Automatische Wortformanalyse für das Spanische

Magisterarbeit

in der Philosophischen Fakultät II
(Sprach- und Literaturwissenschaften)
der Friedrich-Alexander-Universität
Erlangen-Nürnberg

vorgelegt von

Cerstin Elisabeth Mahlow

aus

Greifswald

Inhaltsverzeichnis

Themenstellung

Die automatische Wortformanalyse ordnet Wortformen Informationen über Wortklasse, Grundform und syntaktische Verwendungsmöglichkeiten zu. Auf dieser Grundlage können Komponenten zur syntaktischen und semantischen Analyse entwickelt werden.

In der Implementierungssprache MALAGA wurden bisher Morphologie-Grammatiken für das Deutsche, Italienische, Koreanische und Englische entwickelt.

Um diesen Kanon zu erweitern und vergleichende Aussagen zu morphologischen Eigenschaften verschiedener Sprachen treffen zu können, die auf die Analyse umfangreicher Korpora gestützt sind, soll im Rahmen dieser Magisterarbeit eine Komponente zur automatischen Wortformanalyse für das Spanische entwickelt werden. Dazu sind ein spanisches Grundformlexikon zu erstellen sowie linksassoziative Regeln für Flexion, Derivation und Komposition zu formulieren. Die erstellte Grammatik soll an Korpora getestet werden.

Abkürzungen und Akronyme

DMM	Deutsche MALAGA-Morphologie
EMM	Englische MALAGA-Morphologie
IMM	Italienische MALAGA-Morphologie
SMM	Spanische MALAGA-Morphologie
LAG	Left-Associative Grammar
MALAGA	Programmpaket für die automatisierte Analyse natürlicher Sprache
PSG	Phrasenstrukturgrammatik
RAE	Real Academia Española
DRAE	Diccionario de la Real Academia Española
CREA	Corpus de Referencia del Español Actual
CORDE	Corpus Diacrónico del Español
CRATER	Corpus Resources And Terminology ExtRaction
ITU	International Telecommunications Union
DGLE	Diccionario General de la Lengua Española VOX

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Untersuchungsgegenstand

Der Begriff „das Spanische“ soll hier beschränkt werden auf das „castellano“, das durch die Grammatiken und Wörterbücher der REAL ACADEMIA ESPAÑOLA als solches definiert ist und auf der iberischen Halbinsel gesprochen wird. Regionale Unterschiede werden nicht berücksichtigt.

In Grammatiken für die spanische Sprache werden je nach Autor verschiedene Begriffe für dasselbe Phänomen benutzt. In dieser Arbeit wird für die Terminologie auf die Festlegungen der REAL ACADEMIA ESPAÑOLA zurückgegriffen, wie sie in [RAE 1974] gebraucht werden. Weiterhin werden Grammatiken von JUAN ALCINA FRANCH / JOSÉ MANUEL BLECUA [AlcinaBlecua 1994], EMILIO ALARCOS LLORACH [Alarcos 1994], IGNAZIO BOSQUE MUÑOZ / VIOLETA DEMONTE BARRETO [BosqueDemonte 1999], JACQUES DE BRUYNE [de Bruyne 1993] und JOHANNES THIELE [Thiele 1992] verwendet.

1.2 Teilaufgaben

Um eine Komponente zur automatischen Wortformerkennung zu erstellen, sind zunächst zwei Voraussetzungen zu schaffen. Es wird ein Grammatikformalismus benötigt, der es erlaubt, die Regeln der Morphologie in einer Weise zu formulieren, die die computerlinguistische Umsetzung ermöglicht. Hierfür wird die Linksassoziative Grammatik von ROLAND HAUSSER verwendet.[1] Weiterhin wird eine Implementierungssprache benötigt, die es ermöglicht, die Regeln der Grammatik zu implementieren und die erstellte Komponente an Korpora zu testen. Dazu wird MALAGA verwendet. Dabei handelt es sich um eine Programmiersprache, die die Implementierung von Morphologie- und Syntax-Grammatiken für natürliche Sprachen ermöglicht und eine Programmierumgebung, die die Entwicklung und Anwendung dieser Grammatiken unterstützt. Das Akronym steht für **M**erely **a** **L**eft-**A**ssociative-**G**rammar **A**pplication. MALAGA wurde von BJÖRN BEUTEL und GERALD SCHÜLLER an der Abteilung für Computerlinguistik des Instituts für

[1] Erstmals vorgestellt in [Hausser 1985]. Hier weiter bezeichnet als LAG.

Germanistik der Friedrich-Alexander-Universität Erlangen-Nürnberg entwickelt.[2]

Entsprechend der Phänomene der spanischen Morphologie sind die Anforderungen an die Morphologiekomponente zu formulieren. Diese sind mit Hilfe der Implementierungssprache und dem zugrunde liegenden Grammatikformalismus zu spezifizieren.

Zusätzlich wird ein elektronisches Lexikon des Spanischen benötigt, das aus verschiedenen Ressourcen erstellt wird und dessen Einträge dem Format der Implementierungssprache angepasst werden müssen.

Entsprechend der benutzten Methode der Wortformerkennung sind Regeln zu implementieren. Dabei handelt es sich zum einen um Regeln, die aus dem Lexikon vor der Laufzeit der Analyse-Komponente ein Allomorphlexikon expandieren. Zum anderen sind es Regeln, die die Konkatenation dieser Allomorphe zu wohlgeformten Wortformen des Spanischen steuern.

Die entwickelte Komponente SMM (**S**panische **M**ALAGA-**M**orphologie) ist abschließend an Korpora des Spanischen zu testen, um die Erkennungsrate festzustellen und Mängel der Regeln erkennen und beheben zu können. Diese Korpora sind aus verfügbaren Ressourcen zu erstellen bzw. so zu bearbeiten, dass sie den Eingabebedingungen der Implementierungsumgebung MALAGA genügen.

1.3 Aufbau der Arbeit

Kapitel 2 stellt den verwendeten Grammatikformalismus der Linksassoziativen Grammatik und die Implementierungssprache MALAGA vor. Hier wird insbesondere auf die Elemente eingangen, die zur Wortformanalyse notwendig sind.

Im Kapitel 3 werden Phänomene der spanischen Morphologie dargestellt, die in der Morphologie-Komponente berücksichtigt werden. Dabei wird auf die oben genannten einschlägigen Grammatiken für das Spanische zurückgegriffen.

Im Kapitel 4 werden Erkenntnisse vorgestellt, die andere Systeme der automatischen Analyse des Spanischen liefern. Aus diesen können unter Berücksichtigung der bisher entwickelten Morphologie-Komponenten grundlegende Entscheidungen zur Erstellung des Lexikons und der Regeln abgeleitet werden, die im Kapitel 5 beschrieben werden.

Kapitel 6 beschäftigt sich mit der Erstellung der Grammatik zur Wortformanalyse. Zunächst wird das Lexikon in Bezug auf die Ressourcen und das Format der Lemmata beschrieben. Dann werden die Regeln zur Erzeugung von Allomorphen und zur Konkatenation derselben vorgestellt, die in Konsequenz der Betrachtungen in Kapitel 3 und 4 gewonnen wurden. Hier finden sich auch statistische Ergebnisse, die Aussagen über den Allomorphiequotienten des Spanischen bieten, sowie Beispielanalysen einiger Wortformen.

[2] [Beutel 1999].

Abschließend werden im Kapitel 7 Korpora vorgestellt, mit denen die implementierte Grammatik getestet wurde, sowie die Ergebnisse der Analysen diskutiert.

Die einzelnen Dateien der SMM sind im Anhang und auf der beiliegenden CD-ROM angegeben. Die CD-ROM enthält zusätzlich die Korpora und die Ergebnisse der Korpora-Analysen.

Kapitel 2

Grundlagen

2.1 Linksassoziative Grammatik

Ein elementarer Grammatikformalismus dient der adäquaten Beschreibung linguistischer Fragestellungen. Um diesen für die Computerlinguistik verwenden zu können, muss es weiterhin möglich sein, daraus effiziente Algorithmen für die Datenverarbeitung zu entwickeln. Diese Anforderungen erfüllt die Linksassoziative Grammatik.

Die LAG ist in die ebenfalls von HAUSSER entwickelte SLIM-Sprachtheorie eingebettet (das Akronym steht für **S**urface compositional **L**inear **I**nternal **M**atching) und genügt den Prinzipien der Oberflächenkompositionalität, Zeitlinearität, der Behandlung von Sprachverstehen und -produktion als Sprecher/Hörer-internem Vorgang und dem kontextuellen Abpassen von wörtlichen Bedeutungen auf Verwendungskontexte.[1] Für diese Arbeit sind vor allem die Aspekte der Oberflächenkompositionalität und Zeitlinearität von Bedeutung.

Das Prinzip der Oberflächenkompositionalität wird definiert als:

> An analysis of natural language is surface compositional if it uses only concrete word forms as the building blocks such that all syntactic and semantic properties of complex espression derive systematically from the syntactic category and the meaning$_1$ of their building blocks.[2]

Zeitlinearität wird als die Grundstruktur natürlichsprachlicher Zeichen angesehen:

> The basic structure of natural language signs is their *time-linear order*. This holds for the sentence in a text, the word forms in a sentence, and the allomorphs in a word form. *Time-linear* means: LINEAR LIKE TIME AND IN THE DIRECTION OF TIME.[3]

[1] Diese Forderungen werden allgemein in [Hausser 1989a, S. 13f.] gestellt und in [Hausser 1999, S. 8] als für SLIM gültig postuliert.

[2] [Hausser 1999, S. 80]; [Eine Analyse natürlicher Sprache ist oberflächenkompositional, wenn sie nur die konkreten Wortformen als Bausteine verwendet, so dass alle syntaktischen und semantischen Eigenschaften eines komplexen Ausdrucks systematisch aus den syntaktischen Kategorien und den wörtlichen Bedeutungen der Bausteine abgeleitet werden.], meaning$_1$ steht für die wörtliche Bedeutung einer Wortform.

[3] [Hausser 1999, S. 97]; [Die grundlegende Struktur natürlichsprachlicher Zeichen ist ihre zeit-

Eine LA-Grammatik für eine konkrete Sprache wird durch die Angabe eines Lexikons (bestehend aus Wortformoberflächen und zugeordneten Kategorien), Anfangszuständen, Endzuständen und Regeln (bestehend aus kategorialen Operationen und Regelpaketen) spezifiziert.[4]

2.1.1 Die Regeln der LAG

Im Gegensatz zur Phrasenstrukturgrammatik oder zur Kategorialgrammatik, die das Substitutionsprinzip verwenden, beruht die LAG auf dem Prinzip der möglichen Fortsetzungen. Zu jeder Regel ist jeweils angegeben, mit welchen Regeln fortgefahren werden kann, damit das Resultat im Sinne der Morphologie bzw. Syntax einer Sprache wohlgeformt ist. Die Analyse schreitet dabei von links nach rechts fort und entspricht dem Verlauf der Zeit, ist also zeitlinear. Dies gilt unter der Annahme, dass die Zeit von links nach rechts angetragen wird und dem Schreibverhalten entspricht.

Die Analyse ist oberflächenkompositional, da die Regeln von der konkreten Oberfläche der zu analysierenden Wortform ausgehen.

2.1.2 Wortformerkennung mit LAG

Drei Methoden der automatischen Wortformerkennung können unterschieden werden: Vollform-Methode, Grundform-Methode, Allomorph-Methode.

Die Vollform-Methode beruht auf dem lexical look-up in einem Lexikon, das alle in einer Sprache auftretenden Wortformen enthält. Es müssen alle Komposita, Derivata und flektierten Formen enthalten sein. Wegen der Produktivität der natürlichen Sprachen kann das Lexikon potenziell unendlich groß werden.

Die Grundform-Methode beruht auf der Rückführung der Wortformen auf Sequenzen von Morphemen, die aus den konkreten Allomorphen der zu analysierenden Oberfläche abgeleitet werden. Das Lexikon enthält die Morpheme einer Sprache und weist damit eine begrenzte Zahl von Einträgen auf. Die Konkatenation der ermittelten Morpheme schließt die Analyse ab.

Die Allomorph-Methode verwendet ebenfalls ein Morphemlexikon, aus dem über Allomorphregeln vor der Analyse ein Allomorphlexikon expandiert wird. Dieses ist größer als das Morphemlexikon, weist aber ebenfalls eine begrenzte Zahl von Einträgen auf. Die zu analysierende Wortform wird in Allomorphe zerlegt. Dann werden die Allomorphe mit Hilfe von Kombinationsregeln konkateniert, woraus die Analyse der kompletten Wortform resultiert.

Die Allomorph-Methode bietet zwei Vorteile: Gegenüber der Vollform-Methode wird ein relativ begrenztes Lexikon verwendet, gegenüber der Grundform-Methode wird die konkrete Oberfläche der Wortform genutzt bzw. werden die konkreten Allomorphe konkateniert. Damit ist diese Methode oberflächenkompositional und in

lineare Reihenfolge. Diese gilt für die Sätze eines Textes, die Wortformen eines Satzes und für die Allomorphe in einer Wortform. Zeitlinear heißt: linear wie die Zeit und in der Richtung der Zeit].

[4] Zur algebraischen Definition mittels eines 7-Tupels [Hausser 1999, S. 187].

den Anforderungen an Speicherplatz begrenzt.

HAUSSER entwickelte innerhalb der LAG den LA-MORPH-Ansatz,[5] der auf der Allomorph-Methode beruht. Die Wortformen werden zunächst in die jeweiligen Allomorphe zerlegt, deren kategoriale Eigenschaften mit Hilfe der Regeln konkateniert werden.

Ein Wortanfang (dessen Kategorie entspricht cat_1) wird mit einem nächsten Allomorph (dessen Kategorie entspricht cat_2) zu einem neuen Wortanfang (dessen Kategorie entspricht cat_3) verknüpft, der dann an die möglichen Folgeregeln, die im Regelpaket (rp_i) enthalten sind, übergeben wird. Jede Regel besteht aus dem Regelnamen, der Angabe der kategorialen Operation und dem Regelpaket.[6]

$$r_i\text{: } cat_1\ cat_2 \Rightarrow cat_3\ rp_i$$

Dieser Ansatz wurde in verschiedenen Implementierungen verwirklicht. Die hier genutzte Implementierung ist MALAGA.

2.2 MALAGA

MALAGA wurde für die Implementierung von Morphologie- und Syntax-Grammatiken natürlicher Sprachen entwickelt.[7]

Der zugrunde liegende Grammatikformalismus ist die linksassoziative Grammatik. Bisher wurden mit und für MALAGA Morphologie-Komponenten für die Sprachen Deutsch[8], Italienisch[9], Englisch[10] und Koreanisch erstellt.

Eine Morphologie-Grammatik für eine konkrete Sprache besteht aus einer Lexikondatei der Grundformen, einer Allomorphregel-Datei, einer morphologischen Kombinationsregel-Datei und einer Symboldatei. Die Allomorphregeln steuern die Erzeugung des Allomorphlexikons aus dem Grundformlexikon, die linksassoziativen Kombinationsregeln steuern die Konkatenation der Allomorphe. Alle Dateien sind in der Projektdatei aufgeführt. Innerhalb der Projektdatei können verschiedene Einstellungen vorgenommen werden, die beispielsweise die Ausgabe der Analyseergebnisse steuern.

Um spanische Wortformen mit MALAGA automatisch analysieren zu können, ist es notwendig, ein Lexikon zu erstellen, das spanische Grundformen im MALAGA-Format[11] enthält. Weiterhin müssen Regeln erstellt werden, die zur Generierung des Allomorphlexikons geeignet sind, und Regeln, die die Konkatenation der Allomorphe steuern. Die dabei verwendeten Symbole für Attribute und Werte sind in

[5] [Hausser 1989b].

[6] [Hausser 1989a, S. 35ff.], [Hausser 1999, S. 184f.].

[7] Die folgenden Ausführungen beruhen auf der Dokumentation zu der in dieser Arbeit verwendeten MALAGA-Version 4.3 [Beutel 1999].

[8] [Lorenz 1996], weiterhin bezeichnet als DMM.

[9] [Wetzel 1996], weiterhin bezeichnet als IMM.

[10] [Leidner 1998], weiterhin bezeichnet als EMM.

[11] Genaueres dazu in 6.2.3.

der Symboldatei zu definieren.

Es besteht die Möglichkeit, Multisymbole zu benutzen, die bereits definierte Symbole zusammenfassen. Damit ist es möglich, einer Oberfläche, die verschiedene Kategorien aufweisen kann, ein Symbol zuzuordnen, das alle möglichen Kategorien beinhaltet. So kann ein Multisymbol `singular&plural` definiert werden, das die Symbole `singular` und `plural` umfasst:
`singular&plural:= <singular, plural>;`

Gültige MALAGA-Werte sind die deklarierten Symbole sowie Zeichenketten und Fließkommazahlen. Als komplexe Datentypen können Listen und Verbunde definiert werden, die auch verschachtelt verwendet werden können. Lokale Variablen werden durch vorangestelltes `$` gekennzeichnet.

MALAGA stellt Befehle zum Zugriff auf Werte, zur Verwaltung komplexer Daten (`.`, `+`, `-`, `*`, `/`) und zum Abgleich von Werten (`=`, `/=`, `in`, `~`, `matches`) bereit. Es stehen spezielle Befehle für Bedingungen (`assert`, `require`, `if`, `repeat`), die Auswahl einzelner Werte (`choose`) und die parallele Behandlung von Analysepfaden (`parallel`) zur Verfügung.[12]

2.2.1 Allomorphregeln

Regeln zur Erzeugung von Allomorphen beginnen mit dem Schlüsselwort `allo-rule`, gefolgt vom Regelnamen, einem Parameter und einem Doppelpunkt. Der Parameter bezieht sich auf den einzulesenden Lexikoneintrag. Danach folgen die Anweisungen dieser Regel, die jeweils mit Semikolon voneinander getrennt werden. Mit diesen Anweisungen werden die Einträge des Lexikons einzeln eingelesen und die daraus entwickelten Allomorphe in das Allomorphlexikon eingetragen. Das Ende der Allomorphregel wird durch `end allo-rule;` angegeben.

2.2.2 Kombinationsregeln

Für die Definition der Kombinationsregeln ist zunächst der Startzustand zu definieren. Er besteht aus der Kategorie des leeren Wortanfangs und dem initialen Regelpaket, das spezifiziert, welche Regeln als erste angewendet werden können:
`initial [], rules Regel1, Regel2,... ;`

Jede Kombinationsregel beginnt mit dem Schlüsselwort `combi-rule`, gefolgt vom Regelnamen, den in runden Klammern angegebenen Parametern und einem Doppelpunkt. Die Parameter beziehen sich auf die Kategorien des bisherigen Wortanfangs und des nächsten Allomorphs. Optional können Parameter für die jeweiligen Oberflächen angegeben werden. Es folgen durch Semikolon getrennte Anweisungen, die prüfen, ob Wortanfang und nächstes Allomorph miteinander konkateniert werden können.

Am Ende jeder Regel wird die resultierende Kategorie der Kombination aus Wortanfang und nächstem Allomorph angegeben, die für die im Regelpaket spezifizierten Folgeregeln dann als Wortanfang gilt. Darauf folgt die Kennzeichnung des Endes der Regel:

[12] Zur konkreten Definition der verwendeten Befehle vgl. [Beutel 1999].

```
result $neuer_Wortanfang, rules Folgeregel1, Folgeregel2,... ;
end combi-rule;
```

Die Anzahl und die Form der Kombinationsregeln ist beliebig. Bindend ist die Endregel, durch die der Endzustand definiert wird. Diese Regel liest kein weiteres Allomorph ein, sondern prüft die Wohlgeformtheit des bisher analysierten Wortanfangs. Der Aufruf dieser Regel erfolgt in den Regelpaketen der Kombiregeln. Die Regel beginnt mit dem Schlüsselwort `end-rule`, gefolgt vom Regelnamen, einem in runden Klammern eingeschlossenen Parameter, der sich auf die bisher analysierte Wortform bezieht, und einem Doppelpunkt. Danach folgen Anweisungen. Die Regel endet mit:

```
result $analysierte_Wortform, accept;
end end-rule;
```

Kapitel 3

Spanische Morphologie

3.1 Einführung

3.1.1 Wortklassen

Die Bezeichnungen der Wortklassen des Spanischen variieren bei verschiedenen Autoren. Oft werden Adjektive und Substantive unter dem Begriff Nomen zusammengefasst,[1] da sie gleiche Flexionseigenschaften aufweisen. Hier sollen sie getrennt behandelt[2] und auch jeweils in der Analyse der Wortklasse unterschieden werden. Daneben gibt es im Spanischen Adverben, Artikel, Präpositionen, Konjunktionen, Pronomen, Namen, Akronyme und Interjektionen.

3.1.2 Englische vs. sprachabhängige Terminologie

Innerhalb der Allomorphregeln und der morphologischen Regeln der Englischen MALAGA-Morphologie (EMM) werden englische Termini benutzt. Innerhalb der entsprechenden Regeln für das Italienische (IMM) werden italienische Termini benutzt. Dies lässt darauf schließen, dass die Termini der Sprache benutzt werden, für die die Regeln erstellt werden sollen. Dagegen sind die Termini für die morphologische Analyse des Deutschen (DMM) nicht deutsch, sondern englisch.

Beide Vorgehensweisen, die Verwendung englischer Termini unabhängig von der zu betrachtenden Sprache wie die Verwendung von Termini entsprechend der betrachteten Sprache, sind nachvollziehbar. Belässt man die traditionell in einer Sprache verwendeten Termini in ihrer Form, erleichtert dies die Lesbarkeit für den Sprecher der Sprache. Werden die entsprechenden englischen Termini für jede zu analysierende Sprache verwendet, ergibt sich eine höhere Transparenz beim Vergleich der Regeln und Ausgaben für Wortformen verschiedener Sprachen. Die Analyse einer Wortform einer bestimmten Sprache ist auch für den Benutzer nachvollziehbar, der diese Sprache nicht aktiv beherrscht.

Da der Vorteil der Verwendung englischer Termini den Vorteil der Verwendung von Termini in der zu analysierenden Sprache überwiegt, wurde hier entschieden,

[1] [AlcinaBlecua 1994], [RAE 1974].

[2] Entsprechend [Thiele 1992], [de Bruyne 1993], [Alarcos 1994].

die notwendigen Begriffe und Kategorien vom Spanischen ins Englische zu übertragen.

3.1.3 Gliederung

Morphologische Phänomene können nach verschiedenen Gesichtspunkten betrachtet werden. Dabei handelt es sich um die Prinzipien von Derivation, Komposition und Flexion sowie um die Strukturprinzipien Allomorphie und Konkatenation. Anhand dieser werden die Phänomene der Morphologie des Spanischen beschrieben. Sie werden in einer Form dargestellt, die die Umsetzung in Allomorph- und Kombinationsregeln ermöglicht.

3.2 Derivation

Derivation kann durch Suffigierung, Präfigierung oder durch eine Kombination beider erreicht werden. Im Spanischen ist die Suffigierung produktiver als die Präfigierung. Auch Komposita können deriviert werden. Ebenso können Derivata Komponenten eines Kompositums sein. Wortformen, die durch Derivation entstehen, können flektiert werden.

Es ist möglich, ein Lexem mehrfach zu derivieren. Nach erfolgter Derivation kann es ein weiteres Suffix oder Präfix annehmen (*claro* → *clarísima* → *clarísimamente*). Die innere Struktur einer mehrfach derivierten Wortform ist bei einer linearen Analyse nicht eindeutig zu ermitteln. So kann für *inutilizable* die Analyse zwei Ergebnisse liefern, wobei die Wortform syntaktisch/semantisch verschieden umschrieben werden kann:
[*in* [[[*util*] *iza*] *ble*]] - »que no puede ser utilizado«
[[[*in* [*util*]] *iza*] *ble*] - »que puede ser inutilizado«[3]
Die Auflösung der Ambiguität kann erst durch die semantische Analyse erfolgen.

3.2.1 Derivation mit Präfixen

Die Präfigierung führt nicht zu einer Änderung der Wortklasse der resultierenden Wortform gegenüber der ursprünglichen Wortform. Präfigierung kann nicht nur mit gebundenen, sondern auch mit freien Morphemen erfolgen.[4] Das heißt, dass einige Präfixe auch als Präpositionen oder Adverben vorkommen können (*contra, super, bien, mal*). Entsprechende Wortformen können sowohl als Derivat als auch als Kompositum analysiert werden. Nach ihrer syntaktischen Funktion können präpositionale und adverbiale Präfixe unterschieden werden (Tabelle 3.1[5]).

Mehrere gleiche Präfixe zu Beginn einer Wortform sind ebenso möglich wie die Aufeinanderfolge von adverbialen und präpositionalen Präfixen, verschiedenen präpositonalen Präfixen oder verschiedenen adverbialen Präfixen.

[3] Entnommen aus [BosqueDemonte 1999, S. 4314].
[4] Vgl. [Thiele 1992, S. 14].
[5] Zusammengestellt aus [BosqueDemonte 1999].

Präfix-Morphem	syntaktische Funktion
ante-, con-, contra-, en-, entre-, sin-, sobre-, tras-	präpositional (Präfix-Morphem auch Präposition)
anfi-, anti-, circun-, ex-, extra-, infra-, post-, pro-, sub-, ultra-	(Präfix-Morphem nicht mehr als Präposition existent)
bien-, casi-, mal-, medio-, no-	adverbial

Tabelle 3.1: Unterscheidung der Präfixe nach ihrer syntaktischen Funktion

Semantisch können Lokativ-, Temporal-, Negativ-, Gradations-Präfixe, diatetische und modifizierende Präfixe unterschieden werden. Einige Präfixe werden in mehreren Bedeutungen gebraucht (Tabelle 3.2[6]).

Präfix-Morphem	semantischer Typ
a-, ab-, anfi-, ante-, anti-, circun-, cis-, citr-, con-, contra-, de-, dia-, ecto-, en-, endo-, entre-, ex-, exo-, extra-, hipo-, infra-, intra-, intro-, meta-, para-, per-, peri-, post-, pre-, pro-, retro-, re-, sobre-, sub-, trans-, ultr-	lokativ
ante-, pos-, pre-	temporal
a-, anti-, contra-, des-, in-, no-	negativ
archi-, casi-, entre-, extra-, hiper-, hipo-, infra-, macro-, maxi-, mega-, medio-, micro-, mini-, re-, semi-, sobre-, sub-, super-, ultra-	gradativ
a-, auto-, des-, re-, sobre-, sub-	diatetisch
ambi-, bi-, bien-, centi-, cuatro-, deca-, deci-, dodeca-, enea-, endeca-, equi-, hecto-, hepta-, hetero-, hexa-, homo-, iso-, mal-, mega-, mili-, mini-, mono-, multi-, neo-, octa-, octo-, paleo-, penta-, pluri-, poli-, seudo-, sex-, tetra-, tri-, uni-	modifizierend

Tabelle 3.2: Unterscheidung der Präfixe nach ihrem semantischen Typ

Die Präfigierung erfordert keine Veränderung der Oberfläche des Lexems. Endet das Präfix auf einen Vokal und beginnt die Basis mit einem Vokal, ist die Schreibung mit Bindestrich möglich. Dies ist insbesondere beim Aufeinandertreffen gleicher Vokale der Fall.

3.2.2 Derivation mit Suffixen

Im Gegensatz zur Präfigierung kann die Derivation mit Suffixen zu einer Änderung der Wortklasse der derivierten Wortform führen. Suffixe können daher nach der resultierenden Wortklasse gruppiert werden (Tabelle 3.3[7]).

[6] Zusammengestellt aus [BosqueDemonte 1999].

[7] Zusammengestellt aus [BosqueDemonte 1999]. Für Verb-Suffixe ist jeweils die resultierende Infinitiv-Endung angegeben.

Suffix-Morphem	resultierende Wortklasse
-ada, -ado, -aje, -al, -azo, -ción, -dad, -dero, -do, -dor,-dura, -ería, -ero, -ez, -eza, -ía, -ido, -ío, -ista, -itud, -miento, -ncia, -or	Substantiv
-áceo, -aco, -ado, -al, -án, -anco, -ano, -ar, -ardo, -ario, -arra, -asco, -ata, -átil, -ato, -az, -ble, -bundo, -cio, -de, -dero, -dó, -dizo, -do, -dor, -eca, -eco, -ego, -ejo, -el, -enco, -engo, -eno, -ense, -eño, -eo, -erno, -ero, -és, -esco, -este, -estre, -eta, -eyo, -í, -iaco, -ial, -iano, -ica, -, icio, -ico, -íco, -ícola, -ido, -ién, -iento, -ífero, -ífico, -iforme, -ígero, -igo, -ijo, -il, -ín, -ino, -io, -ío, -iondo, -isco, -ista, -isto, -ita, -izo, -lento, -ndero, -ndino, -ndo, -neo, -no, -nte, ntío, -oidal, -oide, -oideo, -ojo, -ol, -ón -oso, -ota, -ote, -tario, -ticio, -tico, -tivo, -torio, -ucho, -uco, -udo, -uence, -üeño, -ujo, -ulo, -uncho, -uno, -urno, -usco	Adjektiv
-ear, -izar, -ificar, -ecer	Verb

Tabelle 3.3: Gruppierung der Suffixe nach resultierender Wortklasse

Die Derivation, die zu Wortformen der Wortklasse Verb führt, kann auch durch Rekategorisierung erfolgen. Die Stammallomorphe von Adjektiven oder Substantiven werden als Stammallomorphe eines Verbs behandelt und nehmen die Flexionsmorpheme der Verbflexion an, um wohlgeformte Wortformen zu bilden. Meist gehören diese Verben zur ersten Konjugation[8].

