FAU Forschungen, Reihe B, Medizin, Naturwissenschaft, Technik 6

Besim Kabashi

Automatische Verarbeitung der Morphologie des Albanischen

FAU UNIVERSITY PRESS

Besim Kabashi

Automatische Verarbeitung der Morphologie des Albanischen

FAU Forschungen, Reihe B
Medizin, Naturwissenschaft, Technik
Band 6

Herausgeber der Reihe:
Wissenschaftlicher Beirat der FAU University Press

Besim Kabashi

Automatische Verarbeitung der Morphologie des Albanischen

Erlangen
FAU University Press
2015

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Der vollständige Inhalt des Buchs ist als PDF über den OPUS Server der Friedrich-Alexander-Universität Erlangen-Nürnberg abrufbar:
https://opus4.kobv.de/opus4-fau/home

Verlag und Auslieferung:

FAU University Press, Universitätsstraße 4, 91054 Erlangen

Druck: docupoint GmbH

ISBN: 978-3-944057-40-8 (Druckausgabe)
eISBN: 978-3-944057-43-9 (Online-Ausgabe)
ISSN: 2198-8102

Automatische Verarbeitung der Morphologie des Albanischen

Der Technischen Fakultät
der Friedrich-Alexander-Universität
Erlangen-Nürnberg

zur
Erlangung des Doktorgrades
Doktor-Ingenieur (Dr.-Ing.)

vorgelegt von

Besim Kabashi
aus
Istog

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung	:	17.10.2014
Vorsitzende des Promotionsorgans	:	*Prof. Dr.-Ing. habil. Marion Merklein*
Gutachter	:	*Prof. Dr.-Ing. Günther Görz*
		Prof. Dr. Bardhyl Demiraj
		Prof. Dr.-Ing. Elmar Nöth

Zusammenfassung

Dieses Abstract stellt ein System für die automatische Verarbeitung der albanischen Morphologie vor. Das Ziel des Systems ist ein Werkzeug für die Aufgaben der automatischen Verarbeitung der Rechtschreibung, Lemmatisierung, Annotierung der Wortarten und zuletzt, für die vollständige morphologische Analyse der Wortformen zu sein. Das System kann auch im umgekehrten Modus verwendet werden, d. h. Wortformen aus einem gegebenen Lemma und seinen passenden Eigenschaften zu generieren. Das System ist implementiert im Rahmen von XFST (Xerox Finite-State Tools).
Die Haupteigenschaften des Systems sind :

- Linguistische Abdeckung
  Das System deckt die Flexion der albanischen Nomina, Verben, Adjektive, Numeralia, Adverbien und Pronomina ab. Die nicht flektierenden Wortarten werden ebenso abgedeckt.
  - Lexikon
    Das Lexikon besteht aus separaten Teilen, die nach Wortarten aufgebaut sind. Jeder Teil kann unabhängig von den anderen Teilen benutzt werden. Das Lexikon beinhaltet rund 75 000 Einträge. Es beinhaltet auch Eigen- und Einwohnernamen, geographische Namen, sowie andere Namen.
  - Morphologie
    Die Morphologie besteht aus Flexion und Wortbildung. Die Flexion ist gut abgedeckt, während die Wortbildung nur die Hauptklassen im Sinne der Regularität und Häufigkeit abdeckt.
- Technische Eigenschaften
  XFST kann unter den Betriebssystemen Unix/Linux, Microsoft Windows und Mac OS X verwendet werden. Somit ist auch die (übersetzte) Morphologie-Grammatik auf diesen Systemen benutzbar.
- Tests und Erkennungsraten
  Die entwickelte Morphologie wurde mit verschiedenen Testlisten

aus verschiedenen Quellen getestet: (1) Listen extrahiert aus einem langen Romantext, (2) Listen extrahiert aus einem Textkorpus, und (3) per Hand erstellte Listen, welche selten vorkommende Wortformen beinhalten. Die Erkennungsraten sind zwischen 95 % und 98 %.

Mit diesen Eigenschaften kann das Morphologie-Werkzeug in verschiedenen Fällen und für verschiedene Aufgaben in der maschinellen Sprachverarbeitung des Albanischen benutzt werden. Es kann sowohl eigenständig als auch als Teil eines anderen Systems oder Werkzeugs, z. B. als Teil in einer Umgebung für syntaktische Analyse benutzt werden.
Somit ist es das erste umfangreiche und das größte Werkzeug für die maschinelle Verarbeitung der Morphologie des Albanischen.

Abstract

This abstract presents a system for automatic processing of Albanian morphology. The aim of the system is to be a tool for the tasks of automatic processing for the spelling of word forms, lemmatization, POS-tagging, and, at last, for full morphological analysis of word forms. The system can also be used in reverse mode, i. e. to generate word forms from a given lemma and its corresponding attributes. The system is implemented using the XFST (Xerox Finite-State Tools).
The main properties of the system are :

- Linguistic coverage
  The system covers the inflection of Albanian nouns, verbs, adjectives, numerals, adverbs and pronouns. The non-inflectional parts of speech are covered also.
  - Lexicon
    The lexicon consists of separate parts, which have been compiled based on POS. Each part can be used separately from the other parts. The lexicon contains about 75 000 entries. It also contains proper, inhabitant, geographical and other names.
  - Morphology
    The morphology part consists of inflection and word formation. The inflection is well covered, while word formation covers only the main classes in the sense of regularity and frequency.
- Technical properties
  XFST can be used on Unix/Linux, Microsoft Windows, and Mac OS X operating systems. Therefore, the compiled morphology-grammar is also portable to these systems.
- Tests and recognition rate
  The developed morphology is tested against several test lists from different sources : (1) lists extracted from a rich novel text, (2) lists extracted from a text corpus, and (3) lists compiled manually, which contain

rarely occurring word forms. The recognition rate is between 95 % and 98 %.

With these attributes the morphology tool can be used in different cases and for different tasks in natural language processing of Albanian. It can be used both independently and as part of another system or tool, e. g. as part in a framework for syntactic analysis.
So far it is the first rich and the largest tool for automatic processing of the morphology of the Albanian language.

Danksagung

Herrn Prof. Dr.-Ing Günther Görz danke ich für die bereitwillige Betreuung der vorliegenden Arbeit. Im Laufe der Jahre lernte ich sehr viel aus seinen Lehrveranstaltungen, Vorträgen und persönlichen Gesprächen.
Herrn Prof. Dr. Bardhyl Demiraj, München, Herrn Prof. Dr.-Ing Elmar Nöth und Herrn Prof. Dr. Stefan Evert danke ich für wertvolle Gespräche und für zahlreiche Vorschläge, die vorliegende Arbeit in ihrem Werdegang zu verbessern. Dem letzteren danke zusätzlich für die Bereitschaft als Prüfer zu agieren.
Herrn Prof. Dr. Klaus Mayer-Wegener danke ich für die bereitwillige Übernahme des Vorsitzes der Prüfungskommision.
Ein besonderer Dank gilt Frau Prof. Dr. Mechthild Habermann. Ihre Lehrveranstaltungen und Vorträge motivierten mich Sprachwissenschaft zu studieren. Sie gab mir ständig gute Ratschläge über viele Jahre hinweg.
Herr Prof. Dr. Rexhep Ismajli, Prishtina, Herr Prof. Dr. Rami Memushaj, Tirana und Herr Prof. Dr. Marko Snoj, Ljubljana, haben mir zahlreiche lexikalische Einträge zur Verfügung gestellt, so dass für ein Teil des Lexikons, ca. 25%, das manuelle Eintragen der Einheiten gespart werden konnte, wofür ich mich sehr bedanke.
Meiner Kollegin Frau Gabriella Lapesa, sowie meinen Kollegen bzw. ehemaligen Kollegen, Herrn Andreas Blombach M. A., Paul Greiner M. A., Dr.-Ing. Peter Reiß, Dr. Peter Uhrig, insbesondere Thomas Proisl M. A. und Matthias Bethke M. A., alle Universität Erlangen-Nürnberg, danke ich für den wertvollen Gedankenaustausch zum Thema maschinelle Sprachverarbeitung.
Herrn Martin Schmitt danke ich für die Zusammenarbeit zum Thema XFST im Rahmen der Vor- und Nachbereitung der Tutorien zu meinen Lehrveranstaltungen *Einführung in die Grammatikentwicklung* (Sommer 2013) und *Einführung in die maschinelle Sprachverarbeitung* (Winter 2013/-14). In diesem Zusammenhang danke ich auch Herrn Patrick Claus.
Für den Inhalt bin allein ich verantwortlich – selbstverständlich.

Erlangen, Ende Oktober 2014 Besim Kabashi

Meiner lieben Familie.

Inhaltsverzeichnis

1 Einleitung

Da die sprachlichen Daten aus Korpora aus dem tatsächlichen Sprachgebrauch stammen, also nicht einfach zu Anschauungszwecken konstruiert sind, können sie wertvolle Informationen zu einer Sprache liefern bzw. interessante Aufschlüsse über eine Sprache erlauben. So geben sie z. B. genauere Auskunft über die Verbindungen der einzelnen Wörter zueinander, deren Position im Satz, die Konstruktionen und Textsorten, in denen sie vorkommen, sowie ihre Häufigkeit.
Je mehr sprachliches Material ausgewertet wird, desto schlüssigere Daten werden gewonnen, desto höher ist die Wahrscheinlichkeit, dass auch die selten vorkommenden Phänomene abgedeckt und somit berücksichtigt werden können. Doch ist in diesen Fällen die Masse der zu untersuchenden Daten oft derart groß, dass eine manuelle Überprüfung, nicht zuletzt aus zeitlichen Gründen, nicht möglich ist. Es bietet sich daher an, das vorhandene sprachliche Material komplett oder wenigstens teilweise maschinell verarbeiten zu lassen. Um dies zu erreichen, müssen zuerst die entsprechenden Ressourcen und Werkzeuge erstellt werden.[1]

1.1 Die Motivation

Für viele Sprachen, z. B. für das Englische, das Französische, das Deutsche oder das Italienische, gibt es bereits zahlreiche Ressourcen und Anwendungen, die erfolgreich eingesetzt werden. Sie beschleunigen die linguistischen Untersuchungen und sind im Rahmen der maschinellen Sprachverarbeitung (MSV) eine enorme Unterstützung bei der Entwicklung weiterer Komponenten, die zu verschiedenen Zwecken im Bereich der linguistischen Datenverarbeitung eingesetzt werden.
Für das Albanische lässt der Stand der Entwicklung jedoch zu wünschen übrig – umso mehr, als die zunehmende elektronische Kommunikation die fehlenden Ressourcen und entsprechend darauf aufbauenden Komponenten

[1] Mit dem Terminus *Ressource* sind im Folgenden die Hilfs- und/oder Einsatzmittel gemeint, die für die maschinelle Sprachverarbeitung gebraucht werden. Auf das Thema wird genauer im Kapitel 2 eingegangen.

der Systeme für MSV bemerkbar macht. Eine Internetsuche, bspw. eine Suche beim Projekt UNITEX [2], nach einem Lexikon für das Albanische, liefert keine positiven Ergebnisse – dabei ist ein solches Lexikon eine elementare Ressource für die MSV, die in vielen Bereichen unabdingbar ist. Auch weitere Ressourcen, wie bspw. systematisch organisierte morphologische Informationen, die für den Bau einer Komponente für die Wortformanalyse sehr hilfreich wären, sind nicht vorhanden. Diese Informationen, die über mehrere Stellen verteilt sind, die sich nicht gebündelt finden lassen, müssen zuerst gesammelt und systematisiert werden, um sie später leichter auswerten zu können. Die maschinelle Sprachverabeitung für das Albanische muss sich also zuerst mit grundlegenden Fragen und Aufgaben beschäftigen, die größtenteils im Rahmen der Sprachwissenschaft hätten gelöst werden können.
Es stellen sich also folgende Fragen: *Wieso fehlen diese Ressourcen? Wieso ist bis jetzt kein maschinenlesbares Lexikon gebaut worden? Wieso ist keine automatische Morphologie entwickelt worden, welche eingesetzt werden könnte?* Auf diese Fragen gibt es keine einfachen Antworten. In zusammengefasster Form lassen sich jedoch als Erstes politisch-historische Bedingungen sowie wirtschaftliche Interessen erwähnen. Albanien war seit dem Entstehen der Computerlinguistik bis in die 90er Jahre politisch und wirtschaftlich vom Rest Europas und zeitweise sogar vom Rest der Welt isoliert. Den anderen albanischsprachigen Gebieten im südosteuropäischen Raum (die heute etwa die Hälfte der Gesamtgebiete ausmachen) im ehemaligen Jugoslawien und in Griechenland ging es wirtschaftlich zwar besser, politisch aber deutlich schlechter. Das Albanische hatte, mit Ausnahme des Kosovo (Amselfeld), wo es einen hervorgehobenen Status als Amtssprache der föderativen Einheit besaß, den Status einer Minderheitensprache, so wie in Montenegro, Serbien und Mazedonien, oder wurde ganz verboten, so wie in Griechenland.[3] Zu den genannten Umständen kommt die Tatsache, dass an den albanischen Universitäten und anderen wissenschaftlichen Institutionen kein entsprechendes Fach vorhanden war, das sich mit dem Thema befasst hätte – weder in Tirana (Albanien) noch in Prishtina (Kosovo).
Diese „Lücke" lässt sich in Zeiten der rasanten wirtschaftlichen und insbesondere technischen Entwicklungen nicht schnell und nicht leicht füllen. Im Bereich der elektronischen Kommunikation findet die natürliche Sprache immer mehr Anwendung. Einige einfache Anwendungen, z. B. die sogenannten Sprach-Lokalisierungen werden für die Sprachen mit einer großen

[2] Université Paris-Est Marne-la-Vallée; Online: <http://igm.univ-mlv.fr/~unitex/> , letzter Zugriff am 14.7.2014 – im Folgenden Fällen wird er nur als Datum angegeben.

[3] Im Abschnitt 3.1 wird auf den heutigen Status des Albanischen in den genannten Gebieten eingegangen. Es werden weitere sprachspezifische Details vorgelegt.

Zahl von Sprechern seit einer Weile erfolgreich umgesetzt, von mobilen Telefonen bis hin zu Bankautomaten. Das Albanische steht gerade in den Anfängen.
Aus der vorgestellten Lage ergibt sich die Motivation, primäre sprachliche Ressourcen für das Albanische zu entwickeln, insbesondere lexikalische und morphologische, die als Bausteine für maschinelle Sprachverarbeitung verwendet werden können.

1.2 Das Ziel

Da beide Ressourcen, sowohl die lexikographischen als auch die morphologischen, fehlen, ist es entscheidend, beide Teilbereiche auf einmal im Rahmen der vorliegenden Arbeit zu behandeln. Der Weg zum Ziel ist deswegen zweigleisig: ein maschinenlesbares Lexikon zu bauen, sowie ein Morphologie-System zu entwickeln.
Als Erstes wird ein elektronisches maschinenlesbares Lexikon erstellt, das für die Zwecke der maschinellen Sprachverarbeitung geeignet ist und die neuesten wissenschaftlich-technischen Entwicklungen im Bereich der elektronischen Computerlexikographie berücksichtigt.
Anschließend wird das Werk [Kabashi 2003], welches das Verbalsystem behandelt, überarbeitet[4] und mit dem Nominalsystem erweitert. Es wird also die Entwicklung einer – soweit wie möglich – vollständigen Morphologie zum Ziel gesetzt, die als Werkzeug für das morphologische Annotieren (engl. *tagging*) dienen kann.

1.3 Die Aufgaben

Um das genannte Ziel zu erreichen, wird das Lösen einiger Aufgaben notwendig. In erster Linie verteilen sich die Aufgaben auf zwei Themenbereiche:

1. Das Auffinden und Sammeln des linguistischen Materials und des linguistischen Wissens und deren Organisation in einer gebräuchlichen elektronischen Form, die sich für die Zwecke der maschinellen Sprachverarbeitung eignet. Diese Aufgaben sind u. a. die Klassifikation des Wortschatzes, Angabe ausreichender Informationen zur Flexion und ggf. zur Wortbildung, Rektion der Präpositionen, Angaben zur Valenz, welche für die Abstraktionsebene der Syntax notwendig sind,

[4] Ab dem Kapitel 4 wird näher auf diese Einzelheiten eingegangen.

jedoch gewöhnlich im Rahmen der Morphologie-Komponente verarbeitet werden.

2. Die Organisation und Strukturierung der Informationen, sodass sie im Rahmen der MSV implementiert werden können. Es ist das Format der sprachlichen Daten gemeint, die im Rahmen eines Systems oder mehrerer Systeme für maschinelle Sprachverarbeitung verwendet werden. Ebenso ist es notwendig, dass z.B. die linguistischen Daten so organisiert sind, dass ggf. eine Klasse von Wörtern mit einer einzigen Regel verarbeitet werden kann, d. h. Verarbeitung der Informationen in verallgemeinerter Form bzw. mit Hilfe von Ausnahmeregeln.

Die Erstellung des Lexikons und die damit verbundenen Aufgaben machen einen Großteil der Arbeit aus, mindestens die Hälfte. Das Lexikon ist der grundlegende Baustein und eine Voraussetzung für die Entwicklung vieler Ressourcen – auch der Morphologie.
Die Morphologie sollte, aufbauend auf dem erstellten Lexikon, das Taggen bzw. das Analysieren der Flexion und zum Teil der Wortbildung des Albanischen möglich machen, sodass beide Komponenten, sowohl von der linguistischen Seite als auch von der technischen Seite her gesehen, einsetzbar sind.

2 Stand der Forschung

Mit der Entwicklung elektronischer Maschinen, insbesondere des Personal Computers, fand die maschinelle Verarbeitung der Sprache sehr schnell Interesse. Die Automatisierung einiger mit natürlicher Sprache verbundener Aufgaben wurde mit der Zeit zum Ziel.
Ein solches Ziel war mit vielen Schwierigkeiten verbunden, wie z. B. der Darstellung verschiedener Schriftzeichen und der Vorbereitung und Erstellung von Ressourcen, die als Vorlage für gewünschte Anwendungen nötig waren.
Eine der ersten Ressourcen waren die Lexika. Sie wurden mit der Zeit immer größer, sodass deren Strukturierung sich als eine Alternative entwickelte, die viele Vorteile mit sich brachte. Sie deckte zugleich neben einzelnen Wortformen auch die damit verbundenen Prozesse der Flexion und der Wortbildung ab. So entstand ein Teilbereich namens maschinelle Morphologie.[5]
Sie ist heute eines der meistuntersuchten Felder im Bereich der Sprachverarbeitung und zählt zu ihren wichtigsten Themen. Denn sie kann sowohl als selbstständige Komponente eingesetzt werden als auch viele andere Komponenten bedienen, wie etwa eine für Volltextsuche, eine für morphologische Korpusannotation oder eine Syntaxkomponente.
Im Folgenden wird in den wichtigsten Punkten kurz sowohl auf den Stand der Forschung im Allgemeinen eingegangen, d. h. im technischen Sinne, sprachunabhängig, als auch auf den Stand der Forschung im Bereich der Albanologie mit dem Ziel, einen – eher allgemeinen – Überblick zu geben. Zunächst werden im Abschnitt 2.1 die Ressourcen für die maschinelle Sprachverarbeitung vorgestellt. Desweiteren werden im Abschnitt 2.2 lexikalische Ressourcen und maschinelle Lexikographie besprochen. Es folgt das Thema maschinelle Morphologie, (2.3). Danach gilt die Aufmerksamkeit dem Thema albanisch, konkret den Ressourcen (2.4), der maschinellen Lexikographie (2.5) und schließlich der maschinellen Morphologie (2.6). Nach der Einführung dieser Themen wird die Zielsetzung der vorliegenden Arbeit (2.7) hervorgehoben, gefolgt von einer Zusammenfassung des Kapitels (2.8).

[5] Vgl. [Sproat 1992: (§ 1), 1–14] für eine Einführung.

2.1 Ressourcen und maschinelle Sprachverarbeitung

Bis es zu einem fertigen Produkt der linguistischen maschinellen Verarbeitung kommt, werden Daten gebraucht, die in verschiedenen Formen vorliegen oder erst verarbeitet werden müssen. Diese Daten werden Sprachressourcen oder linguistische Ressourcen genannt.
Sie sind durch viele Eigenschaften gekennzeichnet. Demnach werden die Ressourcen auch klassifiziert und verwendet. Zu ihren wichtigsten Typen zählen Texte und Aufzeichnungen gesprochener Sprache, organisiert in Form von Korpora, viele Typen sprachlicher Daten, z. B. die Sammlung des Wortschatzes einer Sprache, je nach Fall deren linguistische Bearbeitung, und schließlich deren Organisation in Form von Lexika. Bei diesen Aufzählungen können diejenigen Ressourcen unterschieden werden, die nicht vorverarbeitet sind (Rohdaten), wie z. B. Texte und solche, die verarbeitet sind, wie z. B. Lexika.
Ein Korpus ist eine Sammlung von gesprochener und/oder geschriebener Sprache, die für linguistische Untersuchungen verwendet werden kann.[6] Da diese Form der Sprache bei der Entstehung nicht (oder sehr wenig) von organisatorischen Faktoren der Datensammlung beeinflusst ist und später für den einen oder anderen Zweck nicht geändert bzw. angepasst werden darf, zählen die Korpora als ursprüngliche Sprachdaten, d. h. als primäre Sprachressourcen.
Sie können in verschiedenen Formen gebraucht werden: nicht verarbeitet, d. h. roh, oder verarbeitet in verschiedenen Hinsichten, d. h. annotiert. Die letzteren sind entweder automatisch verarbeitet mithilfe einer entsprechenden Komponente oder von Fachkräften analysiert. Oft wird in diesen Fällen eine nicht allzu große Menge an Daten sorgfältig, *manuell*, mit den gewünschten Informationen versehen, und dient später dazu die entsprechenden Werkzeuge, bspw. einen Tagger, zu trainieren. Die trainierten Werkzeuge werden dann für die maschinelle Verarbeitung großer Mengen von Daten eingesetzt.
Die ersten elektronischen Korpora, die unter Berücksichtigung der Zwecke der maschinellen Sprachverarbeitung entwickelt wurden, entstanden in den 1960er und 1970er Jahren – das Brown Corpus im Jahr 1964, das LiMaS-Korpus im Jahr 1971. Die wohl bekanntesten Korpora sind die

[6] Eine ähnliche kurze und allgemeine Definition ist auch bei [Kennedy 1987: 1] zu finden. Ausführliche aktuelle deutschsprachige Literatur über Korpora bieten Lemnitzer und Zinsmeister [2010]. Eine umfangreiche Artikelsammlung, die den aktuellen Wissenstand des Fachgebietes auf internationaler Ebene darstellt, ist das Werk [Lüdeling/Kytö], Band 1 [2008], und Band 2 [2009].

folgenden: LOB (Lancaster-Oslo-Bergen Corpus), BNC (British National Corpus)[7], IDS-Korpus/DeReKo (Das Deutsche Referenzkorpus)[8], das über das *Corpus Search, Management and Analysis System* (Cosmas) abgefragt werden kann[9], DWDS (Digitales Wörterbuch der deutschen Sprache)[10], Deutscher Wortschatz (Institut für Informatik, Universität Leipzig)[11], sowie die in Form von Baumbanken geparsten Korpora The Penn Treebank Project[12], TIGER[13] und NEGRA[14]. Im folgenden Listing (2.1) ist der erste Satz des Limas-Korpus angegeben – aus technischen Gründen verteilt auf vier Zeilen. Der Text ist mit keinerlei Informationen versehen, also nicht annotiert.

Listing 2.1: Der Anfang des Limas-Korpuses

```
001  Als es zu schneien aufgehört hatte, verließ Johanna von
002  Rotenhoff, ohne ein rechtes Ziel zu haben, das Gutshaus.
003  Mechanisch einen Fuß vor den anderen setzend, schlug sie den
                 Weg
004  zum verschneiten Park ein. [ ... ]
...
```

Ein solcher Text kann zu verschiedenen Zwecken benutzt werden, wie z. B. zur Erstellung von Frequenzlisten, zur Feststellung der Satzlänge oder der Wortstellung verschiedener Wortformen. Um die Nutzung von solchen Texten zu erleichtern, werden sie mit den jeweiligen Informationen versehen. Eine Suche nach einem Adjektiv wäre nur mithilfe regulärer Ausdrücke möglich, indem bestimmte Endungen wie z. B. „*-lich*+Flexionssuffixe“ abgefangen würden. Allerdings endet natürlich nicht jedes Adjektiv auf „*-lich*+Flexionssuffixe“. Eine Suche nach der Präposition „*auf*“ lieferte als Ergebnis auch die gleichlautende Verbpartikel aus den Fällen, in denen sie getrennt vom zugehörigen Verb vorkommt. Um in diesen Situationen, die sehr zahlreich sein können, das Korpus möglichst schnell und linguistisch zuverlässig abzufragen, bedarf es einer Analyse der Korpusdaten im Voraus.

[7] Vgl. <http://www.natcorp.ox.ac.uk/corpus/>, 22.7.2014.

[8] Vgl. <http://www.ids-mannheim.de/kl/projekte/korpora/>, 22.7.2014. Das Korpus hat über 5,4 Milliarden Textwörter (Stand 29.02.2012).

[9] Vgl. <http://www.ids-mannheim.de/cosmas2/>, 22.7.2014.

[10] Vgl. <http://www.dwds.de/>, 22.7.2014.

[11] Vgl. <http://wortschatz.uni-leipzig.de/>, 22.7.2014.

[12] Vgl. <http://www.cis.upenn.edu/~treebank/>, 22.7.2014.

[13] Vgl. <http://www.ims.uni-stuttgart.de/projekte/TIGER/>, 22.7.2014.

[14] Vgl. <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html/>, 22.7.2014. Bei dieser Auswahl wurden die Ressourcen der deutschen Sprache im Vergleich zu Ressourcen anderer Sprachen bevorzugt.

Das folgende Listing (2.2) zeigt ein Beispiel (aus ca01[15]), wie ein Korpus mit linguistischen Informationen versehen werden kann.

Listing 2.2: Ein Satz aus dem Brown Corpus

```
The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl
said/vbd Friday/nr an/at investigation/nn of/in
Atlanta's/np$ recent/jj primary/nn election/nn
produced/vbd "/" no/at evidence/nn "/" that/cs
any/dti irregularities/nns took/vbd place/nn ./.
...
```

Die Art und die Menge der analysierten Informationen, die einem Korpus hinzugefügt werden kann, ist unbeschränkt. Dennoch haben sich aus praktischen Gründen einige Modelle etabliert, wie z. B. Part-of-Speech-Tagging, kurz POS-Tagging, Chunking und Parsing. Ein Beispiel für POS-Tagging findet sich in Listing 2.2. Dabei wird die Wortart-Information in verschiedenen Formen und Ausführungen angegeben, von der einfachen traditionellen bis zur ausführlichen morphologischen Information mit allen möglichen Kategorien. Der nächste Typ der Annotation ist die morpho-syntaktische, welche mit einem sogenannten Chunker verarbeitet wird. Dabei werden zusammengehörige Einheiten im Satz annotiert, um die Weiterverarbeitung im Bereich Syntax zu erleichtern, indem Teilstrukturen der Sätze für die gesamte Satzverarbeitung vorbereitet werden.[16]

Listing 2.3: Der gechunkte Anfang des Limas-Korpuses (27. Satz)

```
<NC>
Es      PPER    es
</NC>
<VC>
gab     VVFIN   geben
</VC>
kaum    ADV     kaum
<NC>
einen   ART     eine
Wunsch  NN      Wunsch
</NC>
,       $,      ,
<NC>
der     PRELS   die
</NC>
<NC>
den     ART     die
Kindern NN      Kind
</NC>
nicht   PTKNEG  nicht
```

[15] Vgl. <http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/brown_tei.zip>, 24.7.2014.

[16] Die flache Satzverarbeitung in Listing 2.3 hat nicht zwingend nur vorbereitende Funktion. Sie kann gleichzeitig für verschiedene Zwecke dienen.

```
<VC>
erfüllt VVPP    erfüllen
wurde   VAFIN   werden
</VC>
.       $.      .
```

Das nächste Listing zeigt einen geparsten Satz (aus dem Limas-Korpus[17]), wobei zu den Angaben, die auch beim Tagging gemacht werden (Wortform, Lemmatisierung, Wortartspezifizierung und morphologische Angaben), noch Angaben auf Satzebene, wie die syntaktischen Abhängigkeiten der Wörter untereinander (Spalten[18] fünf und sechs), hinzukommen.[19]

Listing 2.4: Der geparste Anfang des Limas-Korpuses (zweiter Satz)

```
1  Mechanisch     Mechanisch  ADJD     Pos                9  MO
2  einen          ein         ART      Acc|Sg|Masc        3  NK
3  Fuß            Fuß         NN       Acc|Sg|Masc        1  MO
4  vor            vor         APPR     _                  3  MNR
5  den            der         ART      Dat|Pl|Neut        4  NK
6  anderen        anderer     ADJA     Dat|Pl|Neut        4  NK
7  setzend        setzend     ADJD     Pos                1  CJ
8  ,              _           $,       _                  9  PUNC
9  schlug         schlagen    VVFIN    3|Sg|Past|Ind      0  ROOT
10 sie            sie         PPER     3|Nom|Sg|Fem       9  SB
11 den            der         ART      Acc|Sg|Masc       12  NK
12 Weg            Weg         NN       Acc|Sg|Masc        9  OA
13 zum            zu          APPRART  Dat|Sg|Neut       12  MNR
14 verschneiten   verschneit  ADJA     Pos|Dat|Sg|Neut   13  NK
15 Park           Park        NN       Dat|Sg|Neut       13  NK
16 ein            ein         PTKVZ    _                  9  SVP
17 .              _           $.       _                  9  PUNC
```

Viele Korpora werden in einer Auszeichnungsprache (engl. Markup Language), vorwiegend XML (eXtensible Markup Language) kodiert, wie z. B. [BNC–XML 2007]. Da XML eine standardisierte Auszeichnungsprache ist, die sehr verbreitet ist und von vielerlei Programmen und Werkzeugen unterstützt wird, haben die damit kodierten Ressourcen viele Vorteile gegenüber denjenigen, die in proprietären oder weniger oft verbreiteten Formaten kodiert sind. Die in XML kodierten Ressourcen können mittels Sprachen

[17] Vgl. <http://www.korpora.org/Limas/> , 24.7.2014.

[18] Der mate-tools-Parser gibt weitere Spalten aus, die redundant zu einander sind und hier aus Platzgründen weggelassen wurden.

[19] Die Ausgabe des Parsers entspricht das Format CoNLL-2009 ([Conference on] Computational Natural Language Learning [2009]), one word per line format (tagged and lemmatized), vgl. <http://ufal.mff.cuni.cz/conll2009-st/task-description.html> , 24.7.2014. Der vorgestellte Satz in Listing 2.3 wurde mit TreeTagger (IMS, Universität Stuttgart), vgl. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>, verarbeitet, der Satz in Listing 2.4 mit Mate-Tools, vgl. <http://code.google.com/p/mate-tools/> , 24.7.2014;

wie z. B. XQuery abgefragt werden und mit verschiedenen Werkzeugen für Auszeichnungssprachen wie z. B. mit DTD (Document Type Definition), XML-Schema usw. insbesondere auf Wohlgeformtheit und Validität geprüft werden. Sie können z. B. mit XSLT (Extensible Stylesheet Language Transformations) verarbeitet, d. h. in verschiedene Formaten transformiert werden.

2.2 Lexikalische Ressourcen und maschinelle Lexikographie

Ein anderer Ressourcentyp sind die lexikalischen Ressourcen. Im Vergleich zu Korpora sind sie weitgehend bis vollständig unabhängig von Quellen, in einer dem Verwendungszweck angemessenen Form organisiert.[20] Sie beschreiben die Lexik (Wortschatz) einer Sprache oder einer Teilsprache bzw. eines Dialektes, indem sie die einzelnen Einheiten selektiv mit den jeweiligen Eigenschaften versehen (z. B. Genus) bzw. mit den jeweiligen Informationen ihrer Zugehörigkeit in eine Gruppe einteilen (z. B. Wortart oder Flexionsklasse).
Unter den vielen möglichen Anwendungsgebieten lexikalischer Ressourcen sind die häufigsten folgende: Orthographie (Rechtschreibung), Silbentrennung (für Zeilenumbrüche im Rahmen der Textverarbeitung), Phonetik und Phonologie (für Text-to-Speech-Systeme; Spracherkennung und -produktion), morphologische Analyse und Produktion (Wortformerkennung und -generierung), syntaktische Analyse und Produktion sowie die Verarbeitung der Sprache hinsichtlich Semantik und Pragmatik (Bedeutungserschließung und Disambiguierung).
Maschinelle Lexikographie[21] beschäftigt sich mit der automatischen bzw. halbautomatischen Erstellung der lexikalischen Daten. Im Vergleich zur traditionellen, herkömmlichen Lexikographie, wo die Daten *händisch* aufgesammelt und bearbeitet werden, ermöglicht die Computerlexikographie eine schnellere und präzisere Untersuchung und Bearbeitung der Daten. Es können größere Quellen untersucht werden. Neben anderen Lexika in erster Linie verschiedene Texte (als Primärquellen); dabei ist natürlich Voraussetzung, dass diese in elektronischer Form zur Verfügung stehen. In den Anfängen, d. h. in den 1970er Jahren, wurden die Daten sogar noch auf Lochkarten gespeichert.[22]

[20] Belege und Verwendungsbeispiele wären Bezüge auf die Quellen.
[21] Vgl. hierzu z. B. [Kunze/Lemnitzer 2007].
[22] Siehe [Hess et al. 1983] und [Schaeder/Willée 1989] für ausführliche Informationen über die Anfänge der maschinellen Lexikographie.

Aufgabe der Computerlexikographie ist die Auffindung der auszuwertenden Quellen, Sammlung der Daten nach vorgegebenen Kriterien sowie die Klassifikation und Organisation der lexikalischen Daten in einem gängigen Datenformat, das unkompliziert von möglichst vielen Systemen der MSV benutzt werden kann. Im besten Fall sind die lexikalischen Daten in einem „offenen“ und standardisierten Format gespeichert.

2.2.1 Orthographische Ressourcen

Ein gutes Beispiel für orthographische Ressourcen ist ein Teil der lexikalischen Datenbank CELEX, die im Max-Planck-Institut in Nijmegen entwickelt wurde. Sie bietet wertvolle Informationen für die Sprachen Englisch, Deutsch und Niederländisch in den Bereichen Phonetik (Kodierung der Aussprache), Orthographie (Rechtschreibung und Silbentrennung, für Lemmata und ihre Flexionsformen), Morphologie (grammatische Angaben, Typ der Konjugation, Deklination, grammatische Angaben der einzelnen Formen sowie Segmentierung der Wortbildung) sowie ein wenig Syntax, wobei Angaben zur Verbvalenz kodiert sind.
Die folgende Abbildung zeigt einen Abschnitt aus der Datenbank CELEX (Datei `german/gol/gol.cd`). In der zweiten Spalte (zweites Feld) sind die Lemmata angegeben, in der vierten die Silbentrennung, in der sechsten Spalte ist z. B. *sprech* (Lemma ohne Infinitivendung *en*) vs. *sprechen* zu finden.

Listing 2.5: CELEX-GOL

```
1
2  37553\sprechen\3598\spre-chen\N\sprech\sprech\N
3
4  37554\Sprechanlage\3\Sprech-an-la-ge\N\Sprechanlage\Sprech-an-la-
                    ge\N
5
6  37555\Sprechchor\18\Sprech-chor\N\Sprechchor\Sprech-chor\N
7
8  37556\Sprecher\479\Spre-cher\N\Sprecher\Spre-cher\N
9
10 37557\sprecherisch\0\spre-che-risch\N\sprecherisch\spre-che-risch\
                    N
11
12 37558\Sprecherziehung\1\Sprech-er-zie-hung\N\Sprecherziehung\
                    Sprech-er-zie-hung\N
13
14 37559\Sprechfunk\8\Sprech-funk\N\Sprechfunk\Sprech-funk\N
15
16 37560\Sprechgesang\1\Sprech-ge-sang\N\Sprechgesang\Sprech-ge-sang\
                    N
17
18 37561\Sprechkunde\0\Sprech-kun-de\N\Sprechkunde\Sprech-kun-de\N
19
20 37562\sprechkundlich\0\sprech-kund-lich\N\sprechkundlich\sprech-
                    kund-lich\N
```

Während Listing 2.5 die orthographischen Eigenschaften der Lemmata darstellt, sind im folgenden Listing 2.6 die Eigenschaften der Wortformen angegeben.[23] Es handelt sich um einen Abschnitt aus der Datenbank CELEX (Datei `german/gol/gow.cd`). Die in dieser Ressource enthaltenen Informationen sind wichtig, da die flektierten Formen der Wörter in einigen Fällen, sowohl was die Rechtschreibung als auch was die Silbentrennung betrifft, anders sein können als die sogenannte Grundform, also der Wörterbucheintrag.

Listing 2.6: CELEX-GOW

```
73411\sprichst\13\37553\sprichst\N
73412\spricht\712\37553\spricht\N
73413\sprechen\1097\37553\spre-chen\N
73414\sprecht\2\37553\sprecht\N
73415\sprachen\230\37553\spra-chen\N
73416\spracht\0\37553\spracht\N
73417\sprechest\0\37553\spre-chest\N
73418\sprechet\0\37553\spre-chet\N
73419\spr"ache\6\37553\spr"a-che\N
73420\spr"achst\0\37553\spr"achst\N
73421\spr"achen\7\37553\spr"a-chen\N
73422\spr"achet\0\37553\spr"a-chet\N
73423\spr"acht\0\37553\spr"acht\N
73424\sprechend\8\37553\spre-chend\N
73425\gesprochen\616\37553\ge-spro-chen\N
112617\sprachst\0\37553\sprachst\N
140462\spreche\91\37553\spre-che\N
140463\sprach\772\37553\sprach\N
153789\spr"achest\0\37553\spr"a-chest\N
155409\sprich\44\37553\sprich\N
```

2.2.2 Morphologische Ressourcen

Die Lexik vieler Sprachen weist Eigenschaften wie Konjugation, Deklination, Steigerung usw. auf. So werden für die Analyse eines Textes, z. B. seiner laufenden Wörter, die in den Sprachen mit „reicher“ Morphologie in vielen Formen erscheinen können, morphologische Ressourcen benötigt. Dieser Ressourcentyp gewinnt an Wert, da er viele morphologische Eigenschaften modelliert.

[23] Ausführliche Informationen über die Datenorganisation und -struktur von CELEX bietet das CELEX-Manual [CELEX 1994].

Eine solche Ressource ist im Listing 2.7 (Ausschnitt aus CELEX, Datei `german/gmw/gmw.cd`) angegeben:

Listing 2.7: CELEX-GWL

```
73411\sprichst\13\37553\2SIE
73412\spricht\712\37553\3SIE
73413\sprechen\1097\37553\13PIE,13PKE,i
73414\sprecht\2\37553\2PIE,rP
73415\sprachen\230\37553\13PIA
73416\spracht\0\37553\2PIA
73417\sprechest\0\37553\2SKE
73418\sprechet\0\37553\2PKE
73419\spraeche\6\37553\13SKA
73420\spraechst\0\37553\2SKA
73421\spraechen\7\37553\13PKA
73422\spraechet\0\37553\2PKA
73423\spraecht\0\37553\2PKA
73424\sprechend\8\37553\pE
73425\gesprochen\616\37553\pA
112617\sprachst\0\37553\2SIA
140462\spreche\91\37553\1SIE,13SKE
140463\sprach\772\37553\13SIA
153789\spraechest\0\37553\2SKA
155409\sprich\44\37553\rS
```

Dabei sind in der letzten Spalte (letztes Feld) die grammatischen Informationen eingetragen, die zu den Wortformen in der zweiten Spalte in der jeweiligen Zeile gehören. Die Zahl 37553, die in jeder Zeile (im vorgestellten Auschnitt) steht, verweist auf das Lemma sprechen.
Eine weitere Ressource dieses Typs ist das DELA-Lexikon (hier für das Deutsche), welches im Folgenden vorgestellt wird:[24]

Listing 2.8: DELA

```
Sehen,.N:aeN:deN:neN
...
sehe,sehen.V:1eGc:1eGi:3eGc
sehe,sehen.V+tr:1eGc:1eGi:3eGc
sehe,sehen.V+refl(d)+tr:1eGc:1eGi:3eGc
sehe,sehen.V+refl(a):1eGc:1eGi:3eGc
sehe,sehen.V+refl(a)+tr:1eGc:1eGi:3eGc
sehe,sehen.V+intr:1eGc:1eGi:3eGc
sehe,sehen.V+intr+refl(a)+tr:1eGc:1eGi:3eGc
...
siehst,sehen.V:2eGi
siehst,sehen.V+tr:2eGi
siehst,sehen.V+refl(d)+tr:2eGi
siehst,sehen.V+refl(a):2eGi
siehst,sehen.V+refl(a)+tr:2eGi
siehst,sehen.V+intr:2eGi
siehst,sehen.V+intr+refl(a)+tr:2eGi
...
```

[24] Vgl. Unitex, <http://www-igm.univ-mlv.fr/~unitex> , 24.7.2014. Das DELA-Lexikon wird automatisch aus den Automaten generiert. Daher können einige redundante Eigenschaften wahrscheinlich nicht vermieden werden.

```
sieht,sehen.V:3eGi
sieht,sehen.V+tr:3eGi
sieht,sehen.V+refl(d)+tr:3eGi
sieht,sehen.V+refl(a):3eGi
sieht,sehen.V+refl(a)+tr:3eGi
sieht,sehen.V+intr:3eGi
sieht,sehen.V+intr+refl(a)+tr:3eGi
...
sehen,sehe.N+FF
sehen,.V:1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+tr:1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+refl(d)+tr:1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+refl(a):1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+refl(a)+tr:1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+intr:1mGc:1mGi:3mGc:3mGi:OI
sehen,.V+intr+refl(a)+tr:1mGc:1mGi:3mGc:3mGi:OI
...
sehenswert,.ADJ:up
```

Ein Ausschnitt des englischen Lexikons von DELA sieht folgendermaßen aus:

Listing 2.9: DELA EN

```
see,.V:W:P1s:P2s:P1p:P2p:P3p
see,.N:s
see-through,.N+XN+z1:s
see-through,.A+z1
see-throughs,see-through.N+XN+z1:p
seeable,.A
seeableness,.N:s
...
saw,see.V:I1s:I2s:I3s:I1p:I2p:I3p
saw,.V:W:P1s:P2s:P1p:P2p:P3p
...
seen,see.V:Kn
...
seeing,see.V:G
seeing,.N:s
...
sees,see.V:P3s
sees,see.N:p
...
```

Dabei sind die Einträge von „see“ angegeben, wobei sowohl entsprechende Verben als auch Nomina aufgeführt sind.
Eine weitere Ressource ist IMSLex (Universität Stuttgart, Institut für maschinelle Sprachverarbeitung);[25] Aus den als Demo-Version zur Verfügung gestellten Ansätzen des IMSLex(ikon) lässt sich ablesen, dass die lexikalischen Einheiten einem Vollformlexikon entsprechen. Jeder Eintrag enthält (1.) eine Vollform einer von drei erwähnten Wortarten, (2.) die Grundform

[25] Vgl. [Fitschen 2004] und <http://www.ims.uni-stuttgart.de/projekte/IMSLex/> sowie <http://www.ims.uni-stuttgart.de/projekte/IMSLex/sample/IMSLexSample.zip> , 25.7.2014.

der flektierten Form, sowie (3.) die Kategorie Wortart (Adj, S und V). Ein konkreter Eintrag sieht wie folgt aus: planst plan V.
In [Fitschen 2004: 75–136] sind die Struktur und Aufbau des IMSLex sowie seine Verwendung beschrieben. Dabei werden anders als in den Demo-Beispielen noch grammatische Kategorien markiert. Ein vorgestelltes Beispiel ist Verbissenheit, vgl. [Op. cit.: 103]. Der Eintrag Verbissenheit wird zusammen mit drei weiteren Einträgen aufgelistet, die seine Generierung ermöglichen, je nachdem zu welchem Zweck die Einträge verwendet werden. Sie sehen aus wie folgt:

Listing 2.10: Ein Beispiel aus IMSLex

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

Ebenso sind in [Op. cit.: 103] eine Reihe von Eigenschaften und Merkmalen des Lexikons beschrieben, auf die hier nicht eingegangen werden kann. Aus dem Beispiel ist zu erkennen, dass neben der Flexionsmorphologie auch Derivation und Komposition behandelt werden.
Eine weitere Art von Ressourcen, die dem Bereich der Morphologie zugerechnet wird, kodiert Struktur und Aufbau der Wörter. Dieser Typ von Ressourcen liefert im konkreten Fall Informationen darüber, wie eine Wortform aufgebaut ist, durch welche Prozesse der Wortbildung sie entstanden ist, und wie sie zerlegt bzw. segmentiert werden kann. Gegebenenfalls sind weitere Informationen möglich. Hierzu bietet CELEX auch ein wertvolles Beispiel. Ein Abschnitt aus der Datei `german/gml/gml.cd` ist im Folgenden (Listing 2.11) dargestellt: Die Felder 9, 10 und 14 geben Informationen über die Struktur der Lemmata und ihrer Wortarten. Das Feld 9 z. B. kodiert den Stamm *sprech* des Verbs *sprechen*, wie er in Zusammensetzungen und Ableitungen vorkommen kann.

