auf

[1] Hierzu rechnen wir auch die elektronisch verfügbaren Versionen traditioneller Wörterbücher.