Verben, die durch eine Kombination aus Präfix und Suffix entstehen, heißen parasynthetische Verben. Die Kombination aus Präfix und Suffix ist kein Circumfix, da die Derivata aus Präfix + Basis und Basis + Suffix ebenfalls wohlgeformte Wortformen sind. Größte Bedeutung haben dabei die Präfixe *a-*, *en-* und *des-*, eine untergeordnete Rolle spielen die Präfixe *con-*, *entre-*, *ex-*, *es-*, *extra-*, *per-*, *pro-*, *re-*, *res-*, *so-*, *sobre-*, *trans-*. Für die ersten beiden Präfixe kann die Basis ein Substantiv oder ein Adjektiv sein, die Endung des Infinitivs ist dann *-ar* oder *-ecer*. Für *des-* sind Substantive, Adjektive und Verb-Stämme als Basis möglich. Ist die Basis ein Verbstamm, kann das Derivat zweifach analysiert werden. Zum einen als Derivat aus Präfix, Basis und Suffix, zum anderen als Derivat aus Präfix und Basis. Im zweiten Fall wird das Derivat aus Basis und Suffix als Basis der Präfigierung betrachtet. Sind in einer Wortform Suffixe und Flexionsmorpheme enthalten, stehen die Flexionsmorpheme rechts von den Suffixen.

Diminutiv-, Pejorativ- und Augmentativ-Suffixe führen nicht zur Änderung der Wortklasse des derivierten Lexems (Tabelle 3.4[9]).

Bei einigen Suffixen ist die Wortklasse des Derivats nicht eindeutig bestimmbar. Hier ist bei der Analyse mit Ambiguität zu rechnen.

Wie oben gezeigt, kann ein Lexem mehrere Suffixe annehmen. Ausgenommen ist dabei die Aufeinanderfolge zweier in ihrer Bedeutung entgegengesetzter Suffixe

[8] Zur Einteilung der Konjugationen und den damit verbundenen Konsequenzen siehe 3.5.

[9] Zusammengestellt aus [BosqueDemonte 1999].

(Diminutiv - Pejorativ).

Suffix-Morphem	Suffix-Art
-ejo, -et, -ete, -ico, -ill, -illo, -ín, -ina, -it, -ito, -uelo	Diminutiv-Suffix
-al, -az, -azo, -ón, -ot, -ote, -udo, -uda	Augmentativ-Suffix
-ac, -aco, -acho, -aj, -ajo, -ales, -alla, -ángano, -ango, -arr, -astre, -astro, -ej, -engue, -ic, -ingo, -ingue, -orio, -orr, -orrio, -orro, -uc, -uco, -uch, -ucho, -uj, -ujo, -ull, -ute, -uza	Pejorativ-Suffix

Tabelle 3.4: Diminutiv-, Pejorativ- und Augmentativ-Suffixe

3.2.3 Interfixe

Die Derivation durch Suffigierung erfordert oft den Einschub eines Interfixes. Suffixe und Interfixe sind interdependent: Bestimmte Interfixe stehen immer vor bestimmten Suffixen, bestimmte Suffixe erfordern den Einschub bestimmter Interfixe (Tabelle 3.5 und 3.6[10]).

3.2.4 Allgemeines

Als Basis eines Derivats können Substantive, Adjektive oder Verben verwendet werden. Ist die Basis ein Verb, kann das Stamm-Morphem oder das um den Themavokal[11] ergänzte Stamm-Morphem verwendet werden. Dies ist abhängig davon, ob das Suffix bzw. das Interfix mit Vokal beginnt. Endet ein Substantiv oder Adjektiv auf Vokal, wird als Basis der Derivation mittels Suffix das um den Vokal verkürzte Allomorph verwendet, wenn das folgende Interfix oder Suffix mit Vokal beginnt. Enthält das Morphem der Basis einen Diphtong, wird meist das monophtongierte Allomorph als Basis zur Derivation verwendet. Bei der linksassoziativen Betrachtung kann also aus der Oberfläche der Basis auf folgende Suffixe geschlossen werden.

Zur Derivation gehört auch die Rekategorisierung. So können die Partizipien der Verben als Adjektiv funktionieren.

3.3 Komposition

Als Komposition bezeichnet man die Verbindung von freien Morphemen oder wohlgeformten Wortformen. Sie nimmt im Spanischen einen geringeren Stellenwert als im Deutschen ein.[12] Es handelt sich in den meisten Fällen um Nominal-Komposition. Substantive, Adjektive, Adverben und Verben können kombiniert

[10] Zusammengestellt aus [BosqueDemonte 1999].
[11] Zur Funktion und Oberfläche des Themavokals siehe 3.5.
[12] [Thiele 1992, S. 95].

Interfix	abhängige Suffixe
-ach-	*-entoina*
-ad-	*-al, -ero, -ijo, -izo, -or, -ura, -uro*
-ag-	*-al, -án, -ero, -ón*
-aj-	*-ero, -eta, -ón, -oso*
-ac-	*-ón*
-al-	*-ache, -ada, -era, -eta, -ías, -ón, -uta*
-all-	*-ón*
-an-	*-ada, -ejo, -era, -oso*
-anch-	*-ín, -ón*
-ancl-	*-ón*
-and-	*-ejo, -ería, -ero, -ón, -ujo, -urria, -usca*
-andr-	*-uca*
-anc-	*-ón*
-ant-	*-ada, -ín, -ina, -ista, -ón*
-ar-	*-acho, -ada, -ajo, -al, -anga, -ano, -asca, -ata, -az, -azo, -eda, -eto, -illa, -ón, -ote, -uco, -ucho, -uto*
-arr-	*-aco, -ada, -ado, -año, -azo, -ero, -eta, -ón, -ucha, -uta*
-at-	*-el, -ero, -ina, -ón, -oso, -e*
-az-	*-án, -ón*
-ed-	*-al, -ero, -or, -izo, -ura, -uro*
-eg-	*-ada, -al, -oso, -ón, -ullo*
-ej-	*-ada, -al, -ero, -ón*
-el-	*-ón*
-ell-	*-ada, -era, -ón*
-end-	*-ero, -ón, -urria*
-ent-	*-in, -ón*
-er-	*-aje, -ano, -eta, -ete, -ón, -udo, -ueca, -uela, -ujo*
-err-	*-ón*
-et-	*-ad, -al, -azo, -ón*
-ec-	*-eja, -ete, -ico, -illo, -ísimo, -ito, -ote, -ucha, -uelo*
-c-	*-ejo, -ete, -eco, -illo, -ísimo, -ito, -ote, -ucho, -uelo*
-ich-	*-ento, -ón, -oso, -uelo*
-id-	*-ero, -izo, -or, -ura*
-ig-	*-ón, -orio, -ueño*
-ij-	*-illas, -ón*
-iqu-	*-era, -ete ón, -oso*
-il-	*-ada, -ero, -ín, -ón*
-ill-	*-ón, -oso*
-in-	*-ada, -ete, -eto, -eso*
-ind-	*-ango*
-ingl-	*-ero*
-ir-	*-ucho, -ujo*
-irr-	*-ito*
-isc-	*-ón, -oso*
-ist-	*-ón*
-it-	*-ajo, -al, -ar, -ero, -ina*
-iz-	*-ada, -al*
-ol-	*-era, -eto, -ina, -ón*
-oll-	*-ón*
-on-	*-eco*
-or-	*-eco, -eto, -ino, -uco*
-orr-	*-ada, -al, -ero, -eta, -ón*
-ud-	*-al*
-uch-	*-ina*
-ug-	*-ada, -azo, -ónera*
-uj-	*-ada, -ón*
-uc-	*-azo, -ón*
-ul-	*-ario, -eco, -ejo, -eque, -ón*
-ull-	*-ada, -ero, -ido, -ista, -ón*
-uñ-	*-ón*
-ur-	*-ucho, -uco*
-urr-	*-an, -illo, -ón*
-usqu-	*-ito*
-uz-	*-ada*
-uzg-	*-ón*

Tabelle 3.5: Abhängigkeit der Suffixe von Interfixen

Suffix	abhängige Interfixe
-aco	*-arr-*
-ache	*-al-*
-acho	*-ar-*
-ado	*-al-, -an-, -ant-, -ar-, -arr-, -ej-, -er-, -et-, -il-, -in-, -iz-, -orr-, -ot-, -ug-, -uj-, -ull-, -uz-, -ar-, -arr-, -ej-, -er-, -et-, -il-, -in-, -iz-, -orr-, -ot-, -ug-, -uj-, -ull-, -uz-*
-aje	*-er-*
-aje	*-er-*
-ajo	*-it-*
-al	*-ach-, -ad-, -ag-, -al-, -ar-, -at-, -az-, -ed-, -eg-, -ej-, -ell-, -et-, -it-, -iz-, -orr-, -ud -*
-ango	*-ar-, -ind-, -urr-*
-año	*-arr-*
-ano	*-ar-, -er-*
-án	*-ag-, -az-*
-ario	*-ul-*
-asco	*-ar-*
-azo	*-ar-, -arr-, -et-, -ot-, -eg-, -uc-*
-az	*-ar-*
-ción	*-ac-, -ag-, -aj-, -al-, -all-, -anc-, -anch-, -acl-, -ant-, -ar-, -arr-, -at-, -az-, -eg-, -ej-, -el-, -ell-, -end-, -ent-, -er-, -err-, -et-, -ic-, -ich-, -ig-, -ij-, -il-, -ill-, -isc-, -ist-, -ol-, -oll-, -orr-, -ot-, -ug-, -uj-, -uc-, -ul-, -ull-, -uñ-, -urr-, -uzg-*
-eco	*-on-, -or-, -ul-*
-eda	*-ar-*
-ejo	*-and-, -c-, -ul-*
-el	*-ach-, -at-*
-ento	*-ach-*
-eque	*-ul-*
-ero	*-ad-, -agu-, -aj-, -al-, -an-, -and-, -arr-, -at-, -ed-, -ej-, -ell-, -end-, -id-, -iqu-, -il-, -ingl-, -it-, -ol-, -orr-, -ot-, -ugu-, -ull-*
-ete	*-iqu-, -in-*
-eto	*-aj-, -al-, -ar-, -arr-, -er-, -in-, -ol-, -orr-*
-ías	*-al-*
-ico	*-ec-, -c -*
-íco	*-ec-, -c-*
-ido	*-ull-*
-ijo	*-ad-*
-illo	*-ar-, -ec-, -c-, -urr-*
-ín	*-ach-, -anch-, -ant-, -at-, -ent-, -it-, -ol-, -or-, -ot-, -uch-*
-ista	*-ant-, -ull-*
-ito	*-ec-, -c-, -irr-, -usqu*
-ón	*-ac-, -ag-, -aj-, -al-, -all-, -anc-, -anch-, -acl-, -ant-, -ar-, -arr-, -at-, -az-, -eg-, -ej-, -el-, -ell-, -end-, -ent-, -er-, -err-, -et-, -ic-, -ich-, -ig-, -ij-, -il-, -ill-, -isc-, -ist-, -ol-, -oll-, -orr-, -ot-, -ug-, -uj-, -uc-, -ul-, -ull-, -uñ-, -urr-, -uzg-*
-orio	*-ig-, -il-*
-or	*-ad-, -ed-, -id-*
-oso	*-aj-, -an-, -at-, -eg-, -ic-, -isc-*
-ote	*-ar-*
-ucho	*-ar-, -arr-, -ir-, -ur-*
-uco	*-andr-, -ar-, -or-, -ur-*
-ud	*-al-*
-ueca	*-er-*
-uello	*-er-, -ez-, -z-, -ich -*
-ueño	*-ig-*
-ujo	*-and-, -er-, -ir-*
-ullo	*-eg-*
-ura	*-ad-, -ed-, -id-*
-uro	*-ad-, -ed-, -id-*
-urria	*-and-, -end-*
-usca	*-and-*
-uto	*-al-, -ar-, -arr-*

Tabelle 3.6: Abhängigkeit der Interfixe von Suffixen

werden. Die Kategorie des Kompositums kann bei der Beteiligung von Substantiven und Adjektiven nicht eindeutig bestimmt werden. Die möglichen Kombinationen und die Wortklassen des jeweiligen Kompositums sind in Tabelle 3.7[13] dargestellt.

Wortklasse der ersten Komponente	Wortklasse der zweiten Komponente	Wortklasse des Kompositums
Substantiv	Substantiv	Substantiv
Substantiv	Adjektiv	Substantiv, Adjektiv
Adjektiv	Adjektiv	Substantiv, Adjektiv
Adjektiv	Substantiv	Substantiv
Verb	Substantiv	Substantiv
Adverb	Adjektiv	Adjektiv
Adverb	Verb	Verb
Adjektiv	Verb	Verb
Substantiv	Verb	Verb

Tabelle 3.7: Möglichkeiten der Komposition von Adjektiven, Substantiven, Verben und Adverben

Meist treten Komposita in der Form auf, dass die beteiligten Wortformen hintereinander ohne Zwischenraum geschrieben werden (*castellano + hablante* → *castellanohablante*, *sordo + mudo* → *sordomudo*). Es ist auch möglich, dass die Wortformen mit einem Bindestrich verbunden werden (*actor + cantante* → *actor-cantante*). Dies ist vor allem dann der Fall, wenn Koordination der beteiligten Wortformen vorliegt. Subordination wird überwiegend in Zusammenschreibung der Komponenten ausgedrückt.

In Fällen, in denen Komposita getrennt geschrieben werden, wie bei mehrstelligen Zahlen (*treinta y uno*), kann dies erst in der Syntax erkannt werden.

3.4 Nominal-Flexion

Flexionsmorpheme treten im Spanischen nur als Suffixe auf. Sie können danach unterschieden werden, ob sie für Substantive und Adjektive oder für Verben verwendet werden. Entsprechend werden Nominal- und Verbal-Flexion unterschieden (auf die Flexion der Verben wird in 3.5 eingegangen). Einige Wortformen anderer Wortklassen können ebenfalls Nominal-Flexionsmorpheme annehmen. Dies betrifft Artikel, Pronomen und Numerale, die ebenfalls in diesem Abschnitt behandelt werden.

3.4.1 Gemeinsamkeiten von Substantiven und Adjektiven

Die Flexionsmorpheme für Substantive und Adjektive kennzeichnen Genus und Numerus. Eine Kasusbestimmung erfolgt nicht.[14] Dabei sind für Numerus Sin-

[13] Zusammengestellt aus [BosqueDemonte 1999, S. 4335f.], [BosqueDemonte 1999, S. 4769].

[14] Es wird nur nach direktem und indirektem Objekt unterschieden. Dies ist erst syntaktisch analysierbar.

gular und Plural möglich. Bei der Unterscheidung der Genera fällt auf, dass im Gegensatz zu anderen Sprachen das Spanische keine Neutrum-Form für Substantive und Adjektive aufweist. Es wird nur nach maskulin und feminin unterschieden. Damit ergeben sich prinzipiell für jedes Substantiv und jedes Adjektiv vier mögliche Formen.

Adjektive weisen in der Regel alle vier Formen auf, Substantive kommen meist nur in der maskulinen oder nur in der femininen Form vor. Ausnahmen bilden Substantive, die Herkunft anzeigen (*el alemán, los alemanes, la alemana, las alemanas*), Bezeichnungen für Tiere (*el león, los leones, la leóna, las leonas*) oder Bezeichnungen für Berufe oder Tätigkeiten (*el monje, los monjes, la monja, las monjas*).

Genus

Das Genusmorphem bildet für Substantive und Adjektive im Maskulinum die Allomorphe *o, e* und im Femininum das Allomorph *a*. Ist das Maskulinum nicht durch Vokal gekennzeichnet, endet das Lexem auf *-d, -l, -n, -r, -s, -z*. Das entsprechende Femininum erhält das Genusallomorph *a* (*el autor - la autora, el marqués - la marquesa*). Ist das Maskulinallomorph *o* oder *e* vorhanden, wird dieses für die feminine Entsprechung der Wortform durch das Femininallomorph *a* ersetzt. Substantive, die nur im Maskulinum oder nur im Femininum auftreten, tragen oft keine Genusmarkierung (*la mujer*) oder der Endvokal entspricht nicht einem Genusallomorph (*la radio, el problema*).

In Fällen von Berufsangaben oder Substantiven, die gesellschaftliche Stellungen ausdrücken, sind zwei feminine Formen möglich. Die eine bezeichnet die Frau von jemandem, der eine bestimmte Stellung innehat. Die zweite Form ist die tatsächliche feminine Form, das heißt, die Frau hat diese Stellung inne (*el médico - la médica, la médico*).

Einige Berufsbezeichnungen weisen auch in der maskulinen Form die feminine Flexionsendung auf. Das tatsächliche Genus ist dann nur am Artikel oder an begleitenden Adjektiven zu erkennen (*el dentista - la dentista*).

Numerus

Der Singular von Nomen ist im Spanischen nicht gekennzeichnet. Er ist durch das Fehlen eines Pluralallomorphs zu erkennen. Das Pluralmorphem weist zwei Allomorphe auf: *s, es*. Das Pluralallomorph *s* wird für alle Substantive und Adjektive benutzt, die auf unbetonten Vokal enden (*alemana - alemanas*). Endet das Wort auf einen betonten Vokal oder auf einen Konsonanten, so wird das Allomorph *-s* verwendet (*alemán - alemanes*).

Wenn bei Komposita die Komponenten fest verschmolzen sind, erhält nur die am weitesten rechts stehende Komponente das Pluralmorphem. Sind die Komponenten

nicht fest verschmolzen,[15] ist es möglich, das Pluralmorphem sowohl nur an die am weitesten links stehende Komponente als auch an alle Komponenten zu hängen. Ist die Art der Verschmelzung der Komponenten nicht eindeutig, sind beide Varianten möglich (*la guardiacivil - las guardiaciviles, la guardia civil - las guardias civiles*).

Ausnahmen

Als Ausnahmen zu den prinzipiell vier möglichen flektierten Formen eines Nomen gelten singularia tantum, pluralia tantum und Formen, die im Singular und Plural (*la crisis - las crisis*) bzw. in der maskulinen und femininen Form die gleiche Oberfläche aufweisen. Letzteres ist vor allem bei Adjektiven möglich (*el libro azul - la sala azul*).

3.4.2 Besonderheiten der Adjektive

Unterscheidung nach Genus-Bildung

In [RAE 1974] werden die Adjektive entsprechend der Bildung von Maskulinum und Femininum unterschieden:

> [...] los que son genéricamente invariables (grupo primero); los que poseen femenino *-a*, masculino *-o* (grupo segundo), y los que tienen un femenino *-a* y un masculino que no es *-o* (grupo tercero).[16]

Diese Kennzeichnung kann zur Unterscheidung des Flexionsverhaltens genutzt werden.

Eine ähnliche Gruppierung ist für Substantive möglich. Substantive, die nur mit einem Genus auftreten, gehören zur ersten Gruppe. Substantive, die in beiden Genera möglich sind, gehören zur zweiten bzw. dritten Gruppe, die jeweils die gleichen Kriterien aufweisen wie die zweite bzw. dritte Gruppe der Adjektive.

Komparation

Im ESBOZO wird die Bildung von Komparativ und Superlativ nicht als Derivation, sondern als Flexion betrachtet, [17] daher erfolgt die Behandlung an dieser Stelle.

Der Komparativ der meisten Adjektive wird morphologisch nicht im Adjektiv selbst gekennzeichnet. Er ist erst in der Syntax zu erkennen, da er durch Voranstellung von *más* gebildet wird. Einige Adjektive haben aus dem Lateinischen die Bildung des Komparativs mit der Endung *-ior* beibehalten. Diese werden oft nicht als Komparativ, sondern als Positiv gebraucht.

Der absolute Superlativ ist morphologisch zu erkennen. Das entsprechende Adjektiv erhält die Endung *-ísimo* (maskulin) bzw. *-ísima* (feminin). Dieses Morphem tritt an die Stelle des Morphems, das im Positiv das Genus kennzeichnet. Auch die

[15] In der Regel ist das Kompositum dann nicht zusammengeschrieben.

[16] [RAE 1974, S. 191]; [... die, die in Bezug auf das Genus unveränderlich sind (erste Gruppe); die, die ein Femininum auf *-a*, Maskulinum auf *-o* aufweisen (zweite Gruppe), und die, die ein Feminium auf *-a* und ein Maskulinum, das nicht auf *-o* endet, aufweisen (dritte Gruppe)].

[17] Vgl. [RAE 1974, S. 198].

Adjektive der ersten Gruppe, die Maskulinum und Femininum nicht explizit kennzeichnen, tragen im absoluten Superlativ die Genuskennzeichnung.

Ausnahmen von dieser Regel bilden einige aus dem Lateinischen entlehnte Adjektive, die den absoluten Superlativ in der lateinischen Form beibehalten haben (*bonísimo*). Der absolute Superlativ kann aber auch regelhaft gebildet werden, beide Formen gelten als wohlgeformt (*bonísimo, buenísimo*).

Der relative Superlativ ist wie der Komparativ erst in der Syntax zu erkennen, da er durch *el más* + Adjektiv gebildet wird.

Ausgenommen von den bisher genannten Regeln zur Komparation sind die Adjektive *bueno, malo, grande* und *pequeño*, die entsprechend Tabelle 3.8 kompariert werden:

Positiv	Komparativ	relativer Superlativ
bueno	*mejor*	*óptimo*
malo	*peor*	*pésimo*
grande	*mayor*	*máximo*
pequeño	*menor*	*mínimo*

Tabelle 3.8: Komparation von *bueno, malo, grande, pequeño*

Apokopen von Adjektiven
Einige Adjektive können in einer verkürzten Form auftreten. Die maskuline Form von *bueno, santo* und *malo* wird zu *buen, san* und *mal*, wenn sie direkt vor einem Substantiv in der maskulinen Singular-Form steht. Die maskuline und die feminine Form von *grande* wird zu *gran*, wenn sie direkt vor einem entsprechenden Substantiv im Singular steht. In beiden Fällen gilt, dass zwischen dem verkürzten Adjektiv und dem betreffenden Substantiv keine weitere Wortform stehen darf. Das gleiche Phänomen tritt bei zwei Ordinalzahlen (siehe 3.4.5) und bei einigen Pronomina (siehe 3.4.4) auf.

3.4.3 Flexion der Artikel
Artikel sind wie Adjektive und Substantive hinsichtlich Genus und Numerus markiert. Hier sind also auch vier Formen möglich (*el - los, la - las*). Im Unterschied zu Substantiven und Adjektiven können Artikel auch im Neutrum auftreten. Dies ist der Fall, wenn ein Artikel vor substantivierten Verben (*lo escribir*) oder vor Adjektiven in der maskulinen Form (*lo bueno*) steht. Damit existieren sechs Formen der Artikel, wobei die Pluralformen von Maskulinum und Neutrum oberflächengleich sind.

3.4.4 Flexion der Pronomina
Über die Wortklasse der Pronomina gibt es in der Literatur geteilte Auffassungen.

DE BRUYNE unterscheidet Personalpronomina, Demonstrativpronomina, Possessivpronomina, Relativpronomina, Interrogativpronomina und Indefinitpronomina.[18]

ALARCOS LLORACH beschreibt dagegen keine Untergruppen von Pronomina, sondern gleichberechtigt neben den anderen Wortklassen sustantivos personales, demostrativos, posesivos, relativos, interrogativos, indefinidos und numerales.[19]

ALCINA FRANCH und BLECUA fassen unter dem Kapitel „El pronombre y el adverbio“ Pronomina und Adverben zusammen. Bei Pronomina unterscheiden sie locativos, personales, posesivos, demostrativos, cuantitativos indefinidos, numerales, enunciativos, interrogativos, exclamativos und temporales.[20] Sie schreiben:

> [Pronombres - C.M.] (a) forman una serie de sistemas morfológicos cerrados; (b) la mayor parte de ellas reciben morfemas de género y número como los nombres; algunas conocen el género neutro; (c) en determinados usos pueden neutralizar la oposición de género en singular; [...] (e) semánticamente, su significado no es pleno hasta que no se les relaciona con el contexto lingüístico o extralingüístico en que son utalizados.[21]

Weiter weisen sie darauf hin, dass die RAE einige Adverben, die Ort und Zeit beschreiben, als Lokativpronomina behandelt.

Besonders in den Fällen von Demonstrativ-, Lokativ- und Temporalpronomina ist eine Unterscheidung von Adverben und Pronomina schwer zu treffen. Einerseits kann es sich um Pronomina handeln, die durch eine Phrase ersetzt werden können. Andererseits kann es sich lediglich um den indexikalischen Gebrauch handeln, dann ist kein Ersatz durch eine Phrase möglich.

Hier soll folgende Lösung realisiert werden: Neben den schon behandelten Wortklassen werden Pronomina und Adverben unterschieden (da die Adverben zu den morphologisch unveränderlichen Wortformen gehören, werden sie in 3.7 behandelt). Innerhalb der Klasse der Pronomina werden Personalpronomina (mit der Unterscheidung nach unbetonten, im Falle der enklitischen Pronomina, und betonten), Possessivpronomina, Lokativpronomina, Demonstrativpronomina, Interrogativpronomina, Relativpronomina und Indefinitpronomina unterschieden. Dabei sollen Indefinitpronomina auch diejenigen Wortformen beinhalten, die keine konkreten Zahlenwerte angeben, sondern nur ungefähre oder nur im Vergleich mit anderen ermittelbare Werte.

---

[18] [de Bruyne 1993].

[19] [Alarcos 1994].

[20] [AlcinaBlecua 1994, S. 594ff.].

[21] [AlcinaBlecua 1994, S. 589f.]; [[Pronomina](a) bilden eine Serie geschlossener morphologischer Systeme; (b) der größte Teil enthält wie Nomen Genus- und Numerus-Morpheme; einige weisen das Genus Neutrum auf; (c) in bestimmten Verwendungsweisen können die Genusunterschiede im Singular aufgehoben sein; [...] (e) semantisch ist ihre Bedeutung ohne den linguistischen oder außerlinguistischen Kontext, in dem sie gebraucht werden, nicht eindeutig.].

Die Personalpronomina für die erste und zweite Person im Singular zeigen keine Genusmarkierung (*yo, me, mí; tú, te, ti*). Alle Pluralformen sowie die dritte Person Singular unterscheiden nach Maskulin und Feminin (*nosotros, nosotras; vosotros, vosotras; ellos, ellas*). Die Unterscheidung erfolgt durch das Maskulinallomorph *o* und das Femininallomorph *a*. Die gleichen Merkmale weisen die Possessivpronomina auf. Neben Informationen über Genus und Numerus des Subjektes enthalten bestimmte Possessivpronomina Informationen über Genus und Numerus des bezeichneten Objektes. So ist an den Formen *mío, mía, míos, mías* abzulesen, dass das Subjekt ein Femininum oder Maskulinum in der ersten Person Singular und das folgende Objekt jeweils ein Maskulinum im Singular, ein Femininum im Singular, ein Maskulinum im Plural und ein Femininum im Plural ist.

Lokativpronomina tragen keine Genus- und Numerus-Markierungen.

Demonstrativpronomina unterscheiden im Genus Maskulin, Feminin und Neutrum sowie im Numerus Singular und Plural. Dabei sind die Pluralformen des Neutrum und Maskulinum oberflächengleich.

Für Interrogativ-, Relativ- und Indefinitpronomina sind drei Gruppen zu unterscheiden: Pronomina, die nicht nach Genus und Numerus differenzieren; Pronomina, die wie die meisten Substantive für Numerus nach Singular und Plural unterscheiden, aber keine Genus-Markierung tragen; Pronomina, die wie die meisten Adjektive für Genus nach Feminin und Maskulin sowie für Numerus nach Singular und Plural differenzieren.

Ob ein Pronomen adverbial oder adjektivisch gebraucht wird, kann erst nach der Untersuchung von Syntax, Semantik und Pragmatik entschieden werden. Um die Anzahl der Analyse-Ergebnisse zu reduzieren, sollen nicht verschiedene Ergebnisse mit der jeweils möglichen Lesart erzeugt werden. Die Wortklasse wird durch das Multisymbol `Pronomen|Adverb` gekennzeichnet. Dies verdeutlicht, dass es sich sowohl um ein Pronomen als auch um ein Adverb handeln kann.

3.4.5 Flexion der Numerale

Unter die Klasse der Numerale fallen hier nur diejenigen Wortformen, die einen konkreten Zahlenwert wiedergeben.

Nicht alle Numerale sind morphologisch unveränderlich. Die Kardinalzahlen für 1 und für die Hunderter passen sich im Genus an die folgende Nominalphrase an. Alle anderen Kardinalzahlen existieren nur in einer Form. Die Entsprechungen für 1 bis 30 bilden jeweils eine Wortform. Die nächstgrößeren Zahlwörter sind Komposita, die nicht zusammengeschrieben werden. Hier kann erst durch syntaktische und semantische Analyse erkannt werden, wie groß die bezeichnete Zahl oder Menge ist.