Listing 2.11: CELEX-GML

```

37553\sprechen\3598\M\1\Y\Y\Y\sprech\V\N\N\N\(sprech)[V]\N\N\N\N\i
                    246\N

37554\Sprechanlage\3\C\1\Y\Y\Y\sprech+Anlage\VN\N\N\N\((sprech)[V
                    ],(((an)[V|.V],(leg)[V])[V])[N])[N]\Y\N\N\N\S3/P
                    3\N

37555\Sprechchor\18\C\1\Y\Y\Y\sprech+Chor\VN\N\N\N\((sprech)[V],(
                    Chor)[N])[N]\N\N\N\N\S1/P1u\N
```

```
37556\Sprecher\479\C\1\Y\Y\Y\sprech+er\Vx\N\N\N\((sprech)[V],(er)[N|V.])[N]\N\N\N\N\S1/P2\N
37557\sprecherisch\0\C\1\Y\Y\Y\sprech+erisch\Vx\N\N\N\((sprech)[V],(erisch)[A|V.])[A]\N\N\N\N\I\N
37558\Sprecherziehung\1\C\1\Y\Y\Y\sprech+Erziehung\VN\N\N\N\((sprech)[V],(((er)[V|.V],(zieh)[V])[V],(ung)[N|V.])[N])[N]\N\N\N\N\S3/P0\N
37559\Sprechfunk\8\C\1\Y\Y\Y\sprech+Funk\VN\N\N\N\((sprech)[V],((funk)[V])[N])[N]\N\N\N\N\S1/P0\N
37560\Sprechgesang\1\C\1\Y\Y\Y\sprech+Gesang\VN\N\N\N\((sprech)[V],((ge)[N|.N],((sing)[V])[N])[N])[N]\Y\N\N\N\S1/P1u\N
37561\Sprechkunde\0\C\1\Y\Y\Y\sprech+Kunde\VN\N\N\N\((sprech)[V],(Kunde)[N])[N]\N\N\N\N\S3/P0\N
37562\sprechkundlich\0\C\1\Y\Y\Y\Sprechkunde+lich\Nx\N\N\N\(((sprech)[V],(Kunde)[N])[N],(lich)[A|N.])[A]\N\N\N\N\I\N
```

Eine lexikalische Ressource, die frei zur Verfügung steht, ist Morphisto.[26] Das Morphisto-Lexikon ist eine Ressource, die in erster Linie der MSV für nichtkommerzielle Zwecke dient. Es baut auf der SMOR-Morphologie[27] auf und kann sowohl für die Analyse als auch für die Produktion von Wortformen eingesetzt werden.
Eine andere lexikalische Ressource, die vom IDS (Institut für Deutsche Sprache, Mannheim) erstellt wurde und kontinuierlich gepflegt und erweitert wird, ist elexiko: *ein Online-Wörterbuch zur deutschen Gegenwartssprache*.[28]

2.2.3 Phonetische Ressourcen

Phonetische Ressourcen haben bei der morphologischen Verarbeitung keine besonders hohe Priorität. Sie sind allerdings für die Kodierung der Sprachausgabe oder das Mappen von einem Sprachsignal in einen Text und mögliche weitere Verwendungen in diesem und ähnlichen Bereichen wichtig, sowie für die Modellierung der Morphologie. Die phonetischen Ressourcen können zur Gewinnung von Informationen über phonotaktische Phänomene verwendet werden und später für die Erkennung okkasionell gebauter Wortformstrukturen eingesetzt werden. Als Beispiel sei im

[26] Vgl. <http://code.google.com/p/morphisto/>, 25.7.2014. Frei i. S. v. Open-source GPL v2; Creative Commons 3.0 BY-SA Non-Commercial license.
[27] Vgl. [Schmid et al. 2004].
[28] Vgl. <http://www.owid.de/wb/elexiko/start.html>, 25.7.2014.

folgenden Listing (2.12) ein Ausschnitt aus CELEX (Datei `german/gpl/gpl.cd`) dargestellt:

Listing 2.12: CELEX-GPL

```
1
2 37553\sprechen\3598\'SprE-x@n\[SprE[x]@n]\'SprEx\[SprEx]\[CCCV[C]VC]\[CCCVC]\SprEx\SprEx
3
4 37554\Sprechanlage\3\'SprEx-&n-la-g@\[SprEx][an][la:][g@]\'SprEx-&n-la-g@\[SprEx][an][la:][g@]\[CCCVC][VC][CVV][CV]\[CCCVC][VC][CVV][CV]\SprEx#an#le:g\SprEx#an#le:g
5
6 37555\Sprechchor\18\'SprEx-kor\[SprEx][ko:r]\'SprEx-kor\[SprEx][ko:r]\[CCCVC][CVVC]\[CCCVC][CVVC]\SprEx#ko:r\SprEx#ko:r
7
8 37556\Sprecher\479\'SprE-x@r\[SprE[x]@r]\'SprE-x@r\[SprE[x]@r]\[CCCV[C]VC]\[CCCV[C]VC]\SprEx+@r\SprEx+@r
9
10 37557\sprecherisch\0\'SprE-x@-rIS\[SprE[x]@][rIS]\'SprE-x@-rIS\[SprE[x]@][rIS]\[CCCV[C]V][CVC]\[CCCV[C]V][CVC]\SprEx+@rIS\SprEx+@rIS
11
12 37558\Sprecherziehung\1\'SprEx-Er-=i-UN\[SprEx][Er][tsi:][UN]\'SprEx-Er-=i-UN\[SprEx][Er][tsi:][UN]\[CCCVC][VC][CVV][VC]\[CCCVC][VC][CVV][VC]\SprEx#Er#tsi:+UN\SprEx#Er#tsi:+UN
13
14 37559\Sprechfunk\8\'SprEx-fUNk\[SprEx][fUNk]\'SprEx-fUNk\[SprEx][fUNk]\[CCCVC][CVCC]\[CCCVC][CVCC]\SprEx#fUNk\SprEx#fUNk
15
16 37560\Sprechgesang\1\'SprEx-g@-z&N\[SprEx][g@][zaN]\'SprEx-g@-z&N\[SprEx][g@][zaN]\[CCCVC][CV][CVC]\[CCCVC][CV][CVC]\SprEx#g@#zIN\SprEx#g@#zIN
17
18 37561\Sprechkunde\0\'SprEx-kUn-d@\[SprEx][kUn][d@]\'SprEx-kUn-d@\[SprEx][kUn][d@]\[CCCVC][CVC][CV]\[CCCVC][CVC][CV]\SprEx#kUnd@\SprEx#kUnd@
19
20 37562\sprechkundlich\0\'SprEx-kUnt-lIx\[SprEx][kUnt][lIx]\'SprEx-kUnt-lIx\[SprEx][kUnt][lIx]\[CCCVC][CVCC][CVC]\[CCCVC][CVCC][CVC]\SprEx#kUnd@#lIx\SprEx#kUnd@#lIx
```

Das nächste Listing (2.13) zeigt die Kodierung der einzelnen Formen eines Lemmas (CELEX, Datei `german/gpw/gpw.cd`). Die phonetischen Eigenschaften variieren oft von Form zu Form innerhalb des Paradigmas eines Lemmas, was die Erstellung dieser Ressource notwendig macht. Aus einem Lemma die Aussprache der Wortformen zu generieren, ist, wenn nicht unmöglich, so doch eine sehr schwierige Aufgabe.

Listing 2.13: CELEX-GPW

```
1 73411\sprichst\13\37553\'SprIxst\[SprIxst]\[CCCVCCC]
2 73412\spricht\712\37553\'SprIxt\[SprIxt]\[CCCVCC]
3 73413\sprechen\1097\37553\'SprE-x@n\[SprE[x]@n]\[CCCV[C]VC]
4 73414\sprecht\2\37553\'SprExt\[SprExt]\[CCCVCC]
5 73415\sprachen\230\37553\'Spra-x@n\[Spra:][x@n]\[CCCVV][CVC]
6 73416\spracht\0\37553\'Spraxt\[Spra:xt]\[CCCVVCC]
```

```
73417\sprechest\0\37553\'SprE-x@st\[SprE[x]@st]\[CCCV[C]VCC]
73418\sprechet\0\37553\'SprE-x@t\[SprE[x]@t]\[CCCV[C]VC]
73419\spraeche\6\37553\'Spr)-x@\[SprE:][x@]\[CCCVV][CV]
73420\spraechst\0\37553\'Spr)xst\[SprE:xst]\[CCCVVCCC]
73421\spraechen\7\37553\'Spr)-x@n\[SprE:][x@n]\[CCCVV][CVC]
73422\spraechet\0\37553\'Spr)-x@t\[SprE:][x@t]\[CCCVV][CVC]
73423\spraecht\0\37553\'Spr)xt\[SprE:xt]\[CCCVVCC]
73424\sprechend\8\37553\'SprE-x@nt\[SprE[x]@nt]\[CCCV[C]VCC]
73425\gesprochen\616\37553\g@-'SprO-x@n\[g@][SprO[x]@n]\[CV][CCCV[␣
    C]VC]
112617\sprachst\0\37553\'Spraxst\[Spra:xst]\[CCCVVCCC]
140462\spreche\91\37553\'SprE-x@\[SprE[x]@]\[CCCV[C]V]
140463\sprach\772\37553\'Sprax\[Spra:x]\[CCCVVC]
153789\spraechest\0\37553\'Spr)-x@st\[SprE:][x@st]\[CCCVV][CVCC]
155409\sprich\44\37553\'SprIx\[SprIx]\[CCCVC]
```

Ein Name, der in diesem Zusammenhang nicht unerwähnt bleiben kann, ist Verbmobil.[29] Es handelt sich dabei um ein System, das in der Lage sein sollte, gesprochene Spontansprache aufzuzeichnen, in maschinenlesbare Form zu überführen und automatische Übersetzungen ins Englische, Deutsche und Japanische zu generieren. Das System erstellt automatisch die Tondaten, welche hier als Ressource betrachtet werden.

2.2.4 Syntaktische Ressourcen

Syntaktische Ressourcen wären in erster Linie Satzmodelle, Muster der Nominalphrasen, Valenzrahmen von Verben zu nennen.[30] Ein Teil dieser Daten, bspw. zur Wortstellung, kann automatisch erhoben werden, wenn ein Textkorpus POS-getaggt ist.
Im Folgenden werden zwei Beispiele vorgestellt: CELEX (Listing 2.14) und VALBU (Listing 2.15 bzw. Listing 2.16).

Listing 2.14: CELEX-GSL

```
37553\sprechen\3598\4\\\N\N\1\5\00000000M;000000000;00N000000;00N␣
    000P00;000000P00;00N0N0000;\I\I\O\I\O\I\O\I\O\\\\
37554\Sprechanlage\3\1\2\\N\N\\\\\\\\\\\\\\\\
37555\Sprechchor\18\1\1\\N\N\\\\\\\\\\\\\\\\
37556\Sprecher\479\1\1\\N\N\\\\\\\\\\\\\\\\
37557\sprecherisch\0\2\\\N\N\\\\\\\\\\\\P\\\
37558\Sprecherziehung\1\1\2\\Y\N\\\\\\\\\\\\\\\\
37559\Sprechfunk\8\1\1\\Y\N\\\\\\\\\\\\\\\\
37560\Sprechgesang\1\1\1\\N\N\\\\\\\\\\\\\\\\
37561\Sprechkunde\0\1\2\\Y\N\\\\\\\\\\\\\\\\
37562\sprechkundlich\0\2\\\N\N\\\\\\\\\\\\P\\\
```

CELEX modelliert die Valenz-Informationen im Vergleich zu anderen Daten in einer fast kryptischen Form, vgl. die erste Zeile in Listing 2.14, den Eintrag

[29] Vgl. <http://verbmobil.dfki.de/> , 25.7.2014 sowie [Wahlster 2000].
[30] Satzbaupläne in [VALBU: 937–1040] und in [Wahrig 1998: 32–35];

37553 (*sprechen*). Die als 00000000M; 000000000; 00N000000; 00N000P00; 000000P00; 00N0N0000; kodierte Information wird mithilfe eines Schlüssels gelesen. Die gesamte Ausdruckkette stellt (mögliche) Valenzen des Verbs zusammen.[31] Der Ausdruck 00000000M z. B. steht für das Beispiel (i. S. v. Satztyp) *Der Bau des Schiffes ist schon weit gediehen.*[32]
Im Gegensatz zu CELEX ist die Valenzangabe der Verben im VALBU ausführlicher und anders aufgebaut. Das folgende Listing (2.15) zeigt einen Ausschnitt des Eintrages für das Verb *sehen* (Alphabetisches Verb-Register). Da das VALBU in erster Linie nicht als Ressource für die maschinelle Sprachverarbeitung, sondern als ein traditionelles Nachschlagewerk gedacht ist, sind die Bezeichnungen und Abkürzungen in einer gängigen Form angegeben.

Listing 2.15: VALBU-L

```
...
sehen   1        NomE  AkkE
sehen   2        NomE  AkkE
sehen   3        NomE  AkkE
sehen   4        NomE  AkkE        AdvE
sehen   5        NomE  AkkE        AdvE1      v  AdvE2
sehen   6  nach  NomE  PräpE[+D]
sehen   7        NomE  AdvE
sehen   8        NomE  VerbE
sehen   9  in    NomE  AkkE        PräpE[+D]
sehen  10  als   NomE  AkkE        PrädE
sehen  11  als   NomE  AkkE        PrädE
sehen  12        NomE  AkkE        PrädE
sehen  13  auf   NomE  PräpE[+A]
sehen  14        NomE  AdvE
sehen  15        NomE  VerbE
sehen  16        NomE  VerbE
sehen  17        NomE  VerbE
...
```

Das nächste Listing (2.16) listet im Gegensatz zu Listing 2.15, d. h. nach Verben wie *sehen*, die Daten nach Valenzmuster bzw. Satzbauplan (Satzmodell-Register). Zum Beispiel ist der Satzbauplan NomE AkkE AdvE1 v AdvE2, bestellen 3, als übergeordneter Plan der Kombination der Ergänzungen NomE AkkE AdvE AdvE zu verstehen. Zu diesem Satzmuster gehört auch der Eintrag sehen 5 in Listing 2.15.

Listing 2.16: VALBU-S

```
...
NomE   AkkE    AdvE              AdvE
NomE   AkkE    AdvE1/AdvE2             lesen       8
NomE   AkkE    AdvE1/AdvE2             zeichnen    9
NomE   AkkE   (AdvE1/AdvE2)           verteilen   5
```

[31] Vgl. CELEX-Manual für das Deutsche (German Linguistic Guide) Seite 189.
[32] Es handelt sich hier um ein Beispiel für einen abstrakten Satztyp.

```
NomE   (AkkE)   AdvE1/AdvE2                kommen      13
NomE    AkkE    AdvE1          v   AdvE2   bestellen    3
NomE   (AkkE)   AdvE1          v   AdvE2   grüßen       4
...
```

Es handelt sich um zwei Register, ein Satzmodell-Register und ein alphabetisches Verb-Register, die dem Leser zwei verschiedene Zugriffsmöglichkeiten auf das in VALBU enthaltene Wissen ermöglichen.
In diesem Zusammenhang kann auch die an der Universität Erlangen–Nürnberg erstellte Erlangen Valency Patternbank[33] genannt werden, welche Valenzmuster des Englischen, versehen mit weiteren Angaben, auflistet. Sie basiert auf den Daten des VDE (*A Valency Dictionary of English*) von [Herbst et al. 2004].
Als nächster Ressourcentyp werden noch Treebanks oder dt. Baumbanken präsentiert. Hierbei handelt es sich um Korpora, deren einzelne Sätze syntaktisch geparst und annotiert wurden, sodass sie als Baumstruktur angezeigt, gespeichert und als solche noch verarbeitet werden können. Für das Deutsche sind Tiger und Negra die bekanntesten Beispiele, vgl. das folgende Listing[34] (2.17), ein Ausschnitt eines geparsten Satzes aus Negra.

Listing 2.17: Negra-TB

```
#BOS 11 2 893259793 1
Kein          PIAT      Neut.Nom.Sg          NK      500
Wunder        NN        Neut.Nom.Sg.*        NK      500
also          ADV       --                   MO      511
,             $,        --                   --      0
daß           KOUS      --                   CP      510
vor           APPR      Dat                  AC      504
rund          ADV       --                   MO      501
acht          CARD      --                   HD      501
Jahren        NN        Neut.Dat.Pl.*        NK      504
die           ART       Def.Fem.Nom.Sg       NK      509
Idee          NN        Fem.Nom.Sg.*         NK      509
entstand      VVFIN     3.Sg.Past.Ind        HD      510
,             $,        --                   --      0
ein           ART       Indef.Neut.Akk.Sg    NK      507
Label         NN        Neut.Akk.Sg.*        NK      507
zu            PTKZU     --                   PM      502
gründen       VVINF     --                   HD      502
,             $,        --                   --      0
das           PRELS     Neut.Nom.Sg          SB      506
klassische    ADJA      Pos.Fem.Akk.Sg.St    NK      505
Musik         NN        Fem.Nom.Sg.*         NK      505
im            APPRART   Dat.Masc             AC      503
weitesten     ADJA      Sup.Masc.Dat.Sg.Sw   NK      503
Sinne         NN        Masc.Dat.Sg.*        NK      503
produziert    VVFIN     3.Sg.Pres.Ind        HD      506
.             $.        --                   --      0
#500          NP        --                   PD      511
```

[33] Online abrufbar unter <http://www.patternbank.uni-erlangen.de/> , 25.7.2014.
[34] Beispiel aus <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/sentno2.html/> , 25.7.2014.

```
#501        AP        --                    NK      504
#502        VZ        --                    HD      508
#503        PP        --                    MNR     505
#504        PP        --                    MO      510
#505        NP        --                    OA      506
#506        S         --                    RC      507
#507        NP        --                    OA      508
#508        VP        --                    OC      509
#509        NP        --                    SB      510
#510        S         --                    SB      511
#511        S         --                    --      0
#EOS 11
```

2.2.5 Semantik-Ressourcen

Die Semantik-Ressourcen bilden im Vergleich zu den bisher vorgestellten Ressourcen einen Typ, der viel mehr Variation erlaubt und dementsprechend schwierig zu erstellen ist. Ein noch vergleichsweise einfach zu erstellender Typ sind die lexikalisch-semantischen Ressourcen. Diese beschreiben und kodieren Informationen auf Wortebene. Es gibt einige Möglichkeiten, diese Daten in Form einer Ressource darzustellen. Die einfachsten Realisierungen sind die Wortassoziationen, wie Synonyme, Antonyme (wenn möglich), Hyponyme, Hyperonyme usw. sowie Kollokationen.[35] Dabei werden alle möglichen Bedeutungen und Kontexte angegeben, so dass beim Einsatz der Ressource die passende Information für den gegebenen Fall ausgewählt werden kann.
Eine Semantik-Ressource, die über die Wortebene (lexikalische Semantik) hinausgeht, etwa die Semantik bestimmter Phrasen oder Satzglieder sowie der ganzen Sätze, wird sehr schnell sehr komplex, da die Kombinatorik der Bedeutung der beteiligten semantischen Einheiten viele Interpretationen erlaubt. Im Vergleich zu lexikalischen und morphologischen Eigenschaften kommen noch weitere Daten hinzu, wie Sprecher und Adressat. Diese Ressourcen werden bislang manuell annotiert.
Im Folgenden sind einige konkrete Ressourcen aufgelistet, die als Semantik-Ressourcen verstanden werden könnten:

- Dornseiff, das Wörterbuch[36], das den deutschen Wortschatz nach Sachgruppen organisiert, bildet auf der einen Seite eine Zuordnung der Einträge in einer Gruppe, vernetzt also den Eintrag mit anderen Stellen, in denen er vorkommt und gibt auf der anderen Seite die Wortassoziationen des jeweiligen Eintrages in Gruppen an.

[35] Vgl. hierzu z. B. [Atkins/Rundell 2008] und [Engelberg 2009].
[36] [Dornseiff et al. 2004], 8. Auflage 2004, 1. Auflage 1934.

- WordNets: WordNet, GermaNet, EuroWordNet, BalkaNet

  WordNet ist eine lexikalisch-semantische Ressource. Es handelt sich um sogenannte Synsets, die die jeweiligen lexikalischen Einheiten sowohl nach Wortart als auch im Kontext (Satzebene) behandeln samt Verweisen auf andere lexikalische Einträge, die eine ähnliche Bedeutung bzw. Verwendung haben.

  GermaNet wurde nach dem Vorbild WordNet gebaut, behandelt aber, wie der Name schon andeutet, den deutschen Wortschatz.

  EuroWordNet baut den Wortschatz von mehreren Sprachen Europas nach dem Vorbild WordNets auf. Es handelt sich um die Sprachen Deutsch, Estnisch, Französisch, Italienisch, Niederländisch, Spanisch und Tschechisch.

  BalkaNet: Es handelt sich um ein WordNet-Projekt für einige Sprachen der Balkan-Halbinsel, nämlich Bulgarisch, Griechisch, Rumänisch, Serbisch, Tschechisch und Türkisch. BalkaNet ist nach dem Vorbild EuroWordNets gebaut.[37].

- SALSA II[38] ist ein per Hand annotiertes Korpus für das Deutsche, das ca. 20 000 verbale und ca. 17 000 nominale Einheiten (engl. instances) beinhaltet. Es handelt sich dabei um Angaben zu semantischen Rollen (engl. semantic roles) im Sinne von FrameNet.

- Als weiteres Modell kann an dieser Stelle [Helbig 2008] genannt werden. Es bietet eine vielseitige und ausführliche Verarbeitung, Annotierung und Darstellung semantischer Informationen für das Deutsche.

2.2.6 Grammatiken

Mit Grammatiken sind die Ressourcen gemeint, welche verschiedene Informationen modellieren, bspw. Grammatiken über die Wortstellung des Deutschen (Topologie), welche bei morpho-syntaktischem Taggen und Parsen verwendet werden können, etwa zur Disambiguierung bestimmter mehrdeutiger Fälle.

[37] Vgl. [Pala 2005]. Das Albanische war nicht Teil des BalkaNet-Projektes.
[38] Vgl. <http://www.coli.uni-saarland.de/projects/salsa/corpus/> , 25.7.2014.

Wertvolle Informationen zu Ressourcen (Korpora, Lexikographie und Morphologie) findet der interessierte Leser in [Lobin/Lemnitzer 2004], [Mitkov 2003: (§ 2)], [Goldsmith 2010] und [Wynne 2005 (*Developing Linguistic Corpora*)].

2.3 Maschinelle Morphologie

Der Teilbereich der Morphologie gilt inzwischen für viele Fachkenner als jener Teilbereich der maschinellen Sprachverarbeitung, der ausreichend untersucht worden ist. Sicher ist jedoch, dass zum Thema maschinelle Morphologie immer noch kontinuierlich Tagungen und Konferenzen stattfinden.[39] Es werden für einige Sprachen maschinelle Morphologien zum ersten Mal erfasst und es zeichnen sich immer wieder neue Ansätze und Methoden für die maschinelle morphologische Verarbeitung der Sprache ab. Diese Entwicklung hat eine sehr große, ja unüberschaubare Anzahl verschiedenster Systeme zur Folge.

2.3.1 Anwendungsgebiete

Die Möglichkeiten der Anwendung einer maschinellen Morphologie sind vielseitig.[40] Sie kann u. a. in den Bereichen eingesetzt werden, welche im Folgenden aufgelistet sind:

- für didaktische Zwecke
  Eine maschinelle Morphologie kann eingesetzt werden, um das Lernen einer Sprache zu erleichtern, indem z. B. das ganze Paradigma eines Lexikoneintrages (der flektierten Wortarten, bspw. eines Verbs) gezeigt, die Grundform einer Wortform angegeben (Lemmatisierung) oder einfach die grammatischen Kategorien einer Wortform gezeigt werden (Analyse/Kategorisierung).

[39] Zum Beispiel die Veranstaltung *State of the Art in Computational Morphology* – Workshop on Systems and Frameworks for Computational Morphology (SCFM), organisiert 2009 von Cerstin Mahlow und Michael Piotrowski am Institut für Computerlinguistik der Universität Zürich, deren Akten als [Mahlow/Piotrowski 2010] publiziert sind. Eine weitere Konferenz zum Thema Morphologie, die 2011 stattfand, ist The eighth Mediterranean Morphology Meeting [MMM 8]), Italien, vgl. <http://www.diplist.it/mmm8/>, 25.7.2014.

[40] Als einführende Literatur zu diesem Thema auf Deutsch bietet sich das fünfte Kapitel („Anwendungen"), Seiten 461–571, in [Klabunde et al. 2004] an.

- für Zwecke der Rechtschreibprüfung
  Ein bekannter Nutzen einer maschinellen Morphologie. Im Vergleich zu gewöhnlicher Rechtschreibprüfungs-Software bietet eine Morphologie zusätzlich noch die grammatischen Eigenschaften der jeweiligen Wortform (Textwort). Ebenso basieren die Korrekturvorschläge einer Morphologie-Komponente auf grammatischen Kriterien, d. h., sie gehen über die sogenannte Levenshtein-Distanz (Minimal Edit Distance) und statistische Verfahren hinaus.

- für Zwecke der Wortform- und Wortbildungsanalyse
  Wortformen werden kategorisiert, segmentiert und lemmatisiert. Ihre Bedeutung kann aus den einzelnen Teilkonzepten erschlossen werden. Die analysierten Daten können sowohl in der Lehre als auch im Sinne der MSV eingesetzt werden.

- zum Taggen von Korpora (maschinelle Sprachverarbeitung)
  Eine der wichtigsten Rollen einer maschinellen Morphologie als Komponente eines komplexeren Systems/Programms. Eine maschinelle Morphologie kann auch als selbständige Komponente eingesetzt werden, um Wortformen eines Korpus zu annotieren.

- für Zwecke des Information Retrieval
  Die Überführung der Wortformen in ihre jeweiligen Grundformen (Lexikoneintrag) oder die Suche nach bestimmten morphologischen Kategorien der Wortformen wären Anwendungen einer maschinellen Morphologie im Bereich des Information Retrieval.[41]

  und

- Beitrag zur Entwicklung anderen MSV-Komponenten
  Auch in diesem Fall ist der Einsatz einer maschinellen Morphologie vielseitig möglich. Eine mögliche Anwendung wäre im Bereich der Textzusammenfassung.

[41] Vgl. hierzu z. B. [Ferber 2002 (§ 4.3)].

2.3.2 Die Anfänge der maschinellen Morphologie

Nach einer experimentellen Anfangsphase entwickelte sich die maschinelle Morphologie rasant. Sie hatte einen schnellen Start, da sie auf den gesammelten, kategorisierten und klassifizierten Daten der maschinellen Lexikographie aufbauen konnte. Die maschinelle Lexikographie[42] und Morphologie erlebten ihre großen Fortschritte in den 1980er und '90er Jahren.[43]
Ein Überblick zum Enwicklungstand der Morphologie Mitte der '90er Jahre ist in der Aufsatzsammlung zu den „Morpholympics“ zu finden, vgl. [Hausser 1996].[44] Dort sind die Systeme (Ansätze) Morphy (W. Lezius), PC-Kimmo (A. Schiller), Morph (G. Hanrieder), Morphix (W. Finkler/O. Lutzy), Plain (H. Visser/H.-D. Koch), LA-Morph (G. Schüller/O. Lorenz), Gertwol (K. Koskenniemi/M. Haapalainen) und Mpro (H. D. Mass) vertreten. Diese Zahl spiegelt das große Interesse an maschineller Morphologie in den 1990er Jahren wider. Obwohl es dabei haupsächlich um die Morphologie des Deutschen ging, könnte in diesem Zusammenhang behauptet werden, dass dieser Stand sich auch auf Sprachen wie Französisch oder Italienisch übertragen lässt, welche eine Morphologie vergleichbarer Komplexität haben.[45]
Diese Entwicklungen, begonnen bzw. intensiviert in den 1970er und '80er Jahren, hoben zugleich die zunehmende Rolle der maschinellen Verarbeitung hervor.
Als sich die Lexika und die Systeme für die morphologische Analyse (und zum Teil auch für die Produktion) etablierten, zeichnete sich eine Schwerpunktverlagerung von den Themen der Lexikographie und Morphologie zu den Themen, die auf ihnen aufbauen, ab, wie morpho-syntaktisches Tagging und syntaktisches Parsing, semantische Verarbeitung, Dialogsysteme, Inhaltsanalyse usw. Doch es wurde weiterhin nach optimierten Lösungen für die maschinelle Morphologie gesucht.
Einen kompakten Überblick über die Entwicklungen in der Computermorphologie bis Ende der 1990er Jahre geben jeweils [Sproat 2000] und [Heid 2003]. Letzterer behandelt zusätzlich auch das Thema Lexikon. [Trommer

[42] Ausführliche Informationen zu den Anfängen der maschinellen Lexikographie im deutschsprachigen Raum bieten [Hess et al. 1983]. Vgl. hierzu auch [Eggers et al. 1980] und [Görz/Paulus 1988].

[43] Ausführliche Informationen zu einigen Morphologiesystemen, die vor Anfang der 1990er Jahre entwickelt und eingesetzt wurden, bieten [Hess et al. 1983], [Schaeder/Willée 1989] und [Hausser 1996].

[44] Dabei handelt es sich um einen Wettbewerb verschiedener Systeme der maschinellen Morphologie, der an der Universität Erlangen-Nürnberg ausgetragen wurde.

[45] Vgl. hierzu [Bátori et al. 1989] zur Entwicklung verschiedener Bereiche der Computerlinguistik bis Ende der 1980er Jahre. Einen guten Überblick bietet auch [Sproat 1992].

2010] gibt einen guten aktuellen Überblick über die verbreitetsten Methoden der maschinellen Morphologie. Einen Überblick über die Lexikographie bzw. maschinelle Lexikographie geben jeweils [Atkins/Rundell 2008], [Fitschen 2004] und [Kunze/Lemnitzer 2007].

2.3.3 Einige Ansätze der maschinellen Morphologie

Im Rahmen der vorliegenden Arbeit wird keine Besprechung und kein Vergleich bereits existierender Systeme angestrebt, sondern es wird im Folgenden ein typen- und eigenschaftenbasierter Überblick gegeben, um die im Rahmen dieser Arbeit entwickelten lexikalischen Ressourcen und die Morphologiekomponente leichter beschreiben zu können.
Im Folgenden wird auf einige dieser Systeme eingegangen, die anhand verschiedener Kriterien ausgewählt wurden, und zwar:

1. Finite-State-Morphology als meistverbreitetes System;
2. LAG/Malaga als ein System, das auf Basis einer anderen Methode arbeitet, und da es eine Zeit lang an der Universität Erlangen–Nürnberg eingesetzt wurde. Im Rahmen von Malaga wurde 2002/3 vom Verfasser der vorliegenden Arbeit das Verbalsystem des Albanischen implementiert;[46]
3. DATR als eine andere Möglichkeit lexikalische Daten zu modellieren, womit auch morphologische Eigenschaften (Flexion und Wortbildung) abgedeckt werden können.

Ein weiteres System, das man hier erwähnen könnte, ist TAGH, das die deutsche Morphologie und Wortbildung behandelt.[47] Im Folgenden wird auf einige andere Systeme gelegentlich eingegangen, um ihre Besonderheiten zu nennen, welche im Kontext der vorliegenden Arbeit von Interesse sind.
Kompakte Informationen über die aktuell wichtigsten Systeme der Morphologie bieten u. a. [Roark/Sproat 2007: 21–136] und [Trommer 2010].

[46] Vgl. hierzu Kapitel 5
[47] Vgl. [Geyken/Hanneforth 2006].

Finite-State-Morphology

Finite-State-Morphology[48] (kurz FSM) ist ein System, welches auf der Arbeitsweise von *endlichen Automaten* (engl. Finite Automata oder Finite State Automata) basiert. „Endliche Automaten sind der einfachste und zugleich verbreitetste Formalismus bei der Modellierung von morphologischen Regelsystemen“ [Trommer 2010: 244].
Unter dem Begriff Endlicher Automat ist ein Mechanismus zu verstehen, mit dem eine Zeichenfolge analysiert werden kann, wobei nur eine korrekte Eingabe den Endzustand erreicht.[49] Es werden hauptsächlich zwei Typen von endlichen Automaten unterschieden, und zwar:

- Nichtdeterministische endliche Automaten (NDEA), engl. Nondeterministic Finite Automata (NFA) und
- Deterministische endliche Automaten (DEA), engl. Deterministic Finite Automata (DFA).

„Der Unterschied zwischen diesen Automatentypen liegt in der Definition der Übergangsfunktion“ [Klabunde 2010: 75 (§ 2.2)]. Bei den NDEA gibt es bei Eingabe eines Zeichens *x* in einem Zustand *T* eine Menge von Nachfolgezuständen, während dies bei den DEA nicht der Fall ist, d. h., bei DEA ist es genau ein Nachfolgezustand.[50]

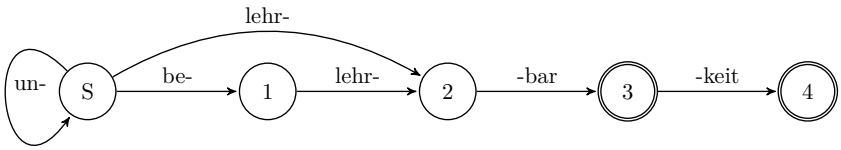
[48] Vgl. [Beesley/Karttunen 2004].

[49] Vgl. [Klabunde 2010: 70–79 (§ 2.2.3)] bzw. [Klabunde 1998]. Für ausführlichere Informationen zum Thema Automatentheorie siehe [Hopcroft et al. 2001].

[50] Vgl. [Klabunde 2010: 70–79 (§ 2.2.3)] für die Definitionen (2.2.6 und 2.2.7) beider Automatentypen sowie weitere Informationen zu DEA und NDEA. Ausführliche Informationen über Endliche Automaten (Finite Automata) bietet [Hopcroft et al. 2001: 37–81 (§ 2)]. Vgl. dort insbesondere die Definitionen 2.2.1 (DFA) und 2.3.2 (NFA). Der interessierte Leser findet bei [Rechenberg 2002] auch ausführliche Informationen zum Thema. Kompakte Informationen in Form von Lexikoneinträgen geben auch [Schneider 1997: 69 ff.] und [Claus/Schwill 2001: 65 f.].

Ein Beispiel für einen DEA, entnommen aus [Klabunde 2010: 74], ist in der Abbildung 2.1 dargestellt.[51]

Abbildung 2.1: Ein Beispiel: *Unbelehrbarkeit*.



Eine Variante der endlichen Automaten, die eine große Bedeutung für die maschinelle Sprachverarbeitung haben, sind die

> Finite State Transducer, kurz FST, auch nur Transducer genannt, auf Deutsch Transduktoren

Sie erzeugen eine Zeichenkette als Ausgabe, wenn die Eingabe erkannt wird, d. h. wenn sie „wohlgeformt“ in Bezug zu einer gegebenen Grammatik ist. Transduktoren können mehrere Symbole bearbeiten, gewöhnlich zwei (Symbolpaare $\langle x, y \rangle$), um in ihre definierten Zustände überzugehen.[52]
Eine für die Umsetzung der FSM nötige Software wird oft aus praktischen Gründen in einen Lexikon- und einen Regelteil aufgeteilt. Beide Teile werden mithilfe eines Compilers in einen ausführbaren Maschinen-Code überführt, der schließlich als Kompilat benutzt werden kann.[53] Für diese Zwecke gibt es inzwischen mehrere Compiler.

Einige davon sind im Folgenden aufgelistet:

1. Xerox Finite-State Tool (Xerox Cooperation, Palo Alto, USA)
2. HFST (Helsinki Finite-State Transducer Technology)[54]
3. S-FST (Stuttgart Finite-State Transducer Tools)[55]
4. foma (Finite-State Compiler and C library).[56]

[51] Ein Beispiel für einen NDEA wäre ein Automat mit den folgenden Zuständen und Übergängen: S→1; 1→2; (1→3); 2→1; 2→3; 3→2; (3→1); 3→T.
[52] Vgl. hierzu [Klabunde 2010: 78 f. (§ 2.2.3)].
[53] In einigen Fällen ist es nicht nötig, den Quellcode zu kompilieren. Er wird interpretiert.
[54] Vgl. <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/> , 27.7.2014.
[55] Vgl. <http://www.ims.uni-stuttgart.de/projekte/gramotron/software/sfst.html> , 27.7.2014.
[56] Vgl. <http://foma.sourceforge.net/dokuwiki/doku.php> , 27.7.2014.

Das verbreitetste unter den aufgelisteten Systemen ist xfst, welches unter Copyright der Xerox Corporation vertrieben wird.[57] Für nicht-kommerzielle Zwecke steht die Software (als ausführbares Programm, ohne Quellcode) zur Verfügung.[58]

Listing 2.18: XFST

```
xfst
Copyright © Palo Alto Research Center 2001-2011
PARC Finite-State Tool, version 2.15.7
(libcfsm-2.25.11) (svn 34269)

Type "help" to list all commands available or
"help help" for further help.

xfst[0]:
```

Zu dieser Software gibt es inzwischen auch Alternativen, nämlich u. a. die Softwarepakete SFST, HFST und foma.
FSA-Systemen können weitere Eigenschaften hinzugefügt werden, insbesondere Gewichtung ist sehr nützlich, eine Art Präferenzsetzung bei Übergängen von einem zum anderen Zustand (state). So können bspw. statistische Beobachtungen als Gewichte modelliert werden, um die Übergänge etwa bei Ambiguität besser zu steuern. Eine freie Software, die diese Eigenschaften bzw. Funktionen implementiert, ist

5. OpenFST (Library for constructing, combining, optimizing, and searching weighted finite-state transducers).[59]

Im Folgenden wird kurz auf die Xerox Finite-State Tools (xfst) eingegangen: Zu dem Softwarepaket, das mit dem Buch [Beesley/Karttunen 2003] geliefert wird oder online zur Verfügung steht, gehören vier Anwendungen, xfst, lexc, lookup und tokenize. Das fünfte Programm, twolc, stellt einen Compiler für den Two-Level-Formalismus von Kimmo Koskenniemi dar.

- xfst ist der Interpreter und Compiler der XFST-Grammatiken sowie eine Programmiersprache. Als Interpreter besitzt xfst verschiedene Funktionen wie words (Anzeige der Pfade bzw. Zeichenkette der Wortformen im Automat), upper-words (Anzeige der Analyse-Seite der Transduktoren), / lower-words (Anzeige der Oberflächenseite der Transduktoren), write (Schreiben eines Netzes [in einer

[57] Siehe hierzu [Beesley/Karttunen 2003] für weitere Informationen.
[58] Z. B. auf der beigelegten CD-ROM zu [Beesley/Karttunen 2003] oder online unter <http://www.stanford.edu/~laurik/fsmbook/home.html> , 27.7.2014.
[59] Vgl. <http://www.openfst.org/twiki/bin/view/FST/WebHome> , 27.7.2014.

Datei]) usw. Zum Kompilieren verfügt xfst über eine Reihe verschiedener Möglichkeiten, wie z. B. das Kompilieren einer Grammatik (xfst -l <grammatik.xfst>), das Interpretieren eines Kompilats (xfst -s <grammatik.xfst>) usw.

- lexc verarbeitet die Lexika und konvertiert sie in Automaten, sodass sie von XFST verwendet werden können.
- lookup dient zur Abfrage der Wortformen in fst-Dateien (Automaten, die vorher aus den kompilierten xfst-Grammatiken, xfst-Projekten, entstanden sind). Neben einzelnen Wortformen bzw. Textwörtern können sie auch in Form von Listen bearbeitet werden.[60]
- tokenize bereitet Texte vor, indem es sie zwecks besserer und leichterer Verarbeitung in Textwörter, d. h. in Wortlisten, eine Wortform pro Zeile, zerlegt.[61]

  Sowohl tokenize als auch lookup sind *Command-line*-Werkzeuge, die Endliche Automaten, die mit lexc und xfst erstellt wurden, verwenden und auf die sie zugreifen.[62]
- twolc ist der Compiler des Two-Level-Formalism, dem ersten Ansatz, der Finite-State Morphology implementierte. Er wurde von Kimmo Koskenniemi (Universität Helsinki) Anfang der 1980er Jahren entwickelt, vgl. hierzu [Koskenniemi 1993].[63]

Die letzten Entwicklungen im Bereich der FST-Familie können in den Sammelbänden [Yli-Jyrä et al. 2006] und [Yli-Jyrä et al. 2010] nachgelesen werden.[64]

[60] In Kapitel 5 werden konkrete Beispiele vorgestellt, vgl. Quellcode 6.1.
[61] Vgl. [Beesley/Karttunen 2003: 422–431].
[62] Vgl. [Beesley/Karttunen 2003: 420].
[63] Der Compiler zu Two-Level Formalism (twolc) ist auf dem Datenträger zu [Beesley/-Karttunen 2003] neben den XFST-Tools enthalten.
[64] Es handelt sich um den *International Workshop* zu *Finite-State Methods and Natural Language Processing*, welcher kontinuierlich seit 1996 stattfindet.

Malaga

Malaga ist ein System für die maschinelle Analyse natürlicher Sprachen. Das Programmpaket besteht aus einer Entwicklungsumgebung und einem Compiler und Interpreter, die zur gleichnamigen Programmiersprache gehören. Es wurde von Björn Beutel an der Universität Erlangen–Nürnberg von 1995 bis 2003 enwickelt.[65] *Malaga* wird auch mit einigen Linux-Distributionen mitgeliefert, etwa mit *openSuSE* (*11.2*).[66] Das Paket arbeitet in groben Zügen nach dem Ansatz der *Linksassoziativen Grammatik* (kurz *LAG*), vertreten von Roland Hausser.[67]

Es folgt eine Analyse der deutschen Verbform *zeigen* mit Malaga:

Listing 2.19: MALAGA

```
malaga project.pro
This is malaga, version 7.12.
Copyright (C) 1995 Bjoern Beutel.
This program is part of Malaga,
    a system for Natural Language Analysis.
You can distribute it
    under the terms of the GNU General Public License.
malaga>
malaga> ma zeigen
Analyses of "zeigen":
1: [base: "zeigen",
    POS: verb,
    subtype: finite,
    subject: pl13,
    tense: present,
    valencies: <<acc, dat>, <acc>>]
2: [base: "zeigen",
    POS: verb,
    subtype: infinitive,
    valencies: <<acc, dat>, <acc>>]
malaga>
```

Der Malaga-Formalismus arbeitet nach dem Prinzip von nichtdeterministischen endlichen Automaten. Im Bereich der Morphologie werden die Allomorphe eingelesen und es wird überprüft, ob sie gemäß ihren morphologischen Kategorien konkateniert werden können. Im gegebenen Fall, d. h., wenn eine mögliche Fortsetzung erlaubt ist, wird das neue Allomorph an das bisher verarbeitete Allomorph bzw. an die bisher verarbeitete Allomorphkette angefügt.

[65] Vgl. <http://home.arcor.de/bjoern-beutel/malaga/> , 28.7.2014. Vgl. dort die Dokumentation zur aktuellen Version (7.12) von Malaga.

[66] Vgl. <http://ftp.hosteurope.de/mirror/ftp.opensuse.org/distribution/11.2/repo/oss/suse/x86_64/> , 28.7.2014, für die Malaga-Softwarepakete.

[67] Zu ausführlichen Informationen zu LAG siehe [Hausser 2001].

DATR

DATR, a language for the lexical knowledge representation[68], ist eine Datenstruktur, die als Vererbungsnetz bezeichnet wird, welche für die Darstellung von lexikalischen Informationen entwickelt wurde.[69]
DATR organisiert das lexikalische Wissen in Knoten. Sie bestehen aus einem Namen, einem Pfad und dem zugewiesenen Wert. DATR verfügt über eine gute Syntax für die Modellierung und Abfrage der lexikalischen Daten. Eine Anfrage wie SPRECHEN: <wortform ind präs sg 1 akv> liefert das Ergebnis s p r e c h e. Dabei ist SPRECHEN: der Name des Knotens während <wortform ind präs sg 1 akv> den Pfad bildet.
Im folgenden Listing (2.20) werden einige Abfragen einer Test-Grammatik für das Deutsche vorgestellt. Sie wurden unter zdatr, einer Implementierung von DATR ausgeführt.[70]

Listing 2.20: Ein Test mit DATR

```
zdatr test.dtr
Enter [Node:<path>], ['quit'] or ['trace#'].
Current Trace Level: 0

#1=> Arbeiten:< form ind präs sg 2 >
Arbeiten:< form ind präs sg 2 >
 = a r b e i t e s t .
Enter [Node:<path>], ['quit'] or ['trace#'].
Current Trace Level: 0

#2=> Arbeiten:< form ind präs pl 3 >
Arbeiten:< form ind präs pl 3 >
 = a r b e i t e n .
Enter [Node:<path>], ['quit'] or ['trace#'].
Current Trace Level: 0

#3=> Arbeiten:< form ind präs pl 2 >
Arbeiten:< form ind präs pl 2 >
 = a r b e i t e t .
Enter [Node:<path>], ['quit'] or ['trace#'].
Current Trace Level: 0

#4=> Buch:< pl dat >
Buch:< pl dat >
 = b ü c h e r n .
Enter [Node:<path>], ['quit'] or ['trace#'].
Current Trace Level: 0

#5=> Buch:< sg gen >
Buch:< sg gen >
 = b u c h e s .
```

[68] Vgl. <http://www.informatics.susx.ac.uk/research/groups/nlp/datr/datr.html> und <http://www.spectrum.uni-bielefeld.de/DATR/index.html> , 28.7.2014.
[69] "DATR is a simple, spartan language for defining nonmonotonic inheritance networks with path/value equations, one that has been designed specifically for lexical knowledge representation." <http://www.informatics.susx.ac.uk/research/groups/nlp/datr/datr.html> , 28.7.2014.
[70] Vgl. <http://www.spectrum.uni-bielefeld.de/DATR/Zdatr/> , 28.7.2014.

Die Modellierung lexikalischer und morphologischer Eigenschaften ist auf fast intuitive Weise möglich. Im folgenden Listing (2.21) wird das Prinzip der Modellierung illustriert:

Listing 2.21: Modellierung der Daten in DATR

```
WORT:
<wortform> == "<stamm>" "<suffix>"
...
<stamm sg> == b u c h
...
<suffix sg gen> == e s
...
Umlaut: <u> ==  ü <>
...
<stamm pl Umlaut> == b ü c h
...
<suffix pl gen> == e r
...
```

Die innere Flexion wie z. B. die Umlautung im Plural, *Buch* in *Bücher* kann ohne Schwierigkeiten in DATR modelliert werden, z. B. durch die Zerlegung der Stämme in Segmente, um sie partiell zu manipulieren bzw. durch Ersetzung der betroffenen Segmente, in diesem Fall *u* durch *ü* (Plural-Stamm).[71]

2.4 Ressourcen für das Albanische

Für das Albanische existieren eine Reihe von Ressourcen, die sowohl deskriptiven als auch präskriptiven Zwecken dienen. Insbesondere die *Akademie der Wissenschaften der Republik Albanien*[72] bzw. das *Institut für Sprach- und Literaturwissenschaft*[73] in Tirana sowie die *Akademie der Wissenschaften und Künste des Kosovo*[74] und das *Albanologische Institut*[75] in Prishtina seien hier erwähnt. Diese Institutionen haben im Laufe der Jahre, ab den 1950er Jahren, eine Reihe an Ressourcen veröffentlicht. Die entsprechenden elektronischen Versionen stehen aber leider nicht zur Verfügung. Dies hängt vielleicht damit zusammen, dass die genannten Institutionen bislang keine Zweige für MSV gegründet haben. Dass diese Ressourcen fehlen verzögert die Arbeit erheblich, denn die Digitalisierung ist mit vielen technischen Problemen sowie finanziellen Ausgaben verbunden. Dazu kommen noch die Urheberrechte der Ressourcen, die ggf. erworben werden müssten.

[71] [Trommer 2010] bietet eine kompakte Beschreibung von DATR auf Deutsch für den interessierten Leser.
[72] Alban. *Akademia e Shkencave e Republikës së Shqipërisë*.
[73] Alban. *Instituti i Gjuhësisë dhe Letërsisë*.
[74] Alban. *Akademia e Shkencave dhe e Arteve e Kosovës*.
[75] Alban. *Instituti Albanologjik*.

Im Folgenden wird auf die einzelnen Ressourcen-Typen kurz eingegangen:

2.4.1 Vorhandene Korpora

Als erster Ansatz könnte hier die digitale Version eines Romans als Teil der European Corpus Initiative/Multilingual Corpus Initiative[76] genannt werden, welche von Aleksander Murzaku bearbeitet wurde. Obwohl sehr weit entfernt von dem, was ein balanciertes Korpus bietet, gab es die Möglichkeit, aus dem elektronischen Text Stichproben des Albanischen zu erstellen, wie etwa Frequenzlisten der Textwörter, Wort- und Satzlänge, sowie die Verteilung der Buchstaben zu berechnen.

Listing 2.22: ECI/MCI Koncert

```
Koncert

Duke hapur derën e apartamentit, ku zilja kishte një copë
here që binte me këmbëngulje, nënqeshja me të cilën Silva
bëhej gati të priste mysafirët e parë, i mbeti në buzë. Në
vend të mysafirëve ajo pa një burrë, që mbante në krahë
një fuçi të rëndë, sipër së cilës dilnin degët e një limoni.
-Familja Gjergj Dibra? - pyeti burri.
-Po, - tha Silva pakëz e hutuar. - Ah, ju keni sjellë këtë
limon për ne?
-E keni porositur, apo jo?
Pa e bërë të gjatë njeriu hyri brenda në korridor.
-Ku do ta vendosni? - pyeti ai me njëfarë padurimi. Ndihej
menjëherë që fuçia ishte e rëndë
-Kujdes! - tha Silva.
- Këtej ju lutem, - dhe hapi derën e njërës prej dhomave.
Njeriu kaloi me hapa të rëndë mes për mes dhomës, për të
dalë në ballkon, derën e të cilit Silva porsa e kishte hapur.
```

Ein balanciertes Korpus für die albanische Sprache, das als Referenzkorpus dienen könnte, fehlt noch. Jedoch gibt es seit jüngster Zeit Initiativen sowohl in Tirana als auch in Prishtina, ein solches Korpus zu erstellen. Man vergleiche hierzu [Arkhangelskij et al. 2012] und [Caka/Caka 2012]. Sie beabsichtigen auch eine morphologische Annotation. Beide Arbeiten stehen noch in einer Anfangsphase. Bis das erste Referenzkorpus des Albanischen zur Benutzung bereitsteht, dauert es sicherlich noch.

[76] Vgl. [ECI/MCI 1994].

2.5 Maschinelle Verarbeitung der albanischen Lexikographie

[ECI/MCI 1994] beinhaltet auch ein Lexikon (oder eine Wortliste) mit 32 000 Einträgen des albanischen Grundwortschatzes, die mit minimalen grammatischen Angaben versehen sind, genauer mit Wortart- bzw. Genusangabe bei Substantiven oder Transitivität bei Verben.

Listing 2.23: Murzakus *Reverse Dictionary of Albanian Language*

```
...
f,dhEnE
adj,dhEnE
part,dhEnE
f,gojEdhEnE
m,kokEdhEnE
adj,dorEdhEnE
adj,faqehEnE
adj,vetullhEnE
f,thEnE
part,thEnE
conj,domethEnE
f,fatthEnE
...
f,mirE
adj,mirE
n,mirE
adv,mirE
adj,gojEmirE
adj,fjalEmirE
adj,kEshillEmirE
adj,orEmirE
adj,dorEmirE
adj,besEmirE
...
trans,pErmbledh
m,bredh
intrans,bredh
trans,dredh
...
```

In der Originalversion der Wortliste stehen C für ç und E für ë, da sich UTF-8 und Unicode damals noch nicht durchgesetzt hatten; stattdessen wurde ASCII bzw. ISO-8859-1 für die Kodierung der Schriftzeichen verwendet.

Eine umkodierte Version des Abschnitts ist im folgenden Listing (2.24) angegeben:

Listing 2.24: Umkodierte Version des Murzakus *Reverse Dictionary* ...

```
...
f,dhënë
adj,dhënë
part,dhënë
f,gojëdhënë
m,kokëdhënë
adj,dorëdhënë
adj,faqehënë
adj,vetullhënë
f,thënë
part,thënë
conj,domethënë
f,fatthënë
...
f,mirë
adj,mirë
n,mirë
adv,mirë
adj,gojëmirë
adj,fjalëmirë
adj,këshillëmirë
adj,orëmirë
adj,dorëmirë
adj,besëmirë
...
trans,përmbledh
m,bredh
intrans,bredh
trans,dredh
...
```

Ebenfalls 1994 erschien das Werk „Rückläufiges Wörterbuch der albanischen Sprache“ von Marko Snoj, welches mehrere grammatische Angaben bietet, im Gegensatz zu der Arbeit Murzakus aber nur in gedruckter Form, wie herkömmliche Wörterbücher, zur Verfügung steht. Betrachtet man das Wörterbuch, insbesondere die Statistiken am Ende, genauer, wird klar, dass es mit computerlinguistischen Methoden erstellt wurde.[77]

Listing 2.25: Snojs *Rückläufiges Wörterbuch der Albanischen Sprache*

```
...
sprov/ój V. tr. ~óva ~úar
konserv/ój V. tr. ~óva ~úar
rezerv/ój V. tr. ~óva ~úar
lex/ój V. tr. ~óva ~úar
rilex/ój V. tr. ~óva ~úar
gux/ój V. intr. ~óva ~úar
harxh/ój V. tr. ~óva ~úar
baz/ój V. tr. ~óva ~úar
...
```

[77] Die Angaben reichen für die Erstellung von Klassen nicht aus, insbesondere nicht für die Konjugation. Die Werke [ECI/MCI 1994] und [Snoj 1994] wurden fast gleichzeitig publiziert, allerdings mit verschiedenen Schwerpunkten. In [Snoj 1994] sind deutlich mehr grammatische Informationen enthalten.

Die umgekehrte rückläufige Version des Wörterbuches sieht wie im folgenden Quellcode (2.26) angegeben aus. In beiden Versionen sind die Akzentzeichen mit angegeben. Einige Zeichen wie z. B. das ë mit ˊ ist nicht in Unicode enthalten und vom Format UTF-8 nicht darstellbar. Stellvertretend wird für Zwecke der maschinellen Sprachverarbeitung *ê* bzw. *Ê* verwendet.

Listing 2.26: Snojs *Wörterbuch der Albanischen Sprache*

```
...
lexím S. m. ~i; Pl. ~e, ~et
lexóhet V. refl., pass.
lex/ój V. tr. ~óva, ~úar
lexúes S. m. ~i; Pl. ~, ~it
lexúesh/ëm Adj. (i), ~me (e)
...
lezhján S. m. ~i; Pl. ~ë, ~ët
lezhján Adj. ~e
lezhján/e S. f. ~ja
...
```

AMM Lex ist das elektronische Wörterbuch, das als Teilarbeit von [Kabashi 2003] entstanden ist. Es beinhaltet die Verben des Albanischen sowie die Indeklinabilia (die unflektierbaren Wortarten) des Albanischen. Das Format entspricht dem der Entwicklungsumgebung und Programmiersprache *Malaga*.

Das Lexikon ist als eine Menge/Liste von Attribut-Werte-Paaren aufgebaut und ermöglicht einen besseren, d. h. gezielten Zugriff im Vergleich zu den bereits erwähnten Lexika (Murzaku und Snoj), auf welche nur über die Oberfläche zugegriffen werden kann.

Listing 2.27: AMM-Lexikon

```
...
[Lemma: "bredh", POS: Verb, Type: Intr., Conjug.: CT16];
[Lemma: "dredh", POS: Verb, Type: Tr., Conjug.: CT16];
[Lemma: "hedh", POS: Verb, Type: TrIntr., Conjug: CT16];
...
[Lemma: "lexoj", POS: Verb, Type: Tr., Conjug.: CT1];
[Lemma: "lidh", POS: Verb, Type: Tr., Conjug.: CT14];
...
```

2.5.1 Vorhandene morphologische Ressourcen

Als morphologische Ressourcen zählen auch die Suffixe des Verbal- und Nominalsystems in ihrer üblichen Klassifikation in Klassen und Unterklassen sowie Wortbildungsmittel, welche in Kapitel 3 (Morphologie des Albanischen), Abschnitt 3.4 behandelt werden.
Zusammen mit dem Lexikon (bzw. den Lexikoneinträgen) bilden diese Mittel ein System, das unter einem als Software implementierten Formalismus läuft und entweder interaktiv bei einer Anfrage eine Wortform produziert bzw. analysiert oder aus dem erstellten System (Modell) ein Vollformlexikon generiert.

2.5.2 Vorhandene syntaktische Ressourcen

Ein syntaktisch getaggtes Korpus, ob handannotiert oder nicht, ist bislang ebenfalls nicht verfügbar. Ebenso gibt es kein Valenzlexikon, kein Lexikon der Konjunktionen und kein Lexikon mit detaillierten Informationen über die Rektion der Präpositionen oder ein Inventar der Satzbaupläne.[78]

2.5.3 Vorhandene semantische Ressourcen

Ein Lexikon mit semantischen Angaben wie [Motsch 2004] oder [Helbig 2008] gibt es für das Albanische noch nicht. Auch eine Ressource vom Typ Ontologie (etwa WordNet, GermaNet und Dornseiff) gibt es noch nicht. Auch BalkaNet beinhaltet Albanisch nicht.[79]
Die Synonyme des Albanischen werden in den Werken [Dhrimo et al. 2002/2007] und [Thomai et al. 2004] behandelt, welche in Form gedruckter Wörterbücher erschienen sind. Theoretische Auseinandersetzungen mit verschiedenen Aspekten dieses Themas sind die Aufsatzsammlungen bzw. Monographien [Thomai 2004] und [Thomai 2005]. [Samara 1985] hat sich mit den Antonymen beschäftigt.

2.6 Maschinelle Verarbeitung der albanischen Morphologie

Auch das Interesse an der maschinellen Verarbeitung der albanischen Morphologie erwachte im Vergleich zu den Nachbarsprachen und anderen Sprachen Europas viel später.

[78] Ein Beispiel für Satzbaupläne wären die Satzmuster für Verben in [Wahrig 1998: 32–35].
[79] Vgl. [Pala 2005].

2.6.1 Erste Ansätze in der albanischen Morphologie

Den nächsten Schritt in die maschinelle Sprachverarbeitung des Albanischen, den ersten im Bereich der Morphologie, macht die Arbeit von Jochen Trommer im Jahre 1997. Er behandelt die Verben des Albanischen im Rahmen der mo_lex, einer Repräsentationssprache für Endliche Automaten (FSA)[80]. Trommer geht auf ihre einzelnen Eigenschaften ein, sowohl bei den regulären Typen als auch bei vielen Ausnahmen. Das System ist leider auf Verben beschränkt und steht als Werkzeug (anwendbares System, Regeln und Lexikon) nicht zur Verfügung.
2002/3 folgt die Arbeit des Verfassers dieser Zeilen, welche ebenso die Morphologie der Verben des Albanischen zum Thema hatte. Das Nominalsystem, d. h. die Substantive, Adjektive, Numeralia, Pronomina und Artikel blieben unbehandelt.[81]

Ein Beispiel aus [Kabashi 2003] unter Verwendung von Version 7.12 von *Malaga* wird im folgenden Listing (2.28) vorgestellt:

Listing 2.28: AMMv

```
malaga almor.pro
malaga> ma ishte
Analyses of "ishte":
1: [Valency: <<Nom>,<>,<Nom,Acc>>,
    Segmentation: "ish<SFX>te",
    WordForm: "ishte",
    BaseForm: "jam",
    POS: Verb,
    VerbType: AuxVerb,
    Conjugation: <[PersonNumber: Sg3,
                   Tense: Imperfect,
                   Mode: Indicative,
                   Voice: Active,
                   Admirative: no]>,
    Terminal: yes,
    AnalysisType: Parsed]
malaga>
malaga> quit
```

Kurz darauf publizierten Jochen Trommer und Dalina Kallulli im Rahmen der LREC 2004 einen Artikel über einen Tagger (Morphological Analyser) für das Albanische. Der Tagger wurde mit Hilfe der Programmiersprache Python und der Datenbank MySQL erstellt. Sein Grundlexikon basiert hauptsächlich auf der Wortliste von [ECI/MCI 1994]. Die Erkennungsrate (Precision/Recall) ist für Tokens 97-98%/94-95% und für Types 96-97%/92-93%.[82]

[80] Vgl. hierzu [Trommer 1997].
[81] Vgl. hierzu [Kabashi 2003].
[82] Vgl. [Trommer/Kallulli 2004] für mehr Details.

Obwohl das Testen des Werkzeugs an kleinen Datenmengen durchgeführt wurde, scheint es ein gutes Werkzeug zu sein, das albanische Texte annotieren könnte.[83]

2007 erstellten Odile Piton, Klara Lagji und Remzi Përnaska ein Werkzeug zur Bearbeitung von Lexika des Albanischen.[84] Die Autoren beabsichtigen Lexikoneinträge samt ihrer Flexionsformen und zum Teil auch Wortbildung mit Hilfe von Finite-State-Transducers und dem Werkzeug *NooJ* zu modellieren. Es ist nicht klar, in welchen Umfang dies gemacht werden sollte bzw. wurde – die Autoren machen dazu keine Angaben. Ebenso wurden dem Leser keine möglichen Erkennungsraten und Testdaten vorgestellt.

2010 präsentierte Arbana Kadriu einen Aufsatz zur Verarbeitung der albanischen Nomina und Verben, vgl. [Kadriu 2010], der auf statistischen Methoden basiert und sie im Rahmen der Two-Level-Modelle realisiert.

[Kadriu 2010] modelliert einen Zwei-Ebenen-Formalismus für die Nomina und Verben des Albanischen. Die verwendete Methode basiert auf maschinellem Lernen, wobei eine bestimmte Zahl an Verb- (4 200) und Substantivformen (100) als Eingabe dient, vgl. 305 ff. und 307 f. Das System wurde im Rahmen von *PC-KIMMO* implementiert. Die Tests wurden an relativ kleinen Datenmengen durchgeführt, für Substantive 856, für Verben 4200 Wortformen.

Eine weitere Arbeit ist die Master-Arbeit von Besmir Hasanaj, vgl. [Hasanaj 2012]. Von ihm wurde Apache OpenNLP[85] (*maximum entropy* und *perceptron*) als Werkzeug für maschinelles Lernen und linguistische Datenverarbeitung (41 f.) verwendet. Das Trainieren des Tools für POS-Tagging wurde an kleinen Datenmengen (3312 Textwörter, 138 Sätze), vgl. S. 42 f., durchgeführt. Die Zeichen ç/Ç und ë/Ë scheinen vernachlässigt worden zu sein, wie aus der Abbildung 5-1 ersichtlich ist, vgl. S. 46. Stattdessen scheinen c/C und e/E benutzt worden zu sein, was höchstwahrscheinlich zur Datenverfälschung im Prozess der Verarbeitung führt.

---

[83] Auf der Internetseite <http://sol.cl-ki.uni-osnabrueck.de/~atag/> , 28.7.2014, der Universität Osnabrück sind einige Komponenten des Taggers vorhanden, allerdings nicht alle, d. h., von einem interessierten Anwender sind sie somit zum Taggen leider nicht einsetzbar. Es sind einige getaggte Texte vorhanden, was hingegen auf den Einsatz des Taggers seitens der Entwicklern hinweist. Wichtige Ressourcen, wie z. B. das Lexikon, fehlen, oder sie stehen auf der angegebenen Adresse nur ansatzweise zur Verfügung, wie z. B. einige Morphologieregeln. Nach Kenntnis des Autors der vorliegenden Arbeit ist die Software in einer anderen bzw. vollständigen Version nicht verfügbar.

[84] Vgl. hierzu [Piton et al. 2007].

[85] Vgl. <http://opennlp.apache.org/> , 28.7.2014.

2.6.2 Stemming

Als Stemming wird die Trennung von Suffixen der Wortformen eines Paradigmas bezeichnet, um den Stamm bzw. die Stämme des Paradigmas zu gewinnen. Stemming wird vor allem im Bereich Information Retrieval verwendet, um die Semantik eines Textes im Rahmen der maschinellen statistischen Sprachverabeitung zu erschließen. Dabei ist die morphologische Information unwichtig oder nimmt eine untergeordnete Rolle ein. Sie wird oft nicht berücksichtigt.
Jetmir Sadiku und Marenglen Biba, vgl. [Sadiku/Biba 2012], stellen eine Arbeit vor, die das sogenannte Stemming behandelt. Die Autoren gehen regelbasiert vor, indem sie die Suffixe abtrennen, um die Stämme zu extrahieren. Sie behandeln auch Ansätze der Wortbildung im Rahmen von Stemming.

2.6.3 Rechtschreibsysteme

In der letzten Zeit zeichnen sich einige Initiativen ab, Rechtschreibsysteme zu entwickeln (engl. spell checker), die jedoch noch in ihren Anfängen stehen. Sie werden daher im Folgenden nicht berücksichtigt. Sie sind nur Listen aus Vollformen von Wörtern, nicht vollständig, und bieten keine oder bei weitem nicht ausreichende lexikalische bzw. morphologische Informationen, was letzten Endes auch nicht ihr Ziel ist. Somit sind sie auch nicht vergleichbar mit einem computerlinguistischen Lexikon bzw. mit einer Morphologie.

2.7 Zielsetzung: Ein einsetzbares Gesamtsystem der Morphologie

Eine maschinelle Morphologie bildet den Kern für viele automatisierte Anwendungen im Rahmen der Sprachtechnologie. Desiderata an ein Morphologie-System sind folgende:

- Das System sollte alle Wortarten behandeln, sodass zumindest die Flexion einer Sprache abgedeckt werden kann. Ein solches System könnte mit Unterstützung einer Lexikon-Komponente vielen Anwendungen dienen. Ein von Zeit zu Zeit erweitertes Lexikon würde auch Entwicklungen bzw. Variationen in Wortschatz und Wortbildung, seien diese zeitlich bedingt oder domänenspezifisch, abdecken. Dieser Stand gleicht fast einem Vollform-Lexikon, das nicht flexibel genug in Hinsicht auf Erkennung von Neologismen und die Segmentierung

von Wortformen (für die Zwecke der Lemmatisierung oder Kategorisierung) wäre.

- Die modernen Systeme der morphologischen Analyse erlauben neben den genannten Eigenschaften auch eine Analyse hinsichtlich der Wortbildung, d. h. der Derivation und Komposition, sowie Hypothesen zu Wortformen, die nicht „bekannt" sind. Die Letzteren werden anhand verschiedener zusätzlicher Kriterien, wie etwa nach Flexionsendung, automatisch analysiert und schließlich kategorisiert, ggf. auch mit einer Markierung versehen, die die hypothesen-basierte Verarbeitung andeutet, um eventuell falsch analysierte Wortformen schnell und gezielt zu suchen bzw. zu identifizieren.

2.8 Zusammenfassung des 2. Kapitels und Schlussbemerkungen

In diesem Kapitel wurden die sprachlichen Ressourcen, insbesondere lexikalische und morphologische Ressourcen sowie einige wichtige Modelle der maschinellen Morphologie vorgestellt. Parallel wurden diese Themen in Zusammenhang mit der maschinellen Verarbeitung des Albanischen behandelt.
In Anbetracht der vorgestellten Ressourcen für das Albanische lässt sich feststellen, dass die Entwicklung der primären Ressourcen für das Albanische die weitere maschinelle Verarbeitung beschleunigen könnte, da viele Ressourcen auf Lexika und Morphologien aufbauen.

3 Die Morphologie des Albanischen

Mit dem Begriff Morphologie einer Sprache wird diejenige Teildisziplin der Grammatik bezeichnet, die sich mit den Eigenschaften der Wörter beschäftigt.[86] Damit sind Eigenschaften wie Wortart (Substantiv, Adjektiv, Verb, Adverb usw.), je nach Fall Deklination (Kasus, Numerus, Genus usw.) oder Konjugation (Person, Tempus, Modus usw.), entsprechende Deklinations- bzw. Konjugationsklassen, Aufbaustruktur der Wörter sowie weitere Eigenschaften gemeint, welche die Wörter näher charakterisieren. Diese Eigenschaften bestimmen die Bedeutung eines Wortes näher und ermöglichen ihm eine flexible Verwendung in einem Kontext, in dem es mit anderen teilnehmenden Wörtern in einer Äußerung bzw. in einem Satz semantisch und funktional kongruiert.

Da das Hauptziel der vorliegenden Arbeit die automatische morphologische Analyse des Albanischen ist, wird in diesem Kapitel ein kurzer Überblick über die Eigenschaften des Albanischen auf der morphologischen Ebene gegeben.[87] In diesem Kapitel wird sehr oft auf weiterführende Literatur verwiesen, da keinesfalls beabsichtigt wird, eine Grammatik bzw. eine Morphologie des Albanischen neu zu schreiben.[88]

Im ersten Abschnitt werden allgemeine Angaben zum Albanischen gegeben: seine Stellung in der indogermanischen bzw. indoeuropäischen Sprachfamilie, kurze Informationen über seine geographische Position sowie seine Dialekte. Als Nächstes wird das Laut- und Schriftsystem des Albanischen vorgestellt, insbesondere die Laute, das Alphabet, einige Informationen über die Rechtschreibung, den Akzent sowie die Phonemveränderungen. Im dritten Abschnitt werden die Wortarten behandelt, wobei sie einzeln und

[86] Siehe z. B. [Matthews 1997], [Bussmann 2002] oder [Glück 2011] für die linguistischen Begriffe, u. a. Wort und Morphologie.

[87] Für weitere Informationen über die Grammatik des Albanischen, die hier nicht berücksichtigt werden können, sei der interessierte Leser vor allem auf das Werk [Buchholz/-Fiedler 1987] verwiesen.

[88] Insbesondere werden die folgenden Werke zitiert: [Kostallari et al. 1984], [Buchholz/-Fiedler 1987], [Buchholz et al. 1993], [Hetzer/Finger 1993], [Morfologjia 1995], [Çeliku et al. 1998] und [Fiedler 2003].

ausführlich beschrieben werden. Schließlich wird auf die wichtigsten Eigenschaften der Wortbildung eingegangen, sowie eine Zusammenfassung des Kapitels gegeben.

3.1 Das Albanische

Die albanische Sprache (alban. *gjuha shqipe* ['ɟuhɑ 'ʃcipɛ]) ist ein Mitglied der indoeuropäischen Sprachfamilie. Sie ist wie das benachbarte Griechische ein Einzel-Mitglied in ihrem Zweig neben den Gruppen, die mehrere Mitglieder besitzen, wie die romanische oder die germanische Sprachgruppe.
Das Albanische wird in Albanien, im Kosovo, im westlichen Teil von Mazedonien, im Südosten von Montenegro, im Südwesten von Serbien, im Nordwesten von Griechenland sowie im Süden von Italien, in vielen Orten, gesprochen. In Albanien und im Kosovo ist es Amtssprache, während es in Mazedonien nur in bestimmten Gebieten und in bestimmten Institutionen als solche auftritt. Albanisch ist die Erstsprache von ca. sieben Millionen Menschen in allen genannten Gebieten. Schätzungsweise wird es von weiteren drei Millionen Menschen außerhalb der genannten Gebiete gesprochen, wie bspw. in den Staaten Westeuropas, in den USA, in Kanada, in der Türkei und in Australien.
Die Bezeichnungen (engl. *Language codes*) für die albanische Sprache nach der ISO (*International Organization for Standardization*) / SIL (*Summer Institute of Linguistics*) sind: sq (ISO 639-1), sqi (T) / alb (B) (ISO 639-2), sqi (ISO 639-3). Der Code für die Hauptsprache (engl. *Macrolanguage*) ist sqi; Die Einzelbezeichnungen (engl. *Individual Codes*) der Hauptdialekte sind: aln für *Gegisch* (engl. *Gheg*), aat für *Arvanitisch*[89] (griech. *Arvanitika*, engl. *Arvanitic*), gesprochen in Griechenland, aae für *Arbërisht*[90] gesprochen in Italien und als für *Toskisch*.[91]

[89] <http://www.ethnologue.com/show_language.asp?code=aat> ,1.4.2012; Es handelt sich um einen albanischen Dialekt, der von der im 14. Jahrhundert nach Griechenland ausgewanderten Bevölkerung stammt. Arvanitisch wird in rund 300 Ortschaften von rund 50 000 Personen gesprochen.

[90] <http://www.ethnologue.com/show_language.asp?code=aae> ,1.4.2012; Es handelt sich um einen albanischen Dialekt, der von der im 15. Jahrhundert nach Italien ausgewanderten Bevölkerung aus Griechenland (Arvaniten) und Süd-Albanien stammt. Arbërisht wird in rund 80 Ortschaften von rund 80 000 Personen gesprochen.

[91] Für weiterführende Informationen zu diesem Thema siehe <http://www.sil.org/iso639-3/codes.asp>, <http://www.sil.org/iso639-3/documentation.asp?id=sqi> (SIL) sowie <http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_de

3.2 Das Laut- und Schriftsystem des Albanischen

Bevor auf die einzelnen Themen der Morphologie eingegangen wird, bedarf es einiger Hinweise zur Phonetik bzw. Phonologie sowie zum Alphabet und zur Rechtschreibung des Albanischen.[92]

3.2.1 Die Laute

Die normierte albanische Sprache hat 36 Laute. Die entsprechenden Lautzeichen nach IPA (International Phonetic Alphabet) sind: a [ɑ], b [b], c [ts], ç [ʧ], d [d], dh [ð], e [ɛ], ë [ə], f [f], g [g], gj [ɟ], h [h], i [i], j [ʲ], k [k], l [l], ll [ɫ], m [m], n [n], nj [ɲ], o [ɔ], p [p], q [c], r [ɾ], rr [r], s [s], sh [ʃ], t [t], th [θ], u [u], v [v], x [dz], xh [ʤ], y [y], z [z] und zh [ʒ], vgl. [Buchholz/Fiedler 1987: 26].
Manche albanische Linguisten, etwa [Memushaj 2010: 116 f. (§ 3.10.2)], transkribieren ‚j' mit [j]. Memushaj [op. cit.] begründet dies mit Beispielen wie *ji* (dt. *sei*) und *bij* (dt. *Söhne*), wo die beiden Phoneme [i] und [j] nebeneinander vorkommen.

3.2.2 Das Alphabet

Die Laute der albanischen Sprache werden durch die 25 Grapheme des lateinischen Alphabets A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, P/p, Q/q, R/r, S/s, T/t, U/u, V/v, X/x, Y/y und Z/z, durch die neun Digraphen DH/Dh/dh, GJ/Gj/gj, LL/Ll/ll, NJ/Nj/nj, RR/Rr/rr, SH/Sh/sh, TH/Th/th, XH/Xh/xh und ZH/Zh/zh sowie durch die zwei mit Diakritika versehenen Zeichen Ç/ç und Ë/ë repräsentiert. Es sind 7 Vokale (a, e, ë, i, o, u sowie y) und 29 Konsonanten (die restlichen Zeichen: b, c, ç, d, f, ..., xh sowie zh). Die Schriftzeichen (Buchstaben) stehen im Albanischen in folgender Reihe: a, b, c, ç, d, dh, e, ë, f, g, gj, h, i, j, k, l, ll, m, n, nj, o, p, q, r, rr, s, sh, t, th, u, v, x, xh, y, z und zh.[93]

---

tail_ics.htm?csnumber=22109> (ISO 639) und <http://www.iso.org/iso/iso_catalogue_tc/catalogue_detail.htm?csnumber=39718>, 1.4.2012

[92] Vgl. [FDSH 1976], [Fiedler 2003], [Kabashi 2009, Alphabet] und [Memushaj 2010].

[93] Dies sei hier aufgrund der Tatsache erwähnt, dass das Alphabet als Schlüssel für die Sortierung dient. Es ist insbesondere für das Lexikon wichtig, denn die Zeichen ë und ç werden bei nicht-albanischer Spracheinstellung falsch sortiert. Sie werden oft am Anfang oder am Ende der anderen Zeichen eingeordnet, bspw. bei englischer oder deutscher Spracheinstellung. Ebenso werden die Digraphen als Untermenge des jeweiligen „ersten Graphems" organisiert, bspw. dh unter d.

3.2.3 Die Rechtschreibung

Die jetzt gültige Rechtschreibregelung ist seit 1972 festgelegt. Sie kodiert die Rechtschreibung anhand der normierten Sprache, welche mehr oder weniger eine „Auswahl“ aus den beiden großen Dialekten des Albanischen, aus dem Gegischen (alban. *gegërisht* [gegisch: *gegnisht*], kurz *geg.*) und aus dem Toskischen (alban. *toskërisht*, kurz *tosk.*) ist, wobei das Toskische überwiegt. Zum Beispiel ist das Wort *zanore* (dt. *der Vokal*) aus dem Gegischen übernommen, während *hëna* (dt. *der Mond*) aus dem Toskischen stammt. Weitere Beispiele wären: *rërë*, tosk. (dt. *Sand*) / *ranor* (dt. *sandig*), *ranishtë*, geg. (dt. *Sandstrand* oder *Sandstreifen*). Die Hauptunterschiede im Bereich der Phonetik bzw. Rechtschreibung und Morphologie liegen im Lautwandel n→r (im sogenannten *Rhotazismus*) und a→ë, vgl. hierzu [Gjinari 2007 und 2008] für ausführliche Informationen.[94] Weitere Unterschiede liegen im Bereich der Syntax bei der Konstruktion des Infinitivs, bspw. *me punue*, geg. vs. *për të punuar*, tosk. (dt. *arbeiten*).
Als Grundlage für die Rechtschreibung dienen vor allem Rechtschreibwörterbücher, wie bspw. das [FDSH 1976], [Fjalori 1980], sowie andere Literatur zur Sprachpflege, bspw. [Kostallari et al. 1984].[95] In der Regel gelten die von der Akademie der Wissenschaften publizierten (Standard-)Werke als präskriptiv bzw. als Richtlinie für die Rechtschreibung und Sprachpflege.

3.2.4 Der Akzent

Durch die Konventionen der Rechtschreibung werden einige Wörter bzw. Wortformen im Albanischen gleich geschrieben (Homographie/Homonymie), bspw. (1) *bar*, ∼*i*, ∼*ëra*, ∼*ërat*; (dt. *Gras*) (2) *bar*, ∼*i*, ∼*e*, ∼*et* (dt. *Lokal*, *Theke* usw.); (3) *bar*, ∼*i*, ∼*na*, ∼*nat* (dt. *Arznei*); (4) *bar*, ∼*i* (Maßeinheit, undekl.); (5) *barí*, ∼*u*, ∼*nj*, ∼*njtë* (dt. *Hirt*); die Form *bar|i* (in 1, 2, 3 und 4) wird wie die Form *bari* (in 5) geschrieben, obwohl es sich um fünf verschiedene Wörter handelt. Daher werden in der Sprachbeschreibungsliteratur oft (insbesondere im Fremdsprachunterricht) die Akzentzeichen verwendet, um den Akzent des Wortes zu markieren. Es werden zu diesem Zweck oft die folgenden Zeichen gebraucht: *á*, *à*, *â* / *é*, *è*, *ê* / *é̋*, *è̋*, *ê̋* / *í*, *ì*, *î* / *ó*, *ò*, *ô* / *ú*, *ù*, *û*, und *ý ỳ* bzw. *ŷ*.[96] Ein illustratives Beispiel wäre *dhe* (dt. *und*)

[94] Diese Phänomene kommen in einigen Wörtern auch im Norm-Wortschatz vor und sind insofern relevant.

[95] Vgl. zu diesem Thema auch die Monographien von Memushaj [2004 und 2010] sowie die Monographie von Dhrimo und Memushaj [2011].

[96] Vgl. hierzu bspw. [Snoj 1995], wo einige dieser Zeichen verwendet wurden. Ebenso wurden im „Standard“-Wörterbuch der albanischen Sprache (Fjalori 2006) die

vs. *dhé* (dt. *Erde*), wo die Akzentsetzung einer schnelleren Disambiguierung dient. Einen Beleg findet man bspw. in [ECI/MCI 1994, Datei `alb01`].
In Zusammenhang mit den Flexionsparadigmen und den Lexikoneinträgen ergeben sich einige Besonderheiten, vor allem bei Verschiebung des Akzentes bei einigen Wortformen, wie in den folgenden Beispielen:[97]

- *njerí*, *njeríu* vs. *njérëz*, *njérëzit* (dt. *Mensch*);
- *lúmë*, *lúmi* vs. *luménj*, *luménjtë* (dt. *Fluss*);

3.2.5 Phonemalternationen

Die folgende Darstellung in Form einer Auflistung soll dazu eine kompakte Übersicht über die Phonemveränderungen im Albanischen geben. Sie macht keine Unterscheidung, ob die entsprechende Änderung im Nominalsystem oder im Verbalsystem vorkommt bzw. welche Eigenschaft oder Kategorie sie repräsentiert. Sie wurde einfach nach den möglichen Änderungen, grob, mit Beispielen, welche hier nicht übersetzt und kategorisiert werden, aufgestellt. Die Phonemveränderungen im Albanischen nach [Memushaj 2010: 130–135 (§ 4.2)] sind:

- Vokalveränderungen:
  - `a`∼`e` [*at*∼*etër*, *krap*∼*krep*, *jam*∼*je*, *marr*∼*merr* und *flas*∼*flet*]
  - `e`∼`i` [*breg*∼*brigje*, *njeh*∼*njihni*, und *sheh*∼*shihni*]
  - `je`∼`i` [*mbjell*∼*mbillni*, *ndjell*∼*ndillni*, und *sjell*∼*sillni*],
  - `o`∼`e` [*njoh*∼*njeh*, *shoh*∼*sheh* und *i␣vogël*∼*të␣vegjël*],
  - `a`∼`ë` [*dhashë*∼*dhënë*, *lashë*∼*lënë* und *thashë*∼*thënë*],
  - `ë`∼`i` [*lë*∼*lini*, *nxë*∼*nxini*, *vë*∼*vini* und *zë*∼*zini*],
  - `a`∼`ë` [*shpargër*∼*shpërgënj* und *i␣madh*∼*të␣mëdhenj*];
  - `e`∼`a` [*rreth*∼*rrathë*, *pëlle*∼*pëlla*, *e␣re*∼*të␣ra* und *them*∼*thashë*];
  - `a`∼`o` [*dal*∼*dola*, *marr*∼*mora* und *flas*∼*fola*];
  - `e`∼`o` [*dredh*∼*drodha*, *hedh*∼*hodha* und *heq*∼*hoqa*];

Lemmata mit Akzentzeichen versehen. Für dialektale Sprachvarianten werden weitere Zeichen verwendet. Ebenso wurden vor der Normierung der Sprache bzw. vor der Normierung des Alphabets (1908) verschiedene Zeichen verwendet, welche hier nicht berücksichtigt werden können. Vgl. hierzu auch die Zeichen in [Fiedler 2003: § 2] und in [Gjinari 2007 und 2008].

[97] Nach [Fiedler 2003].

je∼o [*mbjell∼mbolla, nxjerr∼nxora* und *pjek∼poqa*];

ë∼u [*dhëndër∼dhëndurë*; *vë∼vura* und *zë∼zura*];

e∼ë [*i␣keq∼të␣këqinj*];

- Konsonantenveränderungen:

ll∼j [*akull∼akuj, artikull∼artikuj* und *avull∼avuj*];

l∼j [*kalë∼kuaj, muskul∼muskuj* und *stimul∼stimuj*];

r∼j [*bir∼bij, flamur∼flamuj* und *lepur∼lepuj*];

n∼nj [*balun∼balunj, bërshen∼bërshenj* und *carran∼carranj*];

g∼gj [*djeg∼digj-ni, lëng∼lëngj-e, muzg∼muzgj-e* und *zog∼zogj*];

k∼q [*bujk∼bujq, dushk∼dushq-e, gjak∼gjaq-e* und *ujk∼ujq*];

rr∼r [*bjerr bor-a, marr∼mor-a, nxjerr∼nxor-a* und *tjerr∼tor-a*];

s∼t [*flas∼flet, ngas∼nget* und *zbres∼zbret*];

t∼s [*këput-ja∼këpus-te, mat-ja∼mas-te* und *tret-ja∼tres-te*];

- Andere Veränderungen:

ie∼je [*bie∼bjer, ndiej∼ndje-va* und *zie-j∼zje-va*];

ie∼i [*bie∼bi-ni, ndie-j∼ndi-het* und *shtie∼shti-ni*];

ie∼u [*shpie∼shpu-ra*];

je∼ie [*ndje-u∼u␣ndie, përzje-u∼u␣përzie* und *zje-u∼u␣zie*];

o∼ua [*dorë∼duar, blo-va∼blua-m* und *shpo-va∼shpua-m*];

e∼ye [*derë∼dyer, le-va∼lye-m* und *ngje-va∼ngjye-m*];

ye∼e [*fyell∼fej-e, krye∼kre-rë* und *krye-j∼kre-va*];

ye∼y [*lye-j∼ly-hem, ngjye-j∼ngjy-hem* und *shqye-j∼shqy-hem*];

ua∼o [*ftua∼fto-nj, krua∼kro-je* und *krua-j kro-va*];

ua∼u [*drua-j∼dru-hem, dua∼du-hem* und *qua-j∼qu-hem*];

ë∼♯ [*vegël∼veg♯l-a, vepër∼vep♯r-a* und *i␣vogël∼i␣vog♯l-i*];

a∼♯ [*fukara∼fukar♯-enj* und *qerrata∼qerrat♯-enj*];

ër∼♯ [*drapër∼drap♯-inj* und *gjarpër∼gjarp♯-inj*];

e∼♯ [*lule∼lul♯-ja* und *nuse∼nus♯-ja*];

♯∼n [*arne♯∼arnen-i, bli♯∼blin-i* und *mëti♯∼mëtin-i*];

r∼♯ [*bër-a∼bë♯-më, hyr-a∼hy♯-më* und *zur-a∼zu♯-më*];

♯∼r [*dru♯∼drur-i, kufi♯∼kufiri, ulli♯∼ullir-i* und *zë♯∼zër-i*];

s∼♯ [*ngas∼nga♯-va, shkas∼shka♯-va* und *pres∼pre♯-va*];

♯∼t [*di♯∼dit-a, gogësi♯-j∼gogësit-a* und *ngre♯∼ngrit-a*];

Die folgende Tabelle gibt einen Überblick über die innere Flexion (Lautwandel im Stamm-Morphem) der albanischen Verben:[98]

Tabelle 3.1: Innere Flexion der albanischen Verben

Typ	Lautänderungen		Beispiele
I	-e- / -i-	→	*nd* ez / *nd* izni
II	-e- / -i- / -o-	→	*h* eq / *h* iqni / *h* oqa
III	-a- / -e- / -i-	→	*rr* ah / *rr* eh / *rr* ihni
IV	-a- / -e- -o-	→	*m* arr / *m* err *m* ora
V	-a- / -e- / -i- / -o-	→	*d* al / *d* el / *d* ilni / *d* ola
VI	-a- / -e- -o- / -ua-	→	*d* ashur / *d* esha *d* o / *d* ua
VII	-aj- / -e-	→	*v* aj*ta* / *v* ete
VIII	-ë- / -i- -u-	→	*v* ë / *v* ihet *(u) v* u
IX	-je- / -i- / -o-	→	*mb* jell / *mb* ill / *mb* olla

Zuletzt wird noch die Verteilung der Vokale (=V) und Konsonanten (=K), d. h. Silben, im Albanischen vorgestellt. Die Verteilung sieht wie in den folgenden Beispielen aus, wobei nur bestimmte Fälle – in Anlehnung an [Fiedler 2003] – ausgewählt sind, um dem Leser einen ersten Eindruck zu verschaffen:

Tabelle 3.2: Verteilung der Konsonanten und Vokale (Silben) im Albanischen

Typ	Beispiel	Typ	Beispiel
VK	*ar*	KV	*mi*
V+KV	*ynë* (*y+në*)	KVK	*det*
V+KVK	*iriq* (*i+riq*)	KV+V	*mua* (*mu+a*)
VK+KV	*ulte* (*ul+te*)	KV+KV	*dita* (*di+ta*)
VK+KVK	*ashpër* (*ash+për*)	KV+VK	*dyer* (*dy+er*)
V+V	*ua* (*u+a*)		

[98] Nach [Hetzer/Finger 1993: 190-192. (§ 26.4.2)]. Die ursprüngliche Tabelle wurde mit den Einträgen VI bis IX vom Autor der vorliegenden Arbeit ergänzt.

Eine wie in der Tabelle 3.2 dargestellte Trennung in Vokale und Konsonanten könnte für eine phonotaktische maschinelle Verarbeitung[99] wichtig sein, etwa für eine Erstellung oder Überprüfung der Silbentrennung oder als Hilfe für eine automatische maschinelle Aussprache.

3.3 Die Wortarten

Die meisten Werke, die sich mit der Grammatik des Albanischen befassen, bspw. [Morfologjia 1995: 37], unterscheiden zehn Wortarten und zwar wie folgt: Substantiv (alban. *emri*), Adjektiv (alban. *mbiemri*), Numeral (alban. *numërori*), Pronomen (alban. *përemri*), Verb (alban. *folja*), Adverb (alban. *ndajfolja*), Präposition (alban. *parafjala*), Konjunktion (alban. *lidhëza*), Partikel (alban. *pjesëza*) und Interjektion (alban. *pasthirrma*).
Im Folgenden wird auf die einzelnen Wortarten eingegangen, wobei nur die morphologischen Eigenschaften berücksichtigt werden.

3.3.1 Das Verb

Die morphologischen Kategorien der Verben des Albanischen nach [Buchholz/Fiedler 1987: 272 f. (§ 7.3.6)] sind: Person (alban. *veta*), Numerus (alban. *numri*), Tempus (alban. *koha*), Modus (alban. *mënyra*), Genus verbi (alban. *Diateza*) und Admirativität. [Morfologjia 1995: 291–293 (§ 7.6)] klassifiziert die Kategorie des Admirativs als eine Modusvariante, nämlich als alban. *mënyra habitore*, d. h. ‚*Modus der Verwunderung*', vgl. hierzu auch [Kostallari 1984: 88–136 (§§ 88–114) I. V].

Person: Die Verben im Albanischen besitzen die erste, zweite und dritte Person. Ein Beispiel wäre: alban. *unë shkruaj* (dt. *ich schreibe*), *ti shkruan* (dt. *du schreibst*) usw. Eine Ausnahme bildet hier die Gruppe von Verben, die im normalen Sprachgebrauch nur in der dritten Person Singular vorkommen, wie alban. *veton* (dt. *es blitzt*/*blitzen*). Da in der albanischen Standardsprache ein Infinitiv *per definitionem* nicht vorhanden ist, werden Verben im Wörterbuch oft in der dritten statt in der ersten Person kodiert.

Numerus: Es sind zwei Numeri vorhanden, der Singular (alban. *njëjësi*) und der Plural (alban. *shumësi*). Numerus spielt auch in Verbindung mit anderen morphologischen Kategorien eine wichtige Rolle; z. B. bei der Konjugation der Verben. Im Aorist (s. u.), bspw. beim Verb *punoj*

[99] Wie es der Fall bei CELEX ist, vgl. hierzu Listing 2.12.

(dt. *arbeiten*) wird der Stamm zwischen Singular (*puno-*) und Plural (*punua-*) unterschieden.

Tempus: Die Kategorie Tempus weist bei albanischen Verben die folgenden Typen auf: Präsens (alban. *koha e tashme*), Imperfekt (alban. *k. e pakryer*), Aorist (alban. *k. e kryer e thjeshtë*), Perfekt (alban. *k. e kryer*), Plusquamperfekt (alban. *k. më se e kryer*), Aorist II (alban. *k. e kryer e tejshkuar*), Futur (alban. *k. e ardhme*), Futur Imperfekt (alban. *k. e ardhme e përparme*), Futur Perfekt (alban. *k. e ardhme e së shkuarës*) und Futur Plusquamperfekt (alban. *k. e ardhme e përparme e së shkuarës*), vgl. [Morfologjia 1995: 273–276 (§ 7.3.7)]. Dabei sind die ersten drei Typen Präsens, Imperfekt, Aorist synthetisch, während die restlichen sieben Typen analytisch gebildet werden, nämlich mit einem Hilfsverb (*jam* (dt. *sein*) bzw. *kam* (dt. *haben*)) und einem lexikalischen Verb.

Modus: Die Verben im Albanischen besitzen die folgenden Modi: Indikativ, Konjunktiv, Optativ und Imperativ. Die albanischen Grammatiken (Schulen), vgl. u. a. auch [Ressuli 1985], unterscheiden einen weiteren Typ, nämlich *mënyra habitore*, d. h. ‚*Modus der Verwunderung*', vgl. hierzu Abschnitt Admirativität.[100] Einige Formen im Nicht-Aktiv werden mithilfe der vorangestellten Partikel *u* gebildet, also analytisch.

Genus verbi: Die Kategorie Genus verbi, u. a. auch als Diathese bezeichnet, kommt als Aktiv und als Nicht-Aktiv im Albanischen vor. Sie zeigt bei Verben, in welcher Form eine Handlung bzw. ein Geschehen in sprachlichen Strukturen repräsentiert wird. So kann eine Handlung, falls es das entsprechende Verb erlaubt, im Aktiv und im Nicht-Aktiv, in zwei unterschiedlichen Formen (Perspektiven), ausgedrückt werden. Einige Verben erlauben oder erfordern eine weitere bzw. eine andere Variante, nämlich ein Reflexiv. Die Grammatiken des Albanischen unterscheiden und bezeichnen diese Kategorie samt ihren Formen nicht einheitlich. So spricht [Morfologjia 1995: 270–272 (§ 7.3.5)] zunächst von diateza veprore (Aktiv) und diateza joveprore (Nicht-Aktiv), die in diateza veprore (Aktiv), diateza pësore (Passiv), diateza vetvetore (Reflexiv) und diateza mesore (Medium) weiterspezifiziert werden. [Çeliku et al. 1998: 132 ff. (§ 84)] unterscheiden trajta veprore (Aktiv)

[100] Einige Formen, die in Imperativ Passiv vorkommen, wie z. B. *mëso|h|u|ni* (dt. *lernt euch*) und *shpre|h|u|ni* (dt. *äußert euch*) werden in negierten Sätzen mit *mos* in anderen Formen realisiert, vgl. *mos u mësoni* und *mos u shprehni*.

und trajta pësore (Passiv)[101]. Buchholz und Fiedler [1987: 63, 183–193 (§ 2.10)] verwenden in diesem Zusammenhang die Bezeichnungen Aktiv/Nichtaktiv sowie Reflexiv/Nichtreflexiv.

Admirativität: Die Kategorie Admirativität (Admirativ / Nichtadmirativ) drückt bei Verben die Eigenschaft der Verwunderung aus. In Grammatiken des Albanischen wird die Eigenschaft *Verwunderung* als Modus, *Modus der Verwunderung*, kategorisiert. Buchholz und Fiedler [1987] unterscheiden diese Kategorie von Modus.

Tabelle 3.3 (nach [Buchholz/Fiedler 1993: 697 f.]) gibt die Stammformen der albanischen Verben an. Sie werden in [Buchholz/Fiedler 1987: 88–100 (88), (§ 1.4)] auf 14 reduziert, aufgrund der Tatsache, dass das Wörterbuch ([Buchholz et al. 1993]) erstellt wurde, als die Normsprache (1972) noch nicht „stabilisiert" war. Diese Zahl der Stammformen ist jedoch weiterhin von Vorteil, denn dadurch können zusätzliche Formen des Substandards wie bspw. regionale und dialektale Varianten der einzelnen Wörter berücksichtigt werden, um eine genauere Klassifikation zu erhalten. Tabelle 3.4 (aus [Kabashi 2003], überarbeitet) gibt einen Überblick über das Paradigma des Verbs *punoj* (dt. *arbeiten*):[102]

Einige Besonderheiten der Verben

Wie in anderen indoeuropäischen Sprachen gibt es auch im Albanischen eine bestimmte Zahl der suppletiven Verben (9) und eine Reihe von unregelmäßigen Verben. Diese sind: *është* (dt. *sein*), *ka* (dt. *haben*), *ha* (dt. *essen*), *rri* (dt. *stehen*), *bie* (dt. *bringen*), *bie* (dt. *fallen*), *jap* (dt. *geben*), *sheh* (dt. *sehen*), *vjen* (dt. *kommen*) (suppletiv) und *thotë* (dt. *sagen*), *do* (dt. *wollen*), *lë* (dt. *lassen*), *vdes* (dt. *sterben*), *vete* (dt. *gehen*), *shpie* (dt. *hinbringen*), *shtie* (dt. *hineinbringen*) (nicht suppletiv), vgl. [Bega/Bega 2007: 36 f.].[103]

[101] Einige Verben wie *plakem* (dt. *ich werde alt/ich altere (mich))*, sowie *lahem* (dt. *ich wasche mich*) werden als Passiv bezeichnet, jedoch wird auf ihre besonderen Eigenschaften hingewiesen.

[102] Das Zeichen „|" gibt die Grenze zwischen Stämmen und Flexionssuffixen an. Das Zeichen „¦" grenzt die Hiatus-Elemente von den Flexionssuffixen ab.

[103] Es gibt auch eine bestimmte Zahl Verben, die nur einen Stamm haben, wie z. B. *ha* (dt. *essen*), *këput* (dt. *abtrennen*), *pyes* (dt. *fragen*), *rrit* (dt. *erziehen, wachsen lassen*), *venit* (dt. *welk werden, schlappmachen*) u. a.