Die Ordinalzahlen passen sich in Genus und Numerus an die bezeichnete Nominalphrase an. Wie bei den Adjektiven *grande, bueno, santo* und *malo* ist es für *primero*

und *tercero* möglich, sie in der maskulinen Form im Singular direkt vor maskulinen Substantiven im Singular in einer verkürzten Form zu verwenden (*primer, tercer*).

3.5 Verb-Flexion

Die Flexionsmorpheme der Verben kennzeichnen Person, Numerus, Modus und Tempus. Die Kategorie einer Wortform wird mit den entsprechenden Werten in dieser Reihenfolge angegeben. Die Kodierung von Person und Numerus erfolgt in einem Morphem, beide sind nicht trennbar. Ebenso sind die Informationen zu Modus und Tempus in einem Morphem vereint. Die Unterscheidung von aktivem und passivem Gebrauch des Verbs ist nur in der Syntax möglich.

Das Spanische unterscheidet drei Konjugationen. Die Zuordnung erfolgt nach dem Themavokal, der im Infinitiv zu erkennen ist. Der Infinitiv wird gebildet, indem an den Stamm die Endung *-ar* (erste Konjugation), *-er* (zweite Konjugation) oder *-ir* (dritte Konjugation) als Kombination aus Themavokal und Infinitivallomorph gefügt wird.

ALCINA FRANCH und BLECUA schreiben zur Bildung der flektierten Formen:

> [...] cada forma verbal mantiene el morfema lexemático y cambia los restantes morfemas según una serie finita de posibilidades; vocal temática; morfema auxiliar; morfema concordante (número y persona).[22]

Dabei wird der Stamm als »morfema lexemático« bezeichnet. »Morfema auxiliar« beinhaltet die Kombination von Modus und Tempus. Demnach sind alle flektierten Formen aus Stamm-Morphem, Themavokal, Modus/Tempus-Morphem und Person/Numerus-Morphem aufgebaut, wie in den Tabellen 3.10 bis 3.18 deutlich wird.

3.5.1 Person und Numerus

Für Person werden erste, zweite und dritte Person unterschieden. Für Numerus werden Singular und Plural unterschieden. Daraus ergeben sich sechs Kombinationen von Person und Numerus. Auf die einzelnen Morpheme wird in 3.5.3 eingegangen.

3.5.2 Modus und Tempus

Tabelle 3.9 gibt eine Übersicht über die den Modi Indikativ, Subjunktiv und Imperativ zugeordneten Tempora. Es werden die aus dem [RAE 1974] entnommenen spanischen Termini verwendet.

Insgesamt ergeben sich 17 Kombinationen von Modus und Tempus. Sie weisen alle die oben genannten sechs Formen hinsichtlich Person und Numerus auf. Eine Ausnahme bildet das Präsens des Imperativ, hier ist jeweils nur die zweite Person

[22] [AlcinaBlecua 1994, S. 735]; [jede Verbform weist ein lexematisches Morphem auf und wechselt die übrigen Morpheme entsprechend einer endlichen Serie von Möglichkeiten; Themavokal; Auxiliarmorphem; Konkordanzmorphem (Numerus und Person).].

indicativo	subjuntivo	imperativo
presente	presente	presente
pretérito imperfecto	pretérito imperfecto	
pretérito indefinido		
pretérito perfecto	pretérito perfecto	
pretérito pluscuamperfecto	pretérito pluscuamperfecto	
pretérito anterior		
futuro imperfecto	futuro imperfecto	
futuro perfecto	futuro perfecto	
potencial simple		
potencial compuesto		

Tabelle 3.9: Modi und Zeitformen der Verben

im Singular wie im Plural zu verwenden. Hinzu kommt jeweils eine Form für Infinitiv, Gerundium und Partizip. So ergeben sich für ein Verb 111 flektierte Formen. Die Formen des Futurs im Subjunktiv sind nur noch selten anzutreffen. Da sie in literarischen Texten aber noch verwendet werden, können sie hier nicht vernachlässigt werden.

Bei den Formen des Indikativ Perfekt, Indikativ Plusquamperfekt, Indikativ Präteritum (vorzeitig), Indikativ Futur Perfekt, Indikativ Potenzial II, Subjunktiv Perfekt, Subjunktiv Plusquamperfekt und Subjunktiv Futur Perfekt handelt es sich um zusammengesetzte Formen. Sie bestehen aus einer flektierten Form des Verbes *haber* und dem Partizip des entsprechenden Verbes. In diesen Formen wird das eigentliche Verb nicht morphologisch verändert. So bleiben die übrigen acht einfachen Formen, die in allen sechs verschiedenen Formen von Person und Numerus möglich sind, zwei Formen des Imperativ sowie Infinitiv, Gerundium und Partizip Perfekt für jedes Verb. Das ergibt 53 Formen.

Betrachtet man die Formen hinsichtlich ihrer Oberfläche, reduziert sich die Zahl weiter. Es sind zwar für jedes Verb 53 verschiedene Kategorisierungen möglich, allerdings weisen einige von ihnen die gleiche Oberfläche auf. So kann die Form *cantara* sowohl 1. als auch 3. Person Singular des Subjunktiv Präteritum Imperfekt sein. Zieht man alle ähnlichen Vorkommen für ein Verb in Betracht, reduziert sich die Zahl verschiedener Oberflächen auf 47. Ebenfalls oberflächengleich sind: 1. Person Plural Indikativ Präsens und 1. Person Plural Indikativ Indefinido, 1. und 3. Person Singular Indikativ Präteritum Imperfekt, 1. und 3. Person Singular Indikativ Potenzial, 1. und 3. Person Singular Subjunktiv Präsens sowie 1. und 3. Person Singular Subjunktiv Futur. Die Verwendung von Multisymbolen für die Kategorie der entsprechenden Wortform, wie in 2.2 gezeigt, erlaubt die Zusammenfassung der jeweiligen Formen.

Auch diese ermittelte Zahl ist noch nicht korrekt. Neben dem Phänomen der Oberflächengleichheit bei unterschiedlichen Kategorien ist es möglich, dass zu einer Kategorie verschiedene Oberflächen gehören können, die alle als wohlgeformt gelten. So gibt es im Subjunktiv Präteritum Imperfekt zwei Möglichkeiten, die entsprechenden Formen zu bilden. Das Morphem für Modus und Tempus weist die

Allomorphe *ra* und *se* auf. Damit erhöht sich die Zahl der Formen eines Verbs mit verschiedener Oberfläche um fünf[23] auf 52.

3.5.3 Flektierte Formen

Die Tabellen 3.10 bis 3.19 zeigen die Verbformen Indikativ Präsens, Indikativ Präteritum Imperfekt, Indikativ Indefinido, Indikativ Futur Imperfekt, Indikativ Potenzial, Subjunktiv Präsens, Subjunktiv Präteritum Imperfekt, Subjunktiv Futur Imperfekt, Imperativ Präsens (affirmativ), Infinitiv, Gerundium und Partizip. Diese Formen wurden in 3.5.2 als einfache Formen ermittelt.

Die Segmentierung erfolgt nach Stammallomorph, Themavokal, Modus/Tempus-Allomorph und Person/Numerus-Allomorph. Für jede mögliche Kombination aus Modus und Tempus sind die Formen in der Reihenfolge erste, zweite, dritte Person Singular, erste, zweite, dritte Person Plural angegeben. Für die Formen des affirmativ gebrauchten Imperativs sind die zweite Person Singular und die zweite Person Plural in dieser Reihenfolge angegeben. Die letzte Tabelle 3.19 gibt für Infinitiv, Gerundium und Partizip jeweils Stammallomorph, Themavokal und Auxiliar-Allomorph an. Dargestellt sind als Beispiel für die erste Konjugation *cantar*, für die zweite Konjugation *comer*, für die dritte Konjugation *vivir*. Alle drei sind reguläre Verben. Die Einordnung der einzelnen Allomorphe in die Spalten der Tabellen entspricht [BosqueDemonte 1999, S. 4937ff] mit Ausnahme von Tabelle 3.15. Die dort als Modus/Tempus gekennzeichneten Allomorphe werden als Themavokale behandelt.

Anhand der Tabellen können die Flexionsmorpheme und zugehörigen Allomorphe ermittelt werden. Als Morpheme gelten die drei Themavokale, die sechs Formen für Person und Numerus, die Auxiliarmorpheme für infinite Formen sowie die neun Kennzeichnungen von Modus und Tempus. In einer ergänzenden Spalte ist die komplette Flexionsendung angegeben.

3.5.4 Reguläre und irreguläre Verben

Die in der Literatur als irregulär bezeichneten Verben weichen von der Konjugation der regulären Verben ab. Irregularität tritt nur im Stamm-Morphem auf. Die Oberfläche der jeweiligen Kombination aus Themavokal, Modus/Tempus-Allomorph und Person/Numerus-allomorph für eine Verbform ändert sich nicht. Diese Verben werden danach unterschieden, ob die Veränderungen des Stammes Vokale, Konsonanten oder Vokale und Konsonanten gleichermaßen betrifft. Dabei können innerhalb jeder dieser drei Gruppen verschiedene Verben gruppiert werden. Eine Übersicht der sogenannten „Klassenverben“ gibt Tabelle 3.20[24].

[23] Da 1. und 3. Person Singular dieses Tempus oberflächengleich sind, kommen nur fünf statt sechs Formen hinzu.

[24] Zusammengestellt aus [BosqueDemonte 1999, S. 4952].

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*o*			*o*
	cant	*a*		*s*	*as*
	cant	*a*			*a*
	cant	*a*		*mos*	*amos*
	cant	*á*		*is*	*áis*
	cant	*a*		*n*	*an*
comer	*com*	*o*			*o*
	com	*e*		*s*	*es*
	com	*e*			*e*
	com	*e*		*mos*	*emos*
	com	*é*		*is*	*éis*
	com	*e*		*n*	*en*
vivir	*viv*	*o*			*o*
	viv	*e*		*s*	*e*
	viv	*e*			*es*
	viv	*i*		*mos*	*emos*
	viv	*í*		*s*	*éis*
	viv	*e*		*n*	*en*

Tabelle 3.10: Konjugation von *cantar, comer, vivir* im Indikativ Präsens.

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*a*	*ba*		*aba*
	cant	*a*	*ba*	*s*	*abas*
	cant	*a*	*ba*		*aba*
	cant	*á*	*ba*	*mos*	*ábamos*
	cant	*a*	*ba*	*is*	*abais*
	cant	*a*	*ba*	*n*	*aban*
comer	*com*	*í*	*a*		*ía*
	com	*í*	*a*	*s*	*ías*
	com	*í*	*a*		*ía*
	com	*í*	*a*	*mos*	*íamos*
	com	*í*	*a*	*is*	*íais*
	com	*í*	*a*	*n*	*ían*
vivir	*viv*	*í*	*a*		*ía*
	viv	*í*	*a*	*s*	*íais*
	viv	*í*	*a*		*ía*
	viv	*í*	*a*	*mos*	*íamos*
	viv	*í*	*a*	*is*	*íais*
	viv	*í*	*a*	*n*	*ían*

Tabelle 3.11: Konjugation von *cantar, comer, vivir* im Indikativ Präteritum Imperfekt

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*		*é*		*é*
	cant	*a*	*ste*		*aste*
	cant		*ó*		*ó*
	cant	*a*		*mos*	*amos*
	cant	*a*	*ste*	*is*	*asteis*
	cant	*a*	*ro*	*n*	*aron*
comer	*com*		*í*		*í*
	com	*i*	*ste*		*iste*
	com	*i*	*ó*		*ió*
	com	*i*		*mos*	*imos*
	com	*i*	*ste*	*is*	*isteis*
	com	*ie*	*ro*	*n*	*ieron*
vivir	*viv*		*í*		*í*
	viv	*i*	*ste*		*iste*
	viv	*i*	*ó*		*ió*
	viv	*i*		*mos*	*imos*
	viv	*i*	*ste*	*is*	*isteis*
	viv	*ie*	*ro*	*n*	*ieron*

Tabelle 3.12: Konjugation von *cantar, comer, vivir* im Indikativ Indefinido

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*a*	*ré*		*aré*
	cant	*a*	*rá*	*s*	*arás*
	cant	*a*	*rá*		*ará*
	cant	*a*	*re*	*mos*	*aremos*
	cant	*a*	*ré*	*is*	*aréis*
	cant	*a*	*rá*	*n*	*arán*
comer	*com*	*e*	*ré*		*eré*
	com	*e*	*rá*	*s*	*erás*
	com	*e*	*rá*		*erá*
	com	*e*	*re*	*mos*	*eremos*
	com	*e*	*ré*	*is*	*eréis*
	com	*e*	*rá*	*n*	*erán*
vivir	*viv*	*i*	*ré*		*iré*
	viv	*i*	*rá*	*s*	*erás*
	viv	*i*	*rá*		*irás*
	viv	*i*	*re*	*mos*	*iremos*
	viv	*i*	*ré*	*is*	*iréis*
	viv	*i*	*rá*	*n*	*irán*

Tabelle 3.13: Konjugation von *cantar, comer, vivir* im Indikativ Futur Imperfekt

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*a*	*ría*		*aría*
	cant	*a*	*ría*	*s*	*arías*
	cant	*a*	*ría*		*aría*
	cant	*a*	*ría*	*mos*	*aríamos*
	cant	*a*	*ría*	*is*	*aríais*
	cant	*a*	*ría*	*n*	*arían*
comer	*com*	*e*	*ría*		*ería*
	com	*e*	*ría*	*s*	*erías*
	com	*e*	*ría*		*ería*
	com	*e*	*ría*	*mos*	*eríamos*
	com	*e*	*ría*	*is*	*eríais*
	com	*e*	*ría*	*n*	*erían*
vivir	*viv*	*e*	*ría*		*ería*
	viv	*e*	*ría*	*s*	*erías*
	viv	*e*	*ría*		*ería*
	viv	*e*	*ría*	*mos*	*eríamos*
	viv	*e*	*ría*	*is*	*eríais*
	viv	*e*	*ría*	*n*	*erían*

Tabelle 3.14: Konjugation von *cantar, comer, vivir* im Indikativ Potencial Simple

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*e*			*e*
	cant	*e*		*s*	*es*
	cant	*e*			*e*
	cant	*e*		*mos*	*emos*
	cant	*é*		*is*	*éis*
	cant	*e*		*n*	*en*
comer	*com*	*a*			*a*
	com	*a*		*s*	*as*
	com	*a*			*a*
	com	*a*		*mos*	*amos*
	com	*á*		*is*	*áis*
	com	*a*		*n*	*an*
vivir	*viv*	*a*			*a*
	viv	*a*		*s*	*as*
	viv	*a*			*a*
	viv	*a*		*mos*	*amos*
	viv	*á*		*is*	*áis*
	viv	*a*		*n*	*an*

Tabelle 3.15: Konjugation von *cantar, comer, vivir* im Subjunktiv Präsens

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*a*	*se(ra)*		*ase (ara)*
	cant	*a*	*se(ra)*	*s*	*ases (aras)*
	cant	*a*	*se(ra)*		*ase (ara)*
	cant	*a*	*se(ra)*	*mos*	*asemos (aramos)*
	cant	*a*	*se(ra)*	*is*	*aseis (arais)*
	cant	*a*	*se(ra)*	*n*	*asen (aran)*
comer	*com*	*ie*	*se(ra)*		*iese (iera)*
	com	*ie*	*se(ra)*	*s*	*ieses (ieras)*
	com	*ie*	*se(ra)*		*iese (iera)*
	com	*ie*	*se(ra)*	*mos*	*iesemos (ieramos)*
	com	*ie*	*se(ra)*	*is*	*ieseis (ierais)*
	com	*ie*	*se(ra)*	*n*	*iesen (ieran)*
vivir	*viv*	*ie*	*se(ra)*		*iese (iera)*
	viv	*ie*	*se(ra)*	*s*	*ieses (ieras)*
	viv	*ie*	*se(ra)*		*iese (iera)*
	viv	*ie*	*se(ra)*	*mos*	*iesemos (ieramos)*
	viv	*ie*	*se(ra)*	*is*	*ieseis (ierais)*
	viv	*ie*	*se(ra)*	*n*	*iesen (ieran)*

Tabelle 3.16: Konjugation von *cantar, comer, vivir* im Subjunktiv Präteritum Imperfekt

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant*	*a*	*re*		*are*
	cant	*a*	*re*	*s*	*ares*
	cant	*a*	*re*		*are*
	cant	*a*	*re*	*mos*	*aremos*
	cant	*a*	*re*	*is*	*areis*
	cant	*a*	*re*	*n*	*aren*
comer	*com*	*ie*	*re*		*iere*
	com	*ie*	*re*	*s*	*ieres*
	com	*ie*	*re*		*iere*
	com	*ié*	*re*	*mos*	*iéremos*
	com	*ie*	*re*	*is*	*iereis*
	com	*ie*	*re*	*n*	*ieren*
vivir	*viv*	*ie*	*re*		*iere*
	viv	*ie*	*re*	*s*	*ieres*
	viv	*ie*	*re*		*iere*
	viv	*ié*	*re*	*mos*	*iéremos*
	viv	*ie*	*re*	*is*	*iereis*
	viv	*ie*	*re*	*n*	*ieren*

Tabelle 3.17: Konjugation von *cantar, comer, vivir* im Subjunktiv Futur Imperfekt

Infinitiv	Stammallomorph	Themavokal	Modus/Tempus-Allomorph	Person/Numerus-Allomorph	Flexionsendung
cantar	*cant* *cant*	*a* *a*		 *d*	*a* *ad*
comer	*com* *com*	*e* *e*		 *d*	*e* *ed*
vivir	*viv* *viv*	*e* *i*		 *d*	*e* *id*

Tabelle 3.18: Konjugation von *cantar, comer, vivir* im affirmativ gebrauchten Imperativ Präsens

Infinitiv	Stammallomorph	Themavokal	Auxiliar-Allomorph	Flexionsendung
cantar	*cant* *cant* *cant*	*a* *a* *a*	*r* *ndo* *do*	*ar* *ando* *ado*
comer	*com* *com* *com*	*e* *ie* *i*	*r* *ndo* *do*	*er* *iendo* *ido*
vivir	*viv* *viv* *viv*	*i* *ie* *i*	*r* *ndo* *do*	*ir* *iendo* *ido*

Tabelle 3.19: Formen von *cantar, comer, vivir* im Infinitiv, Gerundium, Partizip

Die Bildung der Allomorphe der jeweiligen Stamm-Morpheme ist semi-regulär bzw. semi-irregulär. Sie kann über Regeln aus der Oberfläche der Grundform abgeleitet oder anhand von Markierungen des Stamm-Morphems erkannt werden:

> *Semi-regular* inflectional paradigm
> The paradigm is represented by one lemma without any special surface markings, from which more than one allomorph is derived [...]

> *Semi-irregular* inflectional paradigm
> The paradigm is represented by one lemma with a special surface marker, from which more than one allomorph is derived [...][25]

Die nach HAUSSER als irregulär bezeichneten Verben umfassen Phänomene für das Stamm-Morphem, die auch unter dem Begriff der Suppletion genannt werden.

> *Irregular* inflectional paradigm
> The paradigm is represented by several lemmata for suppletive allomorps which pass through the default rule [...] The allomorphs serve as input to general combi-rules [...][26]

[25] [Hausser 1999, S. 263]; [*Semi-reguläres* Flexionsparadigma - Das Paradigma wird durch ein Lemma ohne Markierungen der Oberfläche dargestellt, von dem mehr als ein Allomorph abgeleitet wird [...] *Semi-irreguläres* Flexionsparadigma - Das Paradigma wird durch ein Lemma mit markierter Oberfläche dargestellt, von dem mehr als ein Allomorph abgeleitet wird ...].

[26] [Hausser 1999, S. 263], zum Begriff der combi-rules siehe 2.2.2. [*Irreguläres* Flexionsparadigma

Merkmal	Bezeichnung	Beispiel
Vokalalternation	*e - i* *o - u*	*pedir* *podrir*
Diphtongierung	*e - ie* *i - ie* *o - ue* *u - ue*	*tener* *adquirir* *volver* *jugar*
Vokalalternation und Diphtongierung	*e - ie - i* *o - ue - u*	*mentir* *dormir*
Konsonantenalternation	/θ/ - /g/	*hacer*
Konsonanteneinschub	/k/ /g/ *y*	*conducir* *salir* *huir*
Vokal-/Konsonantenalternation	*ab - ep* *ec - ig*	*caber* *decir*
Vokal-/Konsonanteneinfügung	*ig*	*caer*

Tabelle 3.20: Klassenverben

Suppletion tritt im Spanischen für die Verben *ser, estar, haber, dar, ver, saber* und *ir* auf. Alle anderen Abweichungen fallen unter die semi-regulären oder semi-irregulären Verben und weisen daher eine gewisse Regelhaftigkeit auf.

Die Allomorphe, die sich für ein Verb in ihrer Oberfläche von der Oberfläche der Grundform unterscheiden, werden nur in bestimmten Kombinationen aus Modus/Tempus und Person/Numerus verwendet.[27]

Einzelne Verben weichen nur in der Bildung des Partizips von der regulären Konjugation ab (*romper - roto*). Ist neben dem nicht-regulären Partizip auch das reguläre Partizip wohlgeformt, funktioniert das reguläre Partizip als Adjektiv (*freír - frito, freído*).

3.5.5 Enklitische Pronomina

Bildung von Wortformen aus Verben und enklitischen Pronomina

Die enklitischen Pronomina werden als unbetonte Personalpronomina bezeichnet.[28] Diese werden an Infinitiv, Gerundium oder affirmative Imperativ-Formen eines Verbs angefügt. Dabei sind bestimmte Regeln zu beachten.

Bis zu drei enklitische Pronomina können einer Verbform folgen. Enklitische Pronomina werden kategorisiert nach Person und Numerus. Nur die Pronomina der dritten Person werden hinsichtlich ihres Genus (Maskulinum, Femininum, Neutrum) unterschieden (Tabelle 3.21).

- Das Paradigma wird durch mehrere Lemmata dargestellt, die den suppletiven Allomorphen entsprechen. Die Allomorphe dienen als Eingabe für die Kombinationsregel]

[27] Zu einzelnen Bestimmungen vgl. [BosqueDemonte 1999, S. 4953ff.].

[28] Auf Pronomina insgesamt wird in 3.4.4 eingegangen.

Genus	1. Person Singular	2. Person Singular	3. Person Singular	1. Person Plural	2. Person Plural	3. Person Plural
Femininum	*me*	*te*	*la/se*	*nos*	*os*	*las/se*
Maskulinum	*me*	*te*	*le/se*	*nos*	*os*	*les/se*
Neutrum			*lo/se*			*los/se*

Tabelle 3.21: Übersicht der enklitischen Pronomina

Die Verwendung der Pronomina der dritten Person unterscheidet sich regional („loismo“, „leismo“, „laismo“). Hier soll keine Variante bevorzugt werden. Die Morphologiekomponente SMM soll Texte verschiedener Autoren analysieren können, so dass eine Beschränkung zu falschen Analysen führen kann.

Das Subjunktiv Präsens kann in imperativer Bedeutung verwendet werden. Als affirmativer Imperativ sind dabei nur die jeweils erste und dritte Person Singular und Plural gebräuchlich. In verneinter Form werden alle sechs Formen gebraucht. Die affirmativen Formen können enklitische Pronomina aufnehmen.

Für die Bildung von Wortformen aus Verbform und enklitischen Pronomina gilt weiterhin:

- Das *-s* der ersten Person Plural fällt aus, wenn als enklitisches Pronomen *nos* oder *se* folgt.
- Vor enklitischem *os* fällt das *-d* der zweiten Person Plural des Imperativs aus.
- *se* wird für das erste Pronomen nach der Flexionsendung benutzt, wenn zwei Pronomina der dritten Person aufeinander folgen.
- Die Betonung entspricht der Betonung der alleinstehenden Verbform. Es kann also notwendig sein, einen expliziten Akzent zu setzen oder ihn zu entfernen.[29]

Morphologie oder Syntax
Insgesamt stellt sich bei der Betrachtung enklitischer Pronomina die Frage, ob es sich tatsächlich um ein morphologisches Phänomen handelt. Die Verbformen, an die Pronomina angehängt werden können, gehören zwar zum jeweiligen Paradigma des Verbs - die durch die Kombination mit den Pronomina entstehenden Formen aber nicht. Das Entstehen kann weder durch Derivation noch durch Kombination oder Flexion erklärt werden.

Es handelt sich um ein syntaktisches Phänomen: Einem Verb folgt mindestens ein Objekt, das durch ein Pronomen ausgedrückt wird. Dieses Pronomen steht für eine komplexe Nominalphrase. Würde diese statt des Pronomens verwendet werden, erfolgte keine Zusammenschreibung von Verb und Objekt. Es fällt also in den Bereich der Syntax zu prüfen, ob die dem Verb folgende Phrase ein gültiger Valenz-

[29] Näheres dazu in 3.6.3.

füller ist.

Wird die komplexe Nominalphrase durch ein Pronomen ersetzt, ist durch Konvention festgelegt, dass Verb und Pronomen zusammengeschrieben werden. Die morphologische Analyse einer solchen Wortform muss daher als Ergebnis liefern, dass es sich um ein Verb und ein, zwei oder drei Pronomina handelt. Es muss erkannt werden, dass es sich um mehrere Wortformen handelt. Dass dies notwendig ist, zeigt die Möglichkeit, einen negativen Imperativ zu verwenden, also ein Verbot auszusprechen. In diesem Fall stehen die Pronomina vor dem Verb und werden nicht zusammengeschrieben (*¡cuéntamelo! - ¡no me lo cuentes!*)[30].

Bei der Kategorisierung des Verbs muss erkennbar werden, dass es sich um eine Form handelt, die es erforderlich macht, dass Pronomina angehängt werden. Die Kategorisierung der Pronomina muss deutlich machen, dass es sich um enklitische Pronomina handelt.

3.6 Allomorphie

3.6.1 Allomorphie bei Derivation

Bei der Derivation kann Allomorphie im Bereich der Affixe und im Bereich der Lexeme auftreten. Für Präfixe, Interfixe und Suffixe können die jeweiligen Allomorphe, die sich in der Oberfläche vom Morphem unterscheiden, nicht aus dieser abgeleitet werden.

Wie in 3.2.2 dargestellt, erfordert die Derivation mit Suffixen in einigen Fällen eine Veränderung der Oberfläche der Basis. Darunter fällt die Umwandlung von Diphtongen in Monophtonge und bei Basen, die auf Vokal enden, ein zusätzliches Allomorph, dessen Oberfläche um den Endvokal gekürzt ist.

3.6.2 Allomorphie bei Flexion

Allomorphie im Bereich der Flexion tritt vor allem bei Verben auf. Betrifft die Allomorphie die Vokale eines Stamm-Morphems, kann unterschieden werden, ob es sich um Alternation von Vokalen (*e - i (pedir - pide), o - u (poder - pudo)*), Diphtongierung (*e - ie (tener - tiene), i - ie (adquirir - adquiere), o - ue (volver - vuelve), u - ue (jugar - juego)*) oder eine Kombination beider (*e - ie - i (menti - miento - mintió), o - ue - u (dormir - duerme - durmió)*) handelt. Betrifft die Allomorphie die Konsonanten eines Stamm-Morphems, kann unterschieden werden, ob es sich um Alternation von Phonemen (/θ/ - /g/ (*hacer - hago*) oder um den Einschub eines Phonems (/k/ (*conducir - conduzco*), /g/ (*salir - salgo*)) handelt. Betrifft die Allomorphie einen Vokal und einen Konsonanten gleichermaßen, kann unterschieden werden, ob es sich um Alternation (*a - e + b - p (caber - quepa, saber - sepa)*)

[30] Dass die Verbform jeweils eine andere Oberfläche aufweist, liegt an den Regeln für die Bildung des Imperativs, der in der verneinten Form der 2. Person Singular der 2. Person Singular des Subjunktiv Präsens entspricht.

oder die Einfügung von Vokal und Konsonant (*ig (caer - caigo)*) handelt.[31] Einige Verben weisen eine Kombination der genannten Allomorphbildungen auf (*tener: tengo* (Einschub *g*), *tiene* (Diphtongierung *e - ie*)). Dabei handelt es sich um semi-reguläre und semi-irreguläre Allomorphie, die auch für die präfigierten Derivata dieser Verben gilt. Die Bildung der Allomorphe ist aus der Oberfläche der Stamm-Morpheme der entsprechenden Verben ableitbar oder wird aus der Markierung der Oberfläche entnommen.

Flexionsmorpheme der Substantive und Adjektive sowie der Wortklassen, die die gleichen Flexionsmorpheme annehmen können, weisen Allomorphie für das Genus-Morphem des Maskulinum (*o, e* und für das Numerus-Morphem des Plural (*s, es*) auf.

Unter welchen Bedingungen die jeweiligen Allomorphe eines Morphems zu verwenden sind, kann für semi-reguläre und semi-irreguläre Wortformen regelhaft ermittelt werden.