Tabelle 3.3: Die Stammformen der albanischen Verben

Stammformen der albanischen Verben nach [Buchholz et al. 1993]						
I	3. P.	Sg.	Präs.	Indikativ	Nichtadm.	Aktiv
II	1. P.	Sg.	Präs.	Indikativ	Nichtadm.	Aktiv
III	1. P.	Pl.	Präs.	Indikativ	Nichtadm.	Aktiv
IV	2. P.	Pl.	Präs.	Indikativ	Nichtadm.	Aktiv
V	2. P.	Sg.	Präs.	Konjunktiv	Nichtadm.	Aktiv (ohne Partikel *të*)
VI	3. P.	Sg.	Präs.	Konjunktiv	Nichtadm.	Aktiv (ohne Partikel *të*)
VII	1. P.	Sg.	Imperf.	Indikativ	Nichtadm.	Aktiv
VIII	3. P.	Sg.	Imperf.	Indikativ	Nichtadm.	Aktiv
IX	1. P.	Pl.	Imperf.	Indikativ	Nichtadm.	Aktiv
X	3. P.	Sg.	Präs.	Indikativ	Nichtadm.	Passiv/Reflexiv
XI	2. P.	Sg.	Impv.	Indikativ	Nichtadm.	Aktiv
XII	1. P.	Sg.	Aor.	Indikativ	Nichtadm.	Aktiv
XIII	3. P.	Sg.	Aor.	Indikativ	Nichtadm.	Aktiv
XIV	1. P.	Pl.	Aor.	Indikativ	Nichtadm.	Aktiv
XV	3. P.	Sg.	Aor.	Indikativ	Nichtadm.	Passiv/Reflexiv (o. P. *u*)
XVI	2. P.	Sg.	Präs.	Optativ	Nichtadm.	Aktiv
XVII	3. P.	Sg.	Präs.	Optativ	Nichtadm.	Aktiv
XVIII	1. P.	Sg.	Präs.	Indikativ	Admirativ	Aktiv
XIX	Partizip					

Klassifikation der Verben

Um eine Übersicht über die Konjugation der einzelnen Verben zu schaffen, werden die Verben in Klassen organisiert. Das heißt, wenn ein Verb in bestimmten Fällen andere Flexionssuffixe als ein zweites Verb aufweist, wird es entweder einer anderen Klasse oder einer (anderen) Unterklasse zugeschrieben, je nachdem, wie groß die Unterschiede sind. So wird das Wissen über Verben kodiert – sowohl für didaktisch-praktische als auch für wissenschaftliche Zwecke.

Zur Klassifikation der Verben des Albanischen gibt es verschiedene Vorschläge und Modelle, welche vor allem wissenschaftlichen Zwecken dienen, wie u. a. [Morfologjia 1995: 278–285 (§ 7.4.3)], [Buchholz et al. 1993], [Buchholz/Fiedler 1987] und [Memushaj 2003]. Für didaktisch-praktische Zwecke sind in letzter Zeit auch einige Werke entstanden, wovon hier zwei erwähnt werden sollen, nämlich [Munishi 1998] und [Bega/Bega 2007], denn sie weisen eine größere Abdeckung als manche andere Werke

Tabelle 3.4: Die synthetischen Formen des Verbs *punoj*.

	Aktiv									
	Nichtadmirativ									
	Indikativ						Konjunktiv		Optativ	Imperativ
	Präsens		Imperfekt		Aorist		Präsens	Imperfekt	Präsens	Präsens
1. P. Sg.	*puno|j*	II	*puno|ja*	VII	*puno|v¦a*	XII	*të␣puno|j*	*të␣puno|ja*	*puno|fsh¦a*	--/--
2. P. Sg.	*puno|n*		*puno|je*		*puno|v¦e*		*të␣puno|sh* V	*të␣puno|je*	*puno|fsh* XVI	*puno* XI
3. P. Sg.	*puno|n*	I	*puno|nte*	VIII	*puno|i*	XIII	*të␣puno|jë* VI	*të␣puno|nte*	*puno|f¦të* XVII	--/--
1. P. Pl.	*puno|jmë*	III	*puno|nim*	IX	*punua|m*	XIV	*të␣puno|jmë*	*të␣puno|nim*	*puno|fsh¦im*	--/--
2. P. Pl.	*puno|ni*	IV	*puno|nit*		*punua|t*		*të␣puno|ni*	*të␣puno|nit*	*puno|fsh¦i*	*puno|ni*
3. P. Pl.	*puno|jnë*		*puno|nin*		*punua|n*		*të␣puno|jnë*	*të␣puno|nin*	*puno|fsh¦in*	--/--
	Admirativ									
1. P. Sg.	*punua|kam*	XVIII	*punua|kësha*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
2. P. Sg.	*punua|ke*		*punua|kështe*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
3. P. Sg.	*punua|ka*		*punua|kësh*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
1. P. Pl.	*punua|kemi*		*punua|kështim*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
2. P. Pl.	*punua|keni*		*punua|kështit*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
3. P. Pl.	*punua|kan*		*punua|kështin*		--/--		*...␣punuar*	*...␣punuar*	--/--	--/--
	Passiv									
	Nichtadmirativ									
	Indikativ						Konjunktiv		Optativ	Imperativ
	Präsens		Imperfekt		Aorist		Präsens	Imperfekt	Präsens	Präsens
3. P. Sg.	*puno|h¦et*	X	*puno|h¦ej*		*u␣punua*	XV	*...␣punuar*	*...␣punuar*	--/--	--/--
3. P. Pl.	*puno|h¦en*		*puno|h¦eshin*		*u␣punua|n*		*...␣punuar*	*...␣punuar*	--/--	--/--
	Admirativ									
3. P. Sg.	*u␣punua|ka*		*u␣punua|kësh*		--/--		*...␣punuar*	*...␣punuar*	*u␣puno|f¦të*	--/--
3. P. Pl.	*u␣punua|kan*		*u␣punua|kështin*		--/--		*...␣punuar*	*...␣punuar*	*u␣puno|fsh¦in*	--/--
	Partizip									
	punua|r XIX									

auf, die hier nicht berücksichtigt werden können.[104] Nach der Durchsicht dieser Werke fällt es nicht leicht, ein System zu erstellen, das einheitlich wäre. Die meisten Unterschiede liegen in der Klassenzuweisung einzelner Verben, doch es treten auch Schwierigkeiten bezüglich einer Flexionsstelle oder sogar des Nicht-Vorhandenseins einer Form auf.[105]
Die meisten wissenschaftlichen Behandlungen dieses Themas gehen von drei Hauptklassen aus, wobei sie mehrere Unterklassen sowohl in der Breite als auch in der Tiefe aufweisen. Die empirischen Werke ([Buchholz et al. 1993], [Munishi 1998], und [Bega/Bega 2007]) klassifizieren die Verben in „flacher" Form, d. h. Klassen ohne Unterklassen, vgl. hierzu [Buchholz et al. 1993] 55 Klassen, [Munishi 1998] 85 Klassen, und [Bega/Bega 2007] 101 Klassen.[106]

3.3.2 Das Substantiv

Die Substantive (alban. sg. best. *emri*) im Albanischen werden dekliniert.[107] Ihre Merkmale sind Bestimmtheit/Nichtbestimmtheit, Genus, Kasus und Numerus.[108]

Die Bestimmtheit

Die Kategorie der Bestimmtheit (alban. *shquarsia*) bzw. Nichtbestimmtheit (alban. *pashquarsia*), kurz genannt auch *trajtat* (dt. *Formen*), drückt insbesondere die Merkmale der Determiniertheit bzw. Indeterminiertheit aus. Sie kommt in verschiedenen Kommunikationssituationen vor und ist eine komplexe Kategorie.[109] Im Allgemeinen kann ein Vergleich zwischen der

[104] Bei [Buchholz et al. 1993] handelt es sich um eine Auflage, welche ursprünglich 1976 erschien, d. h. die Klassifikation der Verben entspricht der ersten Auflage. Einige Klassen der Verben sind in [Buchholz et al. 1993] in Unterklassen aufgeteilt. Eine grobe Klassifikation von Verben, prototypisch mit einigen Beispielvertretern, findet man auch bei [Sulejmani 1984] und im Anhang von [Fjalori 1984].

[105] Als Beispiel kann die Auflistung der Imperativformen der Verben mit integrierten klitischen Pronomina bei [Munishi 1998] im Kontrast zu [Bega/Bega 2007], wo sie nicht aufgelistet sind, genommen werden. Hier kann leider nicht auf dieses Problem eingegangen werden.

[106] Auf dieses Thema wird in Kapitel 4 (Das Lexikon) näher eingegangen.

[107] Vgl. [Buchholz/Fiedler 1987: 268 ff.] zur Deklination der Substantive und substantivierten Possessiva.

[108] Vgl. hierzu auch [Duden-Grammatik 2009, §§ 197–1162], [Fleischer/Barz 1992], [Motsch 2004] u. a. Literatur über die Morphologie und Wortbildung – des Deutschen.

[109] Für ausführliche Informationen zu diesem Thema siehe [Buchholz/Fiedler 1987: 232–243 (§ III 2.3)] bzw. [Morfologjia 1995: 120–132 (§§ 3.6–3.6)].

Verwendung eines Substantivs mit vorangestelltem Artikel im Deutschen und der bestimmten Form eines Substantivs im Albanischen gezogen werden. Die unbestimmte Form im Albanischen entspräche bei diesem Vergleich der Verwendung eines Substantivs ohne bzw. mit einem unbestimmten Artikel im Deutschen. Ein Beispiel wäre: alban. *lis* unbest. (dt. [*ein*] *Baum*) vs. *lis*i best. (dt. *der Baum*). Das enklitische Formativ i in diesem Beispiel zeigt, dass es zusammen mit dem lexikalischen Morphem eine Einheit bildet.[110] In einigen Beispielen, wie im folgenden alban. *shkoll*ë unbest. (dt. [*eine*] *Schule*) vs. *shkoll*a best. (dt. *die Schule*) wird ein vorhandenes Formativ umgewandelt, in diesem Fall ë in a. Ein weiteres Beispiel wäre alban. *lib*ër unbest. (dt. [*ein*] *Buch*) vs. *lib*ri best. (dt. *das Buch*).[111]

Der Numerus

Der Numerus[112] ist eine semantisch-funktionale Kategorie, die in zwei Formen vorkommt, nämlich im Singular und im Plural.[113] Die Pluralbildung der Substantive im Albanischen ist durch reiche Formen ausgeprägt, sodass deren Klassifikation keine leichte Aufgabe ist.[114]

> Die Pluralbildung erscheint [...] in allen Balkansprachen relativ kompliziert, aber ohne Frage ist sie im Albanischen am kompliziertesten [...]. Man muß sich bei jedem Substantiv den Plural einprägen, wie im Deutschen das Geschlecht der Substantive. [Fiedler 2003: 782]

In der Normsprache (1972) kommen etwa 100 Typen des Plurals vor.[115] Die Bildung des Plurals wird von verschiedenen Regeln bestimmt. Buchholz und Fiedler [1987: 249 ff.] unterscheiden in erster Linie fünf Kriterien (Eigenschaften) als Schlüssel für die Klassifikation der Typen des Plurals, und zwar:

A: Identität von Singular- und Pluralform
nxënës, *-i* vs. *nxënës*, *-it* (dt. *Schüler*); Buchholz und Fiedler [1987:

[110] Die Bezeichnung Formativ wird im Sinne von [Buchholz/Fiedler 1987] verwendet.
[111] Aufgrund dieser Tatsache werden in der Regel in der albanischen Lexikographie, wie etwa im Rechtschreibwörterbuch [Drejtshkrimi 1976], vier Formen angegeben: sg. unbest., sg. best., pl. unbest. und pl. best. Vgl. hierzu auch Kapitel 4.
[112] Vgl. [Fiedler 2005] und [Buchholz/Fiedler 1987: 244–268 (§ III 2.4)].
[113] Vgl. hierzu [Morfologjia 1995: 94–105 (§ 3.5)].
[114] Vgl. hierzu auch Kapitel 5, wo vom Autor der vorliegenden Arbeit eine datenbasierte Klassifikation der Substantive präsentiert wird.
[115] Vgl. hierzu [Buchholz et al. 1993: 652–656].

244–268 (§ III 2.4)] unterteilen die Gruppen weiter in 17 Untergruppen (a–k [m.] und l–q [f.]), wobei jeweils für beide Numeri die entsprechenden Typen angegeben sind. Singular und Plural stimmen in ihren jeweiligen unbestimmten Formen überein, bspw. ist *nxënës* dabei unter c gruppiert und besitzt die Typologie Sg. I a / Pl. V c, 2.[116]

B: Erweiterung der Pluralform

Sg. *punëtor*, *-i* vs. Pl. *punëtorë*, *-t* (dt. *Arbeiter*); Diese Gruppe ist sehr groß. Buchholz und Fiedler [1987: 250–251] unterscheiden zunächst 20 Typen, welche anschließend weiter unterteilt werden. Das angegebene Beispiel entspricht dem Typ 1 a (Maskulina: Pl. -ë) mit den Eigenschaften Sg. I a / Pl. V a, 1.

C: Konsonantenveränderung

Diese Gruppe hat sechs Untergruppen, wobei die ersten fünf je zwei Typen beinhalten, während die sechste nur einen Typ besitzt. Ein Beispiel, entnommen aus der Untergruppe 1 b; Veränderung k→q [nur Maskulina] ist: Sg. *buj*k, *-u* vs. Pl. *buj*q, *-it* (dt. *Bauer*) mit den Klasseneigenschaften Sg. II a / Pl. V c, 1.

D: Vokalveränderung

Ein Beispiel für Vokalveränderung (ua→a, [nur Feminina], in der Untergruppe 3 a) ist: Sg. *gr*ua, *-ja* vs. Pl. *gr*a, *-të* (dt. *Frau*) mit den Klasseneigenschaften Sg. III e, 1 / Pl. V b, 1. Diese Gruppe besteht aus vier Untergruppen und deckt folgende Vokaländerungen ab: a→e [nur Maskulina] *dash* / *desh* (dt. *Hammel*), i→je [nur Maskulina; Heterogenie] *vit* / *vjet* (dt. *Jahr*), ua→u und e→a [nur Feminina] *pëlle*, *-ja* / *pëll*a, *-të* (dt. *Milchschaf*).

E: Erweiterung der Singularform

Bei dieser Gruppe handelt es sich um einen Sonderfall der Vokaländerung ë→0 wie Sg. *dit*|ë, *-a* vs. Pl. *dit*|0, *-ët* (dt. *Tag*), Sg. III a / Pl. V d, 2.[117]

Weitere Typen ergeben sich aus der Kombination der aufgelisteten Kriterien in den Punkten (B) bis (E) untereinander, welche in [op. cit.] mit den Buchstaben (F) bis (L) versehen sind und folgendermaßen heißen:

[116] Hier wird nicht auf die einzelnen Fälle eingegangen, vgl. hierzu für detaillierte Informationen [op. cit.].

[117] Nach der heute gültigen Rechtschreibung (1972) wird der unbestimmte Plural mit ë geschrieben: *dit*|*ë*, *-a*, -ë und *-ët*, vgl. [Dhrimo/Memushaj 2011] und [FDsh 1976].

F: Konsonantenveränderung + Erweiterung der Pluralform (d. h. C+B)

Bei dieser Gruppe handelt es sich sowohl um die Vokalveränderung als auch um die Erweiterung der Pluralform. Es können fünf Fälle unterschieden werden, nämlich k→q + -e, g→gj + -e, ll→j + -e, b→p + -inj und k→q + -ër; Das folgende Beispiel kann als Illustration dienen: *py*ll, *-i* vs. Pl. *py*je, *-t* (dt. *Wald*), Eigenschaften Sg. Ia/Pl. Va,1.

G: Vokalveränderung + Erweiterung der Pluralform (d. h. D+B)

Es handelt sich um eine große Gruppe, welche eine Vielfalt an Eigenschaften und Wörtern hat. Sie wird in 19 Untergruppen aufgeteilt, wobei auch dialektale Wörter sowie solche aus älteren Sprachstufen berücksichtigt wurden. Ein Beispiel aus der Untergruppe 2, e→a + -ë, ist *rreth*, *-i* vs. Pl. *rrath*ë, *-t* (dt. *Kreis*), Eigenschaften Sg. Ia/Pl. Va,1; Vgl. für eine ausführliche Darstellung Buchholz und Fiedler [1987: 253 ff.].

H: Konsonantenveränderung + Vokalveränderung (d. h. C+D)

In diese Gruppe fallen Beispiele wie *pl*ak, *-u* vs. Pl. *pl*eq, *-të* (dt. *Greis*) mit den Klasseneigenschaften Sg. IIa/Pl. Vb,2, wobei die Übergänge a→e und k→q zu erkennen sind; Die Gruppe H besteht, unter der Berücksichtigung des dialektalen Wortschatzes, aus elf Untergruppen.

I: Erweiterung der Pluralform + Erweiterung der Singularform (d. h. B+E)

Bei dieser Gruppe handelt es sich um die Änderung í→-ëz, wobei auch ein Betonungswechsel vorkommt, wie das Beispiel *njerí*, *-u* vs. Pl. *nj*érëz, *-it* (dt. *Mensch*), mit den Eigenschaften Sg. IIb/Pl. Vc,2 zeigt. Diese Gruppe ist klein und besteht aus zwei Mitgliedern, wobei das andere Mitglied aus dem dialektalen Wortschatz stammt.

J: Vokalveränderung + Erweiterung der Singularform (d. h. D+E)

Die Änderungen, die in dieser Gruppe vorkommen, betreffen den Grundwortschatz. Es handelt sich um eine kleine Gruppe aus drei Mitgliedern. Ein Beispiel ist *nat*ë, *-a* vs. Pl. *net*, *-ët* (dt. *Nacht*), a→e und ë→0, mit den Klasseneigenschaften Sg. IIIa/Pl. Vd,2.

K: Vokalveränderung + Konsonantenveränderung + Erweiterung der Pluralform (d. h. D+C+B)

Bei dieser Gruppe handelt es sich um die Änderungen e→i und g→gj+-e, bspw. *breg*, *-u* vs. Pl. *brigj*e, *-t* (dt. *Ufer*) mit den Klasseneigenschaften Sg. IIa/Pl. Va,1. Sie beinhaltet fünf Untergruppen, die jeweils nur aus einem Typ bestehen.

L: Vokalveränderung + Konsonantenveränderung + Erweiterung der Singularform (d. h. D+C+E)
In dieser Gruppe sind Fälle wie *k*a*l*|*ë*, *i* vs. Pl. *k*ua*j*, -t (dt. *Pferd*), enthalten, wobei die Übergänge a→ua, l→j und -ë→0 vorkommen und die Klasseneigenschaften Sg. Ic/Pl. Va,2 besitzen.

M: der Sonderfall Suppletivismus (Suppletion)
Es handelt sich hier um Fälle wie *qengj*/*shtjerra* oder *shqerra* (dt. *Lamm*), wobei die beiden Stämme (Singular und Plural) sich sehr voneinander unterscheiden und etymologisch nicht zusammengehören, vgl. [OP. CIT.: 256].

Der Kasus

Der Kasus ist eine Kategorie des Substantivs, welche in Nominativ, Genitiv, Dativ, Akkusativ und Ablativ realisiert wird. Durch diese Kategorie wird die Kongruenz der Elemente untereinander in der Nominalgruppe hergestellt sowie die Beziehung der Nominalgruppe zur Verbvalenz spezifiziert. Sie trägt dazu bei, dass im Albanischen eine relativ freie Wortstellung möglich ist, denn die Wortformen sind hinreichend genau bestimmt.

Tabelle 3.5: Deklination der Substantive des Typs I

Deklination der Substantive des Typ I				
	Unbestimmt		Bestimmt	
	Singular	Plural	Singular	Plural
Nom.	(*një*)␣*lis*	(*disa*)␣*lis*a	*lis*|i	*lis*a|t
Gen.	{*i, e, të*}␣(*një*)␣*lis*|i	{*i, e, të*}␣(*disa*)␣*lis*a|ve	{*i, e, të*}␣*lis*|it	{*i, e, të*}␣*lis*a|ve
Dat.	(*një*)␣*lis*|i	(*disa*)␣*lis*a|ve	*lis*|it	*lis*a|ve
Akk.	(*një*)␣*lis*	(*disa*)␣*lis*a	*lis*|i	*lis*a|t
Abl.	(*prej*)␣(*një*)␣*lis*|i	(*prej*)␣(*disa*)␣*lis*a|sh	(*prej*)␣*lis*|it	(*prej*)␣*lis*a|ve

Vgl. hierzu auch [ÇELIKU ET AL. 1998: 37–44 (§§ 21–24)]. Laut [OP. CIT. 38, Fußnote 43] besitzt das Substantiv im Albanischen in seiner bestimmten Form keinen Ablativ. [MORFOLOGJIA 1995] und [FJALORI 1984] listen diese Fälle jedoch auf.[118]

[118] Im Rahmen der vorliegenden Arbeit werden die genannten Formen behandelt, d. h., es wird der [MORFOLOGJIA 1995] gefolgt.

Das Genus

Die Kategorie des Genus (alban. *gjinia*) ermöglicht eine zusätzliche grammatische Markierung (Kodierung), welche zusammengehörige Einheiten in der Nominalphrase/im Satz hinsichtlich dieser Eigenschaft kennzeichnet.[119] Das Genus ist eine grammatische Kategorie, durch die die Substantive im Albanischen in Maskulina und Feminina unterteilt werden. In früheren Stadien des Albanischen gab es auch ein Genus Neutrum. Seine Spuren sind in der Norm-Sprache in einigen Fällen noch zu erkennen.[120] Nomina wie *të␣ftohtë|t* (dt. *[die] Kälte*), *të␣zi|të* (dt. *[das] Schwarze*) und *të␣ecur|it* (dt. *[das] Laufen*) sind aus alten Neutra, vgl. [Kostallari 1984: 9–10. (P I)] abgeleitet.[121] Ein Beispiel für die Kategorie Genus in der Norm-Sprache ist im Folgenden angegeben: alban. *student* mask. unbest. (dt. [*ein*] *Student*), alban. *student|i* mask. best. (dt. [*der*] *Student*), alban. *student|e* fem. unbest. (dt. [*eine*] *Studentin*), alban. *student|ja* fem. best. (dt. [*die*] *Studentin*). Aus diesen Beispielen lässt sich erkennen, dass das Genus bei einigen Nomina, die es semantisch erlauben, durch Suffixe gebildet wird. Die größte Zahl der Nomina hat jedoch ein festes Genus, wie z. B. *kohë, -a*, fem. (dt. *[eine] Zeit, [die] Zeit*). Einige Substantive haben im Plural ein anderes Genus als im Singular (*Heterogenie*), vgl. *mal, -i*, mask. sg. vs. *male, -t*, fem. pl. .

Die Artikel-Substantive

Einige Substantive besitzen einen vorangestellten Artikel, wie z. B. alban. *e diel* fem. unbest. (dt. *Sonntag*), oder alban. *të ftohtë* neut. unbest. (dt. *Kälte*). Die Zahl der Artikel-Substantive ist nicht so groß. Es können jedoch aus vielen Verben (zusätzliche) Artikel-Substantive abgeleitet werden.

Deklination der Substantive

Im Albanischen werden auch Personennamen sowie Namen von Städten und Flüssen und alle anderen geographischen Namen wie die Appellativa dekliniert, vgl. *Pejë* Nom. unbest. (Stadtname), *Peja* Nom. best., *Pejës* Gen./Dat. best., *Pejën* Akk. best., *Peje* Gen. unbest. usw.[122] Die Bezeichnung der Bewohner der Ortschaften und der Gebiete ist auch unterschiedlich, vgl.

[119] Vgl. hierzu auch [Buchholz/Fiedler 1987: 203–211 (§ III 2.1)].

[120] Einige Nomina, die im unbestimmten Nominativ auf -ë enden, wie *uj*ë Nom. unbest. → *uj*i Nom. best. (dt. *Wasser*), *uj*in Akk. best.; deuten das (alte) Neutrum an, vgl. [Kostallari 1984: 9–10. (P I)]. Neutra können auch in literarischen Texten vorkommen.

[121] Im [op. cit.] stehen „-“ für Segmentierungszeichen zwischen Stamm und Suffix.

[122] Vgl. [Buchholz/Fiedler 1987: 268–274 (§ III 2.5)].

Pejë → pejan (Bewohner der Stadt *Pejë*); *Korçë → korçar*, *Shkodër → shkodran*, *Tiranë → tiranas*, *Vlorë → vlonjat*, *Delvinë → delvinjot*, *Konispol → konispolat*, vgl. hierzu Fjalori [1984: 1425–1460] und [Dhrimo/Memushaj 2011]. Beide Werke listen die Bezeichnung der Bewohner einiger großer Ortschaften im albanischsprachigen Raum und auch einiger anderer Städte außerhalb dieses Raumes wie *Mynih/-u* (*München*), *Nyrenberg/-u* (*Nürnberg*), *Vjen|ë/-a* (*Wien*), *Lond|ër/-ra* (*London*) auf.[123] Die Ortschaftsnamen der Republik Albanien sind in [Lafe et al. 2002] in Form einer Monographie publiziert worden.[124] Diese Namen müssen im Lexikon in einer Form eingetragen sein, um die weiteren Formen (Bestimmtheit, Kasus und ggf. Numerus) ableiten zu können, bspw. hat *pejan*, Nom. m. sg. unbest., die weiteren Formen *pejane*, Nom. f. sg. unbest., *pejanë*, Nom. m. pl. unbest., *pejane*, Nom. f. pl. unbest., *pejani*, Nom. m. sg. best., *pejanja*, Nom. f. sg. best., *pejanët*, Nom. m. pl. best., *pejanet*, Nom. f. pl. best. sowie die Formen, welche die übrigen Kasus (Gen., Dat., Akk., und Abl.) markieren. Im Normalfall werden diese Namen wie die üblichen Substantive im Lexikon eingetragen, bspw. *Pej|ë* Nom. unbest./*-a*, Nom. best., vgl. hierzu auch Abschnitt 4.3. Die Kasus- und Bestimmtheitsmarkierung kann anhand der folgenden Beispiele gesehen werden: *në Tiranë* unbest. (dt. *in Tirana*), vs. *nga Tirana* best. (dt. *aus Tirana*), *{i, e, të, së} Tiranës* Gen., ohne Artikel auch Dat. und mit bestimmten Präpositionen Abl. (dt. *aus Tirana*); *me Tiranën* Akk. (dt. *mit Tirana*) oder *Birrë e Tiranës* Gen. best. bzw. Abl. ohne Präposition, *Birrë Tirane* Abl. unbest. (dt. *Tiranaer Bier*).

Einige Substantive, die aus der gleichen Wurzel, dem gleichen Stamm abgeleitet sind, folgen zwei oder mehreren Deklinationsmustern aufgrund ihres Genus, wie bspw. *rrjeta* f. (dt. *Netz* [*Fischernetz*]) vs. *rrjeti* m. (dt. *Netzwerk* [*Computernetzwerk*]).[125]

Klassifikation der Substantive

Die albanischen Grammatiken, vgl. hierzu bspw. [Morfologjia 1995: 111–120 (§ 3.5)] unterscheiden hauptsächlich vier Klassen der Substantive, d. h. Deklinationsklassen (alban. *tipat e lakimit*), wobei viele Ausnahmen noch dazu gezählt werden (müssen). Wie bei den Verben werden auch die Substantive vor allem für wissenschaftliche Zwecke behandelt, wie z. B. in [Morfologjia 1995]. Andererseits sind für didaktisch-praktische Zwecke

[123] In Fjalori [1984: 1425–1460] steht als Bezeichnung für München Munich, ∼u, statt der heute verbreiteten Form *Mynih, ∼u*.

[124] Diese Ressource ist leider nicht in elektronischer Form verfügbar.

[125] Vgl. auch dt. die Partikel vs. der Partikel, fachsprachliche Ausdifferenzierung.

keine Werke vorhanden, die jedes bzw. die meisten Substantive, insbesondere diejenigen die zum Grundwortschatz gezählt werden, einer Klasse zuweisen. Als Vergleich könnte man hier die Klassifikation deutscher Substantive in [Wahrig 2012][126] bzw. [CELEX 1994] nennen.
Die meisten Unterschiede zeichnen sich zwischen den beiden Typen in der grammatischen Kategorie Numerus ab, d. h., die Stämme im Singular und im Plural unterscheiden sich in vielen Fällen sehr stark. Die Klassifikation von Substantiven ist umso wichtiger, als im Albanischen auch die Namen (und zwar oft unregelmäßig) dekliniert werden. Tabelle 3.6 zeigt die Artikelendungen der Substantiv–Deklination:

Tabelle 3.6: Artikelendungen der Substantiv–Deklination

Artikelendungen der Substantiv–Deklination										
Typ	I (Sg.)		II (Sg.)		III (Sg.)		IV (Sg.)		V (Pl.)	
	Unb.	Best.	Unb.	Best.	Unb.	Best.	Unb.	Best.	Unb.	Best.
Nom.	–	-i	–	-u	–	-a	–	-t/-të	–	-t/-të/-it
Gen.	-i	-it	-u	-ut	-e	-s/-së	-i	-it	-ve	-ve/-vet
Dat.	-i	-it	-u	-ut	-e	-s/-së	-i	-it	-ve	-ve/-vet
Akk.	–	-in/-në	–	-un/-në	–	-n/-në	–	-t/-të	–	-t/-të/-it
Abl.	-i	-it	-u	-ut	-e	-s/-së	-i	-it	-sh/-ve	-ve/-vet

Stammvariation im Plural ist hier nicht berücksichtigt, bspw. [*dj ̯al¦ ë* | –] sg. ub. Nom. bzw. [*dj ̯em* | –] pl. ub. Nom. usw. vs. [*dj ̯al* | *i*], [*dj ̯al* | *in*], bzw. [*dj ̯em* | *të*] usw. Die markierten Flexionsendungen entsprechen denjenigen, die in der Tabelle 3.6 angegeben sind.

3.3.3 Das Adjektiv

Die Funktion des Adjektivs ist es, das Substantiv näher zu bestimmen.[127] Es kommt mit ihm zusammen in der Nominalphrase oder in substantivierter Form allein vor, wenn das Substantiv aus kontextuellen Informationen bekannt ist, z. B. *Zgjodha të voglin* (dt. *Ich wählte den Kleinen aus*).

[126] Zum Beispiel steht bei dem Lemma *Buch* die Angabe <*n. 12u*>, die auf die Deklinationsklasse des Lemmas hinweist. Diese Klassifikation wurde schon in früheren Ausgaben eingeführt, z. B. in [Wahrig 1998].

[127] Eine ausführliche Beschreibung der Adjektive im Albanischen findet der interessierte Leser u. a. bei [Buchholz/Fiedler 1987: 314–348 (§ III 4)] und bei [Morfologjia 1995: 153–201 (§ IV)].

Die grammatischen bzw. morphologischen Kategorien der Adjektive im Albanischen sind: Numerus, Kasus, Genus, Bestimmtheit/Unbestimmtheit, sowie Graduierbarkeit (Steigerbarkeit/Steigerungsfähigkeit).[128]

Im Folgenden wird auf die wichtigsten Eigenschaften der einzelnen Kategorien eingegangen:

Position in der Nominalphrase: Im Albanischen sind zwei Wortstellungen des Adjektivs möglich: vor dem Substantiv und nach diesem. Diese Eigenschaft ist für die Flexion des Adjektivs von entscheidender Bedeutung, denn in der Position vor dem Substantiv wird das Adjektiv wie ein Substantiv dekliniert.

Zwei oder mehrere Adjektive in koordinierter Beziehung zueinander in post-substantivischer Stellung werden wie im Folgenden (in Tabelle 3.7) dargestellt dekliniert:[129] Bei Artikel-Adjektiven in der post-substantivischen Position wird nur der Artikel dekliniert, vgl. hierzu Tabelle 3.8.

Tabelle 3.7: Deklination des Artikel-Adjektivs

Deklination des Artikel-Adjektivs in post-substantivischer Stellung															
Singular															
		Unbestimmt							Bestimmt						
Nom.		*djal*|ë	i	*urtë*	e	i	*mirë*		*djal*|i	i	*urtë*	e	i	*mirë*	
Gen.	{*i, e, të, së*} *një*	*djal*|i	të	*urtë*	e	të	*mirë*		*djal*|it	të	*urtë*	e	të	*mirë*	
Dat.		*djal*|i	të	*urtë*	e	të	*mirë*		*djal*|it	të	*urtë*	e	të	*mirë*	
Akk.		*djal*|ë	të	*urtë*	e	të	*mirë*		*djal*|in	e	*urtë*	e	të	*mirë*	
Abl.		*djal*|i	të	*urtë*	e	të	*mirë*		*djal*|it	të	*urtë*	e	të	*mirë*	
Plural															
		Unbestimmt							Bestimmt						
Nom.		*djem*|	të	*urtë*	e	të	*mirë*		*djem*|të	e	*urtë*	e	të	*mirë*
Gen.	{*i, e, ...* } *disa*	*djem*|ve	të	*urtë*	e	të	*mirë*		*djem*|ve	të	*urtë*	e	të	*mirë*	
Dat.		*djem*|ve	të	*urtë*	e	të	*mirë*		*djem*|ve	të	*urtë*	e	të	*mirë*	
Akk.		*djem*|	të	*urtë*	e	të	*mirë*		*djem*|të	e	*urtë*	e	të	*mirë*
Abl.		*djem*|ve	të	*urtë*	e	të	*mirë*		*djem*|ve	të	*urtë*	e	të	*mirë*	

[128] In diesem Zusammenhang wird hier die Eigenschaft Position in der Nominalphrase betrachtet, obwohl sie nicht morphologischer Natur ist. Bei der maschinellen Verarbeitung, z. B. der Erkennung des Adjektivs beim Tagging, spielt diese eine wichtige Rolle.

[129] Die Auslassungspunkte („ ... “) in der Tabelle stehen für die aus Platzgründen weggelassenen Genitiv-Artikel *i, e, të* und *së* bzw. Partikel *një* (sg.) und *disa* (pl.), vgl. hierzu auch Tabelle 3.5.

Dabei werden nur die Artikel dekliniert, vgl. für eine tabellarische Übersicht [Buchholz et al. 1993: 662–671] und [Fjalori 1984: 1486–1489].

Graduierung (oder Steigerung) der Adjektive: Das Adjektiv im Albanischen kann im Positiv, Komparativ und Superlativ vorkommen. Der Positiv stellt die Grundstufe dar, bspw. *i mirë* (dt. *gut*). Der Komparativ wird aus einer Partikel (më) und dem Positiv gebildet, bspw. *më i mirë* ← *më* + pos.; Der Superlativ wird ebenso wie der Komparativ aus einer Partikel (shumë) und dem Positiv gebildet, bspw. *shumë i mirë* ← *shumë* + pos.; vgl. [Çeliku et al. 1998: 67–62 (§§ 40–45)]. Einige Adjektive haben (z. B. aus semantischen Gründen) keinen Komparativ oder Superlativ, wie z. B. *i djeshëm* (dt. *gestrig*). Auch durch die Kategorie Bestimmtheit/Nichtbestimmtheit kann die Graduierung ausgedrückt werden, wie das folgende Beispiel zeigt: *Lule e bukur* (dt. *schöne Blume*) [unbest. Positiv]; *Lule më e bukur* (dt. *schönere Blume*) [unbest. Komparativ] *Lulja e bukur* (dt. *die schöne/schönere Blume*) [best. Positiv]; *Lulja më e bukur* (dt. *die schönste Blume*) [best. Superlativ]; vgl. [Buchholz/-Fiedler 1987: 239 (§ 2.3.1.4), ausführlich in 314–348 (§ 4)]. Letzteres Beispiel ist gleichzeitig auch ein Elativ. Eine Unterscheidung ist nur anhand des Kontextes möglich.

Die grammatische Kategorie Numerus drückt wie bei den Substantiven aus, ob Ein- oder Mehrzahl vorhanden ist. Diese Kategorie stimmt mit der gleichen Kategorie bei Substantiven überein (Kongruenz), bspw. *libri i mirë* vs. *librat e mirë* (dt. *Das gute Buch* vs. *Die guten Bücher*) oder *djalë trim* vs. *djem trima* (dt. *Ein tapferer Junge* vs. *Einige tapfere Jungen*).

Ebenso wie der Numerus stellt die Kategorie Genus des Adjektivs eine Kongruenzbeziehung zum zugehörigen Substantiv her, bspw. *Student i zellshëm* vs. *Studente e zellshme* (dt. *Ein fleißiger Student vs. Eine fleißige Studentin*). Diese Kategorie ist bei den Adjektiven in der Regel nicht fest, wie es bei den Substantiven der Fall ist, was bedeutet, dass ein Adjektiv in beiden vorkommen kann, nämlich in Maskulinum und in Femininum. So kann das Artikel-Adjektiv *i zellshëm* (mask.) auch *e zellshme* (fem.) sein. Genauso können auch artikellose Adjektive in den verschiedenen Genera vorkommen, z. B. *dinak* (mask.) (dt. *listig*) vs. *dinak|e* (fem.). Weitere Beispiele wären: *i mirë* (mask.) / *e mirë* (fem.); *i djeshëm* (mask.) / *e djeshme* (fem.); *i kuq* (mask.) / *e kuqe* (fem.) (dt. *rot*). Eine Ausnahme bilden die unregelmäßigen Adjektive wie *i zi* (mask.) / *e zezë* (fem.) (dt. *schwarz*).

Die grammatische Kategorie Kasus wird in der Nominalphrase, wo ein oder mehrere Adjektive beteiligt sind, in zwei Formen ausgedrückt, nämlich entweder beim Substantiv (in der Normalstellung, d. h. Substantiv–Adjektiv) oder beim Adjektiv (in der hervorgehobenen Stellung, d. h. Adjektiv–Substantiv), bspw. *I zellshmi student* bzw. *E zellshmja studente* (Nom.) (dt. *Der fleißige Student* bzw. *Die fleißige Studentin*).[130] Wie beim Substantiv sind auch beim Adjektiv fünf Kasus möglich, der Nominativ, Genitiv, Dativ, Akkusativ und Ablativ, vgl. hierzu auch Tabelle 3.7.

Einige Besonderheiten der Adjektive

Mehrere Artikel-Adjektive ohne den vorangestellten Artikel sind Adverbien, vgl. *i mirë* vs. *mirë* (Grundbedeutung dt. *gut*).
Adjektive besitzen auch zu einigen Substantiven enge Verbindungen, bspw. *trim* (dt. *mutig*, *Held*).[131] In diesem Zusammenhang unterscheidet Fiedler [2003: § 3.3] neben Substantiven und Adjektiven auch die Klasse der Substantiv-Adjektive.
Die Adjektive *i ri* (dt. *jung*, *neu*), *i zi* (dt. *schwarz*), *i lig* (dt. *schwach*), *i vogël* (dt. *klein*), *i madh* (dt. *groß*), *i keq* (dt. *schlecht*) bilden eine Ausnahme bei der Pluralbildung, vgl. [Çeliku et al. 1998: 58–62 (§§ 35–36)]. Das Adjektiv *i vogël* mask. sg. Nom. unbest., *të vegjël* mask. pl. Nom. unbest. vs. *e vogël* fem. sg. Nom. unbest., *të vogla* fem. pl. Nom. unbest. zeigt im Plural einen Konsonantenwechsel innerhalb der Kategorie Genus (g→gj, und zwar im Maskulinum). Bei diesen Lemmata sind im Lexikon vier Formen angegeben, und zwar in sg.m., sg.f., pl.m. und pl.f., vgl. hierzu auch Kapitel 4.
Im Allgemeinen sind Adjektive Simplexe, welche aus einem Stamm sowie Derivations- und/oder Flexionssuffixen bestehen. Sie können aber auch komplexe Wörter im Sinne der Wortbildung sein, wie bspw. das Kompositum *bardhezi* (dt. *schwarzweiß*) [*bardh*|*e*|*zi*, dt. wörtlich: *weiß*|*und*|*schwarz*]. Die verbreitetsten Suffixe bei Artikeladjektiven sind -shëm: *i djeshëm* (dt. *gestrig*); -ueshëm: *i dëgjueshëm* (dt. *hörbar*); -(ë)t(ë): *i drunjtë* (dt. *holzig*); *i gurtë* (dt. *steinig*); -(ë)m(ë): *i jashtëm* (dt. *äußerlich*) und -ë: *i drejtë* (dt. *gerade*, …), vgl. [Morfologjia 1995: 192–196 (§ 4.7.4)]. Beim Beispiel *i drunjtë* (dt. *holzig*), handelt es sich um ein Adjektiv, das aus dem Plural-Stamm gebildet wurde. Dies deutet darauf hin, dass die Wortbildungsprozesse der Adjektive sehr verschieden sind und dass als Basis einer neuen Wortbildung nicht

[130] Nur in bestimmter Form möglich. Der Artikel hat hier in beiden Fällen die Funktion eines Konnektors bzw. wird als solcher betrachtet.
[131] In diesen Fällen darf kein Artikel vorkommen.

zwingend ein Singularstamm vorliegen muss, wenn die semantischen Eigenschaften dies erlauben.

Klassifikation der Adjektive

Die Grammatiken unterteilen die Adjektive in erster Linie in solche mit einem vorangestellten Artikel und in solche, die keinen Artikel haben. Des Weiteren kommen Klassifikationen nach semantischen Kriterien vor, wie z. B. *i madh* (dt. *groß*) vs. *i artë* (dt. *golden*). Formal, anhand der Flexion und des vorangestellten Artikels, werden in erster Linie folgende Typen von Adjektiven unterschieden:[132]

1 *besnik*; *detar*; *dyzanor*; ... (Type 1)

2 *afërt* (*i*, *e*); *artë* (*i*, *e*); ... (Type 2)

3 *afatshkurtër* (dt. *kuzfristig*); *dijeplotë* (dt. *mit bestem Wissen versehen*); *neto* (dt. *Netto*); ...[133] (Type 3)

4 *arritsh|ëm* (*i*), ∼*me* (*e*); ... (Type 4)

5 *epërm* (*i*), ∼*e* (*e*); ... (Type 5)

Aus der obigen Darstellung wird klar, dass es sich morphologisch betrachtet um verschiedene Typen handelt. Dazu kommen noch die irregulären Typen wie *e␣zezë*/*i␣zi* (dt. *schwarze*/*schwarzer*), welche gesondert, nämlich einzeln, behandelt werden müssen.

Artikel

Im Zusammenhang mit den Adjektiven können an dieser Stelle noch einige Eigenschaften des Artikels im Albanischen erwähnt werden. Artikel kommen zusammen mit Substantiven, Adjektiven, Numeralen und Pronomina vor.
Die Ambiguität der Oberfläche bzw. die Multifunktionalität der Artikel i, e, të und së bereiten dem Sprecher/Hörer bzw. Leser/Schreiber oft Schwierigkeiten, vgl. Tabellen 3.8 und 3.9.

[132] Ausgewählt anhand von Adjektiv-Einträgen des Lexikons, das im Rahmen der vorliegenden Arbeit erstellt wurde. Vgl. hierzu auch Kapitel 4.
[133] Obwohl formal einer Gruppe zugeordnet, zeigen diese Adjektive verschiedene Eigenschaften, z. B. wird *neto* nicht dekliniert, während *afatshkurtër*; *dijeplotë* dekliniert wird.

Tabelle 3.8: Deklination des vorangestellten Artikels

Deklination des vorangestellten (d. h. gebundenen) Artikels						
	Unbestimmt			Bestimmt		
	m. sg.	f. sg.	n. sg. ; pl.	m. sg.	f. sg.	n. sg. ; pl.
Nom.	*i*	*e*	*të*	*i*	*e*	*e*
Gen.	*të*	*të*	*të*	*të*	*së*	*të*
Dat.	*të*	*të*	*të*	*të*	*së*	*të*
Akk.	*të*	*të*	*të*	*e*	*e*	*e*
Abl.	*të*	*të*	*të*	*të*	*së*	*të*

Tabelle 3.9: Deklination des selbstständigen Artikels

Deklination des selbstständigen (d. h. ungebundenen) Artikels						
	Unbestimmt			Bestimmt		
	m. sg.	f. sg.	n. sg. ; pl.	m. sg.	f. sg.	n. sg. ; pl.
Nom.	*i*	*e*	*të*	*i*	*e*	*të*
Gen.	*të*	*të* /*së*	*të*	*të*	*të*	*të*
Dat.	*të*	*të* /*së*	*të*	*të*	*të*	*të*
Akk.	*të*	*të*	*të*	*të*	*të*	*të*
Abl.	*të*	*të* /*së*	*të*	*të*	*të*	*të*

Der vorangestellte Artikel dient als Possessivum bei Substantiven: *i*␣*vëlla-i* (= *vëllai*␣*i*␣tij/saj/tyre) [3. P. Sg./Pl. f./m.] (dt. *sein/ihr/... Bruder*), *e*␣*motra* (= *motra*␣*e*␣tij/saj/tyre) [3. P. Sg./Pl. f./m.] (dt. *seine/ihre/... Schwester*).

3.3.4 Das Pronomen

Die Pronomina sind eine geschlossene Wortart. Sie sind aber eine sehr heterogene Menge. Sie können Substantive oder Nominalphrasen ersetzen, da sie einige gleiche grammatische Eigenschaften besitzen, wie Numerus, Kasus und Genus, sowie eine zusätzliche Kategorie, die der Person (bei einigen Pronomina, d. h. Personalpronomina). Durch die Kategorie der Person wird die Kongruenz zu den Verben ermöglicht.[134] Eine weitere Kategorie kommt bei Possessivpronomina vor, nämlich die des Besitzes. Auf der anderen Seite ist die Kategorie der Bestimmtheit von Typ zu Typ unterschiedlich. Es

[134] Substantive werden bei der linguistischen Beschreibung mit der 3. Person bezeichnet.

werden aus morphologischer und semantischer Sicht folgende Typen unterschieden: Personalpronomina (*unë*, *ti*, *ai*/*ajo* ..., siehe Tabelle 3.10), Reflexivpronomina (*vetja*, *vetvetja*, *veten*, *vetveten* ...), Identifizierendes Pronomen[135] (*vetë* (dt. *selbst*)), Reziprokpronomina (*njëri-tjetrit* ...), Possessivpronomina (*imi*, *imja*, ...), Demonstrativpronomina (*ai*, *kjo* ...), Interrogativpronomina (*kush*, *çfarë* ...), Relativpronomina (*i␣cili*, *që* ...), Determinativpronomina (*mbarë*, *tërë* ...), Indefinitpronomina (*një*, *njëri* ...).[136]
Die Personalpronomina, vgl. Tabelle 3.10, zeigen komplexe Eigenschaften. Neben den Hauptformen besitzen sie auch eine *kurze Form* im Akkusativ und im Dativ, sowie eine *reduzierte Form*, wenn sie nach einigen Präpositionen vorkommen.

Tabelle 3.10: Deklination der Personalpronomina

Deklination der Personalpronomina								
	1. Person				2. Person			
	Singular		Plural		Singular		Plural	
Nom.	*unë*	--	*ne*	--	*ti*	--	*ju*	--
Gen.	--	--	--	--	--	--	--	--
Akk.	*mua*	më	*ne*	na	*ty*	të	*ju*	ju
Dat.	*mua*	më	*neve*	na	*ty*	të	*juve*	ju
Abl.	*meje*	--	*nesh*	--	*teje*	--	*jush*	--
	3. Person							
	Singular				Plural			
	Maskulinum		Femininum		Maskulinum		Femininum	
Nom.	*ai*	--	*ajo*	--	*ata*	--	*ato*	--
Gen.	{*i, e, të*}␣*atij*	--	{*i, e, të*}␣*asaj*	--	{*i, e, të*}␣*atyre*	--	{*i, e, të*}␣*atyre*	--
Akk.	*atë* (*të*)	e	*atë* (*të*)	e	*ata* (*ta*).....	i	*ato* (*to*)	i
Dat.	*atij*	i	*asaj*.....	i	*atyre*	u	*atyre*	u
Abl.	*atij* (*tij*)	--	*asaj* (*saj*)	--	*atyre* (*tyre*)	--	*atyre* (*tyre*)	--

Die Kurzformen der Personalpronomina

Neben den Personalpronomina haben sich im Laufe der Sprachgeschichte noch kurze Formen (alban. *trajtat e shkurtra të përemrave vetorë*) entwickelt, welche die normalen Formen in vielen Fällen vertreten oder neben ihnen

[135] Vgl. [Buchholz/Fiedler 1987: 283 (§ III 3.3)]
[136] Ausführliche Informationen zu den Pronomina im Albanischen findet der interessierte Leser u. a. bei [Buchholz/Fiedler 1987: 274–314 (§ III 3)], [Newmark et al. 1982: 261–288] und [Morfologjia 1995: 215–258].

benutzt werden können, vgl. in Tabelle 3.10 die rechte Seite der Spalten, die Formen më, të, na, ju, e, i und u (in Serif-Schrift). Diese Formen werden auch unterschiedlich bezeichnet, bspw. nennen BUCHHOLZ und FIEDLER [1987] sie Objektszeichen, die damit verbundenen Erscheinungen Objektvertretung und Objektverdoppelung. Andere nennen sie klitische Pronomina oder pronominale Klitika.
Die Formen ohne *a*, d. h. *të*, *tij*, *saj*, *ta*, *tyre* und *to*, vgl. in Tabelle 3.10 die in Klammern gesetzten Formen im Akkusativ und Ablativ in der 3. Person, werden nur nach bestimmten Präpositionen verwendet, bspw. nach folgenden: *afër*+ABL, *kundër*+ABL, *mbi*+AKK, *me*+AKK, *prapa*+ABL, *para*+ABL, *përballë*+ABL, *përmes*+ABL, *prej*+ABL, ...[137] Diese Präpositionen regieren die pronominalen Formen, indem sie neben dem Kasus auch bestimmen, welche Form des Pronomens verwendet wird, nämlich diejenige mit *a* oder diejenige ohne.

Amalgamierte Formen

Die pronominalen Klitika (më, të, i, e, u ...) können miteinander kombiniert werden, vgl. (ua, ia, ta, ma, ...). Dabei entstehen aus der Zusammensetzung des Dativs mit dem Akkusativ die in Tabelle 3.11 dargestellten Formen. Amalgamierte Formen von pronominalen Klitika können zusätzlich noch mit Frage- und Negationspartikeln kombiniert werden, wie *m'u* / *ç'm'u* / *s'm'u* in den Sätzen *M'u kujtua kjo gjë.* / *Ç'm'u kujtua kjo gjë.* / *S'm'u kujtua kjo gjë.* (dt. *Diese Sache fiel mir ein.* / *Wieso fiel mir diese Sache ein?* / *Diese Sache fiel mir nicht ein.*) bzw. *t'i* / *ç't'i* / *s't'i* in {*T'i* / *Ç't'i* / *S't'i*} *themi këtij?* (dt. *Wir sagen es ihm.* / *Was sagen wir ihm?* / *Das sollen wir ihm wohl nicht sagen?*), vgl. hierzu auch Abschnitt 3.3.9. Zur Funktion der pronominalen Klitika vgl. [KABASHI 2007]. Ebenso können sie mit der Konjunktivpartikel të kombiniert werden. Die kurzen Formen der Personalpronomina (die klitischen Pronomina) können auch mit der Fragepartikel ç'/Ç' und mit der Negationpartikel s'/S' kombiniert werden. Die Formen (ohne ihre grammatischen Eigenschaften) sehen folgendermaßen aus: *ç'e*, *ç'i*, *ç'ia*, *ç'më*, *ç'ma*, *ç'm'i*, *ç'm'u*, *ç'na*, *ç't'ë*, *ç't'a*, *ç't'i*, *ç't'ia*, *ç't'ju*, *ç't'jua*, *ç't'u*, *ç't'ua*, *ç'ua*, *s'e*, *s'i*, *s'ia*, *s'më*, *s'ma*, *s'm'i*, *s'm'ju*, *s'm'jua*, *s'm'u*, *s'm'ua*, *s'na*, *s't'ë*, *s't'a*, *s't'i*, *s't'ia*, *s't'ju*, *s't'jua*, *s't'u*, *s't'ua*; *s'ua*, *u* (Passiv-Partikel) und *iu* (*i*+ Passiv-Partikel *u*); Die Formen *s'm'ia*,

[137] Dies gilt nicht für Substantive, bspw. *afër tyre (← atyre)* vs. *afër atyre njerëzve* [*afër *tyre njerëzve* ist grammatisch nicht korrekt]. Dies muss beim Tagging berücksichtigt werden. Vgl. hierzu auch Kapitel 4. Diese Präpositionen sind vollständig im Lexikon, das zur Morphologie-Komponente gehört, enthalten.

s'm'ju, *s'm'jua*, *s'm'u*, *s'm'ua* bzw. *s't'ia*, *s't'ju*, *s't'jua*, *s't'u*, *s't'ua* werden kaum verwendet, sind aber theoretisch möglich, genauso wie *s'm'i* bzw. *s't'i*.
Verallgemeinert ausgedrückt, kommt s' zusammen mit Verben vor und bildet das Muster s'+V. Einige Beispiele sind: *s'shkonte* (dt. *es ging nicht*), *s'dua* (dt. *ich will nicht*) und *s'punohej* (dt. *es wurde nicht gearbeitet* bzw. *man konnte nicht arbeiten*);
Eine Ausnahme bilden Fälle wie *s'ëmës* (dt. *seiner/ihrer Mutter*), wobei das *s'* für den Artikel *së* steht, wenn *ë* ausfällt, d. h., statt *së␣ëmës* wird *s'ëmës* verwendet.[138]
Ç' hingegen kommt sowohl mit Verben als auch mit Nomina zusammen vor. Das entsprechende Muster ist ç'+V/+N, wie anhand der folgenden Beispiele gesehen werden kann: *ç'ka?* (dt. *Was gibt es?*/*Was hat er/sie?*), *ç'ndodhi?* (dt. *Was passierte?*), *ç'vend* (dt. *Was für ein Ort/Stelle/Land?*) und *ç'punë* (dt. *Was für eine Angelegenheit/Sache/Arbeit?*)
In den Tabellen 3.11 und 3.12 sind die möglichen Kombinationen der Personalpronomina samt ihren grammatischen Eigenschaften angegeben. In den Grammatiken des Albanischen werden leider nur ihre Grundeigenschaften beschrieben, wie in Tabelle 3.10 (Personalpronomina) angegeben, und nicht die der komplexen Formen ihrer Kombination miteinander und mit den Partikeln të und u.[139]
Die Typen pDA-1' und cTx-1', d. h. die Form *ta*, sind in Hinsicht auf ihre Schreibung mehrdeutig, genauso wie die Typen cTx-1 und pDA-1, d. h. die Form *t'i*. In einem Fall ist *të* Konnektiv (cTx-1 und cTx-1'), im anderen Fall klitisches Pronomen (pDA-1 und pDA-1'). Die Kategorisierung ist für die Zwecke der morphologischen und syntaktischen Beschreibung wichtig, da beide Formen im Verbalkomplex vorkommen können und so entscheidend sind.
In Texten findet man ab und zu auch Formen der Klitika, die von der Norm abweichen, wie bspw. die Form m'e (statt ma) im folgenden Satz: „*Ngaqë secila m'e bukur se tjera qe …*“ (dt. *Weil jede hübscher als die andere war …*), vgl. [Koncert 1994].
Wie in Tabelle 3.11 und in 3.12 zu sehen ist, besitzen die klitischen Pronomina eine sehr hohe Informationsdichte. Die Form *t'ia* und alle Formen, die *i* beinhalten, tragen gegenüber den anderen Formen mehr Eigenschaften, da *i* sowohl im Akkusativ als auch im Dativ, sowohl Femininum als auch Maskulinum, sowohl im Singular als auch im Plural mit all diesen grammatischen Eigenschaften in der dritten Person vorkommt. Diese komplexe Form, extrahiert aus der besprochenen Tabelle, sieht so aus: *të*

[138] Vgl. [Drejtshkrimi 1974: 43 (§ 19 c)].
[139] Für die maschinelle Sprachverarbeitung ist ihre exakte Beschreibung notwendig.

Tabelle 3.11: Kombination der klitischen Pronomina miteinander

Kombination der klitischen Pronomina miteinander					
Dat.		Akk.	D+A	Grammatische Eigenschaften	Typ
më	+	*e*	→ *ma*	[(D|A 1P S)$_m$+(A 3P M|F S)$_e$]	pDA-1
më	+	*i*	→ *m'i*	[(D|A 1P S)+(A 3P M|F P)$_i$] | [(D|A 1P S)+(D 3P M|F S)$_i$]	pDA-1'
na	+	*e*	→ *na␣e*	[(D|A 1P P)$_{na␣}$(A 3P M|F S)]	pDA-2
na	+	*i*	→ *na␣i*	[(D|A 1P P)$_␣$(A 3P M|F P)] | [(D|A 1P P)$_␣$(D 3P M|F S)]	pDA-2
të	+	*e*	→ *ta*	[(D|A 2P S)$_t$+(A 3P M|F S)]	pDA-1
të	+	*i*	→ *t'i*	[(D|A 2P S)+(A 3P M|F P)] | [(D|A 2P S)+(D 3P M|F S)]	pDA-1'
ju	+	*e*	→ *jua*	[(D|A 2P P)$_{ju}$+(A 3P M|F S)]	pDA-1
ju	+	*i*	→ *jua*	[(D|A 2P P)+(A 3P M|F P)] | [(D|A 2P P)+(D 3P M|F S)]	pDA-1
i	+	*e*	→ *ia*	[(D 3P M|F S)+(A 3P M|F S)] | [(A 3P M|F P)+(A 3P M|F S)]	pDA-1
i	+	*i*	→ *ia*	[(D 3P M|F S)+(A 3P M|F P)]	pDA-1
u	+	*e*	→ *ua*	[(D 3P M|F P)$_u$+(A 3P M|F S)]	pDA-1
u	+	*i*	→ *ua*	[(D 3P M|F P)+(A 3P M|F P)] | [(D 3P M|F P)+(D 3P M|F S)]	pDA-1

+ [(D 3P M|F S)+(A 3P M|F S)], *të* + [(A 3P M|F P)+(A 3P M|F S)]] oder *të* + [(D 3P M|F S)+(A 3P M|F P)], d. h., es sind pro Einheit vier[140] einzelne Fälle, also insgesamt zwölf Formen.[141]

Bei Possessivpronomina gibt es noch eine zusätzliche Kategorie, die des Besitzes und die des Besitzers, vgl. [Fjalori 1984: 1475–1515]. Als Beispiel könnte hier erwähnt werden: *imi* (dt. *mein*), mit den Kategorien 1. P. Nom., Besitz: sg. m., Besitzer: sg.; *imja* (dt. *meine*), mit 1. P. Nom., Besitz: sg. f., Besitzer: sg. usw. Es handelt sich um Nominalisierung, vgl. [Morfologjia 1995: 241–243 (§ 6.5.4)].

3.3.5 Numerale

Es werden in erster Linie drei Typen der Numerale unterschieden, nämlich die Kardinalzahlen, die Ordinalzahlen und die Bruchzahlen.[142]

[140] Die zusammengefasste Form [(D 3P M|F S)+(A 3P M|F S)] steht für [(D 3P M S)+(A 3P M S)], [(D 3P M S)+(A 3P F S)], [(D 3P F S)+(A 3P M S)] und [(D 3P F S)+(A 3P F S)].

[141] Die Implementierung der klitischen Pronomina wird in Abschnitt 5.3.5 erläutert.

[142] Vgl. [Buchholz et al. 1993: 676 ff.].

Tabelle 3.12: Kombination von *të* mit den klitischen Pronomina

Kombination von *të* mit den klitischen Pronomina				
të		Kl. Pron.	Grammatische Eigenschaften	Typ
të	+	*më* . . . . . → *të␣më*	[*të␣më*]	cTx-2
të	+	*ma* . . . . . → *të␣ma*	[*të␣ma*]	cTx-2
të	+	*m'i* . . . . . → *të␣m'i*	[*të␣m'i*]	cTx-2'
të	+	*të* . . . . . → *të␣të*	[*të␣të*]	cTx-2
të	+	*ta* . . . . . → *të␣ta*	[*të␣ta*]	cTx-2
të	+	*t'i* . . . . . → *të␣t'i*	[*të␣t'i*]	cTx-2'
të	+	*e* . . . . . → *ta*	[*të*+(A 3P M|F S)]	[1]cTx-1
të	+	*i* . . . . . → *t'i*	[*të*+(A 3P M|F P)|(D 3P M|F S)]	[2]cTx-1'
të	+	*ia* . . . . . → *t'ia*	[[*të*+[(D 3P M|F S)+(A 3P M|F S)]] | [*të*+[(A 3P M|F P)+(A 3P M|F S)]] | [*të*+[(D 3P M|F S)+(A 3P M|F P)]]]	cTx-1''
të	+	*na* . . . . . → *të␣na*	[*të␣na*]	cTx-2
të	+	*na e* . . . . . → *të␣na␣e*	[*të␣na␣e*]	cTx-3
të	+	*na i* . . . . . → *të␣na␣i*	[*të␣na␣i*]	cTx-3
të	+	*ju* . . . . . → *t'ju*	[*të*+(D|A 2P P)]	cTx-1'
të	+	*jua* . . . . . → *t'jua*	[[*të*+[(D|A 2P P)+(A 3P M|F S)]] | [*të*+[(D|A 2P S)+(D 3P M|F S)]]]	cTx-1'
të	+	*u* . . . . . → *t'u*	[*të*+(D 3P M|F P)] cTx-1'	
të	+	*ua* . . . . . → *t'ua*	[[*të*+[(D 3P M|F P)+(A 3P M|F S)]] | [*të*+[(D 3P M|F P)+(A 3P M|F P)]] | [*të*+[(D 3P M|F P)+(D 3P M|F S)]]]	cTx-1'

Kardinalzahlen

Die Kardinalzahlen werden in folgende Typen unterschieden:

- die Zahlen *null* und *eins* bis *neun*, alban. *zero* [0], *një* [1], *dy* [2], *tre* [3], *katër* [4], *pesë* [5], *gjashtë* [6], *shtatë* [7], *tetë* [8] und *nëntë* [9];[143]
- die Zahlen 11 bis 19, alban. *një*|*mbë*|*dhjetë* [11], *dy*|*mbë*|*dhjetë* [12], *tre*|*mbë*|*dhjetë* [13], *katër*|*mbë*|*dhjetë* [14], *pesë*|*mbë*|*dhjetë* [15], *gjashtë*|*mbë*|*dhjetë* [16], *shtatë*|*mbë*|*dhjetë* [17], *tetë*|*mbë*|*dhjetë* [18] und *nëntë*|*mbë*|*dhjetë* [19].[144] Es handelt sich um das Muster {[1–9]}-|mbë¦|dhjetë;

[143] Die Zahl *drei* im Albanischen hat zwei Genera, vgl. alban. *tre* (m.) / *tri* (f.).

[144] Vgl. hierzu [Demiraj B. 1990]. Das gebundene Morphem *mbë* kommt nicht frei vor und wird aus diesem Grund nicht segmentiert. Es kommt nur in diesen Zahlen vor. Vgl. hierzu auch [Demiraj Sh. 1994].

- die Zahlen *zehn*, *zwanzig*, *dreißig* usw. bis *neunzig*, alban. *dhjetë* [10], *një*|¦*zet* (*një·zet*) [20, d. h. (1·20)], *tri*|*dhjetë* [30], *dy*|¦*zet* (*dy·zet*) [40, d. h. (2·20)], *pesë*|*dhjetë* [50], … *nëntë*|*dhjetë* [90].[145] Es handelt sich hier um drei verschiedene Muster: (M-1) dhjetë, (M-2) {një, dy}|¦zet[146] und (M-3) {tri, (katër,) pesë, gjashtë, shtatë, tetë, nëntë}|dhjetë;

- die Zahlen der *hundert*-Reihe werden zusammengeschrieben: (alban. *një*|*qind* [100], *dy*|*qind* [200], *tre*|*qind* [300] usw. bis *nëntë*|*qind* [900]). Es handelt sich hier um das Muster {një, dy, tre, katër, pesë, gjashtë, shtatë, tetë, nëntë}|qind. *qind* kommt nicht allein vor;

- die restlichen größeren Zahlen werden nicht zusammengeschrieben: die *tausend*-Reihe, die *millionen*-Reihe und die *milliarden*-Reihe, vgl. (alban. *një mijë* [1 000], *një milion* [1 000 000], *një miliard* [1 000 000 000] usw.).[147] Es handelt sich hier um das Muster {[1-999]}␣mijë, {[1-999]}␣milion, pl. milionë, {[1-999]}␣miliard, pl. miliardë usw. z. B. {një, dy, tre, katër, pesë, gjashtë, shtatë, tetë, nëntë, dhjetë, njëmbëdhjetë, … shtatëdhjetë, … njëqind␣e␣dymbëdhjetë, … katërqind␣e␣njëzet␣e␣tetë, … nëntëqind␣e␣nëntëdhjetë␣e␣nëntë}␣mijë;

Einige Beispiele wären *dy* [2], *pesëmbëdhjetë* [15], *shtatëdhjetë* [70], *katërqind* [400], *një␣mijë* [1 000], *një␣milion* [1 000 000], *dy␣miliardë* [2 000 000 000] usw. bzw. *njëzet␣e␣pesë* [25], *dyqind␣e␣shtatë* [207], *dyqind␣e␣shtatëdhjetë␣e␣një* [271], *katërqind␣e␣dymbëdhjetë␣mijë␣e␣pesëqind␣e␣shtatëdhjetë␣e␣tetë* [412 578] und *dyqind␣e␣shtatë␣miliardë␣e␣njëqind␣e␣pesëmbëdhjetë␣milionë␣e␣dy␣mijë␣e␣pesë* [207 115 002 005].

[145] Oft wird statt dyzet die gleichwertige Form katërdhjetë verwendet. Wie das gebundene Morphem *mbë* kommt auch *zet* nicht frei vor und wird aus dem gleichen Grund ebenso nicht segmentiert. Es kommt nur in diesen Zahlen vor.

[146] In der Norm-Sprache sind nur die ersten zwei Zahlen als Vigesimalsystem noch vorhanden, während in den Varianten Arvanitisch und Arbëresh immernoch die vollständige Reihe vorhanden ist, vgl. hierzu [Sasse 1990: 136–140] (Arvanitisch: 60=3·20=*tre*|*zét*, 80=4·20=*katər*|*zét*) und [Hamp 2006: 23–29] (Arbëresh: *trɛ*|*zɛt* (3·20) und *katər*|*zɛt* (4·20).

[147] Bei der *zehn*-Reihe wird die feminine Form verwendet, sonst kommt nur die maskuline Form der Zahl drei vor, vgl. tridhjetë [30], treqind [300], tre␣mijë [3 000], tre␣milionë [3 000 000], tre␣miliardë [3 000 000 000] usw.

Ordinalzahlen

Ordinalzahlen haben obligatorischerweise einen vorangestellten Artikel[148] und bis auf *i␣parë* das Suffix *-të* bzw. *-t*, vgl. *i␣dy|të*, *i␣katër|t*. Ein Vergleich zwischen den Kardinalzahlen und Ordinalzahlen wäre: *dy* vs. *i␣dy|të*, *katër* vs. *i␣katër|t*. Die Ordinalzahlen, außer der Zahl *die/der erste*, alban. *e/i␣parë*, werden aus den Kardinalzahlen abgeleitet. Dies kann bei den Zahlen (11.,) 21., 31., … 91. am deutlichsten gesehen werden, vgl. alban. *i␣*një*mbëdhjetë*, *i␣një-zet|e|*një|*të* usw.

Im Folgenden sind die einzelnen Zahlen angegeben: *i␣parë* [1.], *i␣dytë* [2.], *i␣tretë* [3.], *i␣katërt* [4.], *i␣pestë* [5.], *i␣gjashtë* [6.], *i␣shtatë* [7.], *i␣tetë* [8.], *i␣nëntë* [9.], *i␣dhjetë* [10.], *i␣njëmbëdhjetë* [11.], *i␣dymbëdhjetë*, [12.] usw. Es zeichnet sich das folgende Muster ab: {*i*, *e*, *të*} *dy|të*, {*i*, *e*, *të*} *dymbëdhjetë*[149], {*i*, *e*, *të*}*␣njëmij|të*, {*i*, *e*, *të*}*␣njëmilion|të* usw.[150]

Folgendes Beispiel soll die Genera und Bestimmtheit der Ordinalzahlen illustrieren: {*i* [m. sg.], *e* [f. sg.], *të* [m.+f. pl.]}*␣par|ë* [ub.], *i␣par|i* [best. m. sg.], *e␣par|a* [best. f. sg.] und *të␣par|ët* [best. m. pl.] bzw. *të␣par|at* [best. f. pl.] [1.], *i␣dy|të*, *i␣dy|ti*, *e␣dy|ta*, *të␣dy|tët*, *të␣dy|tat* [2.], … *i␣njëzetedy|të*, *e␣njëzete-dy|të*, *të␣njëzetedy|të*, *të␣njëzetedy|ta*, *i␣njëzetedy|ti*, *e␣njëzetedy|ta*, *të␣një-zetedy|tët*, *të␣njëzetedy|tat* [22.] usw.

Weitere Beispiele wären: *i␣shtatëdhjetëenjë|të*, *e␣shtatëdhjetëenjë|të*, *të␣shta-tëdhjetëenjë|të*, *i␣shtatëdhjetëenjë|ti*, *e␣shtatëdhjetëenjë|ta*, *të␣shtatëdhjetëe-një|tët*, *të␣shtatëdhjetëenjë|tat* [71.]. Diese Formen zeigen die gleichen Eigenschaften wie die Artikeladjektive (außer deren Steigerung) und werden als solche behandelt. Der Artikel, vgl. die Belege *e*, *të* und *së*, wird wie in Tabelle 3.8 angegeben dekliniert.[151]

---

[148] Es handelt sich um die Artikel *e*, *i* und *të* (*së*), vier mögliche Formen, die für verschiedene Fälle stehen, vgl. die Tabellen 3.8 und 3.9.

[149] An dem aufgeführten Beispiel ist zu erkennen, dass bei den Zahlen, die auf -të enden (*gjash*të [6], *shta*të [7], *te*të [8], *nën*të [9] und *dhje*të [10]) kein zusätzliches -të vorkommt. Bei der Zahl *pesë* [5] vs. *i␣pes*të [5.] fällt das unbetonte -ë vor -të in der bestimmten Form aus. Die Zahl *katër* vs. *i␣katër*t stellt einen Sonderfall dar. Schließlich werden die Zahlen *njëzet* und *dyzet* in ihren bestimmten Formen mit -ë, d. h. *i␣njëzet*ë bzw. *i␣dyzet*ë, geschrieben. Vgl. auch [Buchholz et al. 1993: 676–680].

[150] Vgl. hierzu auch [Çeliku et al. 1998: 76–82 (§§ 47–52)].

[151] Bei den Formen wie nëntëdhjetë␣e␣nëntë, wo das ursprüngliche ë wie z. B. bei *dhjetë* [10], und die Konjunktion e in Kontaktposition vorkommen, fällt das vorangehende ë aus, vgl. [Drejtshkrimi 1974: § 67 ç].

Bruchzahlen

Die Bruchzahlen bestehen aus einem Zähler und einem Nenner, wobei der Zähler eine Kardinalzahl ist, während der Nenner eine Ordinalzahl ist, vgl. hierzu alban. *një␣e␣katërta* [$\frac{1}{4}$ oder 1/4], *dy␣të␣pestat* [$\frac{2}{5}$ oder 2/5] usw. Am letzten Beispiel ist zu erkennen, dass der Numerus einer Bruchzahl vom Zähler abhängig ist, d. h., außer den Bruchzahlen mit Zähler *një* [1] sind sie im Plural. Während der Zähler nicht dekliniert wird, folgt der Nenner der Deklination eines substantivierten Adjektivs in bestimmter Form im femininen Genus, vgl. im Singular *një␣e␣katërta* [Nom.], {*i*, *e*, *të*}␣*një␣së␣katër-tës* [Gen.], *një␣së␣katërtës* [Dat.], *një␣të␣katërtën* [Akk.] und *një␣së␣katërtës* [Abl.] sowie im Plural *dy␣të␣pestat* [Nom.], {*i*, *e*, *të*}␣*dy␣të␣pestave* [Gen.], *dy␣të␣pestave* [Dat.], *dy␣të␣pestat* [Akk.] und *dy␣të␣pestave* [Abl.]. Die Zahl *drei* [3] kommt als Zähler in femininer Form vor, vgl. alban. *tri␣të␣shtatat* [$\frac{3}{7}$ oder 3/7].
Das Nominalsystem (Adjektive, Artikel, Numerale, Pronomina und Substantive) ist reich an Kategorien und Formen, genauer Kasus (Nominativ, Genitiv, Akkusativ, Dativ und Ablativ), Numeri (Singular und Plural) sowie Bestimmtheit bzw. Unbestimmtheit, also insgesamt 20, sowie Genera (Femininum, Maskulinum und Neutrum).

Im Folgenden werden die weiteren Wortarten des Albanischen behandelt, nämlich Adverb, Interjektion, Konjunktion, Partikel und Präposition, welche allesamt nicht flektiert werden – die Indeklinabilia.

3.3.6 Das Adverb

Adverbien im Albanischen werden nicht flektiert, können aber wie die Adjektive graduiert werden. Sie weisen eine Verwandtschaft zu den Adjektiven auf. Einer der auffälligsten Unterschiede zwischen den zwei Wortarten ist das Vorkommen der vorangestellten Artikel bei den Adjektiven, im Gegensatz zu den Adverbien, wo sie nicht vorkommen, bspw. $e␣mirë_{ADJ}$ vs. $mirë_{ADV}$: *Një orë e mirë* (dt. *Eine gute Uhr*) vs. *Kjo orë punon mirë* (dt. *Diese Uhr arbeitet gut*).[152] Doch nicht jedes Adverb kann durch Voranstellung eines Artikels zu einem Adjektiv überführt werden bzw. kann umgekehrt nicht jedes Artikel-Adjektiv zu einem Adverb werden, wenn sein Artikel weggelassen wird, vgl.

[152] Dies gilt nur für bestimmte Wörter der jeweiligen Wortarten, denn es gibt Adjektive, die keinen vorangestellten Artikel besitzen, sowie Adverbien, die einen vorangestellten Partikel als Bestandteil ihrer Struktur haben.

i␣denjë$_{ADJ}$ (dt. *würdig*), *denjësisht*$_{ADV}$ (dt. *würdevoll*). Eine Form wie *denjë* kommt im Albanischen nicht ohne vorangestellten Artikel vor.
Einige Adverbial-Formen wie *së␣tepërmi*, *së␣pari*, *së␣voni*, *së␣fundi* bestehen aus einem Partikel (*së*) und einer Adverb-, Adjektiv- oder Substantivform, die mit der Grundform nicht zwangsweise übereinstimmt, z. B. *së+voni*, Grundform *vonë* (dt. *spät*).[153] *Së␣voni erdhi e motra, por më␣së␣voni erdhi i vëllai* (dt. *Sehr spät kam die (ihre/seine) Schwester, aber als letztes kam der (ihr/sein) Bruder*). Bei diesen Adverbien handelt es sich um eine Eigenschaft, die Adjektive besitzen, nämlich Graduierbarkeit (oder Steigerbarkeit).

3.3.7 Die Präposition

Die Präpositionen sind Funktionswörter und werden nicht flektiert. Sie spielen in einer Sprache vielfältige Rollen, u. a. besitzen sie die Fähigkeit der Rektion. Die meisten Präpositionen im Albanischen haben Objekte im Ablativ, wie bspw. *prej*. Aber einige sehr häufige Präpositionen haben Objekte im Akkusativ und Nominativ, vgl. hierzu [Newmark 1999: xliii][154] und den Wortschatz in [Hetzer/Finger 1993]. Ausführliche Informationen über die Präposition im Albanischen findet der interessierte Leser bei [Buchholz/Fiedler 1987: 373–384 (§ IV 2)] sowie bei [Samara 1999] und [Thomai 2005].

3.3.8 Die Konjunktion

Die Konjunktionen gehören zu den nichtflektierten Wortarten (Indeklinabilia). Sie weisen ähnliche Eigenschaften wie Partikel und Interjektionen auf. Im Vergleich zu den Interjektionen kommen sie auch als Mittel zur Wortbildung in Frage, bspw. die Konjunktion e bei den Ordinalzahlen. Für die maschinelle Sprachverarbeitung spielen sie im Bereich der Syntax eine wichtige Rolle. Oft wird als erste Charakterisierung die Eigenschaft unterordnend vs. nebenordnend genannt. Konjunktionen haben verschiedene semantische Funktionen, u. a. kausale, konditionale und temporale. Einige Konjunktionen sind zwei- oder mehrteilig, d. h., sie bestehen aus zwei (oder

[153] Lexika, z. B. [Dhrimo/Memushaj 2011] führen diese Form als vó·ni (së) ndajf. an. Solche Formen werden im Lexikon angegeben, denn sie sind mittlerweile nur mit Hilfe von historischem Sprachwissen zu erklären.

[154] "The majority of prepositions in Albanian have ablative case objects, but a few very frequent prepositions me, pa, në, më, për, mbi, nën, ndër, nëpër, as well as phrasal prepositions ending in one of these (e.g., për në, brenda në, tok me, bashkë me), have objects in the accusative case. The prepositions nga and tek (or te) [...] have objects in the nominative case."

mehreren) Wörtern, wie bspw. *ose … ose* (dt. *entweder … oder*), *edhe ashtu edhe kështu* (dt. *sowohl (so) … als auch (so) …*) usw., vgl. hierzu u. a. auch [Çeliku et al. 1998: 271–290 (§§ 244–251)], [Hetzer/Finger 1993 § 29.4.1] und [Buchholz/Fiedler 1987: 385–391 (§ IV 3.1)] sowie Abschnitt 4.7.

3.3.9 Die Partikel

Partikeln bilden, wie die anderen nichtflektierbaren Wortarten, eine geschlossene Klasse. Sie spielen verschiedene, insbesondere syntaktisch-semantische Rollen in der Sprache.[155]
Einige Partikeln, nämlich die Frage- und die Negationspartikeln können mit klitischen Pronomina (auch mit den zweifachen) amalgamiert werden und bilden so gemeinsame Formen. So kann die Fragepartikel ç' als ein Teil der folgenden Formen vorkommen: ç'e, ç'i, ç'më, ç'të, ç'ia, ç'u (Partikel + einfaches klitisches Pronomen bzw. Partikel + Passiv-Partikel) oder als Teil der Formen ç'm'i, ç'm'u, ç't'i, ç't'u, … ç't'iu usw. (Partikel + zweifaches klitisches Pronomen bzw. Partikel + Passiv-Partikel, wobei t' (të) auch Konjunktion sein kann). Ähnlich verhält es sich in der Kombinatorik auch mit der Negationspartikel s' in den Formen s'i, s'e, s'ia, s'na, s'më, s'ta, s'të, s'u bzw. s'm'i, s'm'u, s't'i, s't'u usw.

3.3.10 Die Interjektion

Die Wortart Interjektion umfasst nur eine begrenzte Anzahl an Wörtern und ist daher eine sogenannte geschlossene Klasse. Interjektionen werden nicht flektiert.[156] Ausnahmen sind Fälle wie *djal*|o, sg. (dt. *O Junge*) oder *trim*|o, sg. (dt. *O Held*), wobei die Substantive mit einem Interjektionssuffix versehen werden. Diese kommen jedoch nicht bei allen Substantiven vor, sondern im Wesentlichen nur bei einigen Personenbezeichnungen.[157]

[155] Ausführliche Informationen hierzu bieten u. a. [Buchholz/Fiedler 1987: 392–395 (§ IV 4)] und [Morfologjia 1995: 413–426 (§ XI)].
[156] Ausführliche Informationen hierzu bieten u. a. [Buchholz/Fiedler 1987: 396–402 (§ IV 5.1)] und [Morfologjia 1995: 427–439 (§ XII)].
[157] Es ist jedoch bei der literarischen Verwendung der Sprache nicht ausgeschlossen vgl. etwa die Lyrik von N. Frashëri (1846–1900).

3.3.11 Überblick über die morphologischen Eigenschaften

Bevor auf das Thema Wortbildung eingegangen wird, sei eine Zusammenfassung der morphologischen Eigenschaften des Albanischen in tabellarischer Form gegeben.

Tabelle 3.13: Überblick über die morphologischen Eigenschaften.

	S	Adj	Num	Art	Pron	V	Adv	Konj	Präp	Exkl
Kasus	+	+	+/-	+	+	-	-/+?	-	-	-
Genus	+	+	+/-	+	+	-	-	-	-	-
Bestimmtheit	+	+/-	+/-	-/+	+/-	-	-	-	-	-
Graduierbarkeit	-	+	-	-	-	-	+	-	-	-
Numerus	+	+	+	+	+	+	-	-	-	-
Person	-/+3.	-	-	-	+/-	+	-	-	-	-
Besitz & Besitzer	-	-	-	-/+	-/+	-	-	-	-	-
Tempus	-	-	-	-	-	+	-	-	-	-
Modus	-	-	-	-	-	+	-	-	-	-
Admirativität	-	-	-	-	-	+	-	-	-	-
Diathese	-	-	-	-	-	+	-	-	-	-
+Art/+Part^{mwu}	-/+	+/-	+/-	+	-/+	+(u)	+(së)	-/+	-/+?	-

In der Tabelle 3.13 sind die Klassen innerhalb einer Wortart nicht enthalten. Sie stellen oft starke morphologische bzw. morphologisch-syntaktische Unterschiede zwischen lexikalischen Einheiten dar, wie es bspw. der Fall bei Adjektiven ist.[158]

3.4 Die Wortbildung

Neben der Formenbildung (Deklination und Konjugation) stellt die Wortbildung einen weiteren Bereich der Morphologie dar. Unter dem Begriff Wortbildung werden alle Möglichkeiten verstanden, neue Wörter zu bilden, insbesondere durch Kombination von Wörtern miteinander oder durch Kombination von Wörtern mit Affixen.
Während es sich bei der Flexion um die Darstellung bestimmter grammatischer Eigenschaften desselben Wortes handelt, geht es bei der Wortbildung um die Bildung neuer Wörter, seien diese abgeleitet aus einem

[158] mwu steht für „Multi Word Unit“. Damit sind die vorangestellten Partikeln bzw. Artikel gemeint, welche in der Tabelle in Klammern gesetzt sind. Das Fragezeichen markiert Ausnahmen.

bestehenden Wort oder aus mehreren Wörtern und anderen wortbildenden Mitteln, welche die Sprache bietet.

3.4.1 Mittel und Typen der Wortbildung

Zu den Wortbildungsmitteln zählen – neben dem Wortschatz einer Sprache – Präfixe, Präfixoide, Suffixe, Suffixoide, vorangestellte Artikel und Fugenelemente. Zu den Wortbildungstypen des Albanischen zählen nach [Morfologjia 1995: 58–79]:

- Derivation, alban. *prejardhja*
- Wortbildung durch Hinzufügen von vorangestellten Artikeln, alban. *nyjëzimi*
- Komposition, alban. *kompozimi*
- Zusammenrückung, alban. *përngjitja*
- Konversion, alban. *konversioni*
- Gemischte Modelle, alban. *mënyrat e përziera*.[159]

Zur Beschreibung der Präfixe und Suffixe gelten als Hauptliteraturquelle zwei Werke von Aleksandër Xhuvani und Eqrem Çabej, welche 1962 (Präfixe) und 1975 (Suffixe) publiziert worden sind. Sie werden auch heute noch als Klassiker angesehen, vgl. [Xhuvani/Çabej 1962 und 1975]. Als nächstes kann das Werk von Namik Resuli, vgl. [Ressuli 1985], erwähnt werden. Es widmet der Wortbildung, sowohl der Derivation als auch der Komposition, relativ großen Raum.[160] Das Standardwerk [Morfologjia 1995] behandelt die Wortbildung [43–79 (§ II)] ziemlich kompakt. [Çeliku et al. 1998: 299–371 (§§ 258–395)] behandeln die Wortbildung übersichtlich und erklären diese mit vielen Beispielen. Weitere Werke zur Wortbildung im Albanischen, die erwähnt werden sollen, sind [Hysa 2004], [Thomai 2009: 215–271 (§ IV)], [Buxheli 2007, 2008 und 2009] sowie [Turano 2010]. Eine historische Behandlung dieser Themen bietet [Demiraj Sh. 1994].[161]
Im Folgenden wird auf die einzelnen Punkte näher eingegangen. Die Wortbildungsmittel wurden aus den Quellen [Xhuvani/Çabej 1962 und 1975],

[159] Vgl. insbesondere [Morfologjia 1995: 58–61 (§ 2.1.10)].

[160] Vgl. insbesondere Substantive [136–182 (§§ 246–267)], Adjektiv-Syntagmen [193–196 (§ 287)], [199 f. (§§ 292–294)], Adjektive [213–228 (§§ 314–329)], Verben [599–570 (§§ 692–698)] und Adverbien [671–586 (§§ 718–725)].

[161] Die albanische Morphologie kann im Rahmen der vorliegenden Arbeit nicht aus historischer Perspektive behandelt werden.

[Morfologjia 1995], [Hysa 2004: 203–208], [Buxheli 2008: 351–365] und [Çeliku et al. 1998: 299–371 (§§ 258–395)] entnommen.[162]

Präfixe

In der albanischen Literatur werden unter dem Begriff Präfixe auch Präfixoide eingeordnet, vgl. z. B. [Hysa 2004]. Es werden jeweils spracheigene (indigene) und fremde (exogene) Präfixe und Präfixoide unterschieden. Im Folgenden wird auf die einzelnen Punkte eingegangen:

- Präfixe (alban. *Parashtesat*):
  - Substantive:

    bashkë-, jo-, kundër-, m-, mbas-, mbi-, më-, mi-, mos-, ndaj-, ndën-, ndër-, nën-, pa-, para-, pas-, për-, përmbi-, përtej-, pranë-, prapa-, ri-, sipër-, stër- und tej-. Einige Beispiele wären: *bashkë+punim* (dt. *Zusammenarbeit*), *mbi+ngarkesë* (dt. *Überlast*), *para+thënie* (dt. *Vorwort*) …
  - Adjektive:

    Die Präfixe, die mit Substantiven kombiniert werden können, dienen auch zur Wortbildung der Adjektive, z. B. për-, mbi-, nën-, ndër-, pa-, mos-, ç-/sh-/zh-, stër-, shpër-, tej- usw.

    Beispiele: *i␣drejtë* (dt. *gerade*) + *zh* → *i␣zhdrejtë* (dt. *ungerade*), *i␣afërt* (dt. *verwandter*) + *për* → *i␣përafërt* (dt. *ähnlich* u. ä.).
  - Verben:

    ç-, kaca-, kër-, kërth-, kse-, lë-, m-, mbë-, mbi-, n-, ndaj-, ndër-, nën-, nëpër-, ra-, për-, r(ë)-, ri-, rrë-, s-, sk-, skër-, spër-, shpër-, sh-, shka-, shkala-, shkara-, shkël-, shp(ë)-, stër-, të-, tër-, toro-, vër-, xh-, z-, zdër-, zgër-, zh- und zhdër-. Dazu folgende Beispiele: *mbi+vlerësoj* (dt. *überbewerten*), *për+punoj* (dt. *verarbeiten*) *shpër+ndaj* (dt. *verteilen*) …

- Exogene Präfixe (alban. *Parashtesat me prejardhje të huaj*):
  - Substantive:

    a-, anti-, de-, dis-, dez-, super-, trans-, ultra-, i-, pan-, pro- und eks-. Einige Beispiele: *dis+harmoni* (dt. *Disharmonie*), *super+çmim* (dt. *Superpreis*) …

[162] Die Beispiele werden nur in einzelnen Fällen übersetzt.

– Adjektive:

  Die exogenen Präfixe der Substantive können größtenteils auch mit Adjektiven kombiniert werden, z. B. a-, an-, anti-, auto-, de-, dez-, dis-, in-, inter-, pan-, poli, pro-, super- usw.

  Beispiele: *europian* (dt. *europäisch*/*Europäer*) + *pro-* → *proevropian* (dt. *proeuropäisch*), *moral* (dt. *Moral*) + *i* → *imoral* (dt. *amoralisch* u. ä.).

– Verben:

  a-, de-, dez-, in- ... und andere exogene Präfixe kommen nur in Verbindung mit entlehnten Wörtern vor, wie z. B. *de+kompozoj* (dt. *dekomponieren*), *dez+informoj* (dt. *desinformieren*) usw.

• Präfixoide (alban. *Siparashtesat*):

– Substantive:

  gjysmë-, ish-, krye-, mes-, meso-, ndihmës-, vetë-, zëvendës- usw. Hier wären noch die Präfixoide wie bashkë-, kundër-, para-, pas-, sipër-, një-, dy- usw. zu nennen, welche oft auch als Präfixe gesehen werden.

  Einige Beispiele: *gjysëmpërçues* (dt. *Halbleiter*), *ishkryetar* (dt. *ehemaliger Vorsitzender*), *ndihmësmjek* (dt. *Arzthelfer*).

– Adjektive:

  Die Präfixoide der Substantive passen auch zu den Adjektiven, z. B. kundër-, para-, prapa-, jashtë-, sipër-, gjithë-, krye-, gjysëm-/gjysmë-, shumë- usw.

  Beispiele: *i␣ligjshëm* (dt. *gesetzlicher*) + *kundër* → *i␣kundërligjshëm* (dt. *gesetzwidrig*), *gjuhësor* (dt. *sprachlich*) + *jashtë-* → *jashtëgjuhësor* (dt. *außersprachlich*).

– Verben:

  vetë-, bashkë-, tej-, para-, mirë-, keq- usw.

  Einige Beispiele: *bashkëpunoj* (dt. *zusammenarbeiten*), *parapëlqej* (dt. *bevorzugen*), *keqkuptoj* (dt. *missverstehen*).

• Exogene Präfixoide (alban. *Siparashtesat me prejardhje të huaj*):

– Substantive:

  agro-, auto-, bi-, bio-, centi-, eko-, epi-, filo-, gjeo-, hekto-, hidro-, kilo-, kino-, makro-, mikro-, mili-, mega-, mono-, moto-, neo-, neuro-,

poli-, pseudo-, psiko-, tele- und termo-. Dazu einige Beispiele: *bio+masë* (dt. *Biomasse*), *hidro+teknikë* (dt. *Hydrotechnik*) und *makro+analizë* (dt. *Makroanalyse*).

– Verben:

  Die Zahl der exogenen Präfixoide ist im Albanischen klein. [Buxheli 2008: 352 f.] gibt als exogene Präfixoide im Albanischen nur kara- (türkischer Herkunft), kollo- (griechischer/slawischer Herkunft) und kon- (lateinischer Herkunft) an. Sie sind aber sehr schwach oder gar nicht produktiv, vgl. *kollongjis* (dt. *anheften, anhängen, ...*). In Presse-Texten lassen sich Neologismen wie anti-, auto-, kon-, super- finden, jedoch in erster Linie als Partizipformen. Dazu einige Beispiele: *anti+votojnë* (dt. in Kontexten wie *sie wählen x ab, sie stimmen dagegen ab ...*), *auto+censurohet* (dt. *er/sie übt über sich selbst eine Zensur aus*), *super+informuar* (dt. *superinformiert*) und *super+furnizuar* (dt. *superbeliefert, überversorgt*). Die letzten zwei Beispiele sind Partizipformen, die zusammen mit einem vorangestellten Artikel ein Adjektiv bilden.

Formal: [PREF [LEX flex]]

Präfixe können vor einem Stamm mehr als einmal vorkommen. Belege mit bis zu drei Präfixen wären z. B. *ri|sh|për|ndarja*; *ri|sh|për|bëj*; *ri|sh|për|dredh*; *stër|sh|për|dor*.[163] [Fjalori 1980] bietet als Beispiele die Einträge *rish-përndaj* (Verb), d. i. *ri|sh|për|ndaj* und *rishpërndarja* (Substantiv), d. i. *ri|sh|për|ndarja*; *ndaj* und *ndarja* sind jeweils die Stämme. In [Fjalori 2006] wurden zusätzlich noch die Formen *rishpërndahem*, *rishpërndahet*, welche die reflexive bzw. passive Form des Verbs darstellen, aufgenommen.
Bei einigen Präfixen ist die phonetische Umgebung entscheidend dafür, wie sie kombiniert werden. Die Präfixe ç, s, sh, z und zh tragen (zum Teil nur historisch erklärbar) ähnliche oder gleiche semantische Informationen, vor allem die der Verneinung, Inversion und Intensivierung. Es ergeben sich folgende Gruppierungen: Das Präfix ç kann mit Lexemen kombiniert werden, die mit den Buchstaben (Lauten) {a, e, ë, i, o, u, y und j, l, ll, m, n, nj, r, rr}[164] anfangen. Die weiteren Präfixe verhalten sich wie folgt: s kann mit Lexemen, die mit den Buchstaben (Lauten) {f, k, l, m, n und p, q} anfangen, kombiniert

[163] Gefunden in [Fjalori 1980] und in [Fjalori 2006].

[164] [Drejtshkrimi 1974: 46 f. (§ 23)]. Ausnahme sind die Wortbildungen mit çdo/ç i. S. v. *çdokush*, *çfarë*, *çka*, *diçka*, sowie *shmang*, *shlyej*, *shndërroj*, *shndrit*.

werden, sh entsprechend mit {f, k, p, q, t, th}, z mit {b, d, g, gj, v} und zh mit {b, d, g, gj, v}.[165]

Suffixe

Unter diesem Begriff werden sowohl Suffixe als auch Suffixoide verstanden, vgl. hierzu [Hysa 2004]. Sie werden jeweils in spracheigene (indigene) und fremde (exogene) unterteilt.

- Suffixe (alban. *Prapashtesat*):
  - Substantive:

    -abiq, -ac, -acak, -acan, -acuk,-aj, -ajë, -ak, -al, -alaq, -alec, -alkë, -aluq, -am, -amak, -aman, -an, -anik, -anjos, -aq, -ar, -arak, -are, -ari, -as, -ash, -ashkë, -at, -atar, -atë, -atore, -avac, -azan, -cak, -së, -ç, -çkë, -e, -ë, -ec, -ëni, -ëri, -ës, -esë, -esh, -eshë, -ësi, -ësinë, -ësirë, -ësor, -ez, -ëzi, -i, -ian, -icë, -içkë, -im, -imë, -inë, -iot, -ishte, -ishtë, -it, -jan, -jat, -je, -jot, -kë, -li, -llëk, -lli, -m, -man, -më, -ni, -nik, -njot, -ojs, -ok, -onjës, -or, -ore, -osh, -oshe, -ot, -qar, -ri, -s, -shtë, -shti, -si, -sinë, -sirë, -sor, -tar, -tari, -th, -tinë, -tirë, -tor, -tr, -uc, -ucë, -ue, -ues, -uk, -uqe, -urina, -ush, -ushe, -ushë, -ushkë, -vit, -xhi, -zë und -zi. Einige Beispiele: *ëmbël+sirë* (dt. *Süßigkeit*), *drejt+ësi* (dt. *Gerechtigkeit*) *lug+inë* (dt. *Tal*) und *qytet+ar* (dt. *Bürger*).
  - Adjektive[166]:

    -ak, -anik, -aq, -ar, -ark, -ës, -ët, -ëtë, -m, -ëm, -ëmë, -më, -osh, -tar, -atar, -ëtar, -or, -s, -shëm, -tor, -t, -të, -të. Einige Beispiele: *i␣+lëng+ët* (dt. *flüssig*), *i␣+ar+të* (dt. *golden*) und *i␣+afër+m* (dt. *nahestehend, verwandt u. ä.*).
  - Verben[167]:

    -o, -to, -zo, -ro, -so, -s, -is, und -os. Dazu einige Beispiele: *nder+o+j* (dt. *ehren*), *arsye+to+j* (dt. *begründen, erklären*), *kufi+zo+j* (dt. *begrenzen, beschränken u. ä.*), *flakë+ro+j* (dt. *lodern*), *ngadalë+so+j* (dt. *verlangsamen, verzögern*), *përsëri+s* (dt. *wiederholen*) und *vend+os* (dt. *ablegen, ansiedeln, u. ä.*).

[165] [Drejtshkrimi 1974: 47 f. (§ 24)]. Ausnahmen bilden hier die Verben *zmadhoj* und *zmbraps* samt ihren möglichen Wortbildungen.
[166] Vgl. [Çeliku et al. 1998: 335 ff. (§§ 313–324)].
[167] Vgl. [Çeliku et al. 2011: 349–353. (§§ 325–331)].

– Adverbien[168]:
-isht, -as, -thi, und -çe. Einige Beispiele: *fillim+isht* (dt. *anfänglich*), *fal+as* (dt. *umsonst, geschenkt* u. ä.), *kalim+thi* (dt. *beiläufig, auf der Durchfahrt, flüchtig u. ä.*) und *qen+çe* (dt. *mit großer Anstrengung* u. ä.).