3.6.3 Behandlung von Allographen

Akzentuierung

Die Betonung des Spanischen bietet drei Möglichkeiten. Man unterscheidet:

- palabras agudas: Betonung auf der letzten Silbe (*derivación*)
- palabras llanas: Betonung auf der vorletzten Silbe (*derivaciones*)
- palabras esdrújulas: Betonung auf der vorvorletzten Silbe (*régimen*)

Genügt eine Wortform bestimmten orthographischen Regeln, ist es nicht notwendig, den Akzent explizit zu setzen. Beispielsweise sind Wortformen, die auf *-n, -s* oder Vokal enden, in der Regel palabras llanas. Weicht die Betonung einer Wortform davon ab, muss diese durch die Setzung des Akzentes kenntlich gemacht werden.

Wird die Anzahl der Silben durch Derivation, Komposition oder Flexion verändert, kann es erforderlich sein, einen expliziten Akzent zu entfernen oder auf eine andere Silbe zu verschieben (*derivación - derivaciones; régimen - regímenes*). In diesen Fällen kann das Stamm-Morphem also zwei Oberflächen aufweisen, eine mit und eine ohne Akzent bzw. zwei Oberflächen, die sich durch die Stellung des Akzents unterscheiden. Es handelt sich nicht um Allomorphe im eigentlichen Sinn,[32] sondern um Allographen.

Der explizit gesetzte Akzent kann bei der morphologischen Analyse nicht vernachlässigt werden, da er bedeutungsunterscheidend ist, wodurch sich verschiedene Kategorien ergeben (*que* [Relativpronomen] - *qué* [Interrogativpronomen]; *término*

[31] Siehe auch Tabelle 3.20.

[32] Allomorphe liegen vor, wenn ein Morphem verschiedene Formen annehmen kann, die in der erbwörtlichen oder lehnwörtlichen Herkunft des Morphems begründet sind. Vgl. [Thiele 1992, S. 12].

[Substantiv in der maskulinen Singularform] - *termino* [1. Person Singular Indikativ Präsens des Verbes *terminar*]- *terminó* [3. Person Singular Indikativ Präteritum Indefinido des Verbes *terminar*]).

Um diese Phänomene behandeln zu können, sollen hier die Oberflächen, die sich nur durch das Vorhandensein des Akzentes und nicht in der Kategorie unterscheiden, als Allomorphe angesehen werden.

Erhaltung der Aussprache durch orthographische Veränderung
Einige Phoneme des Spanischen hängen in ihrer orthographischen Entsprechung vom nachfolgenden Vokal ab. Da die Aussprache des Stamm-Morphems durch morphologische Prozesse nicht wechseln soll, ändert sich die Orthographie (Tabelle 3.22).

Phonem	Allograph	Bedingung
/g/	*g*	vor *a, o, u*
	gu	vor *e, i*
/k/	*c*	vor *a, o, u*
	qu	vor *e, i*
/*θ*/	*z*	vor *a, o, u*
	c	vor *e, i*
/x/	*j*	vor *a, o, u*
	g	vor *e, i*

Tabelle 3.22: Phoneme mit mehreren orthographischen Entsprechungen

Wie bei der Setzung expliziter Akzente handelt es sich nicht um Allomorphe, sondern um die Allographen des zugehörigen Phonems. Diese sollen hier gleich behandelt werden und die verschiedenen Gestalten eines Morphems als Allomorphe angesehen werden.

3.7 Morphologisch unveränderliche Wortarten

Präpositionen, Konjunktionen, Interjektionen und Adverben zählen zu den morphologisch unveränderlichen Wortarten. Jedes Morphem weist nur ein Allomorph auf. Präpositionen und Konjunktionen bilden geschlossene Klassen mit begrenztem Umfang.

Die Konjunktionen *y* und *o* bilden eine Ausnahme. Sie weisen jeweils zwei Allomorphe auf: Beginnt die auf *y* folgende Wortform mit dem Phonem /i/, wird *e* statt *y* verwendet. Beginnt die auf *o* folgende Wortform mit dem Phonem /o/, wird *u* statt *o* verwendet.

Die Klasse der Interjektionen ist nicht auf eine bestimmte Zahl begrenzt. Besonders durch Onomatopoetika ist diese Klasse sehr produktiv.

Die Wortklasse der Adverben zählt zu den morphologisch unveränderlichen Wörtern, ist aber sehr produktiv. Es gibt Wortformen, die zur Klasse der Adverben zählen (*bien*). Daneben ist es möglich, Adverben durch Derivation zu bilden. Die Basis ist dabei ein Adjektiv. Die feminine Form erhält das Suffix *-mente* (*clara - claramente*). Die Derivation kann als Basis auch die komparierte Form eines Adjektivs aufweisen. Auch hier muss die feminine Form vorliegen (*clarísimamente*). Die Bildung von Adverben ist ebenfalls durch Rekategorisierung möglich. Hier funktioniert die maskuline Form eines Adjektivs als Adverb (*hablar alto*).

Kapitel 4

Andere Systeme

4.1 Morphologische Analyse des Spanischen im Rahmen von ARIES

In Zusammenarbeit der „Escuela Técnica Superior de Ingenieros de Telecomunicación: Departamentos de Ingeniería de Sistemas Telemáticos y de Matemática Aplicada a las Tecnologías de la Información“ an der „Universidad Politécnica de Madrid“ und des Labors für linguistische Informatik der „Universidad Autónoma de Madrid“ entstand unter dem Namen ARIES eine lexikalische Komponente zur Verarbeitung der spanischen Sprache. Sie umfasst ein großes spanisches Lexikon mit eigener Repräsentationssprache sowie Werkzeuge, um Wortformen zu generieren und morphologisch zu analysieren.[1]

Über die lexikalische Repräsentationssprache schreiben GONZÁLEZ ET.AL.:

> The design of this lexical representation language was influenced by the strong reliance of the Spanish language on inflectional morphology (e.g. 53 simple word forms for verbs, up to 4 forms for nouns and adjectives).[2]

Als Anforderung für die Komponente zur morphologischen Analyse im Rahmen von ARIES gilt die Erzeugung von Allomorphen, um sie dann konkatenieren zu können.

Die Allomorphe werden aus den Lexikoneinträgen des Grundformlexikons mit Regeln, die reguläre Ausdrücke beinhalten, expandiert. Es handelt sich um Ersetzungsregeln, wobei die linke Seite der Regel durch die rechte Seite der Regel ersetzt wird.[3] Dabei werden als Allomorphe auch die Formen eines Morphems behandelt, die aufgrund phonetischer Besonderheiten eine Veränderung in der graphischen Oberfläche aufweisen. Weiterhin sind Lexikoneinträge notwendig, die irreguläre Phänomene abdecken. Es wird eine Symboldatei verwendet.[4]

[1] [ARIES].

[2] [González et al. 1997]; [Das Design der lexikalischen Repräsentationssprache wurde durch die starke Ausrichtung der spanischen Sprache auf Flexions-Morphologie beeinflusst (z.B. 53 einfache Verbformen, bis zu 4 Formen für Substantive und Adjektive).].

[3] [Goñi et al. 1995b].

[4] [González et al. 1997].

Die Lexikoneinträge haben die Form von Attribut-Werte-Paaren.[5] In ARIES werden die Strukturprinzipien Allomorphie und Konkatenation in voneinander getrennten Regeln behandelt.[6] Dabei wird von der konkreten graphischen Oberfläche ausgegangen. Als grundlegender Grammatikformalismus wird die PSG gewählt:

> Ello implica que las reglas que permiten la concatenación de morfemas sean de tipo sintagmático (PSG), y que la gramática empleada sea de Contexto Libre.[7]

Weiterhin wird für die Implementierung Unifikation benutzt.[8]

Die für ARIES implementierten 53 Allomorphregeln erzeugen aus 38 000 Grundformen 465 000 flektierte Formen.[9] Dabei werden 622 Morpheme und 697 lexikalisch unveränderliche Einträge benötigt.[10]

ARIES ermöglicht nur die morphologische Analyse von flektierten Wortformen des Spanischen. Die Erkennung von Derivation und Komposition ist nicht möglich. GOÑI und GONZÁLEZ geben an, dass bis 1995 noch keine formalisierte Beschreibung für Komposition oder Derivation vorlag:

> The treatment of compositional or derivative morphology is not as critical, and, to our knowledge, it does not exists [sic!] any formalized theoretical linguistic description of such phenomena.[11]

Deshalb beschränken sie sich in ihrem morphologischen Parser auf Flexion.

Die in ARIES entwickelten Werkzeuge sollen nach MORENO und GOÑI aber auch für die Behandlung von Komposition und Derivation nutzbar sein.[12] GONZÁLEZ COLLAR ET.AL schreiben dazu:

> Existen diversas causas por las que los procesos morfológicos derivativos son de una complejidad superior a los flexivos, por lo que el procesamiento computacional que aquí se describe no incluirá la derivación, considerando exclusivamente la flexión, si bien, tanto el modelo computacional utilizando, como el procesador morfológico realizado, podrían considerar la derivación si se dispusiese de un modelo lingüístico formalizado.[13]

---

[5] [Goñi et al. 1995a].

[6] GOÑI und GONZÃLEZ beziehen sich dabei explizit auf den LA-MORPH-Ansatz in [Hausser 1989b], Vgl. [Goñi et al. 1995b].

[7] [González Collar et al. 1995]; [Dies impliziert, dass die Regeln, die die Konkatenation der Morpheme erlauben, von syntagmatischem Typ sind (PSG), und dass die implementierte Grammatik kontextfrei ist.].

[8] Ebenda.

[9] [Goñi et al. 1995a].

[10] [González et al. 1997].

[11] Vgl. Fußnote in [Goñi et al. 1995b]; [Die Behandlung der Kompositions- oder Derivations-Morphologie ist nicht so kritisch und es existiert unseres Wissens keine formale theoretische linguistische Beschreibung dieses Phänomens.].

[12] [Moreno et al. 1995].

[13] [González Collar et al. 1995]; [Es existieren verschiedene Gründe dafür, dass die morphologi-

Den Flexionsmorphemen sind Attribut-Werte-Paare zugeordnet, die neben der Kategorisierung Informationen über die Konkatenationsmöglichkeiten enthalten. Alle Flexionsmorpheme, die zu einer konkreten Wortform eines Verbs gehören, werden zusammengefasst. So wird *-abamos* als Beispiel für einen Morphemeintrag angegeben,[14] obwohl die morphologische Analyse noch weitergehen kann und Themavokal *-a-*, Modus/Tempus-Morphem *-ba-* und Person/Numerus-Morphem *-mos* unterscheidet.

4.2 Morphologie-Komponenten für Malaga: DMM, IMM, EMM

Auf die Morphologie-Komponenten für das Deutsche, Italienische und Englische wird in Hinsicht auf Implementierungs- und Strukturierungs-Entscheidungen eingegangen. Sprachabhängige Aspekte werden nur insoweit berücksichtigt, als ähnliche Phänomene im Spanischen auftreten.

4.2.1 DMM

Die Konkatenation der Allomorphe wird von der Startregel, sechzehn Kombinationsregeln und der Endregel gesteuert.[15] Dabei kann jede der Kombinationsregeln zu mehreren Zuständen führen, denen jeweils mehrere Regeln folgen können. Allein die Startregel führt zu fünf Zuständen, die bis zu fünf Folgeregeln aufweisen (Abbildung 4.1).

Der Regelgraph wird damit zu einem höchst komplexen Graph. Welche Kombinationsregeln angesteuert werden, hängt davon ab, ob es sich beim eingelesenen Allomorph um ein zu einem Stamm-Morphem oder zu einem gebundenen Morphem gehöriges Allomorph handelt. Die Kombinationsregeln spiegeln ansatzweise die Prinzipien der Wortbildung wider.

4.2.2 IMM

Die Konkatenation der Allomorphe wird von der Startregel, einer Kombinationsregel und der Endregel gesteuert (Abbildung 4.2[16]).

Die Konkatenation wird über Informationen gesteuert, die in den Allomorpheinträgen enthalten sind. Dort legen Attribut-Werte-Strukturen fest, welche Kategorie ein vorausgehendes Allomorph haben muss bzw. ein nachfolgendes Allomorph haben kann.

---

schen Prozesse der Derivation von einer höheren Komplexität als die der Flexion sind und dass der Prozess, der hier beschrieben wird, die Derivation nicht einschließt, sondern ausschließlich die Flexion behandelt, obwohl das verwendete Modell wie der realisierte Morphologie-Prozessor ebenfalls die Derivation behandeln können, wenn ein formales linguistisches Modell dafür zur Verfügung steht.].

[14] [Goñi et al. 1995b].

[15] Die aktuelle Version der DMM weicht von der in [Lorenz 1996] beschriebenen ab. Aussagen über die Allomorph- und Kombinationsregeln und deren Implementierung beziehen sich auf die DMM4.3.

[16] Nach [Leidner 1998, S. 87] und [Wetzel 1996].

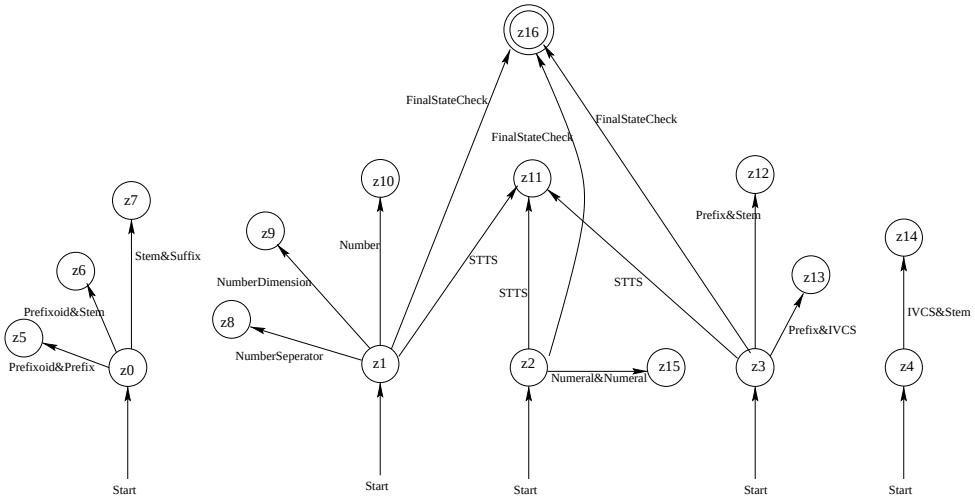

Abbildung 4.1: Ausschnitt aus dem Regelgraph der Kombinationsregeln der DMM

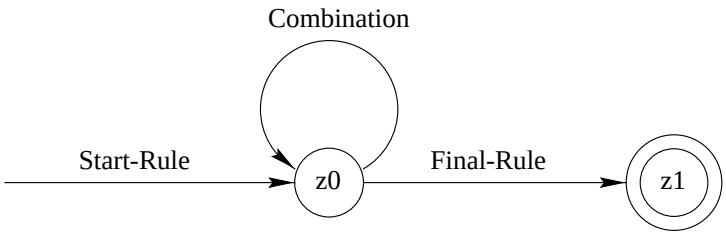

Abbildung 4.2: Regelgraph der Kombinationsregeln der IMM

Vollformen, die für irreguläre Verben schon im Grundformlexikon aufgeführt werden müssen, sind im Eintrag der zugehörigen Grundform enthalten. Verben werden nach Stamm und Flexionsendung segmentiert. Die Flexionsendung eines Verbs gilt als untrennbare Kodierung von Themavokal, Modus, Tempus, Person und Numerus.

4.2.3 EMM

Die Konkatenation der Allomorphe wird von einer Kombinationsregel und der Endregel gesteuert (Abbildung 4.3)[17].

Damit müssen in noch stärkerem Maße als in der IMM in den Allomorpheinträgen Informationen darüber enthalten sein, welche Allomorphe nachfolgen bzw. vorausgehen können. Begründet wird die Entscheidung für eine Kombinationsregel mit der besseren Wartbarkeit der gesamten Komponente. Änderungen, die Allo-

[17] [Leidner 1998, S. 87].

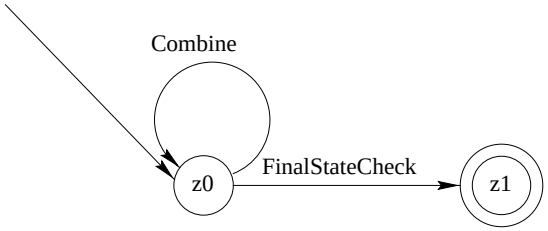

Abbildung 4.3: Regelgraph der Kombinationsregeln der EMM

morphregeln und damit die Attribut-Werte-Struktur der Allomorpheinträge betreffen, müssen nicht in großem Umfang in den Kombinationsregeln nachvollzogen werden.

In der EMM werden regelmäßige (reguläre, semi-reguläre) und unregelmäßige Verben unterschieden (semi-reguläre, irreguläre).

Im Grundformenlexikon werden irreguläre Formen eines Wortes im Eintrag der Grundform aufgeführt.

Kapitel 5

Entscheidungen zur Implementierung der SMM

5.1 Konkatenation der Allomorphe

Zunächst war beabsichtigt, mit den Kombinationsregeln die Prozesse von Derivation, Kombination und Flexion nachzubilden. Der sich daraus ergebende Regelgraph hätte aber ähnliche Ausmaße angenommen wie der der DMM. Darunter leidet die Lesbarkeit des Codes ebenso wie die Wartungsfreundlichkeit. Die SMM weist daher wie die EMM nur eine Kombinationsregel und eine Endregel auf (Regelgraph siehe Abbildung 4.3).

Wie in der IMM wird die Konkatenation der Allomorphe über Informationen gesteuert, die die beteiligten Allomorphe enthalten. Jedes Allomorph enthält ein Attribut `Suc`, das Bestimmungen über mögliche Nachfolger beinhaltet. Damit wird das Prinzip der möglichen Fortsetzungen der LAG explizit erfüllt. Zusätzlich weist jedes Allomorph ein Attribut `Pre` auf, das Bestimmungen über die vorhergehenden Allomorphe bzw. Wortanfänge enthält.

Bei der Konkatenation eines Wortanfangs mit einem nächsten Allomorph wird erst geprüft, ob das Allomorph die Nachfolgebedingungen des Wortanfangs erfüllt. Ist dies der Fall, wird geprüft, ob der Wortanfang die Vorgängerbedingungen des Allomorphs erfüllt. So können Verbstämmen nur Themavokale oder Flexionsendungen folgen. Dies wird über die im Attribut `Suc` des Verbstamms enthaltenen Bestimmungen überprüft. Ist das nächste Allomorph ein Themavokal, so kann dieser nur auf Verbstämme folgen, die aufgrund ihrer Konjugationsklasse einen bestimmten Themavokal erfordern. Dies wird mit Hilfe der im Attribut `Pre` des Themavokals enthaltenen Anforderungen überprüft. Die genaue Funktionsweise dieser Überprüfungen wird in 6.4.1 beschrieben.

Ist der Wortanfang bereits aus mehreren Allomorphen konkateniert, deren letztes ein Interfix ist, muss als nächstes Allomorph ein Suffix folgen. Die Vorgängerbestimmungen des Suffixes bezüglich möglicher Interfixe beziehen sich dann nicht auf den gesamten Wortanfang, sondern auf das Interfix, also das zuletzt eingelesene Allomorph.

Endet ein Wortanfang mit einem der in Tabelle 3.22 dargestellten Phoneme, muss das folgende Allomorph neben den in `Suc` genannten Nachfolgebestimmungen mit einem bestimmten Vokal beginnen. Diese Bedingung ist im Attribut `SucFon` enthalten.

5.2 Erzeugung der Allomorphe aus dem Grundformlexikon

Aus den Einträgen im Grundformlexikon werden über die Allomorphregeln die Einträge des Allomorphlexikons erzeugt. Die Allomorpheinträge enthalten alle kategoralen Informationen, die für die Konkatenation benötigt werden. Jedes Allomorph erhält die Attribute `Suc` und `Pre`, die Bestimmungen über nachfolgende und vorausgehende Allomorphe bzw. Wortanfänge enthalten.

Morpheme, deren Endphonem in Abhängigkeit von folgenden Allomorphen verschiedene Schreibweisen zulässt, erfordern für jede Schreibweise ein eigenes Allomorph. Jedes Allomorph erhält das Attribut `SucFon`, um den Beginn des nächsten Allomorphs festzulegen. Für Morpheme, die abhängig von ihrer Stellung innerhalb der Wortform mit und ohne expliziten Akzent auftreten können, werden zwei Allomorphe mit entsprechender Oberfläche erzeugt. Damit werden ähnlich wie in ARIES Allographen als Allomorphe behandelt.

Die Lemmata des Grundformlexikons enthalten alle Informationen, die zur Erzeugung der Allomorpheinträge notwendig sind. Sind Informationen ableitbar, etwa, dass Verben immer im Infinitiv aufgeführt sind, muss dies im Grundformlexikon nicht vermerkt werden. Sollen die Allomorphe eines Morphems besondere nichtableitbare Eigenschaften, insbesondere bezüglich Vorgängern und Nachfolgern, aufweisen, sind diese dem Eintrag im Grundformlexikon hinzugefügt. Für irreguläre Formen ist das vollständige Paradigma wie in der IMM und der EMM im Eintrag der Grundform aufgeführt.

Reguläre und semi-reguläre Verben werden in einer Regel behandelt, eine zweite erzeugt die Allomorphe für semi-irreguläre Verben. Irreguläre Verbformen sind im Grundformlesikon aufgeführt

5.3 Ausgabe der Analyse-Ergebnisse

Für jede als wohlgeformt kategorisierte Wortform sollen Oberfläche, Grundform, Wortklasse, Segmentierung in Allomorphe und die Struktur der analysierten Wortform entsprechend Abbildung 5.1 angegeben werden. Unter Struktur wird die Zuordnung des entsprechenden Morphems und dessen Wortklasse zu jedem in der Segmentierung ermittelten Allomorph verstanden.

In den bisher implementierten Morphologiekomponenten ist für die Struktur einer Wortform nur die Zuordnung der Morpheme zu den Allomorphen vorgesehen.

Die zusätzliche Angabe der jeweiligen Wortklasse soll Zerlegungen in Allomorphe motivieren, die hinsichtlich der Oberfläche der Allomorphe identisch, hinsichtlich deren Kategorien verschieden sind.

In der Angabe der Segmentierung wird vermerkt, ob zwei Allomorphe durch Prozesse der Derivation, Kombination oder Flexion verbunden sind. Handelt es sich um das Anhängen eines enklitischen Pronomens, die Konkatenation von Ziffern oder von Zahlen und Maßeinheiten, wird dies gesondert gekennzeichnet.

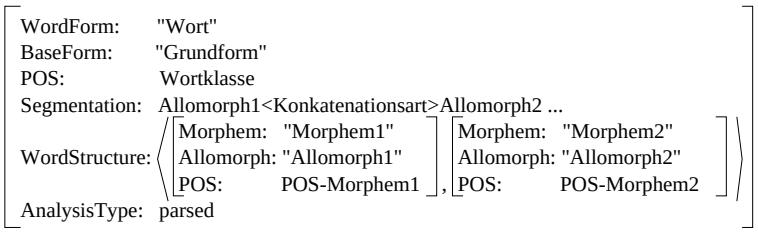

Abbildung 5.1: Struktur der Ergebnisse der Analyse

Weiterhin enthält jedes Analyseergebnis alle kategoralen Informationen, die entsprechend der ermittelten Wortklasse notwendig sind. Für Substantive und Adjektive beispielsweise Genus und Numerus, für Verben Valenz, Tempus, Modus, Person und Numerus. Handelt es sich um ein Derivat, werden semantische und syntaktische Eigenschaften der Präfixe und Suffixe angegeben.

In der DMM, IMM und EMM werden für Wortformen, die ambig sind und mehrere Analysen zulassen, diese Ergebnisse in einer Liste angegeben. Darauf wird hier verzichtet. Eine Ergebnisliste wird für Verbformen mit enklitischen Pronomina erzeugt, da es sich wie in 3.5.5 gezeigt, um mehrere wohlgeformte Wortformen handelt. Das erste Element ist das Analyseergebniss für die Verbform. Die weiteren Listenelemente sind die Analyseergebnisse der jeweils folgenden Pronomina. Beispiele werden in 6.5 gegeben.

Kapitel 6

Die Morphologiekomponente SMM

Die entwickelte Komponente zur morphologischen Analyse spanischer Wortformen besteht aus der Symbol-Datei `smm.sym`, der Allomorph-Datei `smm.all`, der Morphologie-Datei `smm.mor` sowie mehreren Lexikon-Dateien, die in der Datei `smm.lex` zusammengefasst sind. Alle Dateien sind in der Projekt-Datei `smm.pro` aufgeführt. Die Endungen der Dateien (`sym`, `all`, `mor`, `lex` und `pro`) sind durch MALAGA vorgegeben.

6.1 Die Symboldatei

Die in den Allomorph- und Kombinationsregeln verwendeten Symbole für die Bezeichnung von Attributen und Werten werden in die Symboldatei eingetragen. Jedes Symbol steht in einer Zeile, die mit einem Semikolon beendet wird. Wie in 3.1.2 begründet, werden nur englische Begriffe verwendet. Dabei enthalten die Symbole, die für grammatische Kategorien verwendet werden, als Kommentar die spanische Bezeichnung.

Multisymbole sind dadurch gekennzeichnet, dass sie die enthaltenen Symbole durch | verbunden als Namen erhalten.

```
Substantive|Adjective := <Substantive, Adjective>;
```

Symbole, die als Namen andere durch & verbundene Symbole tragen, sind keine Multisymbole. Sie machen deutlich, dass in dem entsprechenden Symbol dessen Bestandteile fest miteinander verbunden sind.

```
Preposition&Determiner;
```

6.2 Das Lexikon

6.2.1 Die Lexikonressourcen

Zunächst war beabsichtigt, das Grundformenlexikon zu verwenden, welches im Rahmen des Projektes ARIES erstellt wurde. Einzelne Komponenten von ARIES werden nur gegen eine Lizenzgebühr zur Verfügung gestellt. Die kostenlose Nutzung für wissenschaftliche Zwecke ist nicht möglich. Die geforderte Lizenzgebühr

des Grundformlexikons[1] war allerdings so hoch, dass dies nicht verwirklicht werden konnte.

Das Lexikon wird daher aus Daten von im Handel erhältlichen oder im World Wide Web verfügbaren einsprachigen spanischen Lexika erstellt. Es handelt sich um das „Diccionario de la lengua española“ der „Real Academia Española“auf CD-ROM[2], das der 21. Auflage der gedruckten Version entspricht, mit 83 014 Lemmata und um das „Diccionario General de la Lengua Española VOX“ mit 95 840 Lemmata.[3] Daraus werden bis auf die Flexionsmorpheme alle Einträge gewonnen. Die jeweiligen Oberflächen und notwendigen Attribute für letztere ergeben sich aus den Anforderungen, die Derivation und Flexion stellen, und werden aus den Angaben in [BosqueDemonte 1999] und [RAE 1974] ermittelt. Affixe, die in beiden Lexika nicht aufgeführt sind, auf die in den verwendeten Grammatiken aber eingegangen wird, werden nachgetragen. Zusätzlich werden Satzzeichen, Zahlen und Maßeinheiten in das Lexikon aufgenommen.

6.2.2 Zusammensetzung des Lexikons

Aus dem DGLE werden 98 964 Einträge gewonnen. Entsprechend der grammatikalischen Kennzeichnung der Wortart im unveränderten Format werden diese nach Wortarten unterschieden. Einträge, die mehrere Wortarten als Kategorie zulassen, werden für jede mögliche Wortart aufgenommen, daher ist die Zahl der gewonnen Einträge größer als die Zahl der im DGLE enthaltenen Lemmata.

Es ergibt sich die in Tabelle 6.1 aufgeführte Zusammensetzung. Entsprechend der Wortarten sind die Einträge in verschiedenen Dateien zusammengefasst, um sie besser verwalten zu können. Artikel und Pronomina befinden sich in einer Datei. Für die Behandlung irregulärer Allomorphie werden die betroffenen Grundformen mit dem vollständigen Paradigma in einer eigenen Datei angegeben.

Bei den Einträgen der Verben werden Formen, die keine Infinitive sind, entfernt. 3408 Verben sind nicht regulär, davon sind 6 irregulär (*estar, ser, haber, ver, ir, saber*). Die anderen 3402 können durch die Auswertung der Konjugations-Markierung 67 verschiedenen Gruppen von Verben zugeordnet werden. Diese Zuordnungen können im Weiteren zur Behandlung der Allomorphie semi-regulärer und semi-irregulärer Verben genutzt werden.

Lemmata, die nur im DRAE enthalten sind, werden in der jeweiligen Lexikon-Datei ergänzt. Dazu werden die Oberflächen aller Einträge als Wortliste formatiert. Diese wird mit der SMM geparst, die bis dahin nur mit den Lemmata des DRAE arbeitet. Alle nicht erkannten Oberflächen führen zu einem zusätzlichen Eintrag. Das DRAE liefert alle notwendigen Informationen zur Kategorisierung der Lemmata.

[1] Die gegen Lizenz verfügbaren Komponenten von ARIES Vgl. [Goñi 1996].
[2] [DRAE 1998], hier weiter bezeichnet als DRAE.
[3] [VOX], hier weiter bezeichnet als DGLE.