- Exogene Suffixe (alban. *Prapashtesat e huaja*):
  – Substantive:
  -acion, -ant, -at, -aturë, -azh, -cion, -çi -er, -ier, -ikë, -ist, -itet, -izëm, -llëk, -urë, und -xhi. Einige Beispiele: *form+acion* (dt. *Formation*), *muzik[·ë]+ant* (dt. *Musikant*), *krahinor+izëm* (dt. *Provinzialismus*), *bank[·ë]ier* (dt. *Bankier*), *bosh+llëk* (dt. *Leere*) und *kandidat+urë* (dt. *Kandidatur*).
  – Adjektive:
  -al, -ian, -ik, -iv und -oz. Einige Beispiele: *muze+al* (dt. *museal*), *laç+ian* (dt. *laçisch* (die Stadt Laç betreffend)), *metod[·ë]ik* (dt. *methodisch*), *impuls+iv* (dt. *impulsiv*) und *vitamin+oz* (dt. *vitaminreich*).

  Verben und Adverbien zeigen keine Tendenz, sich mit fremden Suffixen zu verbinden.

- Suffixoide (alban. *Siprapashtesat*), exogene und indigene zusammen:
  – Substantive:
  -graf, -grafi, -log, -logji und -fil. Einige Beispiele: *leksik+o+graf* (dt. *Lexikograph*), *leksik+o+grafi* (dt. *Lexikographie*), *bio+log* (dt. *Biologe*), *bio+logji* (dt. *Biologie*), *biblio+fil* (dt. *bibliophil*, Adj.).

  Wie bei den exogenen Suffixen zeigen Verben und Adverbien, in diesem Fall aber auch Adjektive, keine Tendenz, sich mit exogenen Suffixoiden zu verbinden.

Formal: [[LEX SUFF] flex]

„Infixe“ und innere Flexion

Als Infixe können im Albanischen die Formen der klitischen Pronomina (Objektszeichen) betrachtet werden, die in den Imperativformen, im Plural,

[168] Vgl. [Çeliku et al. 2011: 353–355. (§§ 332–334)].

bei Verben vorkommen, wie z. B. *lërmani*: *lër*$_{V}$|ma$_{kP.}$|*ni*$_{pl.}$ (dt. *lasst es mir*). Hervorzuheben ist, dass in diesem Zusammenhang in bestimmten Fällen auch unterschiedliche Stämme des Verbs verwendet werden können, vgl. z. B. das Verb *lënë* (dt. *lassen/verlassen*) (1) ohne kl. Pronomina *lër* (sg. impv.) / *li*|*ni* (pl. impv.); (2) mit klitischen Pronomina *lëreni* [169] vs. *lërmani*. Es handelt sich um eine geschlossene Klasse, d. h. eine kleine Menge von Wörtern. Eine besondere Erscheinung sind die Formen cil*i*do / cil*in*do / cil*it*do; cil*a*do / cil*ën*do / cil*ës*do; cil*ët*do / cil*ëve*do; cil*at*do / cil*ave*do.[170] Man könnte von Binnenflexion sprechen.

Die Grammatiken des Albanischen vermeiden aus gutem Grund den Begriff Infix. In der Tat ist diese Bezeichnung mit vielen Fragen verbunden, denn die genannten Wortbildungsmittel haben im Vergleich zu „gewöhnlichen" Affixen/Infixen zum Teil verschiedene Funktionen.

Wenngleich die klitischen Pronomina eine geschlossene Wortklasse darstellen, können sie bei einer unbeschränkten Zahl von Verben vorkommen. Dabei kann aber nach den klitischen Pronomen nur das Plural-Morphem *-ni* stehen. Bei dem Typ cil*i*do wird der Teil *cili* (Pronomen) flektiert, während der Teil *do*, der nach der Flexionsmarkierung unmittelbar vorkommt, nicht flektiert wird.

Formal: [[LEX1- -fLEX1] LEX2]

Lexeme und Morpheme

Die genannten Wortbildungsmittel sind größtenteils solche, die semantische Informationen tragen (*mbi*/*mbi-* (dt. *über*)), ein Teil von ihnen markiert jedoch auch nur grammatische Merkmale wie z. B. Plural (*-ë*, mask. [*punëtor* sg. / *punëtorë* pl.]) oder Genus (*-e*, fem. [*punëtor* mask. / *punëtore* fem.]).

Die Morpheme kommen im Albanischen (1) frei, (2) gebunden oder (3) sowohl frei als auch gebunden vor. Das Morphem *nuk* (dt. *nicht*/*nein*), kommt nur frei vor. Ein Morphem, das nur gebunden vorkommt, ist *-mbë-*. Es ist Bestandteil der Benennung der Zahlen von 11 (*njëmbëdhjetë*) bis 19 (*nëntëmbëdhjetë*). Andere Morpheme wie z. B. *mbi*/*mbi-* und *e*/*-e* kommen sowohl frei als auch gebunden vor, je nachdem welche Funktion sie in einem Kontext übernehmen. Unabhängig von dieser Unterteilung können

[169] Es wird der Stamm *lër* für Pluralformen des Verbs verwendet, wenn klitische Pronomina beteiligt sind. Weitere Formen wären *lëri**ni**, lëria**ni**, ...*, vgl. [Munishi 1998].

[170] Diese Formen entsprechen dem Typ *derjenige, denjenigen* usw. im Deutschen.

Morpheme verschiedene Rollen füllen, z. B. kann *e* als freies Morphem u. a. ein Artikel, ein klitisches Pronomen oder eine Konjunktion sein.[171]

3.4.2 Derivation

Derivation oder *Ableitung* (alban. *prejardhja*) ist der Wortbildungstyp, der im Albanischen, am häufigsten vorkommt. Affixe, d. h. Präfixe, Suffixe und Zirkumfixe (d. h. Prä- und Suffixe), sind die Mittel der Derivation. Im Folgenden wird auf die einzelnen Typen der Derivation eingegangen:

- Präfix-Derivation (alban. *prejardhja parashtesore*). Als Beispiele können die folgenden Wortbildungen dienen: *për|jetoj* (dt. *erleben*) und *stër|gjysh* (dt. *Urgroßvater*), vgl. hierzu [Morfologjia 1995: 61–63 (§ 2.1.11.a)]. Dieser Prozess bewirkt in der Regel keinen Wortartwechsel, sondern nur eine andere Bedeutung des entstandenen Wortes gegenüber dem ursprünglichen Wort/Lexem.

- Suffix-Derivation (alban. *prejardhja prapashtesore*). Als Beispiele können die folgenden Wortbildungen dienen: *letër|si* (dt. *Literaturwissenschaft*) *vul|os* (dt. *stempeln*) und *afr|oj* (dt. *näher bringen*). Dieser Prozess bewirkt in der Regel einen Wortartwechsel. Vgl. hierzu [Morfologjia 1995: 63–67 (§ 2.1.11.b)] und [Ressuli 1985: 138–168 (§ 253)].

- Zirkumfix-Derivation (d. h. gleichzeitige Prä- und Suffix-Derivation) (alban. *prejardhja parashteso-prapashtesore*). Als Beispiele können die folgenden Wortbildungen dienen: *pa|fund|ësi* und *në|punë|s*. Vgl. hierzu [Morfologjia 1995: 67 f. (§ 2.1.11.c)]. Zirkumfix-Derivation ist auch in der Gegenwartssprache produktiv, denn mit ihrer Hilfe werden sowohl okkasionelle Wörter gebildet, etwa in einem Gespräch, als auch in der „geplanten Sprache", etwa neue Termini in der Fachliteratur. Es folgen Beispiele in ausführlicher Darstellung:

  - *përfundoj* (dt. *beenden*) ← për+*fund*+oj
    vs. *[*për* + ~~*fundoj*~~] / *[~~*përfund*~~ + *oj*];
    Muster: për+*N*+oj;
  - *përfundim* (dt. *Abschluss*) ← për+*fund*+im
    vs. *[*për* + ~~*fundim*~~] / *[~~*përfund*~~ + *im*];
    Muster: për+*N*+im;[172]

[171] Vgl. [Xhuvani/Çabej 1962] und [Xhuvani/Çabej 1975].
[172] Die Formen *përfundimisht* und *përfundimtar/e* sind aus *përfundim* abgeleitet.

– *shfletoj* (dt. (*durch*)*blättern*) ← sh+*flet[ë]*+oj
vs. *[*sh* + ~~*fletoj*~~] / *[~~*shflet[ë]*~~ + *oj*];[173]
Muster: sh+*N*+oj;

– *zbukuroj* (dt. *verschönern*) ← z+*bukur*+oj
vs. *[z + ~~*bukoroj*~~] / *[~~*zbukur*~~ + *oj*];
Muster: z+*Adv*/*Adj*+oj.

Die durchgestrichenen Segmente kommen nicht als Wörter/Wortformen vor, was darauf hindeutet, dass die Präfixe und die Suffixe gleichzeitig in einem einzigen Schritt eingesetzt werden. Die aufgelisteten Beispiele sollten nicht mit den Typen *përjetoj* oder *përpunoj* verwechselt werden, die nicht durch Zirkumfigierung gebildet sind, sondern durch mehrstufige Worbildungsprozesse, vgl. *përpunoj* ← për+*punoj* (dt. *verarbeiten*). Vgl. [Çeliku et al. 2011: 355–358 (§§ 335–346)];

Formal: [[CF+ [LEX] +CF] flex]

- Null-Derivation (Konversion) (alban. *prejardhja pa ndajshtesë*). Als Beispiele können die folgenden Wortbildungen dienen:
  – *hap* (dt. *öffnen* [Verb]; *Schritt* [Subst.]);
  hap $_{[V \rightarrow S]}$: Verb → Substantiv;
  – *krip*ë (dt. *Salz* [Subst.]; *salzen* [Verb]);
  krip $_{[S \rightarrow V]: \rightarrow 0}$: Substantiv → Verb;

  Bei diesem Prozess findet ein Wortartwechsel statt, vgl. hierzu [Morfologjia 1995: 68 (§ 2.1.11.ç)] und [Ressuli 1985: 137 f. (§§ 248–252)]. [Morfologjia 1995: 58–61 (§ 2.1.10)] ordnet den Typ *Konversion* im gleichen Rang mit *Derivation*, *Komposition*, *Wortbildung durch Hinzufügen von vorangestelltem Artikel*, *Zusammensetzungen*, und *gemischten Modellen der Wortbildung* ein. Der Unterschied zwischen Null-Derivation [68 (§ 2.1.11.ç)] und Konversion [77 f. (§ 2.1.15)] ist schwer zu erkennen.

- Rückbildung (alban. *prejardhja prapavajtëse*). Als Beispiel kann die folgende Wortbildung dienen: *ftoh* ← *i*␣*ftohtë* (dt. *kalt machen*,

[173] Genauso funktioniert es mit dem Suffix -im bzw. mit Zirkumfix sh- … -im.

erkalten ← *kalter* ...), wobei ein Wortartwechsel stattfindet, von Adjektiv zu Verb, vgl. hierzu [Morfologjia 1995: 67 (§ 2.1.11.b)].[174]

Im Folgenden wird auf die Derivation der Substantive, Adjektive, Verben und der anderen Wortarten im Einzelnen kurz eingegangen.

Substantive

Eine nicht geringe Zahl der Substantive im Albanischen ist aus Verben und Adjektiven abgeleitet. Als Beispiel könnte hier das Substantiv *lexim* (dt. *Lesen*) angebracht werden, das als Ableitung aus dem Verb *lexoj* (dt. *lesen*) gebildet wurde – ebenso *ecje* (dt. *Laufen*) aus *eci* (dt. *laufen*). Im ersten Beispiel wird aus *lexo-* durch Verlust von *-o-* und das Anhängen des Suffix *-im* ein Substantiv, d. h. eine deverbale Substantivierung findet statt.
Ein Beispiel für die Bildung von Substantiven aus Adjektiven wäre *i␣ziu* (dt. *der Schwarze*) ← *i␣zi* (dt. *schwarz*) *+u*, d. h. aus dem Adjektiv *i␣zi* wird durch das Anhängen des Suffix/Morphem *u* ein Substantiv gebildet. Weitere Beispiele sind *i␣ri/+u* (dt. *jung/der Junge*), *i␣lig/+u* (dt. *schwach/der Schwache*), *i␣pasur/+i* (dt. *reich/der Reiche*), *i␣bardh[·ë]/+i* (dt. *weiß/der Weiße*) und *i␣mir[·ë]/+i* (dt. *gut/der Gute*), vgl. [Morfologjia 1995: § 2.1.15 und § 4.6].

Adjektive

Adjektive können aus Substantiven und aus Verben abgeleitet werden. Beispiele für die Ableitung der Adjektive aus Substantiven wären: *fshat+ar* (dt. *Dorf/bäuerlich/dörflich*), *det+ar* (dt. *Meer/das Meer betreffend*), *mal+or* (dt. *Wald/den Wald betreffend*), *mesjet[·ë]+ar* → *mesjetar* (dt. *Mittelalter/-mittelalterlich*) und *muzik[·ë]+al*→ *muzikal* (dt. *Musik/musikalisch*).
Aus den Verben können die Adjektive durch das Anhängen des deverbalen Suffix *-or* abgeleitet werden. Beispiele: *lidh+or*→*lidhor* (dt. *verbindungs-*); *dëft[·oj]+or*→*dëftor* (dt. *hinweisend, zeigend*).[175]

[174] Ein Beispiel, bei dem kein Wortartwechsel stattfindet, ist *zhduk* ← *zhdukem* (← *dukem*) (dt. *vernichten*, ← *ich lasse mich nicht sehen*, ← *ich lasse mich sehen*), Verb, vgl. [Morfologjia 1995: 350].
[175] Vgl. [Çeliku et al. 1998: 337 (§ 315 b)].

Verben

Verben können von Substantiven und Adjektiven mithilfe von desubstantivischen und deadjektivischen Suffixen abgeleitet werden. Aus den Substantiven werden die Verben vor allem durch Anhängen des Suffixes *-o* abgeleitet, wie im folgenden Beispiel: *pun[·ë]+-o→punoj*.[176] Ebenso werden mit dem Suffix *-o* Verben aus Adjektiven abgeleitet, vgl. *i␣holl[·ë]+-o→holloj* (dt. *dünn/verdünnen*).[177] Ausführliche Informationen zu den Suffixen, durch die Verben abgeleitet werden, bietet Buxheli [2008].[178]
Einige Wörter bzw. Wortbildungen lassen sich nur mithilfe historischen Sprachwissens beschreiben, wie bspw. *nguros* (dt. *versteinern*) ← *n-|gur|-os*. Sie werden als Simplizia behandelt, d. h., sie werden nicht segmentiert.[179]

Adverbien

Neben Verben, Substantiven und Adjektiven bzw. Numeralen dienen auch Adverbien zur Bildung von neuen abgeleiteten Wörtern. Es können sowohl Verben als auch Substantive und Adjektive gebildet werden, wie das folgende Beispiel zeigt:[180]

- *afër* (Adv.) (dt. *nah*) → *afr+o* → *afr¦oj* (V) (dt. *nähern*, u. ä.)
- *afër* (Adv.) → *i␣+afër+m* → *i␣afër¦m* (A bzw. S) (dt. *verwandt*, u. ä.)

Irregularitäten bei Derivation

Trotz Ähnlichkeit können sich zwei Wortbildungsprozesse verschieden verhalten, wie in den folgenden zwei Fällen:

- *kujt/-oj/-im/-esë*, d. h. *kujtoj*, *kujtim*, *kujtesë* (dt. *denken/Andenken/Gedächtnis*) und
- *vepr/-oj/-im/*-esë*, d. h. *veproj*, *veprim*, aber nicht **vepresë* (dt. *handeln; wirken/Handlung; Wirkung/**)

Die Formen *veproj* und *veprim* sowie *vepr/-im/-tar/-i* existieren, sind also lexikalisiert, während die mit „*“ markierte Wortform **vepr/esë* nicht existiert

[176] Vgl. [Çeliku et al. 1998: 342 (§ 325 a)].
[177] Vgl. [Çeliku et al. 1998: 343 (§ 325 b)].
[178] Vgl. hierzu u. a. die tabellarische Übersicht auf den Seiten 351–365.
[179] Segmentierung dient in erster Linie dazu, Wissen zu erwerben, um Neologismen zu erkennen. Sie kann aber auch für eine morphologische Analyse eingesetzt werden.
[180] Vgl. [Çeliku et al. 1998: 343 (§ 325 c)].

und keine Bedeutung trägt.[181] Die Wortform *kujtesë* ist jedoch lexikalisiert, trotz der gleichen Struktur, derselben Verbklasse und demselben Suffix.

Abbildung 3.1: der Stamm *pun*

$$
pun_{[.x]}\left\{\begin{array}{l}
+[V]\left\{\begin{array}{ll}
+[V.1][Flex]\left\{\begin{array}{lll} -o_{[cat.a]} & \Rightarrow & pun|o|j \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[V.2][Flex]\left\{\begin{array}{lll} -ua_{[cat.b]} & \Rightarrow & pun|ua|m \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[V.3][Flex]\left\{\begin{array}{lll} -of_{[cat.a]} & \Rightarrow & pun|of|të \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[V.4][Flex]\left\{\begin{array}{lll} -ofsh_{[cat.b]} & \Rightarrow & pun|ofsh|im \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[V.5][Wb]\left\{\begin{array}{lll} -_{[S]}s|o_{[V]} & \Rightarrow & pun|ë|s|o|j\leftarrow[\text{S.2 cat b}] \\ \ldots & \Rightarrow & \ldots \end{array}\right.
\end{array}\right. \\
+[S]\left\{\begin{array}{ll}
+[S.1][Flex]\left\{\begin{array}{lll} -im_{[cat.a]} & \Rightarrow & pun|im|i \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[S.2][Wb]\left\{\begin{array}{lll} pa-\ldots-ë.s.i_{[cat.a]} & \Rightarrow & pa|pun|ës.i|a \\ \ldots-ë.s.i_{[cat.a]} & \Rightarrow & \sout{pun|ës.i|a} \\ & \Rightarrow & drejt|ës.i|a \\ -ë.s.im_{[cat.b]} & \Rightarrow & pun|ës.im|i\rightarrow[V.5] \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[S.3][Wb]\left\{\begin{array}{lll} -.tor_{[cat.b]} & \Rightarrow & pun||tor \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
\ldots
\end{array}\right. \\
+[A]\left\{\begin{array}{ll}
+[A.1][Wb/Flex]\left\{\begin{array}{lll} -ues_{[cat.a]} & \Rightarrow & pun|ues|e \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
+[A.2][Wb/Flex]\left\{\begin{array}{lll} -.tor_{[cat.b]} & \Rightarrow & pun||tor \\ \ldots & \Rightarrow & \ldots \end{array}\right. \\
\ldots
\end{array}\right. \\
\ldots
\end{array}\right.
$$

Abbildung 3.1 zeigt, dass aus dem Stamm *pun* sowohl Verben (V) als auch Substantive (S) und Adjektive (A) gebildet werden können. Ein interessanter Fall ist *papunësi* (dt. *Arbeitslosigkeit*), weil die Form *punësi* nicht existiert,

[181] Man könnte sie anhand der Bedeutung des Stammes interpretieren, sie wäre jedoch auch als okkasionelle Wortbildung höchst seltsam, kaum möglich.

sondern stattdessen *punësim* (dt. *Beschäftigung, Einstellung*). Als Ausgang ist hier das Wort *i␣papunë* (dt. Arbeitsloser) anzunehmen.[182]

3.4.3 Hinzufügen des vorangestellten Artikels

Einige Wortarten bzw. deren Wortformen besitzen die Fähigkeit, in Zusammenschluss mit einem vorangestellten Artikel die Wortart zu wechseln. Verbpartizipien besitzen oft diese Fähigkeit, vgl. hierzu [Morfologjia 1995: 74 f. (§ 2.1.12)] und die folgenden Beispiele:

- $hapur_{PART}$ (dt. *geöffnet*) + $i␣_{PrepART} \rightarrow i␣hapur_{ArtADJ}$ (dt. *geöffnet, offen sein*)
- $të_{PrepART}$ + $ecur_{PART}$+$it_{PostART}$ (dt. *(ge)laufen*) $\rightarrow$ $të␣ecurit_{ArtSUBST}$ (dt. *das Laufen, der Lauf*)

3.4.4 Komposition

Im Vergleich zur Derivation, bei der ein Wort (lexikalisches Morphem) mit Affixen kombiniert wird, geht es bei der Komposition um die Kombination von zwei oder mehreren Wörtern miteinander, sowie je nach Fall auch mit anderen Wortbildungsmitteln, wie Fugenelementen.[183]
Es werden die folgenden Typen der Komposition unterschieden: die Komposita (alban. *fjalët e përbëra*), die Zusammensetzungen (alban. *fjalët e përngjitura*) und die sogenannten Wortgruppen (alban. *lokucione, grupe fjalësh [me vlerën e një fjale që shkruhen ndaras ose me vizë]*).[184]
Die Komposita (alban. *kompozime [përbërje]*) können weiter spezifiziert werden. Es werden folgende Untertypen unterschieden:

- Kopulativ-Komposita (alban. *kompozita këpujore*), wie z. B. *jug|lindje* (dt. *Südosten*); *marrë|dhënie* (dt. *Beziehung* [*nehmen+geben*]) und *vesh|mbathje* (dt. *Kleidung und Schuhwerk*), vgl. hierzu [Morfologjia 1995: 71 f. (§ 2.1.13.a)]. Beide Bestandteile stehen gleichrangig nebeneinander.

[182] Die Wörter *punoj* (dt. *arbeiten*) [←*punë* (dt. *Arbeit*)] und *drejtoj* (dt. *leiten, führen*) [←*drejt* (dt. *Gerade*)], obwohl beide aus der selben Konjugationsklasse, verhalten sich bei der Wortbildung unterschiedlich voneinander, vgl. **punësi* (dt. *Beschäftigung*) vs. *drejtësi* (dt. *Gerechtigkeit*).
[183] Vgl. [Ressuli 1985: §§ 258–267].
[184] Nach [Çeliku et al. 1998: 358–371 (§§ 366–395)].

- Determinativ-Komposita (alban. *kompozita përcaktore*), welche aus zwei Teilen bestehen, wobei der eine den anderen näher bestimmt. Einige Beispiele wären: *krye*|*qytet* (dt. *Hauptstadt*) (Typ 1, der erste Teil bestimmt den zweiten näher) vs. *breg*|*det* (dt. *Küste, Meeresufer*) (Typ 2, der zweite Teil bestimmt den ersten näher), vgl. hierzu [Morfologjia 1995: 72–74. (§ 2.1.13.b)].

- Abkürzungen (alban. *shkurtesat*), wie z. B. ATSH (Agjencia Telegrafike Shqiptare), vgl. hierzu [Morfologjia 1995: 74 f. (§ 2.1.13.c)]. Es werden mehrere Typen unterschieden, solche, die nach Initialen gebildet werden, solche, die nach Silben gebildet werden, sowie irreguläre Typen. Man vergleiche hierzu auch Wortbildungen wie *Tung* (←*Tung*jatjeta) (dt. *ein langes Leben* [Begrüßung]), *Zysha* (←~~Zonj~~-[*y*←*u*]*sha*←Z*onjusha*)[185] (dt. *Fräulein*/*Frau* [Jugend- bzw. Studentensprache: Bezeichnung für eine junge Lehrerin]), *Pafshim* (←Miru*pafshim*) (dt. *Auf Wiedersehen*) und *Natën* (←*Natën* e mirë) (dt. *gute Nacht*).

3.4.5 Zusammenrückungen

Die Zusammenrückungen, alban. *fjalët e përngjitura*, können Substantive, Numerale, Pronomina, Adverbien, Präpositionen, Konjunktionen und Partikel sein. [Morfologjia 1995: 58–61 (§ 2.1.10)] sieht die Zusammensetzung als einen gleichrangigen Typ zur Komposition. In der vorliegenden Arbeit wird die Zusammensetzung als ein Untertyp der Komposition behandelt, wie dies auch [Çeliku et al. 1998: 358 (§ 366)] sieht. Im Folgenden sind die Haupttypen der Zusammensetzungen im Albanischen angegeben:

- Substantive: *thash*|*e*|*theme* (dt. *Geschwafel, Gerüchte*); *gjë*|*e*|*gjëzë* (dt. *Rätsel*);

- Numeralia: *tri*|*dhjetë* (dt. *dreißig*); *dy*|*mbë*|*dhjetë* (dt. *zwölf*);

- Pronomina: *as*|*kush* (dt. *niemand*); *se*|*cili* (dt. *jeder*);

- Adverbien: *as*|*një*|*herë* (dt. *niemals*); *para*|*dite* (dt. *Vormittag*);

[185] *Z* steht für das ursprüngliche Wort. *u* wurde aus phonotaktischen Gründen in *y* umgewandelt. Es handelt sich möglicherweise um Palatalisierung.

3.4.6 Wortgruppen

Den dritten Typ stellen die Wortgruppen dar, die in der albanischen Literatur über Morphologie und Wortbildung Lokucione genannt werden.[186] Somit sind sie einem Wort gleichgesetzt. Sie können als folgende Wortarten vorkommen:[187]

- Substantive: *dita-ditës* (dt. *von Tag zu Tag ∼ mit der Zeit*);
- Numerale: *njëzet␣e␣një* (dt. *einundzwanzig*);
- Pronomina: *njëri-tjetri* (dt. *einander*);
- Verben: *marr␣pjesë* (dt. *teilnehmen*);
- Adverbien: *vende-vende* (dt. *da und dort, an manchen Orten*);
- Präpositionen: *me␣natë* (dt. *noch bei Nacht/sehr früh*);
- Konjunktionen: *kurdo␣që* (dt. *immer wenn*);
- Partikel: *le␣që* (dt. *nicht nur, dass …*);

3.4.7 Besonderheiten der albanischen Rechtschreibung

Bei der Flexion bestimmter Wortformen wie *në␣arrit|të* (dt. *falls er/sie es schafft*); *e␣gjet|të␣e␣mira* (dt. *möge er/sie Glück haben*); *ndrit|të* (dt. *möge er/sie glänzen*); *në␣pyet|të*; (dt. *falls er/sie* (*nach*)*fragt* u. ä.) usw., müssen Grapheme doppelt geschrieben werden, das eine als Teil des Stammes, das andere als Teil des Suffixes. Hier kommt es oft zu Rechtschreibfehlern.
Falls bei einer Komposition einer der Digraphen ll oder rr und l bzw. r in Kontakt miteinander geraten, wird ein Zeichen getilgt, wie im folgenden Beispiel: *për+rreth→për|reth* (dt. *rundherum*), vgl. [Drejtshkrimi 1974 (§ 35 Shën. 1)].[188]
Fugenelemente (alban. *Fonema mbështetëse*)[189] kommen sowohl bei der Derivation als auch bei der Komposition vor, wie z. B. das ë in *komb||ë|tar*

[186] Der Terminus *Lokution*, engl. *Locution* bzw. *Locutionary act*, verwendet von J. L. Austin [1962, *How to do things with words*, Oxford] und J. R. Searle [1969, *Speech acts. An essay in the philosophy of language*, Cambridge] ist hier nicht gemeint.
[187] Die folgenden Beispiele sind aus [Çeliku et al. 2011: 373–379 (§ 388–394)] entnommen.
[188] Vgl. hierzu auch [Kostallari 1984: 169–225. (P III), *Çështje të formimit të fjalëve*];
[189] Nach [Morfologjia 1995].

(dt. *national*), *mal||ë|sor* (dt. *Gebirgsbewohner*), *lavd||ë|roj* (dt. *loben*), *miq||ë|si*) (dt. *Freundschaft*).[190]
Das Verhalten der Fugenelemente lässt sich kaum regelbasiert beschreiben, da sie vor allem phonetisch motiviert sind. Dazu kommen noch historische und regionale Einflüsse. Aus diesen Gründen ist eine Formalisierung dieses Prozesses keine leichte Aufgabe, beinahe unmöglich.

Formal: [LEX-1 [{ FE } LEX-2 / SUFF] flex]

Ein ganz anderer Fall liegt vor, wenn beim Kontakt von *ë* mit einem Vokal das *ë* wegfällt. Ein Beispiel wäre: *bashkautor* ← bashkë+autor (Nomen) (dt. *Koautor*); *i␣gjithanshëm* ← gjithë+anshëm ← gjithë+anë+shëm (Adjektiv) (dt. *allseitig*), d. h. [...ë+a... → ...a...] oder *kokulur* ← kokë+ulur (Adjektiv) (dt. *mit gesenktem Kopf*) [...ë+u... → ...u...].
Nicht immer, wenn im Zuge von Derivation oder Komposition zwei Vokale aufeinandertreffen, fällt der erste der beiden weg. Ausnahmen bilden die folgenden Fälle: *njëanësi* ← një|anësi ← një+anësi (dt. *Einseitigkeit*), *zëëmbël* ← zë|ëmbël ← zë+ëmbël (dt. *mit lieblicher Stimme*).[191] Im Beispiel *i␣dyanshëm* ← i dy+an{ë}+shëm ← i dy+anë+shëm (dt. *zweiseitig*): [...y + a... → ...ya...; ...ë + -shëm → -shëm] liegen zwei verschiedene Phänomene vor, der erhaltene Kontakt ...y+a... sowie der Ausfall von ë. Beim Aufeinandertreffen derselben Buchstaben (Konsonanten oder Vokalen), wie zum Beispiel gojë+ëmbël → *gojëmbël* (dt. *freundlich, flötend*), vgl. [Drejtshkrimi 1974: 22 (§ 5 b shën. 2)], spielt die Betonung die entscheidende Rolle. Im Gegensatz zu *zë* (*zə́*), wo es eindeutig ist, wo die Betonung liegt, nämlich auf dem *ë*, ist beim zweisilbigen Wort *gojë* (*gó·jë*) das *o* betont. Aus diesem Grund fällt bei der Komposition das *ë* in der zweiten Silbe aus.
Bei dem Suffix *-shëm*, das bei der Bildung der Adjektive sehr häufig vorkommt, zeigt sich das gleiche Verhalten: a) i␣për+kohë+shëm → *i␣përkohshëm* (dt. *vorübergehend*), i␣për+ditë+shëm → *i␣përditshëm* (dt. *täglich*), aber i␣zë+shëm → *i␣zëshëm* (dt. *laut, betont*), falls die Betonung auf den Vokal fällt; b) *paanësi* ← pa|anësi ← pa+anësi (dt. *unabhängig von einer Seite*): [...a+a... → ...aa...], krye|engjëll (dt. [der wichtigste] *Engel*), anti|imperialist, jo|organik (dt. *anorganisch*) usw., vgl. [Drejtshkrimi 1974: 41 (§ 18)]. So kann als Regel formuliert werden: Wenn eine Wortform, die auf einen betonten Vokal (Silbe) endet, mit einer anderen Wortform (Komposition) bzw. einem

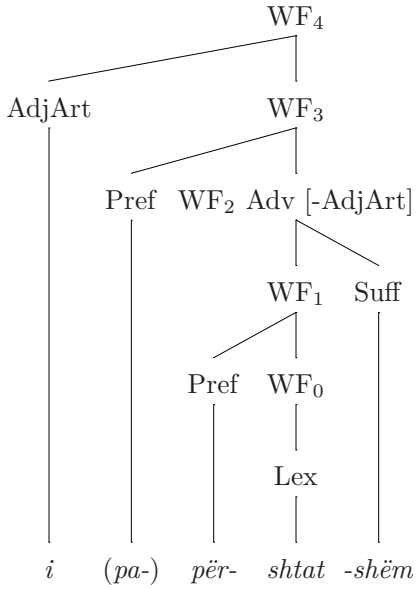
[190] Vgl. hierzu für weitere Beispiele [Memushaj 2004: 129 f.].
[191] Die erste Komponente des Kompositums, also , bzw. *zë*, ist dabei betont, was entscheidend für die Beibehaltung des *ë* ist.

Suffix (Derivation) kombiniert wird und das zweite Element mit einem Vokal anfängt, so fällt der Vokal am Ende des ersten Teils weg.

3.4.8 Zwei Wortbildungsanalysen

Die folgende Abbildung (3.2) zeigt die Analysen der Wortform *i (pa)përshtatshëm* (dt. *unpassender*):

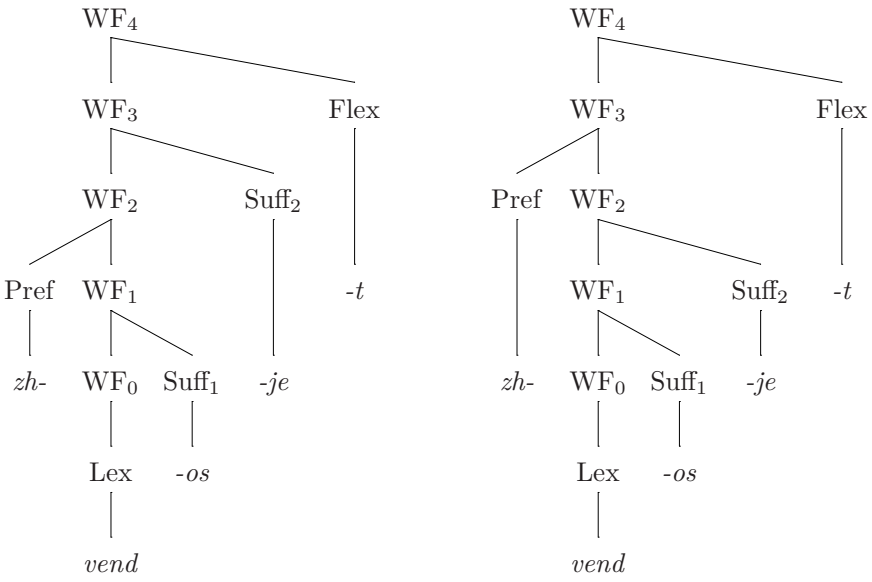
Abbildung 3.2: Morphologische Zerlegung der Wortform *i␣(pa)përshtatshëm*.



In dieser Analyse ist die Flexion nicht abgebildet. Sie käme noch dazu, vgl. *i␣(pa+)për+shtat+shëm* (unbest.) vs. *i␣(pa+)për+shtat+shm+i*$_{Flex}$*;* (best.).
Abbildung (3.3) zeigt zwei verschiedene mögliche Analysen der Wortform *zhvendosjet* (dt. *Versetzungen*). Die beiden Formen unterscheiden sich in einem Punkt, nämlich danach, ob das Präfix *zh-* mit dem Verb kombiniert wird, vgl. Abb. 3.3, links WF_2, oder ob es mit dem Substantiv zusammengesetzt wird, vgl. Abb. 3.3, rechts WF_3, sowie die folgende Darstellung in einer linearen Klammerstruktur:

$$[\,[\,[\text{zh-}_{Prep}\,[\,[\text{vend}_S]\,\text{-os}_V]^{de.S}\,]\,\text{-je}_S]^{de.V}\,\text{-t}_{Flex}] \text{ vs.}$$
$$[\,[\text{zh-}_{Prep}\,[\,[\,[\text{vend}_S]\,\text{-os}_V]^{de.S}\,\text{-je}_S\,]^{de.V}\,]\,\text{-t}_{Flex}];$$

Abbildung 3.3: Morphologische Zerlegung der Wortform *zhvendosjet*.

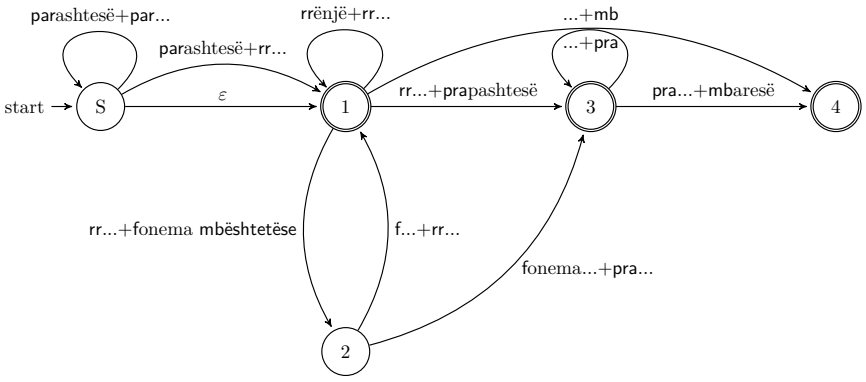


Eine formalisierte Übersicht

Abbildung 3.4 zeigt einen FSA, der grob eine Übersicht über für die Morphologie und Wortbildung des Albanischen geben könnte. Dabei steht parashtesë bzw. kurz par... für Präfix, rrënjë bzw. rr... für Wurzel, prapashtesë bzw. pra... für Wortbildungssuffix, mbaresë bzw. mb... für Flexionssuffix und schließlich fonema mbështetëse, fonema... bzw. f... für Fugenelement(e).
Es handelt sich hier um einen Automaten, der nur die Wortformen behandelt, die keinen vorangestellten Artikel besitzen oder durch diesen gebildet werden, wie z. B. *i*␣*+mirë* → *i*␣*mirë*. Fälle wie *drejt* (dt. *gerade*) → *i*␣*drejtë* (dt. *gerade*, *gerechter*) → *drejtësi* (dt. *Gerechtigkeit*) → *padrejtësi* (dt. *Ungerechtigkeit*) → *padrejtësisht* (dt. *un(ge)rechtlich*) sind nicht berücksichtigt.[192]

[192] Beispiel entnommen aus [Morfologjia 1995: 50].

Abbildung 3.4: Ein FSA für die Morphologie des Albanischen



3.5 Zusammenfassung des 3. Kapitels und Schlussbemerkungen

Es kann abschließend festgestellt werden, dass alle (vier) „Grundprobleme der morphologischen Beschreibung“ nach [Trommer 2010: 239 f.] auch in der albanischen Morphologie vorkommen, nämlich Neutralisierung, Nichtkonkatenativität, Regularitäten und Ausnahmen sowie Allomorphie und Phonologie. Man vergleiche die folgenden Beispiele:

Neutralisierung: *një*, sg. / *disa*, pl. *ditë* (dt. *ein Tag* / *einige Tage*) und *një*, sg. / *disa*, pl. *orë* (dt. *eine Stunde* / *einige Stunden*).

Nichtkonkatenativität: *një krap*, sg. / *disa krep*, pl. (dt. *ein Karpfen* / *einige Karpfen*) und *marr*, 1. Pers. Sg. Ind. Akt. / *merr*, 2./3. Pers. Sg. Ind. Akt.; 2. Pers. Sg. Impv. Akt. (ohne klitische Pronomina) (dt. *nehme*/*nimmst*/*nimm*).

Regularitäten und Ausnahmen: Hier können viele Beispiele vorgeführt werden. Es sei nur eines angebracht: *një natë*, sg. / *disa net*, pl.; Pluralbildung durch Tilgung von Vokalen bildet eine Ausnahme.

Allomorphie und Phonologie: Hier können ebenso viele Fälle festgestellt werden, etwa in den folgenden Beispielen: *një rreth*, sg. / *disa rrathë*, pl. (dt. *ein Kreis* / *einige Kreise*); *një derë*, sg. / *disa dyer*, pl. dt. *eine Tür* / *einige Türen*; und *një dash*, sg. / *disa desh*, pl. (dt. *ein Hammel* / *einige Hammel*).

Die Wortbildung erweist sich als recht kompliziertes Thema, denn es fehlen Modelle der Wortbildungstypen, wie sie für das Deutsche bereits vorhanden sind, etwa bei [Celex 1994].

4 Die lexikalischen Daten

Das Lexikon ist eine der wichtigsten Komponenten eines Systems für maschinelle Sprachverarbeitung. Dafydd Gibbon [2010: 515] drückt dies wie folgt aus: „[D]er unverzichtbare Kern jedes Sprachverarbeitungssystems [...] ist ein Lexikon."
Auch im Rahmen der vorliegenden Arbeit stellt das Lexikon die Grundlage für die Entwicklung einer Komponente für maschinelle morphologische Sprachverarbeitung dar, sowohl für die Analyse als auch für die Produktion. So macht die Sammlung und Aufbereitung von lexikalischen Daten einen erheblichen Teil des Ganzen aus und ist unverzichtbar, um eine hochwertige Verarbeitung zu ermöglichen.
Im ersten Abschnitt 4.1 wird auf die Definition eines Wörterbuches im Allgemeinen sowie im Rahmen der maschinellen Sprachverarbeitung eingegangen. In den darauffolgenden Abschnitten 4.2 bis 4.12 wird auf die Details im Einzelnen eingegangen. Es werden die für jede Wortart (Substantive, Adjektive, Numerale, Pronomina, Verben, Adverbien, Konjunktionen, Präpositionen, Partikeln und Interjektionen) spezifischen Einträge sowie die Sonderfälle (Interpunktionszeichen und Wortbildungsmittel) beschrieben. Ziel ist es, die lexikalischen Daten zu organisieren, um sie später im Rahmen von xfst leichter handhaben zu können.

4.1 Definition

Eine Definition für ein Wörterbuch gibt Hausmann [1985: 369] wie folgt[193]:

> Das Wörterbuch ist eine durch ein bestimmtes Medium präsentierte Sammlung von lexikalischen Einheiten (vor allem Wörtern), zu denen für einen bestimmten Benutzer bestimmte Informationen gegeben werden, die so geordnet sein müssen, daß ein rascher Zugang zur Einzelinformation möglich ist.

[193] Es handelt sich hier um eine Definition eines traditionellen Wörterbuches. Diese Eigenschaften passen genauso auch für maschinenlesbare Wörterbücher (MlW), wofür, beeinflusst vom Englischen, die Bezeichnung Lexikon in der MSV gebräuchlich geworden ist.

Um diese Kriterien zu erfüllen, welche sehr gut auch für die Zwecke der MSV zutreffen, bedarf es:

1. einer guten Organisation und Strukturierung des lexikalischen Materials, so dass es eindeutig angesprochen bzw. auf dieses zugegriffen werden kann *und*

2. einer Möglichkeit das lexikalische Material zu manipulieren, z. B. korrigieren und erweitern

Die anderen zitierten Kriterien, wie Medium und Benutzer, sind vorgegeben. Ein Lexikon, insbesondere sein Format, hängt sehr stark vom System ab, in das es integriert werden soll. Formate in standardisierten Auszeichnungssprachen, wie etwa XML, sind von Vorteil.[194] Dadurch wären die Ressourcen leichter nutzbar.[195]
Im Rahmen der vorliegenden Arbeit werden die lexikalischen Informationen als Erstes in einem freien Format organisiert, um eine einheitliche Struktur der Lemmata zu schaffen. Dieser Schritt ist notwendig, da diese Informationen bei der Erhebung der Daten, etwa aus Texten/Korpora bzw. in ihren ursprünglichen Quellen, nicht einheitlich sind und oft Lücken aufweisen. In einem zweiten Schritt ist es mithilfe der neu gewonnenen Struktur möglich, sie im Rahmen einer Entwicklungsumgebung sowie eines Grammatiksystems ohne Schwierigkeiten einzusetzen.
Als Nächstes wird auf die Eigenschaften der Lemmata einzelner Wortarten eingegangen. Neben deren Struktur werden auch einzelne Besonderheiten beschrieben, die für die MSV, insbesondere für die morphologische Verarbeitung, von Bedeutung sind.

4.2 Verb-Einträge

Die Lexikoneinträge für Verben finden sich in gedruckten Wörterbüchern, etwa in Rechtschreibwörterbüchern, vgl. [FDSH 1976], sowie in einem Teil der grammatischen Angaben in Bedeutungswörterbüchern, vgl. [FJALORI 2006], wie in dem folgenden Eintrag dargestellt:

mend:ój *fol.* ∼óva ∼úar (Lex-V-trad-1)

[194] Ein Vorschlag für die Modellierung der lexikalischen Ressourcen im Albanischen findet sich in [KABASHI 2006].
[195] Vgl. bspw. [LOBIN 2001].

Es handelt sich um das Verb *mendoj*, zu dt. *denken*. Im Albanischen werden die Einträge für Verben entweder in der ersten oder in der dritten Person angegeben, da eine Infinitivform (per definitionem) fehlt. In diesem konkreten Fall handelt es sich um

mend:*ój* (*mendoj*), die Präsens-Form, genauer die 1. P. Sg. Präs. Ind. Akt. Nichtadm., d. h. *ich denke*,

mend:*óva* (*mendova*), die Aorist-Form, genauer die 1. P. Sg. Präs. Aor. Akt. Nichtadm. und

mend:*úar* (*menduar*), die Partizip-Form des Verbs.

Falls ein Verb die 1. und die 2. Person nicht besitzt, wird die 3. statt der 1. Person angegeben. Die Akzentzeichen in den Einträgen werden als zusätzliche Notation für die Betonung der Vokale verwendet, denn das albanische Alphabet, vgl. Abschnitt 3.2.2, besitzt keine Zeichen mit Akzenten.
Im Vergleich zu (Lex-V-trad-1), vgl. FDSH [1976: 382], geben einige Wörterbücher, wie z. B. [FJALORI 2006] ausführlichere Beschreibungen der lexikalischen Einträge. Der besprochene Eintrag sieht in [FJALORI 2006] folgendermaßen aus: *mend/ój jokal., –óva, –úar*. Dabei bedeutet *jokal.* intransitives Verb, also in der gängigen Notation *V. intr.*, während die anderen Angaben – wie oben – die Beugung des Verbs in Aorist (*mend/óva*) und in seiner Partizip-Form (*mend/úar*) markieren.

Das Zeichen „:“ markiert die Stelle, ab der die Beugung des Verb vorkommt. Es wird meistens mit dem Zeichen „/“ angegeben. Das Zeichen „∼“ steht vor dem Teil des Verbs, der sich ändert. SNOJ [1994] nennt zu dem Eintrag die grammatischen Informationen in der Form *V. intr., tr.*, wobei die Wortart angegeben wird. Die Information *V. intr., tr.* hat gegenüber *fol.*, was gleich *V.* ist, den Vorteil, dass die Valenz des Verbs zusätzlich beschrieben wird, wenn auch nur allgemein. Diese Angaben sind für maschinelle Datenverarbeitung im Rahmen der Syntax von großer Bedeutung.
Die traditionellen Formen dienten als Ausgangspunkt für die Aufstellung des Lexikons. Sie wurden in einem nächsten Schritt „normalisiert“, sodass sie wohlgeformt sind, um für die maschinelle Sprachverarbeitung eingesetzt zu werden. Ein solcher Eintrag, der als ein CSV-ähnliches Format strukturiert wurde, d. h. *character-separated values*, mit *n* Feldern und ‚|‘ als Trennzeichen, sieht aus wie folgt:

```
mendoj|5|2|mend/ój|V.|intr.,tr.|∼óva|∼úar|||V-010
```
(Lex-V-mod-1)

Dabei bedeuten die Angaben, d. h. die Felder, die linguistischen Daten, und die weiteren Informationen folgendes:

`mendoj` In diesem Feld steht das Verb in der 1. Person Singular Indikativ Präsens Aktiv Nicht-Admirativ. Eine Ausnahme machen hier einige Verben, die nicht in der 1. und 2. Person vorkommen, wie *vetëtin* (dt. *es blitzt*). Diese Form entspricht der Nennform und ebenso ihre Rechtschreibung (insbesondere ohne Akzente und Segmentierungen). Dieses Feld eignet sich sehr gut als Schlüssel für eine eventuelle Speicherung der lexikalischen Daten in einer Datenbank.

`5` Die Zahl in diesem Feld gibt die Position des Zeichens „`/`“ an (s. u. das Feld `mend/ój|`), aufgezählt von links nach rechts.

`2` Die Zahl in diesem Feld gibt die Position des Buchstabens (Vokals) an, auf den der Akzent fällt (s. u. das Feld `mend/ój|`), wobei hier im Gegensatz zum vorigen Feld von rechts nach links gezählt wird.

`mend/ój` Der Eintrag in diesem Feld stellt die Information dar, die in den ersten drei Feldern angegeben ist. Die Segmentierung, ausgezeichnet mit „`/`“ stellt die wichtigste Information dar, insbesondere in Zusammenhang mit den Feldern `~óva|` und `~úar`.

`V.` In diesem Feld ist die Wortart angegeben. `V.` steht gewöhnlich für Verb. Durch diesen Schlüssel ist es möglich, die Einträge des Lexikons nach Wortarten zu klassifizieren.