Wortklasse	Anzahl	Datei
Substantive	57 627	`substantive.lex`
Adjektive	24 851	`adjective.lex`
Verben	11 873	`verb.lex`
Adverben	2 647	`adverb.lex`
Konjunktionen	33	`conjunction.lex`
Präpositionen	28	`preposition.lex`
Artikel	6	`determiner.lex`
Pronomina	55	`determiner.lex`
Interjektionen	310	`interjection.lex`
Namen	327	`name.lex`
Akronyme	79	`acronym.lex`
Affixe	1128	`affix.lex`
Paradigmen	11	`paradigm.lex`

Tabelle 6.1: Aus dem „Diccionario General de la Lengua Española VOX“ gewonnene Lexikoneinträge

Für Verben ist im DRAE nicht vermerkt, ob diese regelmäßig oder nicht-regelmäßig konjugiert werden. Diese Angabe fehlt nicht nur in der elektronischen, sondern auch in der gedruckten Version. Für Derivata bereits aufgenommener Verben wird die gleiche Konjugation angenommen. Für alle anderen Verben stammt die Information über die Konjugation aus einem im WWW verfügbaren Programm zur Konjugation spanischer Verben.[4]

Konjunktionen, Präpositionen, Pronomina, Interjektionen, Akronyme und Affixe, die in den verwendeten Grammatiken aufgeführt sind, werden nachgetragen, falls sie im DRAE und im DGLE nicht angegeben werden. Ergänzende Angaben, insbesondere zu Affixen, stammen aus [BosqueDemonte 1999].

Zusätzlich wurden Zahlen, Satzzeichen und Maßeinheiten in das Lexikon aufgenommen. Wörter, die sowohl als Adjektiv als auch als Substantiv verwendet werden können und anfangs jeweils in der Adjektiv- und in der Substantiv-Datei aufgeführt waren, sind in der Datei der Substantive zusammengefasst und werden aus der Adjektiv-Datei entfernt.

Durch die beschriebenen Ergänzungen und Korrekturen beträgt die Gesamtzahl der Lexikoneinträge 98 546. Tabelle 6.2 zeigt die Verteilung auf einzelne Wortklassen.

Beide als Ressource verwendeten Lexika geben wohlgeformte Wörter des Spanischen an. Darunter sind viele, die durch Derivation oder Komposition aus anderen Einträgen entstanden sind. Da die SMM auch die Prozesse der Derivation und Komposition umfasst, ist es nicht notwendig, Derivata und Komposita in das Grundformlexikon zu übernehmen. Alle Lexikoneinträge, die durch die SMM als Derivat oder Kompositum mit der gleichen Kategorie wie der entsprechende Eintrag analysiert sind, werden aus dem Lexikon entfernt. Dies betrifft Verben, Adjektive, Substantive und Adverben. Es resultiert ein Lexikon mit 66 103 Einträgen.

[4] [Comp-jugador].

Wortklasse	Anzahl	Datei
Substantive&Adjektive	57 882	`substantive.lex`
Adjektive	21 867	`adjective.lex`
Verben	12 827	`verb.lex`
wc Adverben	2 517	`adverb.lex`
Konjunktionen	26	`conjunction.lex`
Präpositionen	30	`preposition.lex`
Artikel, Pronomina	92	`determiner.lex`
Zahlen, Maßeinheiten	117	`numbers.lex`
Interjektionen	317	`interjection.lex`
Namen	1030	`name.lex`
Akronyme	537	`acronym.lex`
Affixe	1130	`affix.lex`
Flexionsmorpheme	126	`inflection.lex`
Satzzeichen	26	`punctuation.lex`
vollständige Paradigmen	12	`paradigm.lex`

Tabelle 6.2: Verteilung der Lexikoneinträge im vollständigen Lexikon

Die Verhältnisse der Lexika mit einfachen Einträgen für Verben, Adjektive, Substantive und Adverben zu den Lexika, die neben einfachen Einträgen auch Derivata und Komposita enthalten, zeigt Tabelle 6.3.

Wortklasse	vollständig	reduziert	Derivata/Komposita
Verben	12 787	9 026	3 760
Adjektive	21 867	19 618	2 248
Substantive	57 882	32 320	25 561
Adverben	2 517	1 506	1 011

Tabelle 6.3: Vollständige und reduzierte Lexika der Verben, Adjektive, Substantive und Adverben

Das reduzierte Grundformlexikon nimmt weniger Platz ein als das vollständige Grundformlexikon. Es werden absolut weniger Allomorphe erzeugt, so dass das Allomorphlexikon des reduzierten Lexikons weniger Platz benötigt als das des kompletten Lexikons. Werden Wortformen analysiert, die im kompletten Lexikon, nicht aber im reduzierten als Lexem vorhanden sind, reduziert sich die Anzahl der Analyseergebnisse, wenn das reduzierte Lexikon als Grundformlexikon verwendet wird. Ob die ermittelte Struktur der Derivata korrekt ist, muss für jede einzelne Wortform entschieden werden. Die automatische Wortformerkennung zeigt hier Möglichkeiten auf, die in einzelnen Fällen aus semantischen Gründen verworfen werden müssen.

So wird das Substantiv *abrazo* auch als Derivat aus dem Präfix *a-* und dem Substantiv *brazo* analysiert. Dabei handelt es sich bei *abrazo* um eine substantivierte Handlung (Umarmung), während *brazo* ein reines Substantiv ist (Arm). Für das Präfix *a-* sind als Bedeutung Richtung, Entgegengesetztheit, Entzug oder Verursa-

chung möglich.[5] Keine dieser Möglichkeiten in Kombination mit der Bedeutung von *brazo* führt aber zur Bedeutung von *abrazo*. Somit muss die Analyse als Derivat aus Präfix und Substantiv abgelehnt werden.

Die Entfernung der als Derivata oder Komposita kategorisierten Verben, Adjektive, Substantive und Adverben erfolgt automatisch, da die manuelle Behandlung aller Analysen der 95 053 Lexikoneinträge der Verben, Adjektive, Substantive und Adverben im Rahmen dieser Arbeit nicht zu realisieren ist. Es werden daher auch Einträge aus dem Lexikon entfernt, deren Analyse als Derivat oder Kompositum nicht korrekt ist. Wird das reduzierte Lexikon bei der Analyse einer der zugehörigen Wortformen eines solchen Eintrags wie *abrazo* verwendet, ist die Analyse nicht korrekt.

Generell kann die Analyse von Wortformen als Derivata oder Komposita nicht verworfen werden, auch wenn sie Ergebnisse liefert, die der Struktur des Spanischen nicht entsprechen. Die vorgeschlagenen Segmentierungen können auf verschiedene Weisen genutzt werden. So kann es sich um gültige Prozesse von Derivation oder Komposition handeln, oder um Prozesse, die heute nicht mehr gültig sind, aber etymologisch wirksam waren. Um dies entscheiden zu können, ist es sinnvoller, zum Parsen von Wortformen das vollständige Lexikon zu verwenden. Wird eine Wortform dann nur als Derivat oder Kompositum analysiert, ist dies die endgültige Analyse. Ein Ergebnis, das die Möglichkeiten der Derivation oder Komposition wie die Analyse als Lexem offen lässt, ist in der Analyse als Lexem in jedem Fall korrekt. Die Analyse als Derivat oder Kompositum muss unter Vorbehalt akzeptiert werden. Wird das reduzierte Lexikon verwendet, kann die Analyse derselben Wortform als boßes Derivat oder Kompositum nicht als endgültige Analyse betrachtet werden, wie am Beispiel *abrazo* gezeigt.

HAUSSER schlägt als Lösung die Verwendung beider Lexika vor. Im Abgleich der Ergebnisse, die jeweils erzeugt werden, lassen sich dann endgültige Analyse-Ergebnisse ermitteln, die unnötige Ambiguität vermeiden.[6]

6.2.3 Das Format der Lexikoneinträge

Jeder Lexikoneintrag ist in eckige Klammern eingeschlossen, es handelt sich um einen Verbund. Die einzelnen Lexikoneinträge werden durch Semikolon voneinander getrennt. Ein neuer Eintrag beginnt in einer neuen Zeile. Die Elemente des Verbundes sind Attribut-Werte-Paare der Form `attribut: wert`. Die einzelnen Paare werden durch Komma voneinander getrennt.

Für die Lemmata der einzelnen Lexikondateien sind jeweils verschiedene Angaben zur Erzeugung der Allomorphe notwendig. Allen Einträgen gemeinsam ist die Angabe der Oberfläche im Attribut `Lemma` und die Angabe der Wortklasse im Attribut `POS`. Hinsichtlich weiterer Attribut-Werte-Paare unterscheiden sich die Ein-

[5] Vgl. [BosqueDemonte 1999, S. 5036].
[6] [Hausser 1999, S. 267].

träge.

Die Attribute, die ein Lexikoneintrag aufweist, hängen davon ab, welche Informationen einerseits im Ergebnis der Analyse erscheinen sollen und welche Informationen andererseits für die Erzeugung der Allomorphe und für deren Konkatenation benötigt werden. Attribute, die erst durch die Allomorphregeln erzeugt werden, sind in 6.3 beschrieben. Im Folgenden wird die Struktur der Lexikoneinträge und der daraus ableitbaren Informationen an repräsentativen Beispielen für die einzelnen Lexikondateien erläutert.

Substantive

```
[Lemma: "álgebra",
 POS: Substantive,
 Gender: Feminin,
 Group: 1];

[Lemma: "cartógrafo",
 POS: Substantive,
 Group: 23];

[Lemma: "agricultor",
 POS: Substantive,
 Group: 23];

[Lemma: "baobab",
 POS: Substantive,
 Gender: Masculin,
 Group: 1,
 PluralMark: "baobabs"];
```

Für alle Substantive werden im Lexikon die Oberfläche und die Wortklasse `Substantive` angegeben. Kann ein Lexem syntaktisch sowohl als Substantiv als auch als Adjektiv funktionieren, wird ihm die Wortklasse `Substantive|Adjective` zugeordnet. Dabei handelt es sich um ein Multisymbol, das die Symbole `Substantive` und `Adjective` enthält. Entsprechende Lemmata sind in der Datei der Adjektive nicht aufgeführt.

Wie in 3.4.2 begründet, werden zur Flexion der Substantive drei Gruppen unterschieden. Die Zugehörigkeit wird im Attribut `Group` vermerkt. Zur ersten Gruppe gehören alle Substantive, die nur in einem Genus vorkommen (siehe Eintrag für *álgebra*, *baobab)*. Für diese ist im Attribut `Gender` das Genus angegeben. Substantive, die in beiden Genera vorkommen, lassen sich hinsichtlich deren Bildung unterscheiden. Zur zweiten Gruppe gehören alle Substantive, die im Maskulinum das Genusallomorph *o* oder *e* und im Femininum das Genusallomorph *a* tragen (siehe Eintrag für *cartógrafo*). Zur dritten Gruppe gehören alle Substantive, die im Maskulinum kein Genusallomorph und im Femininum das Genusallomorph *a* aufweisen (siehe Eintrag für *agricultor*). Substantive, die zur zweiten oder dritten Gruppe gehören, erhalten für das Attribut `Group` den Wert `23`. Die Oberfläche des Lemmas ist in diesem Fall jeweils die maskuline Form des Substantivs, eine ent-

sprechende Angabe kann entfallen. Die Art der Bildung des Maskulinums kann aus der Oberfläche abgeleitet werden, sodass die zweite und dritte Gruppe nicht gesondert markiert werden muss.

Alle Substantive sind im Singular angegeben, eine entsprechende Markierung ist nicht notwendig. Die Bildung des Plurals erfolgt wie in 3.4.1 beschrieben. Weicht die Pluralform von den dort angegebenen Regeln ab, enthält das Lemma das Attribut `PluralMark`, in dem die Pluralform des Substantivs angegeben ist (siehe Eintrag für *baobab*). Dies ist nur für Substantive der ersten Gruppe möglich.

Adjektive

```
[Lemma: "dialectal",
 POS: Adjective,
 Group: 1];

[Lemma: "dialéctico",
 POS: Adjective,
 Group: 23];

[Lemma: "adjutor",
 POS: Adjective,
 Group: 23];

[Lemma: "bíceps",
 POS: Adjective,
 Group: 1,
 PluralMark: "bíceps"];
```

Für Adjektive werden die Oberfläche und die Wortklasse `Adjective` angegeben. Alle Einträge haben die Oberfläche der maskulinen Form im Singular, diese Angaben können daher entfallen. Die Lexikoneinträge weisen ein zusätzliches Attribut auf, das entsprechend 3.4.2 die Zugehörigkeit zu den drei Gruppen der Genusbildung angibt.

Adjektive mit der Angabe `Group: 1` weisen im Maskulinum und Femininum die gleiche Oberfläche auf (siehe Eintrag für *dialectal*). Adjektive mit der Angabe `Group: 23` gehören wie die Substantive mit der entsprechenden Angabe zur zweiten oder dritten Gruppe. Sie weisen im Femininum das Genusallomorph *a* auf. Adjektive der zweiten Gruppe bilden das Maskulinum mit dem Genusallomorph *o* bzw. *e* (siehe Eintrag für *dialéctico*), Adjektive der dritten Gruppe tragen im Maskulinum kein Genusallomorph (siehe Eintrag für *adjutor*).

Erfolgt die Pluralbildung nicht entsprechend den Regeln in 3.4.1, enthält das Lemma das Attribut `PluralMark` mit der Oberfläche des Plurals (siehe Eintrag für *bíceps*). Dies ist nur für Adjektive der ersten Gruppe möglich.

Die Adjektive *malo, bueno, grande, pequeño, santo* sind wegen der unregelmäßigen Komparation bzw. der Möglichkeit der Bildung von Apokopen im Paradigmenlexikon aufgeführt.

Verben

```
[Lemma: "comer",
 POS: Verb,
 Valencies: <Reflexive,
             Intransitive,
             Transitive>];

[Lemma: "ababillarse",
 POS: Verb,
 Valencies: <Reflexive>];

[Lemma: "huir",
 POS: Verb,
 Valencies: <Reflexive,
             Intransitive>];

[Lemma: "volver",
 POS: Verb,
 Valencies: <Reflexive,
             Intransitive>,
 Participle: <"vuelto">];

[Lemma: "romper",
 POS: Verb,
 Valencies: <Reflexive,
             Intransitive>,
 Participle: <"rompido",
              "roto">];

[Lemma: "abuñolar",
 POS: Verb,
 Valencies: <Transitive>,
 AlloMark: "abuñ{o}l",
 AlloForm: Allo_Norm_oue1];

[Lemma: "decir",
 POS: Verb,
 Valencies: <Reflexive,
             Transitive>,
 AlloMark: "d{ec}",
 AlloForm: Allo_Norm_ecir1,
 P_imp_Sg2: <"di">,
 Participle: <"dicho">];
```

Für Verben wird als Oberfläche des Lemmas der Infinitiv des jeweiligen Verbs gewählt. Jeder Lexikoneintrag enthält die Angabe `Verb` zur Wortklasse und ein Attribut `Valencies`. Dort werden die aus dem DRAE und dem DGLE entnommenen Angaben zum transitiven oder intransitiven Gebrauch aufgeführt. Verben, die eine reflexive Infinitiv-Endung aufweisen, erhalten die Angabe `Reflexive` (siehe Eintrag für *ababillarse*). Verben, die im Infinitiv nicht als reflexiv gekennzeichnet sind, können reflexiv gebraucht werden. Die Valenzliste wird um den entsprechenden Eintrag ergänzt (siehe Eintrag für *comer*).

Reguläre und semi-reguläre Verben weisen keine weiteren Attribute auf (siehe Eintrag für *comer, huir*). Für semi-irreguläre Verben enthält der Lexikoneintrag das Attribut `AlloMark`, dessen Wert die markierte Oberfläche des Verbs ist. Die Buch-

staben, die für die jeweiligen Allomorphe verändert werden, sind in geschweifte Klammern eingeschlossen (siehe Eintrag für *abuñolar, decir*). Das Attribut `Allo-Form` enthält ein Symbol, über das innerhalb der Allomorphregeln die Oberfläche der Allomorphe ermittelt wird (siehe Eintrag für *abuñolar, decir*).

Bei Verben mit unregelmäßiger Partizip-Bildung ist im Attribut `Participle` die Oberfläche des Partizips angegeben (siehe Eintrag für *volver, decir*). Gelten für ein Verb regelmäßige und unregelmäßige Partizip-Bildung als wohlgeformt, sind in diesem Attribut beide Oberflächen angegeben (siehe Eintrag für *romper*).

Verben, für die die Form der zweiten Person Singular des affirmativ gebrauchten Imperativs nicht regelmäßig ist, enthalten im Attribut `P_imp_Sg2` die entsprechende Oberfläche (siehe Eintrag für *decir*).

Die irregulären Verben *ser, estar, haber, dar, ver, saber, ir* sind im Paradigmenlexikon aufgeführt.

Adverben

```
[Lemma: "alrededor",
 POS: Adverb,
 AdverbType: Local|Modal];
```

Für Adverben ist neben der Oberfläche und der Wortklasse `Adverb` die semantische Funktion des Adverbs im Attribut `AdverbType` angegeben (siehe Eintrag für *alrededor*). Als Wert kann `Local, Modal, Temporal` oder ein Multisymbol aus diesen angegeben werden.

Konjunktionen

```
[Lemma: "como",
 POS: Conjunction,
 ConjunctionType: Temporal|Final|Causal];

[Lemma: "o",
 POS: Conjunction,
 ConjunctionType: Disjunctive,
 Allo: <[Surface: "o"],
        [Surface: "u",
         StartNextWord: "/o/"]> ];
```

Für Konjunktionen werden die Oberfläche, die Wortklasse `Conjunction` und die semantische Funktion im Attribut `ConjunctionType` angegeben. Als semantische Funktion sind `Coordination`, `Continuative`, `Distributive`, `Adversative/`, `Consecutive`, `Temporal`, `Final`, `Causal`, `Ilative`, `Copulative`, `Disjunctive`, `Condicional`, `Comparative` oder Multisymbole aus diesen möglich (siehe Eintrag für *como*).

Alle Konjunktionen sind durch die drei genannten Attribute ausreichend gekennzeichnet, da sie unveränderlich sind und jedem Morphem nur ein Allomorph zugeordnet ist. Eine Ausnahme bilden die Konjunktionen *o* und *y*, wie in 3.7 dargestellt. Es wird jeweils das Attribut `Allo` hinzugefügt. Der zugehörige Wert ist eine Liste von Verbunden. Jeder Verbund enthält die Informationen, wodurch sich die einzelnen Allomorphe voneinander unterscheiden. Zum einen ist dies die Oberfläche, die im Attribut `Surface` vermerkt ist. Für das Allomorph, dessen Oberfläche nicht mit der Oberfläche des Morphems identisch ist, wird die Bedingung angegeben, unter der es verwendet wird. Im Attribut `StartNextWord` ist das Anfangsphonem der folgenden Wortform angeführt. Diese Information ist insbesondere für die syntaktische Analyse wichtig.

Präpositionen

```
[Lemma: "a",
 POS: Preposition,
 PrepositionType: Directional|Temporal|Modal|Quantity|Causal|Instrumental];

[Lemma: "al",
 POS: Preposition&Determiner,
 Preposition: [Surface: "a",
               PrepositionType:
                 Directional|Temporal|Modal|Quantity|Causal|Instrumental],
 Determiner: [Surface: "el",
              Gender: Masculin,
              Number: Singular]];
```

Für Präpositionen werden die Oberfläche, die Wortklasse `Preposition` und im Attribut `PrepositionType` die möglichen Bedeutungen angegeben. Für den Präpositionstyp sind `Directional`, `Temporal`, `Modal`, `Quantity`, `Causal`, `Instrumental`, `Final`, `Proximity` oder Multisymbole aus diesen möglich (siehe Eintrag für *a*).

Die Wortformen *al* und *del* sind aus der Verschmelzung einer Präposition mit dem bestimmten Artikel in der maskulinen Singularform entstanden. Folgt auf die Präposition *a* der Artikel *el*, werden beide Wortformen durch die Wortform *al* ersetzt. Folgt auf die Präposition *de* der Artikel *el*, werden beide Wortformen durch die Wortform *del* ersetzt. Für beide Lemmata ist die Wortklasse als `Preposition&Determiner` angegeben. Dabei handelt es sich nicht um ein Multisymbol, sondern um den Hinweis, dass Elemente beider Wortklassen in einer Wortform enthalten sind. Im Attribut `Preposition` sind als Verbund die Informationen zur Präposition enthalten. Angegeben werden die Oberfläche sowie der Präpositionstyp. Im Attribut `Determiner` sind die Informationenen zum Artikel enthalten. vermerkt werden die Oberfläche, Genus und Numerus (Siehe Eintrag für *al*).

Artikel und Pronomina

```
[Lemma: "el",
 POS: Determiner,
 Gender: Masculin,
 Number: Singular]];

[Lemma: "se",
 POS: Pronoun,
 PronounType: PersonalAtonic,
 Pronoun: [Person&Number: Sg3|Pl3,
           Gender: Masculin|Feminin,
           For: <"les", "le", "la", "Usted", "Ustedes">,
           NextWord: <"le", "la", "lo", "las", "los">],
 Suc: <<<POS, Pronoun>, <PronounType, PersonalAtonic>>>,
 Pre: <<<POS, Verb>>>,
 SucFon: l];

[Lemma: "sí",
 POS: Pronoun,
 PronounType: Personal,
 Pronoun: [Person&Number: Sg3|Pl3,
           Gender: Masculin|Feminin,
           LastWord: Preposition]];

[Lemma: "algo",
 POS: Pronoun|Adverb,
 Adverb: [AdverbType: Modal],
 Pronoun: [PronounType: Indefinit]];
```

Artikel, Pronomina sowie Wortformen, die sowohl als Pronomen als auch als Adverb verwendet werden können, sind in einer Datei enthalten.

Für Artikel werden die Oberfläche, die Wortklasse `Determiner`, der Genus und der Numerus angegeben (siehe Eintrag für *el*).

Für Pronomina werden die Oberfläche, die Wortklasse `Pronoun` und im Attribut `Pronountype` die Art des Pronomens angegeben. Als Wert kann `PersonalTonic`, `PersonalAtonic`, `Indefinit`, `Possessive`, `Interrogative`, `Indefinit`, `Relative`, `Demonstrative`, `Local` oder ein Multisymbol aus diesen angegeben werden. Im Attribut `Pronoun` werden in einem Verbund Angaben zu Genus und Numerus aufgeführt. Für Personalpronomina handelt es sich dabei um Angaben zu Person und Numerus sowie Genus. Sind an eine Wortform syntaktische Bestimmungen für vorhergehende oder nachfolgende Wortformen gebunden, sind diese in den Attributen `LastWord` und `NextWord` angegeben (siehe Einträge für *se, sí*). Wird ein betontes Personalpronomen aufgrund nachfolgender enklitischer Pronomina anstelle eines anderen Pronomen verwendet, sind im Attribut `For` in einer Liste die ersetzten Pronomina aufgeführt. Die Bedingung für nachfolgende enklitische Pronomina ist in den Attributen `NextWord` `Suc` und `SucFon` angegeben. Letzteres gibt an, mit welchem Phonem bzw. Buchstaben folgende Allomorphe beginnen. Nachfolgen dürfen nur enklitische Pronomina. Die Bedingung für das vorausgehende Allomorph ist im Attribut `Pre` festgehalten, es sind nur Verben möglich (siehe Eintrag für *sí*). Vorausgehen können auch Verbformen, die bereits enklitische Pronomina aufweisen. Diese sind ebenfalls der Wortklasse Verb zuge-

ordnet. Die Angabe, dass vorhergehende Allomorphe Verben sein müssen, genügt daher.

Für Wortformen, die sowohl Adverb als auch Pronomen sind, werden die Oberfläche und als Wortklasse das Multisymbol `Pronoun|Adverb` angegeben. Für jede Wortklasse werden in einem gleichnamigen Attribut in einem Verbund die entsprechenden Merkmale vermerkt. Für Adverben ist dies der Adverbtyp, für Pronomen mindestens der Pronomentyp und eventuell Genus und Numerus (siehe Eintrag für *algo*).

Ziffern, Numerale und Maßeinheiten

```
[Lemma: "1",
 POS: Number,
 numericValue: 1];

[Lemma: "cero",
 POS: Numeral,
 NumericType: Cardinal];

[Lemma: "uno",
 POS: Numeral,
 NumericType: Cardinal,
 Allo: <[Surface: "uno"],
        [Surface: "una"]>];

[Lemma: "primero",
 POS: Numeral,
 NumericType: Ordinal,
 Allo: <[Surface: "primero"],
        [Surface: "primera"],
        [Surface: "primer",
         NextWord: MasculinSubstantive]>];

[Lemma: "kg",
 POS: Dimension,
 Long: "kilogramo",
 DimensionType: Mass];
```

In der Lexikondatei `numeral.lex` sind arabische Ziffern, Numerale und Maßeinheiten aufgeführt.

Für arabische Ziffern werden die Oberfläche, die Wortklasse `Number` und im Attribut `numericValue` die Ziffer angegeben (siehe Eintrag für *1*).

Für Numerale werden die Oberfläche, die Wortklasse `Numeral` und das Attribut `NumericType` vermerkt. Dieses hat für Kardinalzahlen den Wert `Cardinal` (siehe Eintrag für *cero, uno*) und für Ordinalzahlen den Wert `Ordinal` (siehe Eintrag für *primero*). Sind für ein Numeral mehrere Oberflächen möglich, enthält der Eintrag das Attribut `Allo` dem als Wert eine Liste von Verbunden zugeordnet ist. Jeder Verbund enthält mindestens das Attribut `Surface` mit der Oberfläche des Allomorphs als Wert (siehe Eintrag für *uno*). Für die Ordinalzahlen *primero* und *tercero* ist, wie in 3.4.5 dargestellt, vor maskulinen Substantiven eine apokopierte Form möglich. Dieses Allomorph wird ebenfalls innerhalb eines Verbundes der

Liste für `Allo` beschrieben. Neben der Oberfläche ist die Bedingung der Verwendung im Attribut `NextWord` mit dem Wert `MasculinSubstantive` angegeben (siehe Eintrag für *primero*).

Für Maßeinheiten werden als Oberfläche das entsprechende Akronym und die Wortklasse `Dimension` angegeben. Im Attribut `Long` wird die Langform der Maßeinheit angegeben. Im Attribut `DimensionType` wird die Zugehörigkeit der Maßeinheit zu einer der Angaben `Energy`, `Mass`, `Length`, `Strength`, `Volume`, `Temperature`, `Area` vermerkt (siehe Eintrag für *kg*).

Interjektionen

```
[Lemma: "¡anda!",
 POS: Interjection];
```

Für Interjektionen werden die Oberfläche und die Wortklasse `Interjection` angegeben (siehe Eintrag für *¡anda!*). Interjektionen sind Wortformen, die durch die Satzzeichen für Beginn und Ende einer Exklamation eingeschlossen sind.

Namen

```
[Lemma: "Jesús",
 POS: Name,
 NameType: FirstName,
 Gender: Masculin];

[Lemma: "México",
 POS: Name,
 NameType: Geographic];
```

Für Namen werden die Oberfläche und die Wortklasse `Name` angegeben. Im Attribut `NameType` erfolgt die Unterscheidung nach geographischen Namen, biologischen Namen, Namen für Personen (Vorname oder Nachname) und Namen für Institutionen (siehe Eintrag für *Jesús, México*). Handelt es sich um Vornamen von Personen, wird im Attribut `Gender` das Genus vermerkt (siehe Eintrag für *Jesús*).

Akronyme

```
[Lemma: "Ar",
 POS: Acronym,
 Long: "Argón"];
```

Für Akronyme werden die Oberfläche, die Wortklasse `Acronym` und, soweit bekannt, im Attribut `Long` die Langform angegeben (siehe Eintrag für *Ar*).

Affixe

```
[Lemma: "ab",
 POS: Prefix,
 Allo: <[Surface: "ab"],
        [Surface: "abs"]>,
 PrefixSemType: Local];

[Lemma: "ante",
 POS: Prefix,
 PrefixSynType: Preposition,
 PrefixSemType: Temporal|Local];

[Lemma: "al",
 POS: Interfix,
 SuffixMark: <"ache", "ada", "era", "eta", "ías", "ón", "uta">];

[Lemma: "ías",
 POS: Suffix,
 InterfixMark: <"al">,
 FinalPOS: Substantive];

[Lemma: "ito",
 POS: Suffix,
 InterfixMark: <"ec", "c", "irr", "usqu">,
 FinalPOS: Adjective,
 Allo: <[Surface: "ito",
         FinalGender: Masculin],
        [Surface: "ita",
         FinalGender: Feminin]>,
 SuffixSemType: Diminutive];
```

In der Lexikondatei `affix.lex` sind Präfixe, Interfixe und Suffixe aufgeführt.

Für Präfixe werden die Oberfläche, die Wortklasse `Prefix` und, soweit bekannt, die syntaktische Funktion im Attribut `PrefixSyntype` und die semantische Funktion im Attribut `PrefixSemType` angegeben (siehe Eintrag für *ab*, *ante*). Syntaktisch kann nach adverbialer Funktion und präpositionaler Funktion unterschieden werden. Semantisch kann nach temporaler, lokaler, modaler, gradativer, diathetischer und negierender Funktion unterschieden werden. Für die semantische Funktion sind Mehrfach-Bedeutungen möglich, für die entsprechende Multisymbole definiert sind (siehe Eintrag für *ab, ante*).

Für Interfixe werden die Oberfläche und die Wortklasse `Interfix` angegeben. Gelten für nachfolgende Suffixe Einschränkungen, sind im Attribut `SuffixMark` in einer Liste die Oberflächen der Suffix-Morpheme aufgeführt (siehe Eintrag für *al*).

Für Suffixe werden die Oberfläche und die Wortklasse `Suffix` angegeben. Gelten für vorausgehende Interfixe Einschränkungen, sind im Attribut `InterfixMark` in einer Liste die Oberflächen dieser Interfix-Morpheme aufgeführt (siehe Eintrag für *ías, ito*). Soweit bekannt, werden im Attribut `SuffixSemType` semantische Funktionen des Suffixes angegeben (siehe Eintrag für *ito*). Dabei wird nach Diminutiv-, Augmentativ- und Pejorativ-Suffixen unterschieden. Die Wortklasse eines resultierenden Derivats wird im Attribut `FinalPOS` angegeben. Handelt es sich dabei

um Substantive oder Adjektive, wird im Attribut `FinalGender` das resultierende Genus angegeben. Sind für ein Lemma mehrere Allomorphe möglich, werden die Oberflächen mit den sie unterscheidenden Attributen als Verbunde in einer Liste für das Attribut `Allo` aufgeführt.

Für alle Interfixe, die in der Interfix-Liste eines Suffixes enthalten sind, ist das Suffix in deren Suffix-Liste enthalten. Umgekehrt gilt, dass für alle Suffixe, die in der Suffix-Liste eines Interfixes aufgeführt sind, das Interfix in deren Interfix-Liste enthalten ist.

Flexionsmorpheme

```
[Lemma: "ábamos",
 POS: VerbInflection,
 Structure: <TV&MT&PN>,
 Themevocal: a,
 Category: <[Tense: Imperfect,
             Mood: Indicative,
             Person&Number: Pl1]>,
 Tempus: <Imp_ind_Pl1>];

[Lemma: "emos",
 POS: VerbInflection,
 Structure: <TV&PN>,
 Themevocal: a,
 Category: <[Tense:  Present,
             Mood: Subjunctive,
             Person&Number: Pl1],
            [Tense: Present,
             Mood: Imperative,
             Person&Number: Pl1,
             Sense: Affirmative|Negative]>,
 Tempus: <P_sub_Pl1, P_imp_Pl1>,
 Allo: <[Surface: "emos"],
        [Surface: "émos",
         Suc: <<<POS, Pronoun>, <PronounType, PersonalAtonic>>>,
         Allo_i: <encl2>],
        [Surface: "émo",
         Suc: <<<POS, Pronoun>, <PronounType, PersonalAtonic>, <Person&Number>>>,
         Allo_i: <encl2>]>];

[Lemma: "o",
 POS: NounInflection,
 InflectionType: Gender,
 FinalNumber: Singular,
 Allo: <[Surface: "o",
         FinalGender: Masculin,
         NounFlex: <o|a>],
        [Surface: "e",
         FinalGender: Masculin,
         NounFlex: <e|a>],
        [Surface: "a",
         FinalGender: Feminin,
         NounFlex: <o|a, e|a, c|a>]>];
```

```
[Lemma: "s",
 POS: NounInflection,
 InflectionType: Number,
 FinalNumber: Plural,
 Allo: <[Surface: "s",
         PluralMark: s],
        [Surface: "es",
         PluralMark: es]>];

[Lemma: "ísim",
 POS: NounInflection,
 InflectionType: Comparation];
```

In der Datei `inflection.lex` sind Flexionsmorpheme für die Flexion der Verben, Substantive und Adjektive enthalten.

Zunächst war beabsichtigt, die in den Tabellen 3.10 - 3.19 ermittelten Morpheme für Themavokal, Modus/Tempus und Person/Numerus getrennt zu behandeln. Wie dort gezeigt, entfallen in einigen Formen die Kennzeichnungen von Modus/Tempus oder Person/Numerus. Wegen der Unhandlichkeit dieses Ansatzes wurde entschieden, die drei Bestandteile, die dem Stamm eines Verbes folgen können, als Flexionsendung zusammenzufassen. Daraus wurden 120 Flexionsmorpheme abgeleitet.

Für jedes Verb-Flexionsmorphem sind im Lexikoneintrag die Oberfläche, die Wortklasse `VerbInflection` und im Attribut `Structure` die Struktur des Morphems angegeben. Die Struktur gibt an, ob es sich um Kombinationen von Themavokal, Modus/Tempus-Morphem und Person/Numerus-Morphem oder von Themavokal und Modus/Tempus-Morphem oder von Themavokal und Person/Numerus-Morphem oder von Themavokal und Auxiliar (im Fall von Infinitiv, Gerundium und Partizip) handelt. Im Attribut `Themevocal` wird festgehalten, zu welcher der drei Konjugationen das Flexionsmorphem gehört. Dabei steht `a` für die erste Konjugation, `e` für die zweite Konjugation, `i` für die dritte Konjugation und `ei` dann, wenn sowohl die zweite als auch die dritte Konjugation dieses Flexionsmorphem benötigen.

Im Attribut `Category` werden in einer Liste Verbunde aufgeführt, die die grammatikalische Kategorie angeben, die durch die Verbindung eines Verbstamms mit dem Flexionsmorphem erreicht wird. Im Attribut `Tense` wird das Tempus angegeben. Es wird unterschieden nach `Present`, `Imperfect`, `Indefinite`, `FutureImperfect` und `PotencialSimple`. Im Attribut `Mood` wird der Modus angegeben. Es wird unterschieden nach `Indicative`, `Subjunctive` und `Imperative`. Im Attribut `Person&Number` werden Numerus und Person angegeben. Dabei steht `Sg` gefolgt von einer Ziffer zwischen `1` und `3` für eine der drei Personen im Singular. `Pl` gefolgt von einer Ziffer zwischen `1` und `3` steht für eine der drei Personen im Plural. Im Attribut `Tempus` wird in einer Liste für jeden Verbund des Attributs `Category` ein Symbol angegeben, das für die Kombination aus Tempus, Modus, Person und Numerus eindeutig ist (siehe Eintrag für *ábamos*).

Weist das Morphem mehrere Allomorphe mit unterschiedlicher Oberfläche auf, sind diese im Attribut `Allo` beschrieben. Wie in 3.6.3 begründet, werden auch

Allographen als Allomorphe behandelt. Liegt für eine wohlgeformte Verbform in affirmativen Imperativformen die Betonung im Bereich des Flexionsmorphems, ohne dass ein expliziter Akzent gesetzt ist, kann der Anschluss enklitischer Pronomina die explizite Setzung eines Akzentes erfordern. Für Flexionsmorpheme, die affirmative Imperativformen kodieren und keinen Akzent tragen, sind daher zwei Allomorphe zu erzeugen: Eins ohne Akzent, eins mit Akzent auf dem betonten Vokal. Um die Verwendung beider Formen zu kontrollieren, erhält der Verbund, der das akzentuierte Allomorph beschreibt, zwei weitere Attribute. Dem Attribut `Suc` ist der Verweis, dass nur ein Pronomen aus der Klasse der unbetonten Personalpronomen folgen kann und auch folgen muss, zugeordnet. Im Attribut `Allo_i` wird in einer Liste festgehalten, dass es sich um das Allomorph handelt, auf das enklitische Pronomina folgen müssen (siehe Eintrag für *emos*). Der Wert dieses Attributs wird zur Steuerung der Konkatenation der Allomorphe benötigt.Sind entsprechend der in 3.5.5 angegebenen Regeln veränderte Oberflächen der Flexionsendung vor bestimmten enklitischen Pronomen erforderlich, wird dafür ein zusätzliches Allomorph erzeugt (siehe Eintrag für *emos*).

Die Flexion der Substantive und Adjektive benötigt zwei Flexionsmorpheme. Ein Genusmorphem und ein Numerusmorphem. Für das Genusmorphem ist als Oberfläche die Oberfläche des Genusallomorphs *o* gewählt, als Wortklasse ist `NounInflection` angegeben. Das Attribut `Inflectiontype` erhält den Wert Gender. Der Singular ist für spanische Substantive und Adjektive nicht markiert, eine Wortform, die nur ein Genusallomorph aufweist, ist daher immer im Singular. Das Attribut `FinalNumber` erhält den Wert `Singular`. Im Attribut `Allo` sind in einer Liste die drei Allomorphe als Verbunde mit den Attributen `Surface`, `FinalGender`, `NounFlex` aufgeführt. Damit wird das resultierende Genus festgelegt sowie kodiert, für welche der drei Flexionsgruppen das Allomorph verwendet werden kann. Die Symbole `o|a` und `e|a` geben an, dass das Allomorph für Substantive und Adjektive der zweite Gruppe genutzt wird. Das Symbol `c|a` gibt an, dass das Allomorph für Substantive und Adjektive der dritten Gruppe verwendet wird (siehe Eintrag für *o*).

Für das Numerusmorphem sind Oberfläche und als Wortklasse `NounInflection` angegeben. Das Attribut `InflectionType` erhält den Wert `Numerus`, der resultierende Numerus ist im Attribut `FinalNumber` als Plural vermerkt. Die beiden Allomorphe *s* und *es* sind als Verbunde in einer Liste des Attributs `Allo` angeführt (siehe Eintrag für *s*). Die jeweilige Markierung im Attribut `PluralMark` dient der Steuerung der Konkatenation von Substantiv bzw. Adjektiv und Numerusallomorph.

Wie in 3.4.2 begründet, wird die Komparation als Flexion behandelt. Daher erhält das entsprechende Lemma ebenfalls die Wortklasse `NounInflection`. Das Attribut `InflectionType` hat den Wert `Comparation` (siehe Eintrag für *ísim*).

Satzzeichen

```
[Lemma: "¿",
 POS: Punctuation,
 PunctuationType: BeginQuestionMark];
```

Für Satzzeichen werden die Oberfläche, als Wortklasse `Punctuation` und im Attribut `PunctuationType` eine Beschreibung des Zeichens angegeben (siehe Eintrag für ¿) .

Angabe vollständiger Paradigmen

```
[Lemma: "grande",
 POS: Paradigm,
 ParadigmPOS: Adjective,
 Paradigm: <[Surface: "grande",
             Gender: Masculin|Feminin,
             Number: Singular,
             Comparation: Positive,
             Suc: <<<POS, Substantive|Adjective>>,
                   <<POS, Verb>>>,
             Pre: <<<POS, Substantive|Adjective>>,
                   <<POS, Adverb>>>],
            [Surface: "grande",
             Gender: Masculin|Feminin,
             Number: Singular,
             NextWord: Substantive,
             Comparation: Positive,
             Suc: <<<POS, Substantive|Adjective>>,
                   <<POS, Verb>>>,
             Pre: <<<POS, Substantive|Adjective>>,
                   <<POS, Adverb>>>],
            [Surface: "grandes",
             Gender: Masculin|Feminin,
             Number: Plural,
             Comparation: Positive,
             Suc: <<<POS, Substantive|Adjective>>,
                   <<POS, Verb>>>,
             Pre: <<<POS, Substantive|Adjective>>,
                   <<POS, Adverb>>>],
            [Surface: "mayor",
             Gender: Masculin|Feminin,
             Number: Singular,
             Comparation: Comparative,
             Suc: <<<POS, Substantive|Adjective>>,
                   <<POS, Verb>>>,
             Pre: <<<POS, Substantive|Adjective>>,
                   <<POS, Adverb>>>],
            [Surface: "mayores",
             Gender: Masculin|Feminin,
             Number: Plural,
             Comparation: Comparative,
             Suc: <<<POS, Substantive|Adjective>>,
                   <<POS, Verb>>>,
             Pre: <<<POS, Substantive|Adjective>>,
                   <<POS, Adverb>>>],
            [Surface: "máximo",
             Gender: Masculin,
             Number: Singular,
             Comparation: Superlative,
```

```
                Suc: <<<POS, Substantive|Adjective>>,
                      <<POS, Verb>>>,
                Pre: <<<POS, Substantive|Adjective>>,
                      <<POS, Adverb>>>],
               [Surface: "máximos",
                Gender: Masculin,
                Number: Plural,
                Comparation: Positive,
                Suc: <<<POS, Substantive|Adjective>>,
                      <<POS, Verb>>>,
                Pre: <<<POS, Substantive|Adjective>>,
                      <<POS, Adverb>>>],
               [Surface: "máxima",
                Gender: Feminin,
                Number: Singular,
                Comparation: Superlative,
                Suc: <<<POS, Substantive|Adjective>>,
                      <<POS, Verb>>>,
                Pre: <<<POS, Substantive|Adjective>>,
                      <<POS, Adverb>>>],
               [Surface: "máximas",
                Gender: Feminin,
                Number: Plural,
                Comparation: Superlative,
                Suc: <<<POS, Substantive|Adjective>>,
                      <<POS, Verb>>>,
                Pre: <<<POS, Substantive|Adjective>>,
                      <<POS, Adverb>>>]>];

[Lemma: "ser",
 POS: Paradigm,
 Valencies: <Intransitive>,
 Paradigm: <[Surface: "ser",
             Category: <Infinitive>],
            [Surface: "sér",
             Category: <Infinitive>,
             WellFormed: no,
             Suc: <<<POS, Pronoun>, <PronounType, PersonalAtonic>>>,
             PossibleEnclitics: 1,
             FilledEnclitics: 0],
            [Surface: "siendo",
             Category: <Gerund>],
            [Surface: "siéndo",
             Category: <Gerund>,
             WellFormed: no,
             Suc: <<<POS, Pronoun>, <PronounType, PersonalAtonic>>>,
             PossibleEnclitics: 1,
             FilledEnclitics: 0],
            [Surface: "sido",
             Category: <Participle>],
            [Surface: "soy",
             Category: <[Tense: Present,
                         Mood: Indicative,
                         Person&Number: Sg1]>],
            [Surface: "eres",
             Category: <[Tense: Present,
                         Mood: Indicative,
                         Person&Number: Sg2]>],
            [Surface: "es",
             Category: <[Tense: Present,
                         Mood: Indicative,
                         Person&Number: Sg3]>,
```

```
 Suc: <<<POS, Substantive>>>],
[Surface: "somos",
 Category: <[Tense: Present,
             Mood: Indicative,
             Person&Number: Pl1]>],
[Surface: "sois",
 Category: <[Tense: Present,
             Mood: Indicative,
             Person&Number: Pl2]>],
[Surface: "son",
 Category: <[Tense: Present,
             Mood: Indicative,
             Person&Number: Pl3]>],
```

Für Wortformen, die irreguläre Allomorphie aufweisen (die Verben *ser, estar, haber, dar, ver, saber, ir* und die Adjektive *bueno, malo, pequeño, grande, santo*) werden alle Formen des Paradigmas im Grundformlexikon im Lemma der Grundform angegeben. Die Oberfläche des Lemmas entspricht für Verben dem Infinitiv und für Adjektive der maskulinen Singularform. Als Wortklasse ist `Paradigm` angegeben, um diese Lemmata durch die Allomorphregeln entsprechend behandeln zu können. Die Wortklasse der einzelnen Formen wird entsprechend des Wertes des Attributs `ParadigmPOS` zugewiesen. Die einzelnen Formen sind als Verbunde in einer Liste im Attribut `Paradigm` aufgeführt.

Alle Informationen, die für die Konkatenation bzw. die Analyse der Wortformen notwendig sind, müssen im Lexikon vermerkt werden. Alle Formen sind wohlgeformt, sofern nicht anders angegeben. Für Formen der Adjektive werden Oberfläche, Genus, Numerus, Komparationsstufe, mögliche Nachfolger und Vorgänger angegeben (siehe Eintrag für *grande*). Für Verben werden Oberfläche und Kategorie aufgeführt. Handelt es sich um Formen des affirmativen Imperativs, werden wird jeweils eine zusätzliche Form mit einem expliziten Akzent für den betonten Vokal angegeben. Diese Form ist nicht wohlgeformt. Es können enklitische Pronomina folgen, deren mögliche Anzahl im Attribut `PossibleEnclitics` vermerkt ist. Der Wert des Attributs `FilledEnclitics` ist immer 0. Dieses Attribut wird zur Steuerung der Konkatenation mit enklitischen Pronomina benötigt. Für die dritte Person Singular des Präsens Indikativ sind als mögliche Nachfolger Substantive angegeben (siehe Ausschnitt des Eintrags für *ser*[7]).

6.3 Die Allomorph-Regel

Zur Erzeugung des Allomorphlexikons wird eine Allomorphregel erstellt. Die Behandlung der Grundformeinträge erfolgt abhängig von deren Wortklasse durch spezielle Unterregeln. Weiterhin werden allgemeine Unterregeln erstellt, die wortklassenunabhängig benutzt werden, etwa für die Zuweisung von akzentuierten Vokalen zu nicht-akzentuierten und umgekehrt.

Für alle Lemmata gilt, dass die Oberfläche des Morphems, dem die jeweiligen Allomorphe zugeordnet werden, die Oberfläche des Lemmas ist. Die Informationen,

[7] Hier sind aus Platzgründen nur die Formen des Präsens Indikativ angegeben.

die jedes Allomorph für die Konkatenation und die Bestimmung der Kategorie einer konkatenierten wohlgeformten Wortform aufweist, werden in den entsprechenden Unterregeln zugeordnet. Anschließend wird das Attribut `Lemma` mit der Oberfläche des Grundformeintrags als Wert entfernt. Ein Eintrag im Allomorphlexikon besteht aus der Oberfläche des Allomorphs und der kategorialen Information. Als kategoriale Information gilt das erzeugte Allomorph, dem das Attribut `Allomorph` mit der Oberfläche des Allomorphs als Wert hinzugefügt wird.

Jedes Allomorph enthält mindestens Attribute für die Oberfläche (`Surface`), die Wortklasse (`POS`), die Grundform (`BaseForm`), die Bedingungen für Nachfolger (`Suc`) und die Bedingungen für Vorgänger (`Pre`). Entsprechend der Möglichkeiten der Komposition, die in Tabelle 3.7 dargestellt sind, werden Verben, Substantiven, Adjektiven und Adverben die jeweiligen Vorgänger und Nachfolger für die Komposition zugeordnet.

Entspricht ein Allomorph einer wohlgeformten Wortform, erhält der Allomorpheintrag das Attribut `WellFormed` mit dem Wert `yes`. Handelt es sich nicht um eine wohlgeformte Wortform, hat das Attribut den Wert `no`. Bei der Konkatenation von Allomorphen wird nach jedem Konkatenationsschritt geprüft, ob es sich um eine wohlgeformte Wortform handelt. Allomorphe, deren Konkatenation mit einem Wortanfang zu einer wohlgeformten Wortform führen (etwa Flexionsallomorphe), erhalten daher ebenfalls den Wert `yes` für das Attribut `WellFormed`.

6.3.1 Gemeinsame Regeln

Gemeinsame Regeln werden von den einzelnen Unterregeln unabhängig von der Wortklasse des behandelten Lemmas aufgerufen.

Für jeden Grundformeintrag wird zunächst die Regel `processAllo` aufgerufen. Enthält der Eintrag das Attribut `Allo`, erzeugt sie für jeden Verbund, der in der Liste des Attributs aufgeführt ist, einen Lexikoneintrag, der aus den Attributen des Verbundes und allen Attributen des ursprünglichen Eintrags besteht, das Attribut `Allo` ausgenommen. Damit enthält der Eintrag das Attribut `Surface` mit der Oberfläche des Allomorphs als Attribut. Enthält der Eintrag das Attribut `Allo` nicht, wird er um das Attribut `Surface` ergänzt, das als Wert die Oberfläche des Lemmas erhält. Dieses Attribut ist also in jedem Lexikoneintrag enthalten und mit der Oberfläche des Morphems oder eines Allomorphs initialisiert.

Die Regel `accent` dient der Zuweisung von akzentuierten und nicht-akzentuierten Vokalen. Der Regel wird ein Parameter übergeben. Ist der Parameter ein Vokal ohne Akzent, wird der Vokal mit Akzent zurückgegeben. Ist der Parameter ein Vokal mit Akzent, wird der Vokal ohne Akzent zurückgegeben.

6.3.2 Substantive und Adjektive

Substantive, Adjektive und Lemmata, die als Wortklasse das Multisymbol `Substantive|Adjective` aufweisen, werden gleich behandelt, da sie die gleichen Informationen beinhalten.

Zunächst werden dem Lexikoneintrag drei Attribute hinzugefügt. Das Attribut `FinalPOS` erhält als Wert die Wortklasse des Lemmas. Dieses wird für die Derivation und Kombination benötigt. Das Attribut `Suc` wird als leere Liste initialisiert. Das Attribut `Pre` enthält eine Liste mit Angaben über Vorgänger von Substantiven oder Adjektiven. Möglich sind Präfixe, wohlgeformte Substantive, Adjektiveund Satzzeichen für Komposita, die aus wohlgeformten Wortformen, verbunden durch Bindestrich, bestehen.

Die weitere Behandlung richtet sich danach, ob die Lemmata bezüglich der Flexion der ersten oder zweiten bzw. dritten Gruppe zugeordnet sind.

Gehört das Lemma zur ersten Gruppe, wird das Attribut `WellFormed` mit dem Wert `yes` hinzugefügt, um deutlich zu machen, dass es sich bereits um eine wohlgeformte Wortform handelt. Für Adjektive der ersten Gruppe entspricht die Oberfläche sowohl dem Femininum als auch dem Maskulinum. Der Wert für `Gender` erhält das Multisymbol `Masculin|Feminin`. Substantive der ersten Gruppe sind nur in einem Genus möglich, das bereits im Grundformlexikon vermerkt ist. Diese Angabe wird übernommen.

Ist im Grundformeintrag eine besondere Pluralform vermerkt, wird unterschieden, ob die Pluralform mit der Oberfläche des Lemmas identisch ist. In diesem Fall erhält das Attribut `Number` das Multisymbol `Singular|Plural` als Wert. Sind beide nicht identisch, werden zwei Allomorphe erzeugt. Eines erhält die Oberfläche des Lemmas mit dem Wert `Singular` für das Attribut `Numerus`, das andere erhält die Oberfläche der Pluralform mit dem Wert `Plural` für das Attribut `Numerus`. Alle anderen Informationen bezüglich Genus, Grundform und Wohlgeformtheit sind identisch. Anschließend wird der Vermerk über die Pluralform entfernt und die Liste der möglichen Nachfolger ergänzt. Möglich sind Verben, Adverben, Substantive, Adjektive, Präfixe, denen eine Basis folgt, und Bindestrich für Komposita sowie Interfixe und Suffixe für Derivata. Endet die Oberfläche auf einen Vokal, wird ein zusätzliches Allomorph erzeugt, das als nicht-wohlgeformt gekennzeichnet ist und als Oberfläche die um den Vokal verkürzte Oberfläche erhält. Nachfolgen können dann nur Suffixe und Interfixe, die mit einem Vokal beginnen, wie in 3.2.4 gezeigt. Weitere Veränderungen erfolgen für diese Grundformeinträge nicht.

Ist keine Pluralform vermerkt, erhält der Lexikoneintrag den Wert `Singular` für den Numerus. Für diese Lemmata sind weitere Unterscheidungen möglich.

Endet die Oberfläche mit einem Vokal, werden zwei Allomorphe erzeugt. Ein Allomorph erhält als Oberfläche die Oberfläche der Grundform inklusive End-Vokal. Das Allomorph gilt als wohlgeformte Wortform. Die Pluralbildung erfolgt mit dem Allomorph *s*. Als nachfolgende Allomorphe sind nur das Pluralallomorph *s* für die

Flexion, der Bindestrich und Präfix mit nachfolgender Basis für die Komposition und Interfix oder Suffix für die Derivation möglich. Das Allomorph der um den End-Vokal gekürzten Oberfläche ist nicht wohlgeformt. Nachfolgende Suffixe und Interfixe müssen mit Vokal beginnen.

Endet die Oberfläche mit einem betonten Vokal, gefolgt von einem Konsonanten, werden zwei Allomorphe erzeugt. Ein Allomorph ist die wohlgeformte Wortform mit betontem Vokal, dem nur der Bindestrich für Komposita folgen kann. Das zweite Allomorph ist die Wortform mit dem entsprechenden unbetonten Vokal, dem das Pluralallomorph, Substantive, Adjektive, Verben und für Derivation Interfix oder Suffix folgen können.

Endet die Oberfläche auf *-n* oder *-s*, ist die vorletzte Silbe betont. Durch Anfügen des Pluralallomorphs *-es* ändert sich die Silbenzahl, die Betonung läge jetzt auf der ursprünglich letzten Silbe. Um die Betonung auf der ursprünglich vorletzten Silbe zu belassen, muss ein expliziter Akzent gesetzt werden. Es werden also ein Allomorph ohne expliziten Akzent zur Komposition und Derivation sowie ein Allomorph mit Akzent zur Flexion erzeugt.

Endet die Oberfläche auf einen anderen Konsonanten, ausgenommen *n* und *s*, und trägt die vorletzte Silbe einen Akzent, muss diese Oberfläche für die Pluralbildung erhalten werden. Für Derivation und Komposition wird dagegen die Oberfläche ohne den Akzent verwendet. Es werden ein Allomorph mit explizitem Akzent zur Flexion und ein Allomorph ohne Akzent zur Derivation und Komposition erzeugt.

Für Lemmata, deren Oberfläche auf eines der Phoneme aus Tabelle 3.22 endet, werden Allomorphe erzeugt, die jeweils mit einer der Entsprechung des Phonems enden. Die jeweils folgenden Allomorphe müssen mit dem entsprechenden Vokal beginnen. Nur das Allomorph, dessen Nachfolger mit *e* oder *i* beginnen müssen, ist für die Flexion mittels des Pluralallomorphs *-es* geeignet.

Endet die Oberfläche auf einen anderen Konsonanten und ist kein expliziter Akzent vorhanden, wird für die Pluralbildung das Allomorph *-es* verwendet. Derivation und Komposition kann auf diese Allomorphe uneingeschränkt angewendet werden. In diesem Fall ist die letzte Silbe betont. Durch das Anfügen des Pluralallomorphs *-es* verändert sich die Silbenanzahl. Die Betonung der ursprünglich letzten, jetzt vorletzten Silbe bleibt erhalten, da die Wortform nun auf *-s* endet und auf der vorletzten Silbe betont wird.

Für Substantive und Adjektive der zweiten und dritten Gruppe ist im Grundformlexikon eine gemeinsame Kennzeichnung vorgesehen. Die Oberflächen von Lemmata der zweiten Gruppe werden um das Genusallomorph *-o* bzw. *-e* gekürzt. Die entsprechenden Allomorpheinträge werden als nicht-wohlgeformt gekennzeichnet. Nachfolgende Allomorphe müssen mit Vokal beginnen. Möglich sind dabei die Genusallomorphe sowie Interfixe und Suffixe.

Substantive und Adjektive, die nicht auf *-o* oder *-e* enden, gehören zur dritten

Gruppe. Ihre Behandlung richtet sich wie die der ersten Gruppe nach der Oberfläche der Grundform. Es werden die gleichen Unterscheidungen getroffen wie oben beschrieben. Das Allomorph, dessen Oberfläche der des Grundformeintrags entspricht, gilt als wohlgeformt. Angefügt werden können das Genusallomorph für die feminine Form, das Pluralallomorph sowie Allomorphe zur Derivation und Komposition.

Für Adjektive wird ein zusätzliches Attribut `AdjectiveComparation` eingefügt. Es erhält zunächst den Wert `Positive`. Dieser kann im Verlauf der Konkantenation von Allomorphen geändert werden. Um dies zu ermöglichen, wird der Nachfolgerliste der Hinweis auf das Anfügen des Komparationsallomorphs hinzugefügt.

Sind Angaben zu Genus und Numerus in den Allomorpheinträgen vorhanden, werden diese jeweils in die Attribute `FinalGender` und `FinalNumber` übernommen. Die Angaben werden zur Derivation und Komposition benötigt.

6.3.3 Verben

Auf Lexikoneinträge, die Verben sind, werden zwei Unterregeln angewandt. Zunächst wird der Themavokal ermittelt und damit die Konjugationsklasse zugewiesen sowie die Anzahl der möglichen enklitischen Pronomina bestimmt. In einer zweiten Regel wird die Allomorphie behandelt. Abhängig davon, ob die Oberfläche markiert ist, handelt es sich um semi-irreguläre oder um reguläre bzw. semireguläre Verben.

Die Infinitiv-Endung aus Themavokal und *r* wird zur Bestimmung der Konjugationsklasse verwendet. Als Oberfläche des Allomorphs wird der Stamm (Lemma-Oberfläche um Infinitiv-Endung gekürzt) genutzt. Abhängig von den Angaben im Attribut `Valencies` wird die Anzahl der möglichen enklitischen Pronomina bestimmt. Dabei gilt: Ist keine Angabe zum reflexiven Gebrauch des Verbs gegeben, wird ein möglicher reflexiver Gebrauch dennoch berücksichtigt. Die Anzahl der enklitischen Pronomina, die folgen können, ist daher um eins größer als der Wert in `PossibleEnclitics`.