`intr.,tr.` Die Wortart wird nach bestimmten Eigenschaften spezifiziert. In diesem Fall handelt es sich um ein sowohl intransitiv (`intr.`) als auch transitiv (`tr.`) verwendbares Verb. In diesem Feld können auch Angaben wie `refl.`, für Reflexiv, oder `pass.`, für Passiv, vorkommen.

`~óva` Diese Angabe, zusammengesetzt mit dem Teil vor dem Zeichen „`/`“ stellt die Aorist-Form des Verbs dar.

`~úar` In gleicher Form wie im vorigen Feld wird die Information für die Partizip-Form des Verbs gezeigt.

`<Kein Eintrag im Feld>` In diesem Feld, d. h. im neunten Feld, steht die Information über die Partikel `u` einiger reflexiver Verben in Aorist-Form.

<Kein Eintrag im Feld> Das zehnte Feld wurde für spezielle Markierungen freigelassen.

`V-010` In diesem Feld ist die Verbklasse angegeben. Sie wurde automatisch anhand der vorgestellten Eigenschaften und Informationen erstellt.[196]

Diese Angaben entsprechen denen einer minimalen Darstellung pro Eintrag. Sie haben eine einheitliche, normalisierte Struktur. Somit ist sichergestellt, dass für jede Wortart, in diesem Fall für jeden Verb-Eintrag, eine eindeutige Information in jedem Feld steht. Diese Einträge können problemlos um neue Felder erweitert werden und ggf. systematisch in neue Strukturen umgewandelt werden.
Ein optimaler Eintrag für diese Zwecke könnte folgendermaßen aussehen:[197]

Listing 4.1: Lexikon der Verben

```
[LEX: "mendoj",
 POS: Verb,
 TYP: V-010,
 SEG: <ACN: "mend/ój", ACN-NUM:-2, ALT-SEG:5>,
 GRA: <AOR: "~óva", INF:"~úar">,
 ALM: <"o", "ua", "of", "ofsh">,
 SPE: <HYP: "men-doj">,
 PHO: <SAMPA: "mEndOj", CV: "[CVC][C][VC]">,
 VAL: { {},
        {PP_që, PP_të, PP_se},
        {kështu, ashtu, ndryshe, njësoj, ... } },
 ...
];
```

Um einen solchen Eintrag zu erstellen, wären allerdings weitere Informationen vonnöten, die in Wörterbüchern nicht vorhanden sind bzw. nur sporadisch von Fall zu Fall angegeben sind. Interessant wären die Angaben unter dem Attribut `ALM`, die zusammen mit dem Teil vor dem Zeichen „`/`“ die Allomorphe des Verbs bilden. Ebenso von Nutzen wäre eine Silbentrennung, wie unter `SPE`. Weiterhin wären Aussprache-Informationen (`SAMPA`, *Speech Assessment Methods Phonetic Alphabet* und `CV`, Darstellung des Lemmas

[196] Sie wurden nach den für die jeweilige Wortart charakteristischen Feldern absteigend sortiert. Das heißt, dass die Gruppe der Einträge, die am meisten Mitglieder hat, mit der Klasse *Wortart*`-001` versehen wurde, die zweitgrößte Gruppe mit *Wortart*`-002` usw. Die Verben werden bspw. nach den Eigenschaften *Wortart*, *Valenz*, *Aorist-* und *Partizip-Feld* sortiert.

[197] Es handelt sich hier um einen Entwurf eines proprietären Formats. ‚< ... >‘ stehen für einen Wert, der mehrere AWPe in Form einer Liste organisiert, ‚{ ... }‘ für einen Wert, der mehrere Elemente in Form von Listen beinhaltet.

zerlegt in Konsonanten (C) und Vokale (V)) wie unter PHO von Vorteil, genauso wie ausführliche Angaben zur Valenz unter dem Attribut VAL.
Besondere Aufmerksamkeit verdienen Verb-Einträge, wie etwa die der Hilfsverben *kam* (dt. *haben*) und *jam* (dt. *sein*), die suppletive Formen besitzen und in dem gezeichneten Schema untergebracht werden müssen bzw. untergebracht sind. Bei diesen Verben sind in den jeweiligen Aorist- bzw. Partizip-Form-Feldern die vollen Suppletiv-Formen angegeben, für *kam* die Formen *pata* und *pasur*, für *jam* die Formen *qeshë* und *qenë*. Ebenso müssen Verben, wie z. B. *mbetem* refl./pass., bzw. *mbes* (dt. *zurück bleiben* u. ä.) gesondert behandelt werden. Diese weisen nämlich ein sogenanntes defektives Paradigma auf, worin einige Formen nicht vorkommen. Dieses Verb listet Snoj [1994] als *mbe/s V. intr. ∼ta, ∼tur*. Dhrimo/Memushaj [2011] geben die folgenden Informationen: *mbe/s jokal. ∼ta, ∼tur; ∼tem; ∼ste, ∼sësh (të), ∼së (të), ∼tkësh, ∼tsha, ∼tsh, ∼ttë.* In anderen Beschreibungen wie in Munishi [1998] und Bega/Bega [2007] ist das Verb nur in seiner passiven/-reflexiven Form aufgelistet. In Texten bzw. Textkorpora lassen sich Belege des Verbs auch im Aktiv finden. So sind Formen mit *-s*, wie *mbes*, die nicht selten vorkommen, berücksichtigt und werden ins Lexikon aufgenommen und im Rahmen von xfst verarbeitet.

4.3 Substantiv-Einträge

Im Vergleich zu den Verb-Einträgen sehen die Strukturen der Substantiv-Einträge aufgrund der Tatsache, dass Substantive andere morphologische Kategorien aufweisen, anders aus:

arr̈ë *f. sh.* ∼a, ∼a, ∼at (Lex-S-trad-1)

So wie die Verb-Einträge wurden auch die Substantiv-Einträge für die Zwecke der maschinellen Sprachverarbeitung „normalisiert“. Die Version des obigen Eintrages (Lex-S-trad-1) sieht wie folgt aus (CSV, 13 Felder):

```
arrë|1|1|árr/ë|S.|f.|∼a|∼a|∼at||||S-001
```
(Lex-S-mod-1)

Selbst innerhalb der Substantive variiert die Form der Einträge. Einige Substantive, wie z. B. *e␣hënë* (dt. *Montag*) aus *ditë␣e␣hënë*, bestehen aus einem Artikel und einem lexikalischen Wort. Ebenso bestehen einige aus Adjektiven bzw. Adverbien abgeleitete Substantive wie *të␣afërmit* (dt. *die Angehörigen*), aus zwei Teilen. Sie sind lexikalisiert und daher in dieser Form

in die Wörterbücher aufgenommen worden, vgl. z. B. [FDSH 1976], [SNOJ 1994] oder [FJALORI 2006], und werden auch mit ihrer „neuen“ Wortart markiert.
Im Folgenden werden die einzelnen Felder in (Lex-S-mod-1) erklärt:

`arrë` In diesem Feld wird das Wort in seiner der orthographischen Regelung für das Norm-Albanische entsprechenden korrekten Form geschrieben. Es wird auch als Hauptreferenz (Schlüssel) verwendet. Die grammatischen Merkmale sind Nominativ Singular unbestimmt.

`1` Wie bei Verb-Einträgen wird in diesem Feld der Akzent markiert, d. h., der Vokal, auf den er fällt, wird mit einem zusätzlichen Zeichen versehen, s. u. das Feld `árr/ë`.

`1` Auch dieses Feld hat den gleichen Status wie bei den Verb-Einträgen. Es wird die Segmentierung des Lemmas markiert, d. h., die Stelle, ab der die Alternation des Stammes bzw. des Stammes und der Suffixe anfängt.

`árr/ë` In diesem Feld wird der Akzent (Betonung) durch ein Akzentzeichen (Betonungszeichen), „´“ über dem `a`, markiert sowie die Segmentierung des Wortes (vom ersten Feld) angegeben. Die Segmentierung des Wortes, angedeutet durch das Zeichen „`/`“, spielt hier eine entscheidende Rolle und hängt mit den Angaben in den Feldern zusammen, welche die Flexionsmerkmale beschreiben, hier die Felder ∼`a` (Feld 7), ∼`a` (Feld 8) und ∼`at` (Feld 9).

`S.` `S.` steht für *Substantiv* und markiert so die Wortart des Lemmas.

`f.` In diesem Feld wird der Genus des Substantivs markiert. Es kommen die drei Genera, `m.` (Maskulin), `f.` (Feminin) und `n.` (Neutrum) vor.

`∼a` In diesem Feld wird die Form des Substantivs in seiner bestimmten Form im Nominativ Singular angegeben. Es wird nur der Teil der alterniert angegeben, also der Teil nach dem Zeichen „`/`“. Dies ist dann als `arr/` und ∼`a`, d. h. `arra` zu lesen. Die unbestimmte Form ist dem vierten Feld (`árr/ë`), d. h. `arrë`, zu entnehmen.

`∼a` Die unbestimmte Form des Substantivs im Nominativ Plural wird in diesem Feld angegeben. Die Form wäre, wie im vorigen Feld, `arr/` und ∼`a`, d. h. `arra`.

`~at` Die vierte Form, die Substantive markiert, nämlich der bestimmte Nominativ Plural, wird in diesem Feld angegeben. Da sich der Plural der Substantive oft sehr stark vom Singular unterscheidet, ist die Angabe dieser Formen sehr wichtig. Die „komplettierte“ Form wäre `arrat`, d. h. `arr/` und `~at`.

<Kein Eintrag im Feld> In diesem Feld, d. h. im zehnten Feld, wird der Artikel des Substantivs im Singular angegeben. Es kommen die Artikel *i* für Maskulin (`m.`), *e* für Feminin (`f.`) und *të* für Neutrum (`n.`) vor. Für das Substantiv *i␣afërm* wäre in diesem Feld `i` eingetragen.

<Kein Eintrag im Feld> In diesem Feld, d. h. im elften Feld, wird der Artikel des Substantivs im Plural angegeben. Es kommt der Artikel `të` vor, der sowohl für Maskulin und Feminin als auch für Neutrum und Plural aller drei Genera steht.

<Kein Eintrag im Feld> Dieses Feld, das Zwölfte, ist wie bei den Verb-Einträgen für besondere Angaben reserviert.

`S-001` Hier wird die Flexionsklasse des Wortes vom ersten Feld innerhalb der Wortart angegeben, in diesem Fall des Substantivs, daher das Symbol `S`. Die Klasse entspricht den morpho-syntaktischen Eigenschaften des Substantivs, insbesondere denen der Deklination. In den Fällen, in denen assoziierte Partikel vorhanden sind, werden sie auch berücksichtigt.

Wie in Abschnitt 3.3.2 erläutert, werden im Albanischen auch Namen wie andere Substantive dekliniert, also sind Kasus durch Suffixe markiert. Sie werden in diesen Listen nur beschränkt aufgenommen, da ihre Zahl theoretisch unbegrenzt groß ist.

4.4 Adjektiv-Einträge

Die Lexikoneinträge für Adjektive bilden eine Klasse, die aufgrund der verschiedenen Typen der Adjektive im Vergleich zu den Lexikoneinträgen der Nomina und Verben sehr heterogen ist, vgl. hierzu auch Abschnitt 3.3.3. Der am häufigsten vorkommende Typ ist der folgende:

tásh|ëm (i)/~me (e) Adj. (Lex-Adj-trad-1)

Der Eintrag wird als *i␣táshëm*, m. bzw. *e␣táshme*, f. entschlüsselt. Die Markierung der Alternationsgrenze (Segmentierung) und die Angabe des variierenden Segments für das Genus Feminin (in diesem Fall ∼*me*) ermöglicht eine Art „Zusammenfassung“ der Formen der beiden Genera. Diese Art der Informationskodierung hat sich in der Tradition der herkömmlichen gedruckten Wörterbücher als praktisch erwiesen – bekanntlich hauptsächlich deshalb, weil der zur Verfügung stehende Platz sehr begrenzt ist.
Die verschiedenen Formen der traditionellen Adjektiveinträge werden in ein Format überführt, welches deren zum Teil sehr unterschiedliche Eigenschaften widerspiegelt. Das Format verfügt für alle Eigenschaften der Adjektive über ein entsprechendes Feld, das in Abhängigkeit vom Typ gefüllt oder nicht gefüllt wird.

`tashëm|2|2|tásh/ëm|Adj.||i|∼me|e||||||Adj-004` (Lex-Adj-mod-1)

`tashëm` Das erste Feld ist, wie bei Verben und Substantiven, in gleicher Form für das Lemma reserviert.

`2` Im zweiten Feld wird in Form einer Zahl die Position des Buchstabens und des Vokals angegeben, auf den der Akzent fällt.

`2` In Analogie zu den Verben und Substantiven wird hier die Position angegeben, ab dem die Alternation des Lemmas innerhalb des Flexionsparadigmas erscheint. Das heißt, zwei Buchstaben (ab dem Zeichen „/“ im Feld 4) werden durch die Angaben in den korrespondierenden Feldern, in diesem Fall durch die aus dem Feld 7, ersetzt, vgl. *i␣tásh*ëm, m. vs. *e␣tásh*me, f.

`tásh/ëm` Dieses Feld hat die gleiche Funktion wie bei Verben und Substantiven. Es fasst die ersten drei zusammen und markiert das Lemma mit zusätzlichen Zeichen.

`Adj.` Die Wortart belegt dieses Feld.

`<Kein Eintrag im Feld>` Dieses Feld, das Sechste, wurde freigelassen. Adjektive, die nicht flektiert werden, wie *neto* (dt. *netto*), können dort markiert werden.

`i` In diesem Feld wird der Artikel des Adjektivs eingetragen, falls es ein maskulines Genus besitzt. Die Angaben in diesem Feld werden mit den Angaben in Feld 3 kombiniert, d. h. es sollte `i␣tásh`ëm resultieren.

`~me` In diesem Feld steht die Alternation der femininen Form, falls diese existiert, d. h. *tásh*`me`.

`e` Hier steht der Artikel des Adjektivs für die feminine Form des Lemmas, falls sie existiert. Im konkreten Fall ist es `e␣`*tásh*`me`.

<*Kein Eintrag im Feld*> In diesem Feld, d. h. im zehnten Feld, wird die Alternation derjenigen Lemmata eingetragen, für die gewöhnlich vier Formen angegeben werden, wie *i␣keq* (dt. *schlechter*), s. u. für weitere Sonderfälle. Für den Eintrag *i␣keq* wäre in diesem Feld *~ëqíj* eingetragen. Der komplette Eintrag in traditioneller Form wäre: Sg. m. *i␣keq* / f. *e␣keqe*, Pl. m. *të␣këqinj* / f. *të␣këqija*; Der normalisierte Eintrag ist weiter unten, als (Lex-Adj-mod-2) bezeichnet, angegeben.

<*Kein Eintrag im Feld*> Hier, im elften Feld, steht der Artikel des Adjektivs für die maskuline Form des Lemmas im Plural, falls sie existiert. Im konkreten Fall ist es `të`. Die komplettierte Form wäre `të␣`*k*`ëqíj`, s. u. (Lex-Adj-mod-2).

<*Kein Eintrag im Feld*> Für die Lemmata, für die gewöhnlich vier Formen eingetragen werden, steht in diesem Feld, d. h. im zwölften Feld, die Alternation für die feminine Form des Lemmas im Plural, entsprechend (Lex-Adj-mod-2), s. u., wäre dies `~ëqíja`.

<*Kein Eintrag im Feld*> Das vorige Feld wird durch einen Artikel ergänzt, der in diesem Feld, d. h. ins 13. Feld, eingetragen wird. Im Fall von (Lex-Adj-mod-2), s. u., wäre der Artikel `të`, d. h., beide Felder zusammen bilden die Form `të␣`*k*`ëqíja`.

<*Kein Eintrag im Feld*> Dieses Feld, d. h. das 14. Feld, wurde für besondere Markierungen, wie bei Verben und Substantiven, freigelassen.

`Adj-004` In diesem Feld wird die Flexionsklasse des jeweiligen Lemmas innerhalb der Wortart eingetragen.

Im Vergleich zu dem Eintrag *i␣táshëm*, `m.` bzw. *e␣táshme* `m.`, weisen einige Adjektive in ihren vier Formen Singular/Plural Feminin/Maskulin sehr starke Abweichungen voneinander auf, so dass ihre Formen als Ganzes eingetragen werden. Einige Beispiele sind die folgenden:[198]

[198] Vgl. hierzu [Hetzer/Finger 1993: 52].

Sg. m. *i␣ri* / f. *e␣re*, Pl. m. *të␣rinj* / f. *të␣reja*;
Sg. m. *i␣zi* / f. *e␣zezë*, Pl. m. *të␣zinj* / f. *të␣zeza*;
Sg. m. *i␣madh* / f. *e␣madhe*, Pl. m. *të␣mëdhenj* / f. *të␣mëdha*;
Sg. m. *i␣lig* / f. *e␣ligë*, Pl. m. *të␣ligj* / f. *të␣liga*;

Diese Einträge haben die Form wie der folgende, bereits diskutierte, Fall:

```
keq|2|2|k/éq|Adj.||i|∼éqe|e|∼ëqíj|të|∼ëqíja|të||Adj-022
```

(Lex-Adj-mod-2)

4.5 Pronomina-Einträge

Die lexikalischen Einträge der Pronomina sind, je nach Typ, sehr unterschiedlich voneinander. Eine Systematik wie bei den drei größeren Wortarten, Verben, Substantive und Adjektive zu schaffen, lohnt sich für Pronomina nicht. Es bleibt nur übrig, eine sehr heterogene Klasse zu bilden. In einigen Fällen ist es am leichtesten, alle Formen eines Pronomen gleich beim Kompilieren des Lexikons als „volle Formen“ einzutragen. Die Erstellung einer Grammatik, die diese Formen generiert oder produziert, ist angesichts der wenigen Formen nicht lohnend. Pronomina bilden als Wortart eine geschlossene Klasse, d. h., es ist eine berechenbare Zahl der Pronomina, welche die Sprache besitzt und diese wird, synchron betrachtet, nicht erweitert oder verringert.
Die einfachste Lösung ist es, die lexikalischen Einträge der Pronomina in Form eines Vollformlexikons, d. h. als analysierte Einträge, zu erstellen. Dabei muss das Format der Einträge die Eigenschaften der Untergruppen beschreiben können und dennoch (möglichst) einheitlich für alle Einträge sein.
Eine mögliche Form, sie darzustellen, z. B. als Attribut-Werte-Paare (AWP), wäre wie im folgenden Quellcode 4.2, wo einige Personal-Pronomina modelliert sind:

Listing 4.2: Lexikon der Personal-Pronomina

```
1
2 [Lemma:"unë", POS:Pron, Type:PersPron, Nom, Sg, P1 ];
3 [Lemma:"mua", POS:Pron, Type:PersPron, Acc, Sg, P1, pC:"më" ];
4 [Lemma:"meje", POS:Pron, Type:PersPron, Dat, Sg, P1, pC:"më" ];
5  ...    ...   ...   ...   ...   ...   ...
6 [Lemma:"atyre", POS:Pron, Type:PersPron, Dat, Pl, P3, pC:"u", ␣
                    Gender:M+F ];
7  ...    ...   ...   ...   ...   ...   ...
```

Die Bezeichnung der Attribute und ihrer Werte, wie `pC: "më"`, sind selbsterklärend. In Anführungszeichen sind die konkreten Formen gesetzt, anders als die grammatischen und formalen Eigenschaften. Fälle wie z. B. das Possessiv-Pronomen `"i saj"`, zu dt. *ihr*, haben besondere zusätzliche Eigenschaften (Genus des Besitzes). Sie werden nach dem gleichen Schema behandelt.

In der „normalisierten" Form sähen die Einträge aus der Abbildung 4.2 wie folgt (Listing 4.3) aus:

Listing 4.3: Lexikon der Personal–Pronomina (2)

```
unë|0|0|unë|Pron.|Pers.NomS1|||||Pron-001
unë|0|0|unë|Pron.|Pers.NomS1|||||Pron-001
mua|0|0|unë|Pron.|Pers.AccS1|pC:më||||Pron-001
mua|0|0|unë|Pron.|Pers.DatS1|pC:më||||Pron-001
meje|0|0|unë|Pron.|Pers.AblS1|||||Pron-001
... ... ... ... ... ... ...
atyre|0|0|ata|Pron.|Pers.DatP3m|pC:u||||Pron-001
atyre|0|0|ato|Pron.|Pers.DatP3f|pC:u||||Pron-001
atyre|0|0|ata|Pron.|Pers.GenP3m||pA:i,e,të|||Pron-001
atyre|0|0|ato|Pron.|Pers.GenP3f||pA:i,e,të|||Pron-001
atyre|0|0|ata|Pron.|Pers.AblP3m|||r:tyre||Pron-001
atyre|0|0|ato|Pron.|Pers.AblP3f|||r:tyre||Pron-001
tyre|0|0|ato|Pron.|Pers.AblP3f||||ff:atyre|Pron-001
tyre|0|0|ata|Pron.|Pers.AblP3m||||ff:atyre|Pron-001
... ... ... ... ... ... ...
```

Anders als bei den Einträgen der Verben, Substantive und Adjektive ist hier jeweils im zweiten und im dritten Feld der Wert 0 eingetragen. Dies bedeutet, dass kein Akzent markiert ist (Feld 2), sowie dass es sich um volle Formen handelt (Feld 3). Bei den letzten zwei Einträgen handelt es sich um kurze Formen der Personalpronomina, die im Kasus Ablativ ohne ein vorangesetztes *a* erscheinen, vgl. auch *atë* → *të*, *asaj* → *saj*, *atij* → *tij*, *ata* → *ta*, *ato* → *to*. Sie werden doppelt eingetragen, um sowohl die Form ohne *a* als auch die Form mit *a* als Schlüssel zu haben, bspw. sowohl für die Analyse als auch für die Produktion.
Die Erklärung der Felder in Kurzform: In Feld 4 wird die Grundform eingetragen, in Feld 5 die Wortart, in Feld 6 die Art der Pronomina und ihre grammatischen Eigenschaften, in Feld 7 die kurze, d. h. klitische, Form des jeweiligen Pronomens, in Feld 8 die Artikel im Genitiv (überflüssig), in Feld 9 die Form „ohne *a*", in Feld 10 die volle Form, falls als Lemma (Feld 1) eine kurze Form „ohne *a*" angegeben ist. Das Feld 11 ist für besondere Markierungen freigelassen. Schließlich steht das Feld 12 für die Angabe der Wortart zur Verfügung.

4.6 Numeral-Einträge

Die Wortart der Numeralia bildet aufgrund ihrer Eigenschaften eine offene Klasse mit einer begrenzten Zahl an zugrundeliegenden Lemmata. Die „Grundwörter“ werden in das Lexikon aufgenommen und entsprechend ihrer jeweiligen Eigenschaften beschrieben, die „verbleibenden“ werden einer Grammatik überlassen, welche sie analysiert und generiert.
Die Kardinal- und Ordinalzahlen, die ins Lexikon aufgenommen werden, bilden nur eine kleine Zahl. Die Kardinalzahlen sind die folgenden: *zero* [0], *një*, ..., *nëntë* [1–9], *njëmbëdhjetë*, ..., *nëntëmbëdhjetë* {1, 2, 3, ... 9}0, *njëqind*, ... *nëntëqind* {1, 2, 3, ... 9}00, *mijë*, *milion*, *miliard*, *bilion*, *biliard*, *trilion*, *triliard*, *kuadrilion*, *kuadriliard* usw. Sie sind in Abbildung 4.4 (Lexikon) aufgelistet.

Listing 4.4: Lexikon der Numerale

```
"zero"|NC|...|0 ;
"një"|NC|...|1 ;
"dy"|NC|...|2 ;
"tre"|NC|Gender:M|...|3 ;
"tri"|NC|Gender:F|...|3 ;
"katër"|NC|...|4 ;
"pesë"|NC|...|5 ;
...
"nëntë"|NC|...|9 ;
"dhjetë"|NC|...|10 ;
"njëmbëdhjetë"|NC|...|11 ;
...
"nëntëmbëdhjetë"|NC|...|19 ;
"njëzet"|NC|...|20 ;
"tridhjetë"|NC|...|30 ;
"dyzet"|NC|...|40 ;
"pesëdhjetë"|NC|...|50 ;
"gjashtëdhjetë"|NC|...|60 ;
...
"nëntëdhjetë"|NC|...|90 ;
"njëqind"|NC|...|100 ;
...
"nëntëqind"|NC|...|900 ;
"mijë"|NC|...|x.000 ;
...
"milion"|NC|...|x.000.000 ;
...
"miliard"|NC|...|x.000.000.000 ;
...
```

In diese Liste gehören auch Zahlen wie *e␣dhjeta* (dt. *zehntel*), *e␣qindta* (dt. *hundertstel*), *e␣mijëta* (dt. *tausendstel*), *e␣milionta* (dt. *millionstel*), *e␣miliardta* (dt. *milliardstel*) usw.

Die Ordinalzahlen, wie in Abschnitt 3.3.5 ausführlich beschrieben, kommen stets zusammen mit einem vorangestellten Artikel vor.

Listing 4.5: Lexikon der Ordinalzahlen

```
1
2 "i pari"|ON|DeclT: ... ;
3 "e para"|ON|DeclT: ... ;
4 "të parët"|ON|DeclT: ... ;
5 "të parat"|ON|DeclT: ... ;
6 "i dyti"|NC|DeclT: ... ;
7 "e dyta"|NC|DeclT: ... ;
8 "të dytët"|NC|DeclT: ... ;
9 "të dytat"|NC|DeclT: ... ;
10 ...
```

Ordinalzahlen müssen formalisiert werden, um durch die Morphologie automatisch bearbeitet werden zu können, vgl. *katërqind␣e␣dymbëdhjetë␣mijë␣e␣pesëqind␣e␣shtatëdhjetë␣e␣tetë* [412 578], *dyqind␣e␣shtatë␣miliardë␣e␣njëqind␣e␣pesëmbëdhjetë␣milionë␣e␣pesë* [207 115 000 005] bzw. *i␣, e␣, të␣njëmilion|të*, im Abschnitt 3.3.5, wobei sowohl die Flexion als auch die Zusammen- vs. Getrenntschreibung die beiden Typen unterscheidet.[199] Als Konjunktion wird nur *e*, vergleichbar mit *und* im Deutschen, verwendet.
Auch unbestimmte Formen müssen in das Lexikon eingetragen werden. Man vergleiche hierzu Formen wie „*i␣, e␣, të␣parë* " i. S. v. „*është␣hera␣e␣parë*", zu dt. *es ist das erste Mal* oder *për␣herë␣të␣parë*, zu dt. *zum ersten Mal*.

4.7 Einträge der Konjunktionen

Wie in Abschnitt 3.3.8 ausführlich erklärt, kommen Konjunktionen sowohl als ein- als auch als mehrteilige Wörter vor. Die einfachen Konjunktionen werden wie die üblichen einfachen Lexikon-Einträge behandelt, da sie keine Flexion aufweisen. Für die lexikalische Beschreibung und insbesondere für die Formalisierung der mehrteiligen Konjunktionen wird im Lexikon eine Art Valenz der Zugehörigkeit benutzt, um die Teile der Konjunktion zusammenzuhalten. Sie werden entsprechend annotiert – wie die übliche Valenz-Annotation, damit sie bei maschineller Sprachverarbeitung als mehrteilige Elemente erkannt werden. Ihre Abhängigkeit von einander bleibt im Rahmen der Syntax zu behandeln.

Die einfachen Einträge werden im Lexikon (lexc) wie im Quellcode 4.6 dargestellt:

[199] Beispiel-Einträge: *një, dy, ... i pari, i dyti, ...* und *njëri*.

Listing 4.6: Abschnitt aus Einträgen der Konjunktionen im lexc-Format

```
[ ... ]
e:e Conjunctions;
edhe:edhe Conjunctions;
gjersa:gjersa Conjunctions;
gjithsaherë:gjithsaherë Conjunctions;
haj:haj Conjunctions;
hoj:hoj Conjunctions;
ja:ja Conjunctions;
jomë:jomë Conjunctions;
jose:jose Conjunctions;
josemë:josemë Conjunctions;
kinse:kinse Conjunctions;
kish:kish Conjunctions;
ku:ku Conjunctions;
kur:kur Conjunctions;
kurse:kurse Conjunctions;
[ ... ]
```

Für sie ist nur die Wortart angegeben. Eine Bereicherung um grammatische (syntaktisch-semantische) Informationen ist problemlos möglich.[200]

4.8 Einträge der Präpositionen

Die Präpositionen sind eine geschlossene Wortart und bilden somit eine kleine Menge von Einträgen. Sie besitzen jedoch die Eigenschaft der Rektion, vgl. Abschnitt 3.3.7.

Die Einträge der Präpositionen werden im Lexikon wie im Quellcode 4.7 dargestellt:

Listing 4.7: Lexikon der Präpositionen

```
...
andej|0|0|andej||Abl|1|Prep-001|
anekënd|0|0|anekënd||Abl|1|Prep-001|
anembanë|0|0|anembanë||Abl|1|Prep-001|
anës|0|0|anës||Abl|1|Prep-001|
brenda|0|0|brenda||Abl|1|Prep-001|
brendapërbrenda|0|0|brendapërbrenda||Abl|1|Prep-001|
bri|0|0|bri||Abl|1|Prep-001|
buzë|0|0|buzë||Abl|1|Prep-001|
...
me|0|0|me||A|1|Prep-001|
mes|0|0|mes||Abl|1|Prep-001|
më|0|0|më||A|1|Prep-001|
mënjanë|0|0|mënjanë||Abl|1|Prep-001|
midis|0|0|midis||Abl|1|Prep-001|
ndaj|0|0|ndaj||Abl|1|Prep-001|
ndanë|0|0|ndanë||Abl|1|Prep-001|
ndër|0|0|ndër||A|1|Prep-001|
```

[200] Vgl. [Morfologjia 1995] und [Çeliku et al. 1998] zu den syntaktisch-semantischen Eigenschaften der Konjunktionen im Albanischen.

```
ndërmjet|0|0|ndërmjet||Abl|1|Prep-001|
në|0|0|në|InDet|A|1|Prep-001|
nën|0|0|nën|InDet|A|1|Prep-001|
nëpër|0|0|nëpër|InDet|A|1|Prep-001|
nëpërmes|0|0|nëpërmes||Abl|1|Prep-001|
nëpërmjet|0|0|nëpërmjet||Abl|1|Prep-001|
nga|0|0|nga|Det|N|1|Prep-001|
ngjat|0|0|ngjat||Abl|1|Prep-001|
pa|0|0|pa|InDet|A|1|Prep-001|
para|0|0|para||Abl|1|Prep-001|
pas|0|0|pas||Abl|1|Prep-001|
për|0|0|për|InDet|A|1|Prep-001|
...
tatëpjetë|0|0|tatëpjetë||Abl|1|Prep-001|
te|0|0|te||N|1|Prep-001|
tej|0|0|tej||Abl|1|Prep-001|
tejendanë|0|0|tejendanë||Abl|1|Prep-001|
tejpërtej|0|0|tejpërtej||Abl|1|Prep-001|
tek|0|0|tek||N|1|Prep-001|
teposhtë|0|0|teposhtë||Abl|1|Prep-001|
tutje-tehu|0|0|tutje-tehu||Abl|1|Prep-001|
veç|0|0|veç||Abl|1|Prep-001|
për në|0|0|për në||A|2|Prep-002|
brenda në|0|0|brenda në||A|2|Prep-002|
tok me|0|0|tok me||A|2|Prep-001|
bashkë me|0|0|bashkë me||A|2|Prep-002|
...
```

Einige Einträge (vgl. Zeilen 41–44) bestehen aus mehreren Teilen (engl. *Multi-Word Units*). Sie werden besonders markiert, wie die mehrteiligen Konjunktionen. Dieser Typ bildet im Gegensatz zu den anderen Präpositionen eine Klasse, deren Größe nicht klar definiert ist. Dies bedeutet auch, dass der Status einiger solcher mehrteiliger Wörter strittig ist.

4.9 Einträge der Adverbien

Die Einträge der Adverbien sind einfach und bestehen aus zwei Haupttypen, dem einteiligen Typ und dem zweiteiligen Typ. Der Letztere besitzt einen vorangestellten Artikel, der nicht dekliniert wird: den Artikel *së*, wobei der Hauptteil des Adverbs markiert wird, vgl. `së␣`*afër*`mi` (Zeile 5 in der Abbildung 4.8) oder `së␣`*von*`i`.

Listing 4.8: Lexikon der Adverbien

```
...
afër|1|0|áfër|Adv.|||||Adv-001
afërisht|5|0|afërísht|Adv.|||||Adv-001
afërmendësh|6|0|afërméndësh|Adv.|||||Adv-001
afërmi|1|0|áfërmi|Adv.|art_së||||Adv-002
afërsisht|6|0|afërsísht|Adv.|||||Adv-001
...
```

Für die Adverbien, die gesteigert werden können, ist eine Behandlung mithilfe einer Grammatik möglich, vgl. Abschnitt 5.8. Dabei ist eine Modellierung und die Verarbeitung einer Struktur wie *më␣së␣voni* möglich.

4.10 Einträge der Partikeln

Die Einträge für Partikeln bilden eine kleine Menge. Sie sind einfache Einträge ohne grammatische Informationen, können aber jederzeit um neue Informationen ergänzt und erweitert werden.

Listing 4.9: Lexikon der Partikel

```
...
ja|2|0|já|Part.|||Part-001
jo|0|0|jo|Part.|||Part-001
kinse|2|0|kínse|Part.|||Part-001
ku|0|0|ku|Part.|||Part-001
kund|2|0|kúnd|Part.|||Part-001
le|0|0|le|Part.|||Part-001
lum|2|0|lúm|Part.|||Part-001
lumthi|2|0|lúmthi|Part.|||Part-001
madje|5|0|madjé|Part.|||Part-001
...
```

Einige Partikeln wie *nuk* (dt. *nicht*), *jo* (dt. *nein*), und *as* (dt. *auch nicht*, *nicht mal*, *überhaupt nicht*) usw. haben eine besondere Rolle, da sie bestimmte Funktionen haben, im konkreten Fall Negation, und im Vergleich zu den anderen Partikeln von größerer Bedeutung sind.

4.11 Einträge der Interjektionen

Die Einträge der Interjektionen bilden eine kleine Menge. Sie sind nicht mit viel Information versehen, wobei die Wortart die Wichtigste ist. Im folgenden Listing (4.10) sind einige Einträge angegeben.

Listing 4.10: Lexikon der Interjektionen

```
...
ah|1|0|áh|Interj.||Itj-001
aha|3|0|ahá|Interj.||Itj-001
alilluja|6|0|alillúja|Interj.||Itj-001
alo|3|0|aló|Interj.||Itj-001
aman|3|0|amán|Interj.||Itj-001
ani|3|0|aní|Interj.||Itj-001
...
```

Die Einträge sind minimal mit Informationen versehen, da sie sich mehr durch semantische Eigenschaften unterscheiden als durch grammatische. Diese Beschreibung entspricht den Angaben zu den jeweiligen Interjektionen in Wörterbüchern des Albanischen.

4.12 Andere Einträge

Eine nicht geringe Zahl von Einträgen passt nicht in die Hauptgruppen des Lexikons. Die amalgamierten Formen der klitischen Pronomina, Frage- und Negationspartikeln, für sich und als Kombinationen miteinander, sowie einige andere Einträge, u. a. einige Adverbien wie *së tepër|mi*, *së par|i*, *së von|i*, *së fund|i*[201], werden gesondert behandelt, vgl. hierzu auch die Abschnitte 3.3.4 und 3.3.9.

4.12.1 Einträge der Artikel

Wie in Kapitel 3 oft angesprochen, spielen die Artikel eine wichtige und sogar entscheidende Rolle, wie z. B. bei Adjektiven oder Pronomina.
Da sie flektiert werden, aber doch eine geschlossene Klasse bilden, werden sie in Form eines Vollformlexikons erfasst. Eine Grammatik für die Modellierung ihrer Flexion wäre nicht lohnenswert.

Listing 4.11: Lexikon der Artikel

```
"i"|Art.|GRA:m.N.sg.|...;%Pron.D.sg./A.pl.%
"e"|Art.|GRA:f.N.sg.|...;%Pron.A.sg.%
"së"|Art.|GRA:f.G.sg.|...;%së +Adv.%
"të"|Art.|GRA:N.pl.|...;%Pron.D./A.sg %Konj.%
"u"|Art.|GRA:D.pl.|...;%POS:PassArt.%
...
```

4.12.2 Einträge der Flexionssuffixe und Wortbildungsmittel

Im Folgenden, Listing 4.12, ist das Format des Lexikons der Suffixe angegeben. Dabei ist unter der Nummer (1) ein traditioneller Eintrag angegeben, der die Suffixe für den Kasus Nominativ im Singular unbestimmt, Singular bestimmt, Plural unbestimmt und Plural bestimmt auflistet. Ergänzt wurden diese Segmente[202] für die restlichen Kasus, nämlich Akkusativ, Dativ und

[201] Vgl. [Dhrimo/Memushaj 2011] für Adverb-Einträge wie z. B. *së voni*, der als vó·ni (së) *ndajf.* angegeben ist.

[202] Es handelt sich hier um „Segmente“, wie in den Abschnitten 4.2, 4.3, 4.4 und 4.6, erklärt.

Ablativ für alle vier Formen, d. h. Singular und Plural jeweils in bestimmter und unbestimmter Form, wie unter Nummer (2) zu sehen ist.

Die Einträge aus Tabelle 3.5 können auch in der folgenden Form (Abbildung 4.1) organisiert werden:[203]

Abbildung 4.1: Klassifikation der Flexionssuffixe am Beispiel von *lis*/∼*i*.

```
lis/+{ [0,i,i,0,i]; [a,ave,ave,a,ave]; [i,it,it,i,it]; [at,ave,ave,at,ave] };
lis/+{ [Suff-S1u]; [Suff-P1u]; [Suff-S1b]; [Suff-P1b] }→[S-011];
lis[S1u/P1u/S1b/P1b]/+{ [f_S1u]; [f_P1u]; [f_S1b]; [f_P1b] }→[S-011];
```

Ein anderes Beispiel, das die Organisation der Suffixe illustriert, ist das folgende:

Listing 4.12: Format des Lexikon der Suffixe

```
1
2 (1)  S-001=>[ adresë|adrés/ë|f.|~a|~a|~at ];
3
4 (2)  S-001=>{ f. adresë:adrés/ ~ë,  ~e,   ~e,   ~ë,  ~e,
5                                ~a,  ~ës,  ~ës,  ~ën, ~e,
6                                ~a,  ~ave, ~ave, ~a,  ~ave,
7                                ~at, ~ave, ~ave, ~at, ~ave };
```

Das Beispiel stellt eine besondere Information dar, es trägt die Information der Segmentierung, besprochen auch in den Abschnitten 4.2 (Verbeinträge), 4.3 (Substantiveinträge), 4.4 (Adjektiveinträge) und 4.6 (Numeraleinträge) sowie in besonderen Fällen bei anderen Wortarten.
Im Listing 4.12 werden die Suffixe in zwei Formen dargestellt, und zwar unter (1), wie sie in Wörterbüchern gewöhnlich angegeben werden, und in (2) in ihrer vollständigen Form, d. h. in ihrem vollständigen Paradigma. Die Suffixe ∼ë, ∼a, ∼a, ∼at, dargestellt in (1), bzw. die erste Spalte, dargestellt in (2), entsprechen der *initialen* (repräsentativen) Form des Paradigmas, jeweils den Nominativ. Die Stelle ab *ë* bzw. ab *a* markiert die Alternation (Beugung). In gleicher Weise könnte auch das Beispiel *lis*/*lisi*, vgl. die Abbildung 4.1, betrachtet werden, wobei das ∼*a* bzw. ∼*a*... die Plural-Formen markieren.

[203] S1u bedeutet Klasse 1 der Singular-Suffixe in unbestimmter Form, entsprechend steht P1b für Klasse 1 der Plural-Suffixe in bestimmter Form. Diese Vorgehensweise findet man auch bei [Heringer 2010]. Die Daten (Stämme und Suffixe der Verben) in [Kabashi 2003] wurden ebenso in ähnlicher Form kategorisiert, vgl. hierzu Abschnitt 5.1.

4.12.3 Einträge der zusätzlichen Zeichen

Zusätzlich zu den „normalen“ Einträgen, die in traditionelle Wörterbücher und Lexika aufgenommen sind, müssen in ein Lexikon für computerlinguistische Zwecke auch Zahlen, Sonderzeichen[204] und andere wichtige Zeichen wie Interpunktionszeichen aufgenommen werden.

Listing 4.13: Lexikon der Interpunktionszeichen

```
1
2| "."|IP|"Pikë"|... ;
3| "?"|IP|"Pikëpyetje"|... ;
4| "!"|IP|"Pikëçuditje"|... ;
5| ";"|IP|"Pikëpresje"|... ;
6| ":"|IP|"Dy pika"|... ;
7| ","|IP|"Presje"|... ;
8| "'"|IP|"Apostrof"|... ;
9| "-"|IP|"Vizë lidhëse"|... ;
10| "+"|IP|"Plus"|... ;
11| "*"|IP|"Yll"|... ;
12| "/"|IP|"Thyes"|... ;
13| "="|IP|"Baras"|... ;
14| "("|IP|"Kllapë hapëse"|... ;
15| ")"|IP|"Kllapë mbyllëse"|... ;
16| ...
```

Ein anderer Typ von Lexikon-Einträgen sind die Ziffern. Sie sehen wie im Quellcode 4.14 aus. Dabei steht NV für Numeric value.

Listing 4.14: Lexikon der Ziffern

```
1
2| "0"|NV|"Zero"|... ;
3| "1"|NV|"Një"|... ;
4| "2"|NV|"Dy"|... ;
5| "3"|NV|"Tre m. / Tri f."|... ;
6| "4"|NV|"Katër"|... ;
7| "5"|NV|"Pesë"|... ;
8| "6"|NV|"Gjashtë"|... ;
9| "7"|NV|"Shtatë"|... ;
10| "8"|NV|"Tetë"|... ;
11| "9"|NV|"Nëntë"|... ;
```

4.13 Morpho-syntaktische Kategorisierung der Wortarten

Sowohl in gesprochener als auch in geschriebener Sprache kommen Wortarten in verschiedenen Sätzen und in verschiedenen Kombinationen vor.

[204] Bspw. Minus- oder Pluszeichen aus Unicode.

Dabei ist es oft der Fall, dass die gleiche Wortart, bspw. Adjektive, in verschiedenen Fällen verschiedene Eigenschaften aufweisen, sodass eine Unterklassifizierung nötig ist. Ein einfaches Beispiel ist der Vergleich der Konjugationen der Verben *sein*, *geben* und *schreiben*. Obwohl sie der gleichen Wortart angehören, weisen sie verschiedene Eigenschaften wie z. B. Stammmuster und Flexionsparadigma auf. Aus diesem Grund werden sie klassifiziert, etwa um sie beim Fremdspracherlernen besser zu bemerken. Eine Subkategorisierung ist wichtig für ihre Modellierung im Rahmen der MSV.
Solche Informationen sind jedoch nicht gegeben oder sie sind nur allgemein (in traditioneller Form) in Grammatiken zu lesen, wobei ein illustratives Beispiel angeführt wird. Wenn doch alle Wörter modelliert werden sollen, wie es der Fall für die Zwecke der MSV ist, erfährt man meist, dass oft mehr Informationen erwünscht werden, als vorhanden sind. Jedes einzelne Lemma innerhalb der Wortart systematisch in Klassen einzuordnen, bildet eine nicht zu unterschätzende Herausforderung.
Wenn von der traditionellen Informationsstruktur in den Wörterbüchern des Albanischen ausgegangen wird, ergibt sich die Möglichkeit, eine weitestgehend in großen Maßen automatische Klassifikation der einzelnen Lemmata in ihre Subtypen zu klassifizieren. Eine voll automatische Klassifikation ist leider nicht möglich, da die lexikalischen Angaben in den üblichen Wörterbüchern, z. B. vier Suffixe für die Substantive, vgl. alban. *zë* (dt. *Stimme*), keine eindeutige Klassifikation erlauben. Dies kann anhand des folgenden Beispiels, Listing 4.15, veranschaulicht werden.

Listing 4.15: Klassen innerhalb einer Wortart

```
1
2| tigër|2|2|tig/ër|S.|m.|~ri|~ra|~rat||||S-024
3| urdhër|1|2|urdh/ër|S.|m.|~ri|~ra|~rat|||3|S-024
4| zë|2|0|zë|S.|m.|~ri|~ra|~rat||||S-199
```

Das Suffix-Muster `|S.|m.|~ri|~ra|~rat|` ist gleich für alle drei Einträge, doch der dritte Eintrag unterscheidet sich von den ersten zwei Einträgen. Das Lemma `zë` hat die folgenden Suffixe `-`, `~ri`, `~ri`, `-`, `~ri`, `~ri`, `~rit`, `~rit`, `~rin`, `~rit`, `~ra ...` Anders als das Lemma `urdhër`, welches die folgenden hat: `~ër`, `~ri ...` Obwohl es verschiedene Suffixe sind, stimmen sie in den vier Stellen überein, eine Tatsache, die eine Ambiguität darstellt und so die Klassifikation erschwert.

4.14 Ein Vollformlexikon für Testzwecke

Nachdem die linguistischen Daten kategorisiert und systematisiert wurden, liegt es nahe, ein morphologisches Lexikon zu erstellen, das zum Testen oder ggf. zum Annotieren verwendet werden kann.
Das im Folgenden vorgestellte Vollformlexikon wurde automatisch aus dem linguistischen Wissen (klassifizierte Wortarten in Gruppen, deren Flexionsmerkmale und deren Paradigmen) und einem Lexikon (mit eindeutig organisierten und klassifizierten Informationen) generiert.[205]

Listing 4.16: Suchen im Vollformlexikon der Nomina

```
search_ff-lexicon_LEMMA.sh "gur"

gur/gur/S-002-M_NS-;S-002-M_AcS-
guri/gur/S-002-M_GS-;S-002-M_DS-;S-002-M_AbS-;S-002-M_NS+
gurin/gur/S-002-M_AcS+
gurit/gur/S-002-M_GS+;S-002-M_DS+;S-002-M_AbS+
gurë/gur/S-002-M_NP-;S-002-M_AcP-
gurësh/gur/S-002-M_AbP-
gurët/gur/S-002-M_NP+;S-002-M_AcP+
gurëve/gur/S-002-M_GP-;S-002-M_DP-;S-002-M_GP+;S-002-M_DP+;S-002-M
                    _AbP+


search_ff-lexicon_WORD-FORM.sh "gurëve"

gurëve/gur/S-002-M_GP-;S-002-M_DP-;S-002-M_GP+;S-002-M_DP+;S-002-M
                    _AbP+
```

Die Angabe der Gruppe (`S-002` in Listing 4.16) hilft bei der eventuell später folgenden Fehlersuche (*debugging*). Sie kann einfach und problemlos ein- und ausgeschaltet werden. Für die Erstellung eines Vollformlexikons für maschinelle Sprachverarbeitung etwa wäre sie unnötig.