Abschließend wird das Allomorph als nicht-wohlgeformt gekennzeichnet. Für die Derivation ist die Angabe `FinalPOS: Verb` notwendig. Als Nachfolger werd zunächst der Themavokal des Verbs, Suffixe und Interfixe eingetragen. Angaben zu nachfolgenden Flexionsallomorphen werden, abhängig von der Allomorphie des Verbs, in der folgenden Unterregel ermittelt. Als Vorgänger für Verbstämme sind Präfixe, Adverben, wohlgeformte Substantive und Adjektive sowie der Bindestrich möglich.

Enthält die Oberfläche des Grundformeintrags eine Markierung, handelt es sich um semi-irreguläre Allomorphie. Dies wird im Allomorpheintrag vermerkt. Zur Ermittlung der Oberfläche der jeweiligen Allomorphe eines semi-irregulären Verbs dient eine Tabelle, in der die Veränderung gegenüber der Oberfläche des Grundformeintrags sowie die Tempora, in denen das jeweilige Allomorph verwendet

wird, angeführt sind. Neben Allomorphen können so auch die Allographen für Verbformen mit enklitischen Pronomina ermittelt werden. In einigen Fällen ist dort eine dritte Angabe enhalten. Im Attribut `Allo_i` wird angegeben, ob besondere Formen der Flexionsallomorphe erforderlich sind. Dies ist insbesondere für die Konkatenation von Verbformen, denen enklitische Pronomina folgen, aus Gründen der Akzentsetzung notwendig. Möglich sind Veränderungen der Oberfläche an einer oder an zwei Stellen des Verbs. Dementsprechend haben einige Tabelleneinträge nur das Attribut `Vowel`, dem die Oberfläche des oder der Vokale folgt. Andere weisen ein zusätzliches Attribut `Consonant` auf, dem die Oberfläche der neben Vokalen zu verändernden Konsonanten folgt. Sind markierter Vokal und Konsonant nicht durch weitere Buchstaben getrennt, wie in *p{od}er*, ist nur ein Attribut angegeben.

Sind Angaben über unregelmäßige Partizipien oder Imperativ-Formen im Grundformeintrag vorhanden, wird für diese ein eigener Allomorpheintrag erzeugt. Aus der Nachfolgeliste des ursprünglichen Allomorpheintrags wird die Angabe zum Partizip-, oder Imperativ-Flexionsmorphem entfernt. Partizipien können auch als Adjektive funktionieren. Es wird ein Eintrag mit der Partizip-Oberfläche und der Wortklasse Adjektiv erzeugt, der der zweiten bzw. dritten Gruppe zugeordnet und an die Unterregel zur Behandlung von Substantiven und Adjektiven übergeben wird.

Für Verben, die keine Allomorphie-Markierung tragen, wird die Konjugationsart zunächst als semi-regulär festgelegt und in die Nachfolgeliste die Allomorphe zur Verbflexion aufgenommen. Welche der Flexionsallomorphe folgen können, wird in Abhängigkeit der Allomorphie der Verben bestimmt. Für semi-reguläre Verben sind innerhalb der Unterregel Muster angegeben. Passt die Oberfläche eines Verbs auf eines der Muster, erfolgt die Erzeugung der Allomorphe entsprechend der Bestimmungen des Musters. Passt die Oberfläche des Verbs auf keines der Muster, handelt es sich um ein reguläres Verb. Für diese müssen Allographen mit akzentuiertem Stammvokal für Verbformen erzeugt werden, die das Anhängen enklitischer Pronomina erlauben. Die Konjugationsart wird als regulär gekennzeichnet. Wie für semi-irreguläre Verben werden für Einträge, die Partizip oder Imperativ als unregelmäßige Form aufführen, weitere Allomorphe erzeugt.

6.3.4 Adverben

Für Adverben ist der jeweilige Allomorpheintrag der um die Allomorphoberfläche, Vorgänger und Nachfolger ergänzte Grundformeintrag. Als Vorgänger sind Präfixe möglich. Als Nachfolger können Affixe, Verben, Adjektive und der Bindestrich auftreten. Der Vermerk über die Wohlgeformtheit wird hinzugefügt.

6.3.5 Ziffern, Numerale und Maßeinheiten

Für Ziffern, Numerale und Maßeinheiten wird zunächst die Unterregel `processAllo` aufgerufen. Jeder Eintrag erhält den Hinweis, dass es sich um wohlgeformte Wortformen handelt. Anschließend werden die Nachfolger- und Vorgänger-Listen angegeben. Auf Ziffern können Ziffern, Maßeinheiten, Zeichen und Abkürzungen folgen und ihnen vorausgehen. Maßeinheiten müssen Ziffern vorausgehen, Nachfolger sind nicht möglich. Numeralen können Zeichen folgen und vorausgehen.

6.3.6 Pronomina, Artikel, Präpositionen, Konjunktionen, Interjektionen, Namen, Akronyme und Satzzeichen

Die Grundformeinträge von Pronomina, Artikeln, Präpositionen, Konjunktionen, Interjektionen, Namen, Akronymen und Satzzeichen werden unverändert ins Allomorphlexikon übernommen. Sind keine Angaben zu Nachfolgern oder Vorgängern enthalten, werden die entsprechenden Listen als leere Listen gekennzeichnet. Damit sind weder Derivation noch Kombination möglich. Alle Allomorphe sind wohlgeformte Wortformen, daher erhält das Attribut `WellFormed` den Wert `yes`.

6.3.7 Themavokale

Themavokale können nach Verbstämmen und vor Affixen, Substantiven, Adjektiven, Verben und Adverben auftreten. Entsprechende Angaben werden den Attributen `Pre` und `Suc` zugeordnet. Zusätzlich erhalten die Allomorpheinträge den Vermerk, dass resultierende Wortformen nicht wohlgeformt sind.

6.3.8 Affixe

Für Affixe werden Präfixe, Interfixe und Suffixe innerhalb der Unterregel getrennt behandelt. Um welches Affix es sich handelt, gibt die Wortklasse an. Auf alle wird zuerst die Unterregel `processAllo` angewandt.

Die Einträge für Präfixe werden um die Angabe zu Vorgänger, Nachfolger und die Nicht-Wohlgeformtheit der resultierenden Wortform ergänzt. Als Vorgängerallomorph ist ein weiteres Präfix oder im Fall von Komposition ein wohlgeformtes Verb, Substantiv, Adjektiv oder Adverb möglich. Als Nachfolger sind Verben, Substantive, Adjektive, Adverben oder weitere Präfixe möglich. Da die Schreibung mit Bindestrich bei aufeinanderfolgenden gleichen Präfixen möglich ist, kann ebenfalls der Bindestrich folgen bzw. vorausgehen.

Für Interfixe sind als Nachfolger nur Suffixe möglich. Vorgänger können Verben, Substantive, Adjektive und Adverben sein. Wortanfänge, die aus Basis und Interfix konkateniert sind, sind nicht wohlgeformt. Diese Informationen werden für jeden Interfix-Eintrag ergänzt. Sind im Grundformeintrag Angaben zu Nachfolgern gemacht, werden diese auch übernommen.

Wird ein Wortanfang mit einem Suffix konkateniert, ist die resultierende Wortform wohlgeformt. Die Einträge werden um diesen Vermerk, die Nachfolgerliste mit Verben, Substantiven, Adjektiven und Adverben sowie die Vorgängerliste mit Verben, Substantiven, Adjektiven, Adverben und Interfixen ergänzt. Sind im Grundformeintrag Angaben zu Vorgängern enthalten, werden diese hinzugefügt. Für Suffixe, die als resultierende Wortklassen Substantive oder Adjektive angeben, sind als nachfolgende Allomorphe auch die Genus-, Plural- oder Komparationsallomorphe möglich. Dabei sind diese Suffixe jeweils der ersten, zweiten oder dritten Gruppe zugeordnet.

6.3.9 Flexionsmorpheme

Für Verb-Flexionsmorpheme wird zunächst die Unterregel `processAllo` aufgerufen. Dem Allomorpheintrag werden dann die Vorgängerliste und die Nachfolgerliste hinzugefügt. Als Vorgänger kommt nur ein Verb in Frage, das einen Themavokal aufweist, der mit dem des Flexionsallomorphs identisch ist. Ist im Eintrag das Attribut `Allo_i` vorhanden, wird dessen Wert in die Bestimmung zum Vorgänger übernommen, ist es nicht vorhanden, wird es mit dem Wert `no` ergänzt, um deutlich zu machen, dass es sich um das von der Morphemoberfläche nicht abweichende Allomorph handelt, welches für entsprechende Verbstämme, die dies erfordern, verwendbar ist. Sind im Grundformeintrag die Attribute `Suc` und `Pre` enthalten, werden diese übernommen. Als resultierende Wortklasse wird `Verb` gesetzt.

Für Flexionsmorpheme zur Flexion der Substantive und Adjektive bezüglich Genus und Numerus werden den Einträgen der Verweis auf die resultierende Wohlgeformtheit und als Vorgängerbestimmung Substantive und Adjektive hinzugefügt. Die Numerusallomorphe können als Nachfolger Substantive, Adjektive, Verben, Adverben, den Bindestrich oder Affixe annehmen. Die Genusallomorphe erhalten den Hinweis, dass als Pluralallomorph nur *s* folgen kann. Folgende Allomorphe können das Komparationsallomorph, der Bindestrich, Verben, Adverben, Substantive, Adjektive oder Affixe sein. Das Komparationsallomorph erfordert als Vorgänger ein Adjektiv und als Nachfolger ein Genusallomorph. Daher wird eine resultierende Wortform nicht als wohlgeformt markiert. Um die Komparationsstufe der resultierenden Wortform ändern zu können, wird dem Attribut `FinalComparation` der Wert `Superlativ` zugeordnet.

6.3.10 Paradigmen-Einträge

Sind vollständige Paradigmen im Grundformlexikon angegeben, wird für jeden Verbund in der Liste des Attributs `Paradigm` ein wohlgeformtes Allomorph erzeugt. Für Verbformen, denen enklitische Pronomina folgen, sind Allographen im jeweiligen Paradigma der Verben angegeben. Diese sind nicht wohlgeformt. Der entsprechende Eintrag enthält für das Attribut `WellFormed` den Wert `no`. Diese Information wird in den Allomorpheintrag übernommen.

Die Wortklasse jedes Allomorphs wird aus dem Attribut `ParadigmPOS` des Grundformeintrags übernommen. Sind keine Angaben zu Nachfolgern enthalten, kann kein Allomorph folgen - dem Attribut `Suc` wird die leere Liste zugewiesen. Sind keine Angaben zu Vorgängern enthalten, werden für Verbformen entsprechend 3.7 Substantive, Adjektive und Adverben sowie zusätzlich Präfixe bestimmt. Für Adjektive sind nur Präfixe möglich.

6.3.11 Allomorphie-Quotient

Der Allomorphie-Quotient spiegelt das Verhältnis der Anzahl der Grundformeinträge und der Anzahl der Allomorpheinträge wider. Er berechnet sich demnach aus:

$$\frac{\text{Anzahl der Allomorphe} - \text{Anzahl der Grundformen}}{\text{Anzahl der Grundformeinträge}} \times 100\%$$

Für das vollständige Lexikon werden aus 98 546 Grundformeinträgen 168 394 Allomorpheinträge erzeugt. Das entspricht einem Allomorphie-Quotienten von 70,88%. Für das reduzierte Lexikon werden aus 66 103 Grundformeinträgen 107 730 Allomorpheinträge erzeugt. Das entspricht einem Allomorphie-Quotienten von 62,97%. Beide Werte sind im Vergleich zu den Allomorphiequotienten, die von den MALAGA-Morphologiekomponenten für Deutsch (31%)[8], Englisch (8,94%)[9] oder Italienisch (37%)[10] erreicht werden, sehr hoch. Ein Grund liegt in der Behandlung der Allographen des Spanischen als Allomorphe, was durch die Implementierungssprache erzwungen ist. Wie in 3.6.3 begründet, handelt es sich dabei nicht um Allomorphe. Eine Aussage über den tatsächlichen Grad der Allomorphie des Spanischen kann aus den angegebenen Werten daher nicht abgeleitet werden.

Werden alle Elemente der Allomorphregel entfernt, die Allographen erzeugen, und beide Varianten des Lexikons erneut in ein Allomorphlexikon überführt, ergeben sich für das vollständige Grundformlexikon 153 039 Allomorpheinträge. Das entspricht einem Allomorphie-Quotienten von 55,3%. Für das reduzierte Lexikon werden 99 297 Allomorpheinträge erzeugt, das entspricht einem Allomorphie-Quotienten von 50,19%. Beide Werte sind weiterhin nicht identisch, liegen aber näher beieinander und ergeben einen Durchschnittswert von 52,75%.

6.4 Die Kombinations-Regeln

Die allgemeine Form der Kombinations-Regeln ist in 2.2.2 beschrieben. Wie in 5.1 begründet, beinhaltet die Morphologiedatei der SMM nur eine Kombinationsregel und eine Endregel zur Überprüfung, ob ein Endzustand erreicht ist. Als Folgeregel des initialen Startzustands kann daher nur die Regel `Concat` folgen.

[8] [Hausser 1999, S. 268].
[9] [Leidner 1998, S. 120].
[10] [Wetzel 1996, S. 44].

6.4.1 Die Regel `Concat`

Beim Durchlauf der Kombinationsregel wird unterschieden, ob diese zum ersten Mal durchlaufen wird oder ob bereits ein Allomorph eingelesen wurde.

Wird die Regel das erste Mal durchlaufen, ist der Wortanfang leer. Das einzulesende Allomorph bildet den Beginn des Wortes, das analysiert wird. Allomorphe, die ein Interfix, Suffix, eine Flexionsendung oder ein Themavokal sind, müssen ausgeschlossen werden, da eine wohlgeformte Wortform nicht mit diesen beginnen kann. Weitere Einschränkungen gelten für das erste Allomorph nicht. Als neuer Wortanfang wird der Lexikoneintrag des Allomorphs übernommen. So bleiben alle Informationen für eine Überprüfung der Konkatenierbarkeit mit einem nächsten Allomorph erhalten, insbesondere die Bestimmungen in `Suc` und `SucFon`. Anschließend wird die in 5.3 beschriebene Ausgabestruktur mit Hilfe der Attribute `WordForm`, `BaseForm` `POS`, `Segmentation` und `WordStructure` angelegt.

Wird die Regel ein weiteres Mal durchlaufen, muss überprüft werden, ob das aktuelle Allomorph mit dem vorliegenden Wortanfang konkateniert werden kann. Um dies zu kontrollieren, wurden bei der Erzeugung der Allomorphe durch die Allomorphregel jedem Eintrag die Attribute `Suc`, `Pre` und eventuell das Attribut `SucFon` mitgegeben. Ist `SucFon` im Wortanfang vorhanden, wird geprüft, ob der Beginn des nächsten Allomorphs die dort angegebene Bedingung erfüllt. Ist dies nicht der Fall, wird die Konkatenation abgebrochen.

Die Bestimmungen zur Nachfolge eines Wortanfangs sind in einer Liste als Wert des Attributs `Suc` enthalten. Ist die Liste leer, darf kein Allomorph folgen und die Konkatenation wird abgebrochen. Zur Überprüfung auf Verträglichkeit von Wortform und nächstem Allomorph wird aus der Nachfolgeliste jeweils ein Element ausgewählt. Für folgendes Beispiel:

```
Suc: <<<POS, NounInflection>, <InflectionType, Gender>>, <<POS, Suffix>>>
```

enthält die Nachfolgeliste also zwei Elemente:

```
<<POS, NounInflection>, <InflectionType, Gender>>
<<POS, Suffix>>
```

Das nachfolgende Allomorph muss mindestens die Bedingungen eines Elementes erfüllen, um konkateniert werden zu können. Jedes Element ist eine Liste, die mindestens ein Element umfasst. Für das erste Nachfolgeelement sind dies die Elemente:

```
<POS, NounInflection>
<InflectionType, Gender>
```

Dabei handelt es sich wiederum um Listen, die aus genau zwei Elementen bestehen. Das erste Element ist jeweils ein Attribut, dass im Eintrag des nächsten Allomorphs enthalten sein muss. Das zweite Element ist jeweils der Wert, den das Attribut aufweisen muss. Da als Werte auch Listen von Werten oder Multisymbole möglich sind, wird überprüft, ob das zweite Element und der Wert des Attributs im Allomorpheintrag kongruent sind. Kongruenz bedeutet, dass beide in mindestens einem Wert übereinstimmen müssen.

Auf vergleichbare Weise erfolgt der Abgleich der Vorgängerbedingungen des Allomorphs, die in dem Attribut `Pre` angeführt sind, mit dem Wortanfang. Der Wert

des Attributs ist eine ebenso strukturierte Liste wie die Nachfolgeliste des Wortanfangs.

Ist das nächste Allomorph ein gültiger Nachfolger des Wortanfangs und der Wortanfang ein gültiger Vorgänger des Allomorphs, können beide konkateniert werden.

Wie in 3.2.3 dargestellt, sind Interfixe und Suffixe interdependent. Ist also das letzte Allomorph des Wortanfangs ein Interfix und das nächste Allomorph ein Suffix, wird überprüft, ob das Suffix in der Suffix-Liste des Interfixes und das Interfix in der Interfix-Liste des Suffixes aufgeführt sind. Haben beide Überprüfungen ein positives Ergebnis, erfolgt die Konkatenation, sonst wird der Pfad abgebrochen.

Können Wortanfang und nächstes Allomorph konkateniert werden, müssen alle kategorialen Informationen des resultierenden Wortanfangs und die für weitere Konkatenationen benötigten Informationen im neuen Wortanfang vermerkt werden. Abhängig von den Wortklassen von Wortanfang und nächstem Allomorph wird die Konkatenierungsart (Flexion - `<FLX>`, Kombination - `<CO>`, Derivation - `<DV>`, Anhängen enklitischer Pronomina - `<ENCL>`, Konkatenation von Ziffern - `<NUM>` oder Anhängen einer Maßeinheit an eine Ziffer - `<NDM>`) bestimmt. Der Wert für `WordForm` wird um das Allomorph ergänzt. Der Wert für `Segmentation` wird um die Konkatenierungsart und um das Allomorph erweitert. In die Liste des Attributs `WordStructure` wird ein neues Element als Verbund eingetragen, der Informationen zur Oberfläche des Morphems, zur Oberfläche und zur Wortklasse des Allomorphs enthält.

Sind im Allomorph Angaben zur resultierenden Wortklasse, zum resultierenden Genus oder Numerus enthalten, werden diese in die entsprechenden Attribute des neuen Wortanfangs übernommen. Besitzt das Allomorph Angaben zur Wohlgeformtheit, werden diese übernommen.

Ändert sich für Adjektive durch Flexion die Komparationsart, wird diese im Wortanfang geändert. Ist im Allomorph eine Angabe zum Themavokal enthalten, die im Wortanfang nicht vermerkt ist, wird diese übernommen. Dies ist insbesondere bei Kombination oder Derivation mit Präfix der Fall.

Für Verb-Flexionsallomorphe wird die Strukturbeschreibung des Flexionsmorphems in den Wortanfang übernommen. Ist das nächste Allomorph kein Flexionsallomorph, wird der Wert für `BaseForm` um das Allomorph ergänzt.

Die Anzahl der enklitischen Pronomina, die einer Verbform folgen können, ist für jedes Verb durch die Allomorphregel im Attribut `PossibleEnclitics` festgehalten. Für jedes Pronomen, das konkateniert wird, wird dieser Wert um eins verringert. Vor jeder Konkatenation wird geprüft, ob noch Pronomina folgen dürfen. Da, wie in 6.3.3 beschrieben, die Zahl der möglichen Pronomina um eins größer ist als der Wert des Attributs `PossibleEnclitics`, muss dieser Wert größer oder gleich null sein, um ein weiteres enklitisches Pronomen anfügen zu können. Wird ein enklitisches Pronomen konkateniert, erhält der Wortanfang ein zusätzliches Attribut

`EncliticalPronouns`. Der Wert des Attributs ist eine Liste von Verbunden, die jeweils alle Informationen über das Pronomen beinhalten. Aus dieser Struktur kann in der Endregel das Analyseergebnis als Liste aus analysierter Verbform und analysierten Pronomina erzeugt werden, wie in 5.3 beschrieben.

Am Ende jedes Regeldurchlaufs werden die Bestimmungen über die Vorgänger des Allomorphs entfernt und als Nachfolgebestimmungen des neuen Wortanfangs die Nachfolgebestimmungen des Allomorphs übernommen. Um beim nächsten Regeldurchlauf die Vorgängerbestimmungen des nächsten Allomorphs nicht nur mit dem Wortanfang, sondern wenn notwendig auch mit dem zuletzt eingelesenen Allomorph vergleichen zu können, wird im Attribut `LastMorphem` die Oberfläche des dem Allomorph zugeordneten Morphems und im Attribut `LastPOS` die dem Allomorph zugeordnete Wortklasse festgehalten.

6.4.2 Die Endregel `FinalStateCheck`

Die Regel prüft, ob die Allomorphe der Wortform vollständig konkateniert wurden und ob es sich um eine wohlgeformte Wortform handelt. Dazu muss die Wortform das Attribut `WellFormed` mit dem Wert `yes` aufweisen. Affixe, Flexionsendungen und Themavokale werden nicht als wohlgeformte Wortform akzeptiert. Für die Ausgabe des Ergebnisses werden aus der bisherigen Analyse die Werte für Wortklasse, Genus, Numerus, Konjugationsart, Valenzen, Verbkategorie, Struktur der Flexionsendung, Themavokal, Grundform, Segmentierung, Wortstruktur, Prefixe, Suffixe, nachfolgende bzw. vorausgehende Wortformen und Wohlgeformtheit übernommen, soweit diese Angaben jeweils vorhanden sind. Für Verbformen mit enklitischen Pronomina wird eine Liste erzeugt, deren erstes Element die Verbform-Analyse aufzeigt. Das zweite Element ist die Liste der enklitischen Pronomina, wobei jedes Pronomen kategorisiert wird.

6.5 Analysen mit SMM

An einigen Beispielen sollen die Ergebnisse verdeutlicht werden, die SMM beim Parsen von Wortformen liefert. Zum Parsen wird das vollständige Lexikon verwendet.

Wie in 3.6.3 an den Beispielen *término, termino, terminó* erläutert, ist die Setzung eines expliziten Akzentes entscheidend für die Kategorie und die Bedeutung eines Wortes.

Die Wortform *término* wird als maskulines Substantiv im Singular analysiert. Die Wortform wurde nicht aus verschiedenen Allomorphen konkateniert. Die Oberfläche des Morphems stimmt mit der Oberfläche des Allomorphs überein (siehe Abbildung 6.1).

Für *termino* ergeben sich zwei Analysen, zum einen als Verb, zum anderen als deriviertes Adjektiv. Das maskuline Adjektiv *termino* ist vom Substantiv *termo* ab-

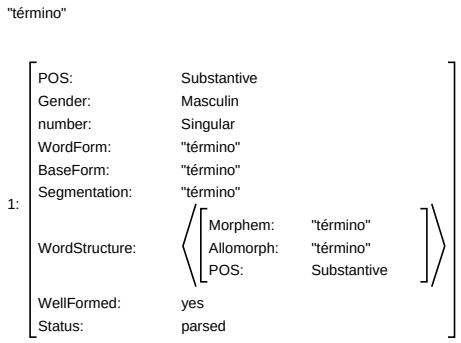

Abbildung 6.1: Analyse für *término*

geleitet. Das Suffix ist ein Diminutiv-Suffix. Da es mit einem Vokal beginnt, ist das Allomorph des Substantivs das um den Endvokal gekürzte Allomorph. Die vorgeschlagene Analyse ist nach den Prinzipien der Derivation des Spanischen formal korrekt, semantisch ist sie nicht sinnvoll.
Die Verbform *termino* ist Teil des Paradigmas des regulären Verbs *terminar*. Das Verb kann intransitiv und reflexiv verwendet werden. Die Flexionsendung besteht nur aus dem Themavokal. Die daraus ableitbare Kategorie der Verbform ist die erste Person Singular des Präsens Indikativ (siehe Abbildung 6.2).

Die Wortform *terminó* gehört ebenfalls zum Paradigma des Verbs *terminar*. Es ist die Form der dritten Person Singular des Indefinido Indikativ. Die Flexionsendung wird als Allomorph für die Modus/Tempus-Kennzeichnung analysiert (siehe Abbildung 6.3).

Die Analyse der Wortform *cuéntamelo* ist eine Liste mehrerer Elemente. Wie in 3.5.5 und 5.3 verdeutlicht, werden Analysen, die eine affirmative Imperativ-Form mit enklitischen Pronomina ermitteln, als Liste aus den Analysen der Verbform und den Pronomina dargestellt. Die Verbform gehört zum Paradigma des semiirregulären Verbes *contar*, die Flexionsendung entspricht dem Themavokal *a*. Das Verb ist sowohl intransitiv als auch transitiv zu verwenden. Als Kategorie wird die zweite Person Singular des Präsens Imperativ in affimativer Bedeutung angegeben. Im Attribut `NextWord` ist vermerkt, dass enklitische Pronomina folgen.

Das zweite Element der Ergebnisliste ist die Liste der enklitischen Pronomina. Für jedes Pronomen ist die Oberfläche, die Wortklasse, Person und Numerus, Genus und die Kategorisierung als unbetontes Personalpronomen angegeben (siehe Abbildung 6.4).

In 3.2 wurde auf verschiedene Möglichkeiten der Analyse von *inutizable* eingegangen. Die SMM ermittelt die Wortklasse Adjektiv, für Numerus Singular und für Genus sowohl Femininum als auch Maskulinum. Die Analysen unterscheiden

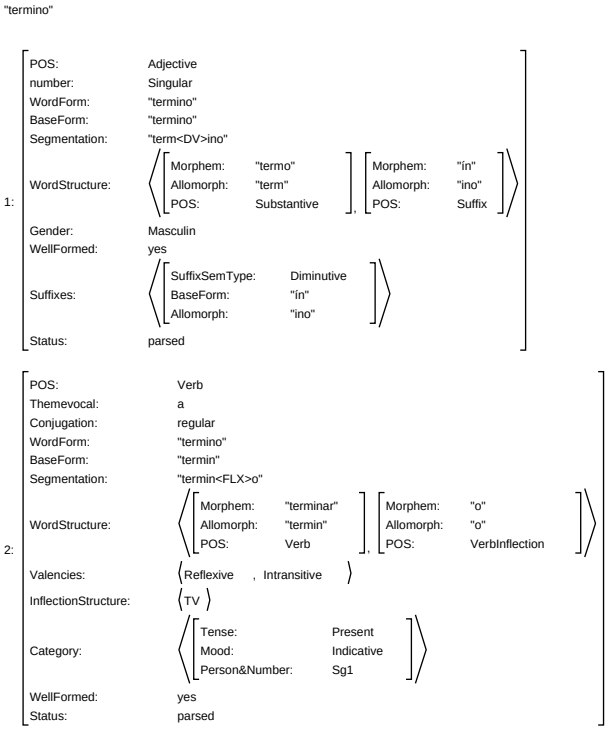

Abbildung 6.2: Analyse für *termino*

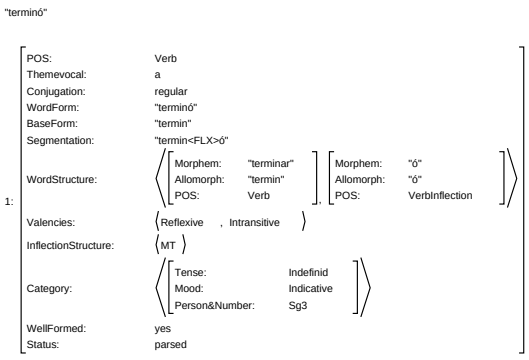

Abbildung 6.3: Analyse für *terminó*

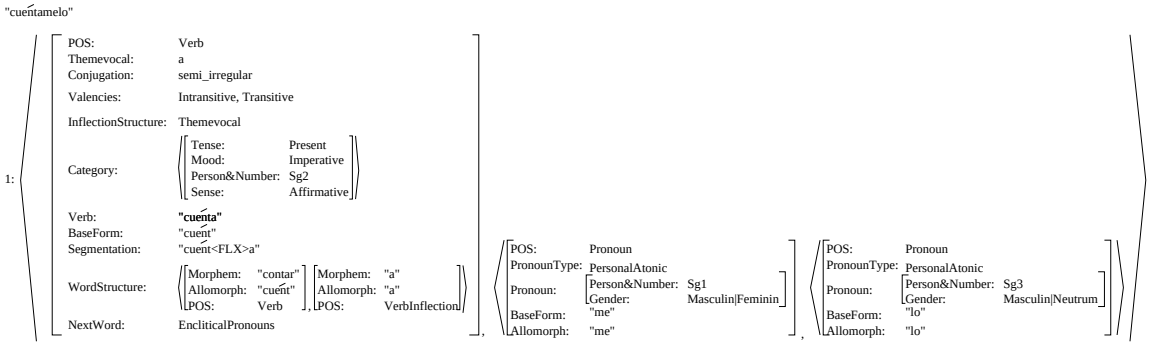

Abbildung 6.4: Analyse für *cuéntamelo*

sich hinsichtlich der Zerlegung in Allomorphe und damit hinsichtlich verwendeter Suffixe. Diese bestimmen unter anderem die Semantik der jeweiligen Lesart. In 3.2 wurden zwei Zerlegungen vorgestellt:
[*in* [[[*util*] *iza*] *ble*]]
[[[*in* [*util*]] *iza*] *ble*]
Die SMM ermittelt entsprechend Abbildung 6.5 die Zerlegungen:
in<DV>*utilizable*
inutil<DV>*iza*<DV>*ble*
in<DV>*util*<DV>*iza*<DV>*ble*
in<DV>*ut*<DV>*il*<DV>*iza*<DV>*ble*
in<DV>*ut*<DV>*il*<DV>*iza*<DV>*ble*
Dabei ist *in* jeweils ein negierendes Präfix, *ble* eine Suffix, das zur resulierenden Wortklasse Adjektiv führt und die Ambiguität des Genus bewirkt. *iza* ist ebenfalls Suffix. Die beiden letzten Zerlegungen sind identisch, werden nur die Oberflächen betrachtet, *il* kann aber sowohl als Suffix als auch als Interfix funktionieren (siehe Abbildung 6.5).

"inutilizable"

1:
```
POS:          Adjective
WordForm:     "inutilizable"
BaseForm:     "inutilizable"
Segmentation: "in<DV>utilizable"
WordStructure: ⟨ [Morphem: "in"  Allomorph: "in"  POS: Prefix] , [Morphem: "utilizable"  Allomorph: "utilizable"  POS: Adjective] ⟩
Prefixes:      ⟨ [PrefixSemType: Negative  BaseForm: "in"  Allomorph: "in"] ⟩
Gender:       Masculin|Feminin
number:       Singular
WellFormed:   yes
Status:       parsed
```

2:
```
POS:          Adjective
WordForm:     "inutilizable"
BaseForm:     "inutilizable"
Segmentation: "inutil<DV>iza<DV>ble"
WordStructure: ⟨ [Morphem: "inútil"  Allomorph: "inutil"  POS: Adjective] , [Morphem: "izo"  Allomorph: "iza"  POS: Suffix] , [Morphem: "ble"  Allomorph: "ble"  POS: Suffix] ⟩
Suffixes:      ⟨ [BaseForm: "izo"  Allomorph: "iza"] , [BaseForm: "ble"  Allomorph: "ble"] ⟩
Gender:       Masculin|Feminin
number:       Singular
WellFormed:   yes
Status:       parsed
```

3:
```
POS:          Adjective
WordForm:     "inutilizable"
BaseForm:     "inutilizable"
Segmentation: "in<DV>util<DV>iza<DV>ble"
WordStructure: ⟨ [Morphem: "in"  Allomorph: "in"  POS: Prefix] , [Morphem: "útil"  Allomorph: "util"  POS: Substantive|Adjective] , [Morphem: "izo"  Allomorph: "iza"  POS: Suffix] , [Morphem: "ble"  Allomorph: "ble"  POS: Suffix] ⟩
Prefixes:      ⟨ [PrefixSemType: Negative  BaseForm: "in"  Allomorph: "in"] ⟩
Suffixes:      ⟨ [BaseForm: "izo"  Allomorph: "iza"] , [BaseForm: "ble"  Allomorph: "ble"] ⟩
Gender:       Masculin|Feminin
number:       Singular
WellFormed:   yes
Status:       parsed
```

4:
```
POS:          Adjective
WordForm:     "inutilizable"
BaseForm:     "inutilizable"
Segmentation: "in<DV>ut<DV>il<DV>iza<DV>ble"
WordStructure: ⟨ [Morphem: "in"  Allomorph: "in"  POS: Prefix] , [Morphem: "uta"  Allomorph: "ut"  POS: Substantive] , [Morphem: "il"  Allomorph: "il"  POS: Interfix] , [Morphem: "izo"  Allomorph: "iza"  POS: Suffix] , [Morphem: "ble"  Allomorph: "ble"  POS: Suffix] ⟩
Prefixes:      ⟨ [PrefixSemType: Negative  BaseForm: "in"  Allomorph: "in"] ⟩
Suffixes:      ⟨ [BaseForm: "izo"  Allomorph: "iza"] , [BaseForm: "ble"  Allomorph: "ble"] ⟩
Gender:       Masculin|Feminin
number:       Singular
WellFormed:   yes
Status:       parsed
```

5:
```
POS:          Adjective
WordForm:     "inutilizable"
BaseForm:     "inutilizable"
Segmentation: "in<DV>ut<DV>il<DV>iza<DV>ble"
WordStructure: ⟨ [Morphem: "in"  Allomorph: "in"  POS: Prefix] , [Morphem: "uta"  Allomorph: "ut"  POS: Substantive] , [Morphem: "il"  Allomorph: "il"  POS: Suffix] , [Morphem: "izo"  Allomorph: "iza"  POS: Suffix] , [Morphem: "ble"  Allomorph: "ble"  POS: Suffix] ⟩
Prefixes:      ⟨ [PrefixSemType: Negative  BaseForm: "in"  Allomorph: "in"] ⟩
Suffixes:      ⟨ [BaseForm: "il"  Allomorph: "il"] , [BaseForm: "izo"  Allomorph: "iza"] , [BaseForm: "ble"  Allomorph: "ble"] ⟩
Gender:       Masculin|Feminin
number:       Singular
WellFormed:   yes
Status:       parsed
```


Abbildung 6.5: Analyse für *inutilizable*

Kapitel 7

Korpora

7.1 Auswahl und Beschaffung von Korpora

Für die spanische Sprache wurden in den letzten Jahren zwei große Korpora zusammengestellt. Zum einen handelt es sich um ein Referenzkorpus der spanischen Gegenwartssprache (Corpus de Referencia del Español Actual), entwickelt vom „Instituto de Lexicografía“ der Real Academia Española in den Jahren 1994 bis 2000.[1] Zum anderen handelt es sich um ein diachronisches Korpus (Corpus Diacrónico del Español), entwickelt ebenfalls vom „Instituto de Lexicografía“ der Real Academia Española im gleichen Zeitraum wie CREA, das Texte von den Anfängen der spanischen Schriftsprache bis 1975 enthält.[2]

CORDE enthält 70 000 000 Wortformen, CREA enthält 200 000 000 Wortformen. Auf beide kann über das Internet zugegriffen werden.[3] Auf eine elektronische Anfrage an die Abteilung für Lexikographie der Real Academia Española vom 21.7.2000 nach der Möglichkeit, beide Korpora für diese Arbeit nutzen zu können, erfolgte die Antwort, dass wegen ungeklärter Rechte weder Teile noch die Korpora als Ganzes zur Verfügung gestellt werden könnten:

> Lamentamos no poder dar curso a su petición, ya que, en estos momentos estamos tratando de resolver la cuestión de los derechos de los textos y, mientras tanto, y hasta que ese complicado tema no se solucione, no podemos hacer ningún tipo de cesión total o parcial del corpus.[4]

Sollte in absehbarer Zeit diese Frage gelöst werden und Zugriff auf die Korpora damit möglich sein, kann die entwickelte Komponente SMM am CRAE und an Texten des 20. Jahrhunderts aus dem CORDE getestet werden.

Eine elektronische Anfrage an die Mailing-Liste „Corpora“ vom 17.1.2000 erbrachte mehrere Hinweise auf Korpora. Darunter waren die schon genannten CRAE

[1] [CREA]. Der Korpus wird hier weiter bezeichnet als CREA.

[2] [CORDE]. Das Korpus wird hier weiter bezeichnet als CORDE.

[3] [Corpus].

[4] E-mail von María José Gil vom 26.7.2000; 15:02:00 (GMT) an die Autorin; [Wir bedauern, Ihrer Bitte nicht nachkommen zu können; da wir im Moment versuchen, die Frage der Rechte an den Texten zu klären, und solange dieses komplizierte Thema keine Lösung hat, können wir in keiner Form des Zugriff auf Teile oder das gesamte Korpus ermöglichen.].

und CORDE sowie das Projekt CRATER (**C**orpus **R**esources **A**nd **T**erminology **E**xt**R**action)und eine Zusammenstellung von Korpora von Mark Davies.[5]

Mark Davies, Professor für spanische Linguistik an der Illinois State University, gibt auf seiner Web site unter anderem an, ein Korpus aus spanischen Zeitungen mit 35 000 000 Wortformen, ein Korpus gesprochener Sprache von 457 Sprechern aus vierzehn lateinamerikanischen und spanischen Städten mit 2 500 000 Wortformen, ein Korpus aus 356 Kurzgeschichen des zwanzigsten Jahrhunderts mit 1 000 000 Wortformen selbst erstellt zu haben. Auf eine elektronische Anfrage vom 1.8.2000 an ihn, das Korpus der spanischen Zeitungen und der Kurzgeschichten nutzen zu dürfen, folgte eine ähnlich gelagerte Antwort wie von der RAE:

> The major problem with the use of the corpus is that it contains copyrighted material, which is why up until now I've only used it for my own research. I'm afraid I'd run into some real legal problems if I released it to anyone else, even with assurances that it would'nt be further distributed, etc., just in case somehow someone did get a hold of a portion of it. [6]

Auf das Korpus des Projektes CRATER konnte erfolgreich über das Internet zugegriffen werden. Es wird in 7.1.1 näher vorgestellt.

Ein zweites Korpus wurde aus einem spanischsprachigen Text gewonnen, der innerhalb des Projektes Gutenberg im Internet erhältlich ist. Darauf wird in 7.1.2 näher eingegangen.

7.1.1 CRATER

Innerhalb des Projektes CRATER wurden Werzeuge und Ressourcen für multilinguistische Korpus-Arbeit entwickelt. Teilnehmer des Projektes sind die Lancaster University (Großbritannien) als Koordinator, die Firma Computers, Communications and Visions (Frankreich) und die Universidad Autónoma de Madrid (Spanien). Sie arbeiten mit IBM-Frankreich und der Abteilung „Escuela Técnica Superior de Ingenerios de Telecomunicación“ an der „Universidad Politécnica de Madrid“ zusammen.[7] Mit JOSÉ CARLOS GONZÁLEZ CRISTÓBAL und AMALIO F. NIETO SERRANO waren daran auch zwei Mitarbeiter des Projektes ARIES beteiligt.

Als Korpus wird das Korpus „International Telecommunications Union“ (ITU) verwendet, das parallele Texte in Englisch, Spanisch und Französisch enthält. Dabei handelt es sich um Texte, die technische Fragen behandeln. Unter anderem wurde

[5] E-mail Susana Sotelo Docó vom 17.1.2000 16:51:32 (MET) und e-mail von Dorothee Graf vom 17.1.2000 16:51:55 (MET) an die Autorin.

[6] E-mail von Mark Davies vom 3.8.2000 17:31:18 (EST); [Das Hauptproblem beim Benutzen des Korpus ist, dass es copyright-geschütztes Material ist, was der Grund dafür ist, dass ich es bis jetzt nur für meine eigenen Forschungen verwendet habe. Ich fürchte, ich würde rechtliche Probleme bekommen, wenn ich es an irgendwen weitergeben würde (selbst mit der Versicherung, dass es nicht weiterverbreitet wird) und jemand irgendwie einen Teil davon zu fassen bekommen würde.].

[7] [CRATER].

ein POS-Tagger für das Spanische entwickelt. Mit diesem Tagger wurde der spanische Teil des ITU-Korpus bearbeitet und anschließend manuell korrigiert. Die korrigierte Version ist über den Server der Universidad Autónoma de Madrid[8] zugänglich und kann online benutzt werden.[9] Die im Rahmen des Projektes erstellten Reporte sind ebenfalls auf dem Server abgelegt, auf sie konnte zum Zeitpunkt der Erstellung dieser Arbeit aber noch nicht zugegriffen werden.

Das handgetaggte Korpus (hier weiter bezeichnet als CRATER-Korpus) umfasst 478 825 laufende Wortformen. Durch Abtrennen der Tags, die sich an die jeweilige Worform nach einem Unterstrich anschließen, wird eine Wortliste erstellt, die als Eingabe für die SMM nutzbar ist. Die akzentuierten Vokale *á*, *é*, *í*, *ó*, *ú* sowie *ü* und *ñ* sind durch `&aacute; &eacute; &iacute; &oacute; &uacute; &uuml; &ntilde;` kodiert. Sie können mit Hilfe eines Perl-Skripts durch die entsprechenden Oberflächen ersetzt werden. Die Unique-Liste enthält alle verschiedenen Wortformen, das heißt, tritt eine Wortform im Korpus mehrfach auf, ist sie in der Unique-Liste nur einmal vorhanden. Sie umfasst 19 224 Wortformen.

7.1.2 Korpus aus dem „Projekt Gutenberg“

Das Projekt Gutenberg existiert seit 1971, begründet durch Michael Hart von der Firma Xerox. Literarische Texte in elektronischer Form werden in ASCII-Format zugänglich gemacht, um hardware- und software unabhängigen Zugriff zu ermöglichen. Die Auswahl der Texte erfolgt zufällig nach dem Geschmack der Mitarbeiter des Projekts. Veröffentlicht werden Texte, für die keine Rechtsansprüche des Autors mehr bestehen. 50 Jahre nach dem Tod eines Autors ist dies der Fall.[10] Darin liegt die Ursache, dass die meisten Texte vor der zweiten Hälfte des zwanzigsten Jahrhunderts verfasst sind. Der überwiegende Teil der Texte sind englischsprachig. Spanische Texte stammen von Cervantes bzw. von anonymen Verfassern des sechzehnten und siebzehnten Jahrhunderts. Diese sind für die Analyse durch die SMM nicht geeignet. Es fand sich nur ein Text neueren Datums, eine Übersetzung des Berichts über den Atombombenabwurf auf Hiroshima und Nagaski.

Mittels anonymen ftp auf den Server der University of North Carolina wurde die ZIP-komprimierte Version des Textes „Los Bombardeos Atómicos de Hiroshima y Nagasaki“ heruntergeladen. Der Text ist eine Übersetzung des englischen Textes „The Atomic Bombings of Hiroshima and Nagasaki“. Als Autor des ursprünglichen Textes ist „United States. Army. Corps of Engineers. Manhatten District“ angegeben.[11]

Der Text hat 25 857 laufende Wortformen, die Unique-Liste umfasst 4 464 Wortformen. Dabei sind Lokutionen nicht als eine Wortform behandelt worden.

[8] Anonymes ftp auf <ftp://ftp.lllf.uam.es/pub/corpus/ITU_spanish_hand-corrected.tar.gz>.
[9] <http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html>.
[10] [Projekt Gutenberg Hintergrund].
[11] [Projekt Gutenberg Catalog].

7.2 Parsen der Korpora

In den Tabellen 7.1 und 7.2 sind die Ergebnisse der Analyse für das CRATER-Korpus und das Korpus aus dem Projekt Gutenberg dargestellt. Für die Liste der laufenden Wortformen wie für die Unique-Liste wird angegeben, wieviele Wortformen die Liste enthält. Da jede der Listen einmal unter Verwendung des reduzierten und einmal unter Verwendung des vollständigen Lexikons geparst wurde, wird jeweils aufgeführt, wieviele Wortformen erkannt wurden, welcher Erkennungsquote dies entspricht und wieviele Wortformen pro Sekunde (WF/s) geparst wurden.

Wortliste	Wort-formen	Erkannt	Quote	WF/s	Ergeb-nisse/WF
laufende WF CRATER	478 825	446 529	93,26	51	3.514
unique WF CRATER	19 223	14 896	77,49	30	5,944
laufende WF PGB	25 831	25 070	97,05	57	3,252
unique WF PGB	4 463	4 268	95,63	21	5,754

Tabelle 7.1: Ergebnisse der Korpus-Analysen mit vollständigem Lexikon

Wortliste	Wort-formen	Erkannt	Quote	WF/s	Ergeb-nisse/WF
laufende WF CRATER	478 825	446 382	93,22	60	2,921
unique WF CRATER	19 223	15 113	78,62	30	5,024
laufende WF PGB	25 831	25 156	97,39	51	2,875
unique WF PGB	4 463	4 307	96,50	25	4,919

Tabelle 7.2: Ergebnisse der Korpus-Analysen mit reduziertem Lexikon

Durchschnittlich werden 90,85% der Wortformen beim Parsen mit vollständigem Lexikon und 91,43% der Wortformen beim Parsen mit reduziertem Lexikon erkannt. Die durchschnittliche Analysegeschwindigkeit liegt bei 41,5 WF/s (vollständiges Lexikon) bzw. 36,25 WF/s (reduziertes Lexikon).

Die getaggte Version des CRATER-Korpus behandelt Lokutionen als eine Wortform. Da dies innerhalb der SMM nicht vorgesehen ist, müssen alle diese Formen als unbekannt klassifiziert werden. Eine vollständige Erkenngsquote ist daher nicht zu erwarten. Da es sich um technische Texte handelt, sind zahlreiche Wortformen aus Buchstaben und Ziffernkombinationen (*1C4, B-7/T.60*) zusammengesetzt. Diese werden nicht erkannt.

Insgesamt sind die Erkennungsquote sowie die Anzahl der ermittelten Analysen für jede Wortform zufriedenstellend. Hochgradige Ambiguität liegt offensichtlich nicht vor. Die Analysegeschwindigkeit ist nicht sehr hoch. Ein Grund liegt in der Vielzahl der Suffixe und Interfixe des Spanischen, die für mögliche Zerlegungen zunächst angenommen, dann aber wieder verworfen werden.

Kapitel 8

Zusammenfassung und Ausblick

Entsprechend der Themenstellung der Arbeit wurde die MALAGA-Morphologie-Komponente SMM zur Analyse spanischer Wortformen entwickelt. Zunächst wurde ein großes Grundformlexikon aus umfangreichen Ressourcen zusammengestellt. Dann wurden Regeln entwickelt, die aus dem Grundformlexikon ein Allomorphlexikon generieren und Regeln, die die Konkatenation der Allomorphe steuern. Die Komponente wurde an zwei großen Korpora des Spanischen getestet.

Um natürlichsprachliche Texte analysieren zu können, sollte ein zugrundeliegendes Lexikon neben den üblichen Wortklassen Namen und Akronyme beinhalten. Die Zahl der im vorliegenden Lexikon enthaltenen Namen und Akronyme ist relativ gering. Sie sollten in weiterführenden Arbeiten ergänzt werden. Wenn auf das Referenz-Korpus der spanischen Gegenwartssprache in absehbarer Zeit ein Zugriff möglich wird, können die dort enthaltenen Namenslisten in das Lexikon integriert werden.

Auf das Problem der Analyse von Wortformen, die nach den formalen morphologischen Prinzipien die Analyse als Derivat oder Kompositum zulassen, wurde bereits eingegangen. Alle Grundformeinträge, die sowohl als Lexem, als auch als Derivat bzw. Kompositum analysiert werden, sollten in folgenden Arbeiten manuell daraufhin überprüft werden, ob tatsächlich Derivation und Komposition möglich sind. Hier konnte nur die maschinelle Bearbeitung durchgeführt werden, die keine Prüfung einzelner Fälle zulässt.

Die SMM analysiert spanische Wortformen nach den Prozessen der Kombination, Derivation und Flexion. Die durchschnittlich geringe Anzahl der Analysen einer Wortform zeigt, dass Flexion, Derivation und Kombination kontrolliert durch die SMM vollzogen werden. Für Komposita, deren Wortklasse morphologisch nicht eindeutig zu bestimmen ist, werden alle Möglichkeiten angegeben. Weiterführende Arbeiten zur Syntax und Semantik des Spanischen sollten sich mit der Auflösung dieser Ambiguitäten beschäftigen. Für die Flexion kann die Generierung nicht wohlgeformter Wortformen ausgeschlossen werden, da diese über Markierungsattribute gesteuert wird und eindeutige Regeln innerhalb der Sprache existieren.

Es werden alle wohlgeformten Derivata analysiert. Kontrolliert erfolgt die Kon-

katenation von Interfix und Suffix für diejenigen Interfixe und Suffixe, die nach [BosqueDemonte 1999] interdependent sind. Die Unterdrückung von Analyseergebnissen, die Derivata kategorisieren, die nicht wohlgeformt sind, gestaltet sich schwierig. Dies ist nicht nur ein Problem der SMM, sondern allgemein in der Beschäftigung mit der spanischen Morphologie anerkannt.

> [...] el establecimiento de reglas seguras es ciertamente complicado. Porque, al contrario de lo que sucede en el nivel sintáctico, los criterios para determinar si un derivado es agramatical o no, no son tan estables como los que deciden las condiciones de aceptabilidad o inaceptabilidad de una oración.[1]

Die Vielzahl der Interfixe und Suffixe, die die spanische Sprache bietet und die relative Offenheit der Konkatenation von Basis und Interfix bzw. Basis und Suffix führt in der Analyse einer Wortform zu relativ vielen Ergebnissen, die als wohlgeformt kategorisiert werden. Die im Vergleich zur EMM geringere Anzahl der Wortformen, die durchschnittlich in einer Sekunde geparst werden, resultiert aus der großen Anzahl der Analysen, die zunächst durch Derivation möglich scheinen, aber im Verlauf der Konkatenation verworfen werden.

Die Analyse der Korpora CRAE und CORDE sollte dazu genutzt werden, für die Derivation eine einschränkendere formale Beschreibung zu finden, so dass stärkere Kontrolle möglich ist. Da beide Korpora im Moment wegen urheberrechtlicher Fragen nicht zur Verfügung stehen, muss diese Aufgabe von Folge-Arbeiten erfüllt werden.

[1] [BosqueDemonte 1999, S. 4653]; [... die Etablierung von sicheren Regeln ist kompliziert. Denn im Gegensatz zur Syntax sind die Kriterien zur Entscheidung, ob ein Derivat agrammatisch ist oder nicht, nicht so stabil wie diejenigen Kriterien, die über die Akzeptanz einer Aussage entscheiden.]

Literaturverzeichnis

[AlcinaBlecua 1994] Juan Alcina Franch/José Manuel Blecua: Gramática española. 9. Auflage, Barcelona 1994.

[Alarcos 1977] Emilio Alarcos Llorach: Gramática estructural. 2. Auflage, Madrid 1977.

[Alarcos 1994] Emilio Alarcos Llorach: Gramática de la lengua española. 4. Auflage, Madrid 1994.

[VOX] Diccionario General de la Lengua VOX [online], 6. Juli 2000 [zitiert am 8.8.2000; 16:03], erhältlich im WWW: <http://www.anaya.es/diccionario/diccionar.htm>.

[ARIES] ARIES Natural Language Tools. Natural Language Processing Group (UPM - UAM) [online], [zitiert am 25.2.2000; 11:47], erhältlich im WWW: <http://www.mat.upm.es:80/ãries/>.

[Beutel 1999] Björn Beutel: Dokumentation für Malaga 4.3 [online], 18.8.1999 [zitiert am 21.2.2000; 13:31], erhältlich im WWW: <http://www.linguistik.uni-erlangen.de/Malaga.de.html>.

[BosqueDemonte 1999] Ignazio Bosque Muñoz/Violeta Demonte Barreto (Hrsg.): Gramática descriptiva de la lengua española, Madrid 1999.

[de Bruyne 1993] Jacques de Bruyne: Spanische Grammatik, Tübingen 1993.

[CREA] CREA [online], zitiert am [16.8.2000; 13:10], erhältlich im WWW: <http://www.rae.es/NIVEL1/CREA.HTM>.

[CORDE] CORDE [online], zitiert am [16.8.2000; 13:10], erhältlich im WWW: <CORDE http://www.rae.es/NIVEL1/CORDE.HTM>.

[CRATER] CRATER [online], 7.5.1996 [zitiert am 11.8.2000; 14:46], erhältlich im WWW: <http://elvira.lllf.uam.es/ fernando/projects/CRATER.html>.

[Corpus] Tipología y características de los corpus [online], [zitiert am 10.8.2000; 17:46], erhältlich im WWW: <http://www.cervantes.es/internet/acad/oeil/Oeiltipo.htm>.

[Davies] Mark Davies: Corpora [online], [zitiert am 11.8.2000; 14:13], erhältlich im WWW: <http://mdavies.for.ilstu.edu/personal/texts.htm>.

[DRAE 1998] Diccionario de la lengua española, edición en cd-rom, Versión 2.0, 21. Auflage, Madrid 1998.

[Comp-jugador] Daniel M. Germán: Web comp-jugador [online], 19.2.1998 [zitiert am 8.8.2000; 17:27], erhältlich im WWW: <http://aries17.uwaterloo.ca/lando/verbos/con-jugador.html>.

[Goñi 1996] José Miguel Goñi: Description of ARIES Natural Language Tools [online], 20.6.1996 [zitiert am 9.2.2000; 12:10], erhältlich im WWW: <http://www.mat.upm.es/ aries/description.html>.

[Goñi et al. 1995a] José Miguel Goñi/José C. González/Antonio Moreno: ARIES: A Lexical Platform for Engineering Spanish Processing Tools. To appear in the journal Natural Language Engineering.

[Goñi et al. 1995b] José Miguel Goñi/José C. González: A framework for lexical representation. In: Proceedings of AI'95: Fiftieenth International Conference. Language Engineering '95, pp. 243 - 252. Montpellier, June 27 - 30, 1995.

[González Collar et al. 1995] Angel Luis González Collar/José Miguel Goñi Menoyo/José Carlos González Cristóbal: Un analizador morfológico para el castellano basado en chart. In: Actas de la VI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'95), Alicante 1995.

[González et al. 1997] José C. González/José M. Goñi/Antonio F. Nieto/Antonio Moreno: The ARIES toolbox: a continuing R+D effort. In: International Workshop on Spanish Language Processing Technologies, Santa Fe 1997.

[Conzález Ollé et al. 1992] Fernando González Ollé/Manuel Casado Velarde: Wortbildungslehre/Formación de palabras, In: Lexikon der romanistischen Linguistik, Bd. 6, Tübingen 1992.

[Halm 1992] Wolfgang Halm: Das spanische Verb, 8. Auflage, Ismaning 1992.

[Hausser 1985] Roland Hausser: Left-Associative Grammar and the Parser NEWCAT, Center for the Study of Language and Information, Stanford University, 1985.

[Hausser 1989a] Roland Hausser: Computation of Language, An Essay on Syntax, Semantics and Pragmatics in Natural Man-Machine Communication, Berlin; New York 1989.

[Hausser 1989b] Roland Hausser: Principle of Computational Morphology, Center for Machine Translation, Carnegie Mellon University, 1989.

[Hausser 1999] Roland Hausser: Foundations of Computational Linguistics, Berlin; Heidelberg; New York 1999.

[Hönigsperger 1992] Astrid Hönigsperger: Flexionslehre/Flexión. In: Lexikon der romanistischen Linguistik, Bd. 6, Tübingen 1992.

[Leidner 1998] Jochen Leidner: Linksassoziative morphologische Analyse des Englischen mit stochastischer Disambiguierung. Magisterarbeit. Friedrich-Alexander-Universität Erlangen-Nürnberg, Abteilung für Computerlinguistik 1998.

[Lorenz 1996] Oliver Lorenz: Automatische Wortformenerkennung für das Deutsche im Rahmen von Malaga. Magisterarbeit. Friedrich-Alexander-Universität Erlangen-Nürnberg, Abteilung für Computerlinguistik 1996.

[Martín 1992] María Antonia Martín Zorraquino: Partikelforschung/Partículas y modalidad. In: Lexikon der romanistischen Linguistik, Bd. 6, Tübingen 1992.

[Monedero et al. 1995] Juan Monedero/José C. González/José Miguel Goñi/Carlos. A. Iglesias/ Amalio F. Nieto: Obtención automática de marcos de Subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. In: XI Conferencia de la Sociedad Española para el Procesamiento de Lenguaje Natural, SEPLN-95, Bilbao 1995.

[Moreno et al. 1995] Antonio Moreno/José Miguel Goñi: GRAMPAL: A morphological model and processor for Spanish implemented in Prolog. In: 1995 Joint Conferencee on Declarative Programming (GULP-PRODE'95), Marina di Vietri 1995.

[Projekt Gutenberg] Project Gutenberg [online], [zitiert am 11.8.2000; 15:05], erhältlich im WWW: <http://promo.net/pg/>.

[Projekt Gutenberg Hintergrund] History and Philosophy of Project Gutenberg [online], [zitiert am 11.8.2000; 15:09], erhältlich im WWW: <http://promo.net/pg/history.html>.

[Projekt Gutenberg Catalog] Catalog Search „ Los Bombardeos Atomicos de Hiroshima y Nagasaki in Spanish“ [online], [zitiert am 11.8.2000; 17:53], erhältlich im WWW: <http://promo.net/cgi-promo/pg/t9.cgi?entry=2367>.

[RAE 1974] Real Academia Española (Comisión de Gramática): Esbozo de una nueva gramática de la lengua española, 2. Auflage Madrid 1974.

[Thiele 1992] Johannes Thiele: Wortbildung der spanischen Gegenwartssprache, Leipzig; Berlin; München 1992.

[Wetzel 1996] Christian Wetzel: Erstellung einer Morphologie für Italienisch in Malaga, Studienarbeit im Fach Informatik. Friedrich-Alexander-Universität Erlangen-Nürnberg 1996.