Listing 4.17: Suchen im Vollform-Lexikon der Verben

```
search_ff-lexicon_LEMMA.sh "lejoj"

lejo/lejoj/V-003_2P.Sg.Ipv.Prs.Act.Adm-
lejofsh/lejoj/V-003_2P.Sg.Opt.Prs.Act.Adm-
lejofsha/lejoj/V-003_1P.Sg.Opt.Prs.Act.Adm-
lejofshi/lejoj/V-003_2P.Pl.Opt.Prs.Act.Adm-
lejofshim/lejoj/V-003_1P.Pl.Opt.Prs.Act.Adm-
lejofshin/lejoj/V-003_3P.Pl.Opt.Prs.Act.Adm-
lejoftë/lejoj/V-003_3P.Sg.Opt.Prs.Act.Adm-
lejohej/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejohem/lejoj/V-003_1P.Sg.Ind.Prs.Pas.Adm-
lejohemi/lejoj/V-003_1P.Pl.Ind.Prs.Pas.Adm-
lejohen/lejoj/V-003_3P.Pl.Ind.Prs.Pas.Adm-
```

[205] Das Vollformlexikon dient als Zwischenschritt, d. h. zu Testzwecken während der Entwicklung der maschninellen Morpholgiekomponente. In diesem Sinne ist es nicht Ziel oder Teilziel der vorliegenden Arbeit.

```
lejoheni/lejoj/V-003_2P.Pl.Ind.Prs.Pas.Adm-
lejohesh/lejoj/V-003_2P.Sg.Ind.Prs.Pas.Adm-
lejohesha/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejoheshe/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejoheshim/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejoheshin/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejoheshit/lejoj/V-003_1P.Sg.Ind.Ipf.Pas.Adm+
lejohet/lejoj/V-003_3P.Sg.Ind.Prs.Pas.Adm-
lejohu/lejoj/V-003_1P.Pl.Ipv.Prs.Pas.Adm-
lejohuni/lejoj/V-003_1P.Pl.Ipv.Prs.Pas.Adm-
lejoi/lejoj/V-003_3P.Sg.Ind.Aor.Act.Adm-
lejoj/lejoj/V-003_1P.Sg.Ind.Prs.Act.Adm-;
                  V-003_1P.Sg.Sbj.Prs.Act.Adm-
lejoja/lejoj/V-003_1P.Sg.Ind.Ipf.Act.Adm-;
                  V-003_2P.Sg.Ipv-ja.Prs.Act.Adm-
lejojani/lejoj/V-003_2P.Pl.Ipv-ja.Prs.Act.Adm-
lejoje/lejoj/V-003_2P.Sg.Ind.Ipf.Act.Adm-;
                  V-003_2P.Sg.Ipv-je.Prs.Act.Adm-
lejojeni/lejoj/V-003_2P.Pl.Ipv-je.Prs.Act.Adm-
lejoji/lejoj/V-003_2P.Sg.Ipv-ji.Prs.Act.Adm-
lejojini/lejoj/V-003_2P.Pl.Ipv-ji.Prs.Act.Adm-
lejojmë/lejoj/V-003_1P.Pl.Ind.Prs.Act.Adm-;
                   V-003_1P.Pl.Sbj.Prs.Act.Adm-
lejojnë/lejoj/V-003_3P.Pl.Ind.Prs.Act.Adm-;
                   V-003_3P.Pl.Sbj.Prs.Act.Adm-
lejoju/lejoj/V-003_2P.Sg.Ipv-ju.Prs.Act.Adm-
lejojua/lejoj/V-003_2P.Sg.Ipv-jua.Prs.Act.Adm-
lejojuani/lejoj/V-003_2P.Pl.Ipv-jua.Prs.Act.Adm-
lejojuni/lejoj/V-003_2P.Pl.Ipv-ju.Prs.Act.Adm-
lejojë/lejoj/V-003_3P.Sg.Sbj.Prs.Act.Adm-
lejoma/lejoj/V-003_2P.Sg.Ipv-ma.Prs.Act.Adm-
lejomani/lejoj/V-003_2P.Pl.Ipv-ma.Prs.Act.Adm-
lejomi/lejoj/V-003_2P.Sg.Ipv-mi.Prs.Act.Adm-
lejomini/lejoj/V-003_2P.Pl.Ipv-mi.Prs.Act.Adm-
lejomë/lejoj/V-003_2P.Sg.Ipv-me.Prs.Act.Adm-
lejomëni/lejoj/V-003_2P.Pl.Ipv-me.Prs.Act.Adm-
lejon/lejoj/V-003_2P.Sg.Ind.Prs.Act.Adm-;
                  V-003_3P.Sg.Ind.Prs.Act.Adm-
lejona/lejoj/V-003_2P.Sg.Ipv-na.Prs.Act.Adm-
lejonani/lejoj/V-003_2P.Pl.Ipv-na.Prs.Act.Adm-
lejoni/lejoj/V-003_2P.Pl.Ind.Prs.Act.Adm-;
                   V-003_2P.Pl.Sbj.Prs.Act.Adm-;
                   V-003_2P.Pl.Ipv.Prs.Act.Adm-
lejonim/lejoj/V-003_1P.Pl.Ind.Ipf.Act.Adm-
lejonin/lejoj/V-003_3P.Pl.Ind.Ipf.Act.Adm-
lejonit/lejoj/V-003_2P.Pl.Ind.Ipf.Act.Adm-
lejonte/lejoj/V-003_3P.Sg.Ind.Ipf.Act.Adm-
lejosh/lejoj/V-003_2P.Sg.Sbj.Prs.Act.Adm-
lejova/lejoj/V-003_1P.Sg.Ind.Aor.Act.Adm-;
                   V-003_1P.Sg.Ind.Aor.Pas.Adm-
lejove/lejoj/V-003_2P.Sg.Ind.Aor.Act.Adm-;
                   V-003_2P.Sg.Ind.Aor.Pas.Adm-
lejua/lejoj/V-003_3P.Sg.Ind.Aor.Pas.Adm-
lejuaka/lejoj/V-003_3P.Sg.Ind.Prs.Act.Adm+
lejuakam/lejoj/V-003_1P.Sg.Ind.Prs.Act.Adm+
lejuakan/lejoj/V-003_3P.Pl.Ind.Prs.Act.Adm+
lejuake/lejoj/V-003_2P.Sg.Ind.Prs.Act.Adm+
lejuakemi/lejoj/V-003_1P.Pl.Ind.Prs.Act.Adm+
lejuakeni/lejoj/V-003_2P.Pl.Ind.Prs.Act.Adm+
lejuakësh/lejoj/V-003_3P.Sg.Ind.Ipf.Act.Adm+
lejuakësha/lejoj/V-003_1P.Sg.Ind.Ipf.Act.Adm+
lejuakëshe/lejoj/V-003_2P.Sg.Ind.Ipf.Act.Adm+
lejuakëshim/lejoj/V-003_1P.Pl.Ind.Ipf.Act.Adm+
lejuakëshin/lejoj/V-003_3P.Pl.Ind.Ipf.Act.Adm+
lejuakëshit/lejoj/V-003_2P.Pl.Ind.Ipf.Act.Adm+
lejuam/lejoj/V-003_1P.Pl.Ind.Aor.Act.Adm-;
                   V-003_1P.Pl.Ind.Aor.Pas.Adm-
lejuan/lejoj/V-003_3P.Pl.Ind.Aor.Act.Adm-;
```

```
                    V-003_3P.Pl.Ind.Aor.Pas.Adm-
lejuar/lejoj/V-003_Part
lejuat/lejoj/V-003_2P.Pl.Ind.Aor.Act.Adm-;
                    V-003_2P.Pl.Ind.Aor.Pas.Adm-


search_ff-lexicon_WORD-FORM.sh "lejohem"

lejohem/lejoj/V-003_1P.Sg.Ind.Prs.Pas.Adm-
```

Die im Beschreibungsfeld mit `Ipv-…` versehenen Einträge sind Formen des Imperativs mit klitischen Pronomina, die gemäß der albanischen Orthographieregelung zusammengeschrieben werden. [Kostallari et al. 1984: 88 ff. (P.1.: §§ 80–144)], z. B., geben nur die Formen der 2. Person Singular und der 2. Person Plural Imperativ (*lejo* und *lejo-ni* (dt. *erlauben*)) an. [Munishi 1998] listet zusätzlich die folgenden Formen: *lejo-më*, *lejo-ma*, *lejo-mi*, *lejo-ja*, *lejo-ji*, *lejo-ju*, *lejo-jua*, *lejo-më-ni*, *lejo-ma-ni*, *lejo-mi-ni*, *lejo-je-ni*, *lejo-ja-ni*, *lejo-na-ni*, *lejo-ji-ni*, *lejo-ju-ni* und *lejo-jua-ni*, was alle möglichen Fälle der klitischen Pronomina sind, welche mit dem Verb zusammengeschrieben werden, vgl. [op. cit. 24], das Verb *mbuloj* (dt. *abdecken*), das wie *lejoj* konjugiert wird (Muster 3).
Bei einigen Verben werden bestimmte Formen des Imperativs mit klitischen Pronomina selten verwendet, wie z. B.:

lan (dt. *waschen*) (3.P. …): *laj* (2.P.Sg.) / *lani* (2.P.Pl.); mit klitischen Pronomina *laje* (2.P.Sg.) / *lajeni* (2.P.Pl.); reflexiv: *lahu* (2.P.Sg.) / *lahuni* (2.P.Pl.);

rrit (dt. *erziehen, vergrößern, groß werden (lassen) u. ä.*) (3.P. …): *rrit* (2.P.Sg.) / *rritni* (2.P.Pl.); mit klitischen Pronomina *rrite* (2.P.Sg.) / *rriteni* (2.P.Pl.); reflexiv: *rritu* (2.P.Sg.) / *rrituni* (2.P.Pl.);

zë (dt. *fangen, streiten, …*) (3.P. …): *zër* (2.P.Sg.) / *zini* (2.P.Pl.); mit klitischen Pronomina *zërë[e]* (2.P.Sg.) / *zëreni* (2.P.Pl.); reflexiv: *zihu* (2.P.Sg.) / *zihuni* (2.P.Pl.);[206]

Ein Vollformlexikon kann oft nur mit enormen Schwierigkeiten erweitert werden, was einen großen Nachteil bedeutet. Auch Neologismen und okkasionelle Wortverwendungen können nicht erkannt werden.

[206] In Form eines kurzen Satzes: *Hiqmu sysh/qafe*, deutsch wortwörtlich etwa: *Geh mir aus den Augen/vom Hals*, d. h. *Geh mir aus dem Weg*.

4.15 Zusammenfassung des 4. Kapitels und Schlussbemerkungen

Die Sammlung der lexikalischen Daten und ihre Klassifikation machten einen großen Teil der vorliegenden Arbeit aus. Die Einteilung der Daten in Klassen innerhalb der Wortarten wird oft unterschätzt. Diese Klassifikation, d. h. die Zuordnung der Lemmata zu ihren jeweiligen Allomorphie- bzw. Flexionsklassen, stellt die entscheidende Information für die spätere Modellierung und Implementierung der maschinellen Morphologie dar.

5 Maschinelle Verarbeitung der Morphologie des Albanischen

Sobald die lexikalischen Daten in einer strukturierten und klassifizierten Form vorliegen, ist ein Grundstein für ein Morphologieprojekt gelegt. Ein Vollformlexikon, als eine Form der strukturierten, klassifizierten und analysierten lexikalischen Daten, wie in Kapitel 2 und 4 besprochen, kann für Zwecke der MSV eingesetzt werden, hat jedoch viele Einschränkungen. Um diese zu umgehen, bräuchte man ein Morphologie-System. Damit wäre es mittels Regeln, welche sich auf die lexikalischen Informationen stützen, möglich Wortformen zu erkennen, die nicht im Lexikon enthalten sind. Unter den vielen Systemen, die für die Implementierung einer Morphologie einer natürlichen Sprache existieren, fiel die Wahl auf XFST (Xerox Finite State Tools). Sie geht auf die in Kapitel 2, insbesondere in Abschnitt 2.3.3, genannten Eigenschaften von XFST zurück.
In diesem Kapitel wird näher auf die Implementierung der Morphologie des Albanischen eingegangen. Zunächst wird im ersten Abschnitt (5.1) AMMv beschrieben, eine Implementierung der Morphologie der Verben des Albanischen. Im zweiten Abschnitt (5.2), Organisation des Morphologie-Systems, wird ein Gesamtüberblick über die implementierte Morphologie gegeben. Es folgen Abschnitte, in denen Verben (5.3), Substantive (5.4), Adjektive (5.5), Numeralia (5.6), Pronomina (5.7), Adverbien (5.8) und schließlich die unflektierten Wortarten, die sogenannten Indeklinabilia (5.9) behandelt werden. Außerdem werden in Abschnitt 5.10 noch zusätzliche Erweiterungen erläutert. Nach der Beschreibung der einzelnen Wortarten wird die Wortbildung im Rahmen von XFST (5.11) beschrieben. Als Nächstes wird auf die Organisation der Grammatik samt ihrer Komponenten (5.12) sowie ihrer Eigenschaften eingegangen (5.13). Zum Schluss (5.14) werden eine Zusammenfassung des Kapitels und Schlussbemerkungen gegeben.

5.1 AMMv

Die Albanian Malaga-Morphology for Verbs, kurz AMMv, wurde als Magisterarbeit[207] entwickelt und implementiert. Das System, das Verben und Indeklinabilia behandelt, wurde so weit entwickelt, dass damit Formen der Verben einzeln oder in Korpora identifiziert und analysiert (getaggt) werden konnten.[208] Das System war so konzipiert, dass das Nominalsystem auf dieser Basis aufgebaut werden könnte, sobald die lexikalischen Daten und die Grammatiken entwickelt wären. Ein einfaches Lexikon hätte als Grundlage für das Nominal-System dienen sollen. Doch dem Lexikon für das Nominal-System des Albanischen fehlten wichtige Informationen, wie zur Deklination der Substantive, der Adjektive und Numeralia. Auch wichtige Informationen über das Nominalsystem, Kasus, Genus und insbesondere Numerus der Substantive fehlten. So war es keinesfalls möglich, diesen Teilbereich der Morphologie im Rahmen der Magisterarbeit zu behandeln.
Im Folgenden wird das Verbalsystem, entwickelt im Rahmen von *Malaga*, erklärt. Tabelle 5.1 zeigt die linguistische (Morfologjia) und die traditionelle Wörterbuch-Trennung (Buchholz et al.) der Wörter.

Tabelle 5.1: Trennung der Wortformen in Stamm- und Suffixmorphem.

punoj: akt., n.admir., ind., aor.		
	Morfologjia	Buchholz et al.
1. P. Sg.	*puno*|*va*	*pun*|*ova*
2. P. Sg.	*puno*|*ve*	*pun*|*ove*
3. P. Sg.	*puno*|*i*	*pun*|*oi*
1. P. Pl.	*punua*|*m*	*pun*|*uam*
2. P. Pl.	*punua*|*t*	*pun*|*uat*
3. P. Pl.	*punua*|*n*	*pun*|*uan*

Auf dieser Basis wurden die Segmente (Buchholz et al.) mit den grammatischen Eigenschaften, wie in Abbildung 5.1 dargestellt, versehen. So war eine direkte Konkatenation der beiden Teile (Segmente) des Quasi-Stamms, z. B. *pun* und des Quasi-Suffix o*va* möglich. Für die Erkennung der so entstandenen Wortformen wäre dieser Schritt ausreichend. Diese Form der Konkatenation ermöglicht eine Analyse der Wortformen wie mit einem Vollformlexikon.

[207] Vgl. [Kabashi 2003].
[208] Vgl. [Op. Cit.: 75–80].

Abbildung 5.1: Buchholz/Fiedler-Methode der Konkatenation.

$$({}^{i})\ pun_{[CT1]} + \begin{cases} -oj_{[CT1]} & \Rightarrow & pun|\texttt{o}j_{(Sg1PresIndNAdAkv)} \\ -on_{[CT1]} & \Rightarrow & pun|\texttt{o}n_{(Sg23PresIndNAdAkv)} \\ \ldots & \Rightarrow & \ldots \\ -uan_{[CT1]} & \Rightarrow & pun|\texttt{ua}n_{(Pl2AorIndNAdAkv)} \\ \ldots & \Rightarrow & \ldots \end{cases}$$

Die Segmente in Abbildung 5.1, die hervorgehoben sind, d. h. o und ua, auf der rechten Seite des Zeichens „|“ , gehören, linguistisch gesehen, nicht zu den Suffixen, sondern bilden zusammen mit dem Teil vor dem Zeichen „|“ den Stamm bzw. die Stammallomorphe.
Im nächsten Schritt, wie in Abbildung 5.2 gezeigt, werden die Segmente rechts des Zeichens „|“ gruppiert. [CT1] steht dabei für CunjunctionType, dt. *Konjugationstyp* und dient dazu, die Lemmata innerhalb der Wortart anhand ihrer Flexionseigenschaften zu klassifizieren.

Abbildung 5.2: Trennung der Alternationssuffixe in Gruppen (1).

$$({}^{ii})\ pun_{[CT1]} + \begin{cases} \begin{cases} -oj_{[CT1]} & \Rightarrow & pun|\texttt{o}j \\ -on_{[CT1]} & \Rightarrow & pun|\texttt{o}n \\ \ldots & \Rightarrow & \ldots \end{cases} \\ \begin{cases} -uan_{[CT1]} & \Rightarrow & pun|\texttt{ua}n \\ \ldots & \Rightarrow & \ldots \end{cases} \\ \{\ \ldots \ \Rightarrow \ \ldots \end{cases}$$

In Abbildung 5.3 sind alle vier möglichen Gruppen der Teilsegmente vom Verb *punoj* bzw. [CT1] gezeigt. Die Segmente zwischen den beiden Zeichen „|“ bilden zusammen mit dem Segment *pun* den Stamm. Das heißt das bisherige Segment o*j*, vgl. Abbildung 5.1 wird zerlegt, wobei „|o“ mit dem Segment *pun* amalgamiert wird und *j* als Suffix, im linguistischen Sinne, übrig bleibt. Ziel ist es, linguistisch motivierte Morpheme zu bauen, die sowohl bei der Konkatenation transparent sind, als auch bei den Prozessen der Wortbildung als Bausteine dienen können.

Abbildung 5.3: Trennung der Alternationssuffixe in Gruppen (2).

$$
(^{iii})\ pun_{[CT1]} + \left\{ \begin{array}{ll}
+ \left\{ \begin{array}{lcl} \texttt{o}|j_{[CT1]} & \Rightarrow & pun|\texttt{o}|j \\ \texttt{o}|n_{[CT1]} & \Rightarrow & pun|\texttt{o}|n \\ \ldots & \Rightarrow & \ldots \end{array} \right. \\
+ \left\{ \begin{array}{lcl} \texttt{ua}|n_{[CT1]} & \Rightarrow & pun|\texttt{ua}|n \\ \ldots & \Rightarrow & \ldots \end{array} \right. \\
+ \left\{ \begin{array}{lcl} \texttt{of}|t\ddot{e}_{[CT1]} & \Rightarrow & pun|\texttt{of}|t\ddot{e} \end{array} \right. \\
+ \left\{ \begin{array}{lcl} \texttt{ofsh}|a_{[CT1]} & \Rightarrow & pun|\texttt{ofsh}|a \\ \ldots & \Rightarrow & \ldots \end{array} \right.
\end{array} \right.
$$

Vor der Weiterverarbeitung werden die Segmente mit den nötigen Informationen versehen, um keine Information beim Spalten bzw. Zusammensetzen zu verlieren. Dies ist in Abbildung 5.4 dargestellt, vgl. die mit [_o] bzw. mit [_ua] markierten Stellen. Von den ursprünglichen Segmenten rechts vom Zeichen „|“ , d. h. o*j*, o*n*, ..., ua*n* usw., vgl. Abbildung 5.3, werden das o bzw. ua usw. abgespalten und dem Segment *pun* hinzugefügt, wobei beide Seiten mit den entsprechenden Informationen zu versehen sind.

Abbildung 5.4: Trennung der Alternationssuffixe in Segmente und Suffixe.

$$
(^{iv})\ pun_{[CT1]} + \left\{ \begin{array}{ll}
\leftarrow \texttt{o}_{[\_o]} + \left\{ \begin{array}{lcl} -j_{[CT1][\_o]} & \Rightarrow & pun|\texttt{o}|j \\ -n_{[CT1][\_o]} & \Rightarrow & pun|\texttt{o}|n \\ \ldots & \Rightarrow & \ldots \end{array} \right. \\
\leftarrow \texttt{ua}_{[\_ua]} + \left\{ \begin{array}{lcl} -n_{[CT1][\_ua]} & \Rightarrow & pun|\texttt{ua}|n \\ \ldots & \Rightarrow & \ldots \end{array} \right. \\
\leftarrow \ldots + \left\{ \begin{array}{lcl} \ldots & \Rightarrow & \ldots \end{array} \right.
\end{array} \right.
$$

Bei diesem Schritt ergibt sich die Situation, dass ein Suffix, wie *n* in Abbildung 5.5, aus verschiedenen Gruppen kommt. Um Informationsverlust zu

vermeiden, wird in diesem Fall das Suffix mit den ursprünglichen grammatischen Eigenschaften aus allen Gruppen versehen, d. h., in dem Fall erhält *n* die Informationen Sg2... und Sg3... aus der Gruppe [_o] und Pl3... aus der Gruppe [_ua] usw., die in Multikategorien aufgelistet werden.

Abbildung 5.5: Konkatenation der Stämme mit den Suffixen.

$$(1') \; puno_{[Stem][CT1\_o]} + \begin{cases} -j_{[CT1\_o,...]} & \Rightarrow & puno|j \\ -n(^{*}) & \Rightarrow & puno|n \longleftarrow 1\ (Sg2,...) \\ -n(^{*\prime}) & \Rightarrow & puno|n \longleftarrow 1'\ (Sg3,...) \\ -jmë & \Rightarrow & puno|jmë \\ ... & \Rightarrow & ... \end{cases}$$

$$(1'') \; puna_{[Stem][CT1\_ua]} + \begin{cases} -m_{[CT1\_ua,...]} & \Rightarrow & punua|m \\ -t & \Rightarrow & punua|t \\ -n(^{**}) & \Rightarrow & punua|n \longleftarrow 2 \\ ... & \Rightarrow & ... \end{cases}$$

In Abbildung 5.6 sind diese Multikategorien am Beispiel des Suffix *n* zu sehen. Mit der Markierung dieser Kategorien ([_o], Sg2... usw.) wird die Kombination festgestellt. *n* kann nur mit Morphemen mit diesen Kategorien kombiniert werden. Die entsprechenden grammatischen Informationen sind beim Suffix gespeichert.

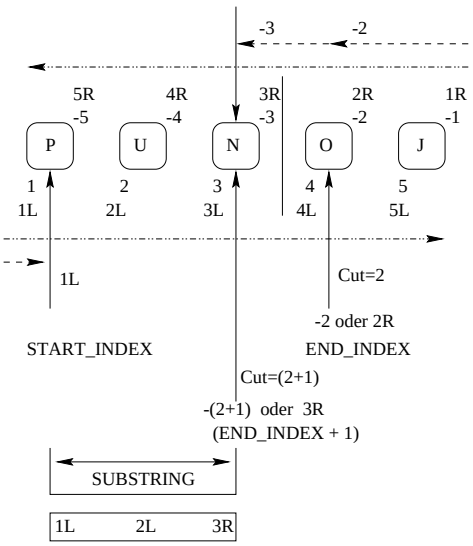
Abbildung 5.6: Multikategorien der Suffixe.

$$(2) \; n_{[Suffix][<[CT1\_o],[CT1\_ua],...>]} \rightarrow \begin{cases} n_{[CT1\_o]}(^{*}) & \longleftarrow 1 \\ n_{[CT1\_o]}(^{*\prime}) & \longleftarrow 1' \\ n_{[CT1\_ua]}(^{**}) & \longleftarrow 2 \\ ... & ... \end{cases}$$

Einen Vergleich mit dem Deutschen soll das Beispiel *schreiben* ermöglichen. Die Formen des Präsens sind (*ich*) *schreib*|*e*, (*du*) *schreib*|*st*, (*er*, *sie*, *es*) *schreib*|*t*, (*wir*) *schreib*|*en*, (*ihr*) *schreib*|*t* und (*sie*) *schreib*|*en*, die Formen des Präteritums (*ich*) *schrieb*|0, (*du*) *schrieb*|*st*, (*er*, *sie*, *es*) *schrieb*|0, usw., wobei der Stammvokal umgelautet wird. Nach dem obigen Vorgehen wären der Ausgangspunkt Stämme, die segmentiert wie folgt aussähen: *schr*|*eiben* und

schr|*ieben*, vgl. 5.1 und 5.2. Als nächstes wäre es nötig, sie in *schr*|*eib*|*en* und *schr*|*ieb*|*en* umzuwandeln, vgl. 5.3 und 5.4. Dabei wären die Stämme *schrieb* und *schreib* und die Suffixe gewonnen, welche entsprechend grammatisch zu markieren wären, vgl. 5.5 und 5.6. Zum Beispiel würde *en* mit Informationen versehen werden, die angeben, mit welchen Stämmen es kombiniert werden könnte und mit welchen grammatischen Eigenschaften es versehen wird, d. h. *schreib* mit Infin., 1./3.P.Pl.Ind.Präs.Akt., *schrieb* mit 1./3.P.Pl.Ind.Prät.Akt., usw. Ebenso wären die Stämme und die Suffixe, falls möglich, als Multikategorien kodiert, um sie besser benutzen zu können bzw. einen besseren Überblick über sie zu bekommen. Die Segmentierung, welche rein graphembasiert ist, wird in eine linguistisch motivierte Schreibung umgewandelt, indem auf der einen Seite die Allomorphstämme gebildet werden, auf der anderen Seite die Suffixe.
In Abbildung 5.7 ist schematisch dargestellt, wie der erste Schritt dieser Operationen im Rahmen von *Malaga*, u. a. mithilfe von *Substring*, realisiert wird. Vgl. hierzu auch die Abbildungen 5.1 und 5.2.

Abbildung 5.7: Substring in Malaga.



Um einen Eindruck zu gewinnen, wie dies in Malaga implementiert wurde, ist in Listing 5.1 ein Ausschnitt des Malaga-Codes, welcher die Stammallomorphe bildet, angegeben. Die Variable Cut trägt die Information, welche

Teile von *rechts* nach *links*, vgl. Abbildungen 5.1 bis 5.6, übertragen werden. Diese Information ist üblicherweise bei jedem Lexikoneintrag vorhanden.

Listing 5.1: Substring in MALAGA

```
IF Cut IN $LexEntry

THEN $Stem :=
       substring($LexEntry.Lemma, 1L, -($LexEntry.Cut + 1));

ELSE $Stem :=
       $LexEntry.Lemma;

END IF;
```

Beim Verb *punoj* hat die Cut-Variable den Wert 2, d. h. sie enthält das Suffix *oj*, das wie oben beschrieben verarbeitet wird – also zuerst in *puno* und *j* aufgeteilt und schließlich kombiniert wird.
In Abbildung 5.8 ist in allgemeiner Form die Konkatenation für den Stamm *punua* und das Suffix *n* dargestellt.

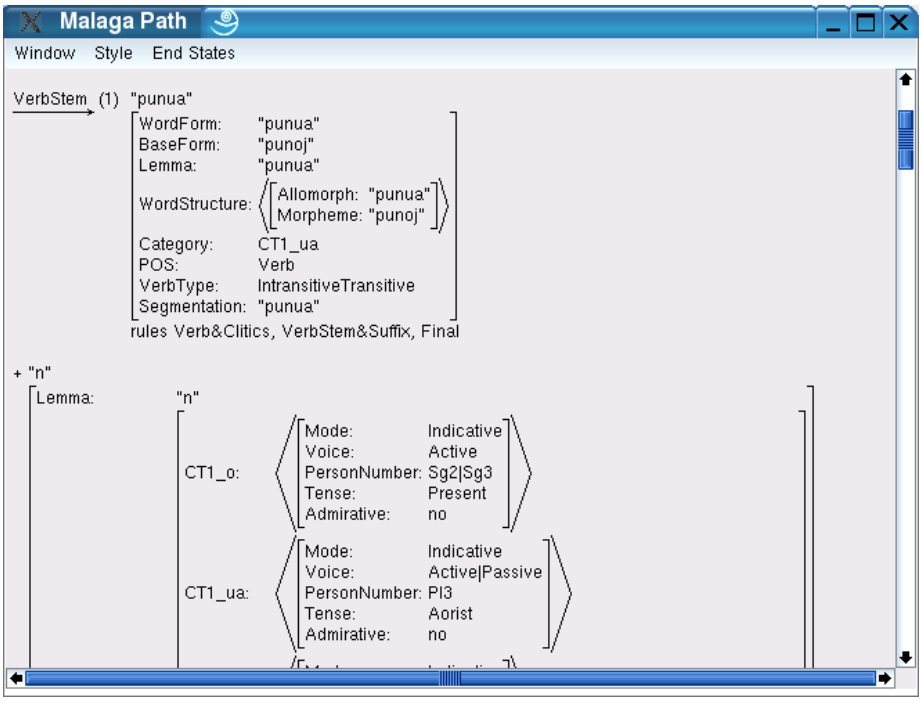
Abbildung 5.8: Allgemeine Form der Konkatenation.

$$
\begin{aligned}
& \textit{VerbStem} \left[\begin{array}{l} \texttt{Surface}\text{: Allomorph}_{[\textit{VerbStem}]} \\ \texttt{Rule}\text{: VerbStem} \\ \texttt{Category}\text{: CT\_x}_{[\textit{VerbStem}]\ (\ [\textit{Suffix}]' \mid [\textit{Clitic}]' \mid \textit{Final}\ )} \\ \texttt{Following}\text{: VerbStem\&Suffix, Clitic, NI, Final} \end{array}\right] \\
+ \quad & \textit{Suffix} \left[\begin{array}{l} \texttt{Surface}\text{: Allomorph}_{[\textit{Suffix}]} \\ \texttt{Rule}\text{: VerbStem\&Suffix} \\ \texttt{Category}\text{: CT\_x}_{[\textit{Result}][\textit{Suffix}]'} \\ \texttt{Following}\text{: Final} \\ \texttt{Result}\text{: Result}_{[CT\_x]} \end{array}\right] \\
\rightarrow \quad & \textit{Final} \left[\begin{array}{l} \texttt{Surface}\text{: WordForm}_{[\textit{Allomoph}_{[\textit{VerbStem}]}+\textit{Allomoph}_{[\textit{Suffix}]}]} \\ \texttt{Category}\text{: CT}_{[\textit{Result}]} \\ \texttt{Rule}\text{: Final,} \\ \texttt{AnalysisType}\text{: Parsed} \end{array}\right]
\end{aligned}
$$

Dabei ist zu sehen, dass der Pfad durch die Regeln (vgl. das Attribut Rule), die folgen können (Following), und die Kategorie der Allomorphe (Category) bestimmt wird. Wenn die Fortsetzung mit dem Suffix, definiert in VerbStem,

möglich ist, werden noch die Kongruenzen, definiert in Category, überprüft. Bei einer Übereinstimmung werden die Allomorphe zu einer Wortform konkateniert. Attribut-Wert-Paare, die nur temporäre Funktion hatten, werden gelöscht.
Abbildung 5.9 zeigt einen Screenshot aus der Entwicklungsumgebung von Malaga. Zu sehen ist die Konkatenation des Stamms *puno* und des Suffix *n*. Der Schlüssel der Kombination ist die Kategorie CT1_ua, die bei beiden Elementen vorhanden sein muss.

Abbildung 5.9: Ein Konkatenationsschritt unter MALAGA.



Für AMMv wurden die Informationen zu Verbklassen aus [BUCHHOLZ ET AL. 1993] entnommen. Als Lexikon-Grundlage dienten vom Autor der vorliegenden Arbeit selbst erstellte lexikalische Daten sowie einige von Marko Snoj zur Verfügung gestellte lexikalische Daten, die er für die Erstellung des Werks *Rückläufiges Wörterbuch der albanischen Sprache*[209] vorbereitete.

[209] Die Daten stammten aus dem Jahr 1993.

Sie stimmen größtenteils mit den Daten im später (1994) publizierten Werk überein.

5.2 Organisation des Morphologie-Systems

Das Morphologie-System wurde nach dem Prinzip „so einfach wie möglich“ konzipiert und entwickelt. Dies kann in kompakter Form folgendermaßen zusammengefasst werden:

- Das Morphologiesystem orientiert sich an den traditionellen Wortarten, so wurde z. B. für Verben eine Verb-Grammatik erstellt, für Substantive eine Substantiv-Grammatik usw.
- Die Dateistruktur entspricht der Gruppierung in Wortarten und deren Subklassifizierung.
- Ausnahmen wurden soweit wie möglich in gleicher und ähnlicher Form wie die Hauptphänomene der Wortarten und deren morphologische Eigenschaften modelliert und behandelt.
- Ebenso entsprechen die Ergänzungen, wie z. B. die Behandlung der Namen oder Interpunktionszeichen, in ihrer Modellierung den anderen Teilgrammatiken in ihrer Lexikon- und Regelorganisation.
- Das Lexikon entspricht dem Wortschatz eines Universalwörterbuches mit einem Umfang von ca. 50 000 Einträgen.
- Es wird neben der Flexion auch die Ableitung (Derivation) als Prozess der Wortbildung behandelt. Modelliert werden die produktiven Haupttypen.
- Es wurde neben den lexikalischen Einträgen auch ein Hypothesensystem entwickelt, das für unbekannte Wortformen mögliche Analysen vorschlägt.
- Es wird ein Testsystem gebaut bzw. werden die Tools, die das XFST-Paket zur Verfügung stellt, benutzt, um das erstellte Morphologie-System zu testen.

5.3 Verben im Rahmen von XFST

Nachdem das Ziel der Entwicklung des Morphologie-Systems eine vollständige Abdeckung sowie ein einheitliches System war, wurde AMMv beiseite gelassen. Die Verben wurden neu organisiert und im Rahmen von XFST implementiert. Somit wird eine Kompatibilität mit den zahlreichen lexikalischen Einträgen der Wortarten Substantiv, Adjektiv, Numeral und Adverb, die zum ersten Mal implementiert werden, erreicht.

5.3.1 Reorganisation der lexikalischen Daten der Verben

Die Implementierung der Verben im Rahmen von XFST erforderte eine Anpassung der lexikalischen Daten. Um einige untergegangene Fehler in AMMv nicht in die Reimplementierung mitzuschleppen, wurden die lexikalischen Daten der Verben neu be- und verarbeitet. Als Erstes wurden ihre Klassen neu berechnet. Die Klassen-Informationen der Verben ermöglichen und erleichtern ihre Modellierung allgemein, wie auch im Rahmen von XFST.

Listing 5.2 zeigt die Klassifikation der Verben (1. Spalte, V-*nnn*) anhand von Eigenschaften der Valenz sowie einiger Suffixe (2. Spalte). Zum Beispiel wurden die Suffixe |∼a|∼ur je nach Valenzangabe itr., tr., tr., itr., refl. usw. eingruppiert, vgl. V-003, V-006, V-011 und V-014.

Listing 5.2: Verbalklassen

```
1  V-001   tr.|~óva|~úar          (1')
2  V-002   refl., pass.||
3  V-003   tr.|~a|~ur             (2')
4  V-004   pass.||
5  V-005   refl.||
6  V-006   refl.|~a|~ur           (2'')
7  V-007   intr.|~óva|~úar        (1'')
8  V-008   tr., intr.|~óva|~úar
9  V-009   intr.|~ói|~úar
10 V-010   intr., tr.|~óva|~úar   (1''')
11 V-011   intr.|~a|~ur           (2''')
12 V-012   refl.|~óva|~úar
13 V-013   tr.|~va|~rë
14 V-014   tr., intr.|~a|~ur      (2'''')
15 V-015   ||
16 V-016   tr.|~éva|~ýer
17 V-017   intr.|~i|~ur
18 V-018   refl.|~|~ur
19 V-019   refl.|~úa|~úar
20 V-020   intr.|~va|~rë
21 ...     ...
```

5.3.2 Aufbau der Verb-Grammatik

Wie in Abschnitt 5.2 angedeutet, wurde das Projekt in erster Linie auf der Grundlage der Wortarten aufgebaut. Die Grammatik besteht aus Teilgrammatiken, die für Teilaufgaben zuständig sind. So deckt die Teilgrammatik der Verben die Verben ab, in erster Linie deren Flexion. Die Verbgrammatik besteht aus zwei Typen von Dateien, den LEXC-Dateien, die für das Lexikon gedacht sind, sowie den XFST-Dateien, die für Regeln gedacht sind. Letztere verarbeiten die lexikalischen Daten.

5.3.3 Einträge der Verben in LEXC

Für jede Verbklasse gab es zunächst eine eigene Datei. Zum Beispiel waren die Verben, welche die Eigenschaften tr.|∼óva|∼úar besitzen, in einer Datei aufgelistet und entsprechend ihrer Flexion und anderer Eigenschaften aufgebaut. In einer zweiten Version wurden die Lexikoneinträge der Verben in einer einzigen Datei umformatiert und angepasst. Diese Version hat gegenüber der ersten Version den Vorteil, dass im Falle einer gewünschten Änderung bzw. Wartung nur eine Datei zum Bearbeiten geöffnet werden muss, was außerdem einen kompakten Überblick ermöglicht.

Listing 5.3 zeigt Ausschnitte aus der LEXC-Datei der Verben:

Listing 5.3: Lexikoneinträge der Verben in LEXC.

```
Multichar_Symbols
...
+V
+Pres +Impf +Aor
+Ind +Subjv +Impv +Opt
+NonAdm +Adm
+Act +Pass
+Part
...
+pC
...
+A_kam +A_jam
...
+A_null
...
+Neg
+Which
+What
...
LEXICON Root

jam+A_jam+V C_jam;
kam+A_kam+V C_kam;
...
vetëton+A_mbul+V C_agullohet;
veton+A_mbul+V C_agullohet;
vezullon+A_mbul+V C_agullohet;
xixëllon+A_mbul+V C_agullohet;
xixëmon+A_mbul+V C_agullohet;
```

```
xixon+A_mbul+V C_agullohet;
...
nuhat+A_null+V C_rrit;
nxit+A_null+V C_rrit;
padit+A_null+V C_rrit;
pahit+A_null+V C_rrit;
...

LEXICON C_jam

!+Part^0 #;

! Verb_Endungen_CT-001_ja_jam
+1P+Sg+Ind+Pres+Act+NonAdm^m #;
+3P+Pl+Ind+Pres+Act+NonAdm^në #;
...
! Verb_Endungen_CT-001_je_jam
+2P+Sg+Ind+Pres+Act+NonAdm^0 #;
+1P+Pl+Ind+Pres+Act+NonAdm^mi #;
+2P+Pl+Ind+Pres+Act+NonAdm^ni #;
...
! Verb_Endungen_CT-001_është_jam
+3P+Sg+Ind+Pres+Act+NonAdm^0 #;
...

LEXICON C_kam

+Part^ur #;

! Verb_Endungen_CT-001_ka_kam
+1P+Sg+Ind+Pres+Act+NonAdm^m #;
+3P+Sg+Ind+Pres+Act+NonAdm^0 #;
+3P+Pl+Ind+Pres+Act+NonAdm^në #;
...
! Verb_Endungen_CT-001_ke_kam
+2P+Sg+Ind+Pres+Act+NonAdm^0 #;
+1P+Pl+Ind+Pres+Act+NonAdm^mi #;
+2P+Pl+Ind+Pres+Act+NonAdm^ni #;
...
```

In den Zeilen 1–18 in Listing 5.3 ist ein Abschnitt angegeben, in dem die Symbole definiert sind. Sie dienen zur Modellierung der grammatischen Eigenschaften der Wortformen. Das Hauptlexikon wird in Zeile 20 definiert. Die Lexikoneinträge sind ab Zeile 22 aufgelistet, wobei sie in Gruppen aufgeteilt sind. Die Hilfsverben (Zeilen 22-23) bspw. sind mit anderen Eigenschaften versehen als bspw. die Verben in Zeile 25 oder in 32. Die Symbole `+A_jam` und `C_jam` in Zeile 25 bedeuten, dass die Stammallomorphe des Verbs durch Allomorphieregel `+A_jam` gebildet werden, während die Flexionssuffixe durch das `C_jam`-Teillexikon (Zeilen 38–53) definiert werden. Die einzelnen Formen des Flexionsparadigmas werden je nach Stamm definiert, vgl. *je*, Zeilen 47-49, Suffixe `0`, *mi* und *ni*, bzw. *është*, Zeile 52, kein Suffix. Die Allomorphieregeln werden in einer separaten Datei, vgl. Listing 5.4, behandelt.

5.3.4 Regeln für Verben in XFST

In der Regeldatei (grammar.xfst) werden die LEXC-Dateien eingebunden. Dazu kommen auch Dateien, die zwecks besseren Überblicks und leichterer Modellierung ausgelagert wurden, wie die Variablen (var.xfst) oder die Allomorph-Modellierung (allo.xfst). Die Ausgabe der Kompilierung (Übersetzung) der Verbgrammatik wird in eine Datei (verbs.fst) geschrieben. Sie kann (mit dem Befehl xfst –s verbs.fst) auch als Kompilat ausgeführt werden.

Listing 5.4: Allomorphieregeln beim Hilfsverb *jam*.

```
read regex


{jam} -> {ja}    || _ "+A_jam" "+V" "+1P" "+Sg" "+Ind"
                      "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {ja}    || _ "+A_jam" "+V" "+3P" "+Pl" "+Ind"
                      "+Pres" "+Act" "+NonAdm"  .o.

{jam} -> {je}    || _ "+A_jam" "+V" "+2P" "+Sg" "+Ind"
                      "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {je}    || _ "+A_jam" "+V" ["+1P"|"+2P"] "+Pl"
                      "+Ind" "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {je}    || _ "+A_jam" "+V" ?^2 "+Subjv" "+Pres"
                      "+Act" "+NonAdm" .o.

{jam} -> {është} || _ "+A_jam" "+V" "+3P" "+Sg" "+Ind"
                      "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {ish}   || _ "+A_jam" "+V" ?^2 "+Ind" "+Impf"
                      "+Act" "+NonAdm" .o.

{jam} -> {qe}    || _ "+A_jam" "+V" ?^2 "+Ind" "+Aor"
                      "+Act" "+NonAdm" .o.

{jam} -> {qenë}  || _ "+A_jam" "+V" "+Part" .o.

{jam} -> {qen}   || _ "+A_jam" "+V" ?^2 ["+Pres"|"+Impf"]
                      "+Act" "+Adm" .o.

{jam} -> {ji}    || _ "+A_jam" "+V" "+2P" ?^1 "+Impv"
                      "+Act" "+NonAdm" .o.

{jam} -> {qofsh} || _ "+A_jam" "+V" ["+1P"|"+2P"] ?^1
                      "+Opt" "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {qofsh} || _ "+A_jam" "+V" ?^1 "+Pl" "+Opt"
                      "+Pres" "+Act" "+NonAdm" .o.

{jam} -> {qof}   || _ "+A_jam" "+V" "+3P" "+Sg" "+Opt"
                      "+Pres" "+Act" "+NonAdm" ;
```

Die Allomorphieregeln dienen dazu die Allomorphie (Alternation im Stamm) bei Verben und Substantiven zu behandeln. `jam` wird zu `je` umgewandelt, falls die Bedingung(en), markiert durch `|| _` und die Eigenschaften `+A_jam +V` ... (Zeilen 10–11, 13–14 bzw. 16–17), jeweils erfüllt werden.

5.3.5 Implementierung der klitischen Pronomina

Formen wie *lexojeni*, dt. *lest es*, *dorëzohuni*, dt. *ergebt euch/ergeben Sie sich* usw. stellen eine Besonderheit in der albanischen Verbmorphologie dar. Wie in den Abschnitten 3.3.4, 4.5, 4.12 und 4.14 erläutert, können Verbformen im Modus Imperativ auch klitische Pronomina enthalten.

In Listing 5.5 sind einige Abfragen der Verbformen mit und ohne klitische Pronomina an die XFST-Verbgrammatik aufgeführt.

Listing 5.5: Klitische Pronomina unter xfst

```
xfst[1]: up dorëzoj
dorëzoj+V+1P+Sg+Subjv+Pres+Act+NonAdm
dorëzoj+V+1P+Sg+Ind+Pres+Act+NonAdm

xfst[1]: up dorëzoni
dorëzoj+V+2P+Pl+Ind+Pres+Act+NonAdm
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm

xfst[1]: up dorëzo
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm

xfst[1]: up dorëzohem
dorëzoj+V+1P+Sg+Ind+Pres+Pass+NonAdm

xfst[1]: up dorëzohuni
dorëzoj+V+2P+Pl+Impv+Pres+Pass+NonAdm

xfst[1]: up dorëzohu
dorëzoj+V+2P+Sg+Impv+Pres+Pass+NonAdm

xfst[1]: up dorëzomi
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+më_+_i+1P+Sg+Dat+3P+Masc+
                    Pl+Acc

xfst[1]: up dorëzomini
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm+pC+më_+_i+1P+Sg+Dat+3P+Masc+
                    Pl+Acc

xfst[1]: up dorëzoma
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+më_+_e+1P+Sg+Dat+3P+Masc+
                    Sg+Acc
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+më_+_e+1P+Sg+Dat+3P+Fem+Sg
                    +Acc

xfst[1]: up dorëzomani
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm+pC+më_+_e+1P+Sg+Dat+3P+Masc+
                    Sg+Acc
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm+pC+më_+_e+1P+Sg+Dat+3P+Fem+Sg
                    +Acc

xfst[1]: up dorëzojua
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+ju_+_e+2P+Pl+Dat+3P+Masc+
                    Sg+Acc
```

```
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+ju_+_e+2P+Pl+Dat+3P+Fem+Sg
                    +Acc
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+ju_+_i+2P+Pl+Dat+3P+Masc+
                    Pl+Acc
dorëzoj+V+2P+Sg+Impv+Pres+Act+NonAdm+pC+ju_+_i+2P+Pl+Dat+3P+Fem+Pl
                    +Acc

xfst[1]: up dorëzojani
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm+pC+ju_+_e+2P+Pl+Dat+3P+Masc+
                    Sg+Acc
dorëzoj+V+2P+Pl+Impv+Pres+Act+NonAdm+pC+ju_+_e+2P+Pl+Dat+3P+Fem+Sg
                    +Acc
...
```

Wie in Listing 5.5 zu sehen ist, sind die Analyseausgaben für einzelne Wortformen unterschiedlich lang. Die kurzen Analysen, vgl. Zeilen 1–19, entsprechen der üblichen Reihung der Eigenschaften, Grundform, +Wortart, +Person, +Numerus, +Modus, +Tempus, +Genus verbi, +Admirativität. Bei den langen Analysen, vgl. Zeilen 21–56, sind zusätzlich dazu noch die Eigenschaften der klitischen Pronomina vorhanden, die im Anschluss auf das Merkmal +NonAdm folgen. Die klitischen Pronomina sind mit dem Symbol +pC gekennzeichnet. Dem Symbol folgen die Eigenschaften des klitischen Pronomens. Bei einfachen klitischen Pronomina folgen Kasus, Numerus und Person, ggf. auch Genus. Bei komplexeren klitischen Pronomina sind die jeweiligen Bestandteile angegeben, wie z. B. +m'i+më+i, gefolgt von Kasus, Numerus und Person, ggf. auch Genus. +m'i+më+i steht für das klitische Pronomen m'i, welches ein Ergebnis der Kombination von më und i ist.

5.3.6 Verben mit einer oder mehreren vorangestellten Partikeln

Die Modellierung der Verbformen mitsamt ihren Partikeln ist möglich, etwa *të␣punojë* bzw. *u␣punua*, vgl. Listing 5.6.

Listing 5.6: Verben mit vorangestellten Partikeln

```
xfst[1]: up të punoj
punoj+V+1P+Sg+Subjv+Pres+Act+NonAdm
xfst[1]: up të punosh
punoj+V+2P+Sg+Subjv+Pres+Act+NonAdm
xfst[1]: up të punojë
punoj+V+3P+Sg+Subjv+Pres+Act+NonAdm
xfst[1]: up të punojmë
punoj+V+1P+Pl+Subjv+Pres+Act+NonAdm
xfst[1]: up të punoni
punoj+V+2P+Pl+Subjv+Pres+Act+NonAdm
xfst[1]: up të punojnë
punoj+V+3P+Pl+Subjv+Pres+Act+NonAdm

xfst[1]: down punoj+V+1P+Sg+Subjv+Pres+Act+NonAdm
të punoj
xfst[1]: down punoj+V+2P+Sg+Subjv+Pres+Act+NonAdm
të punosh
```

```
xfst[1]: down punoj+V+3P+Sg+Subjv+Pres+Act+NonAdm
të punojë
xfst[1]: down punoj+V+1P+Pl+Subjv+Pres+Act+NonAdm
të punojmë
xfst[1]: down punoj+V+2P+Pl+Subjv+Pres+Act+NonAdm
të punoni
xfst[1]: down punoj+V+3P+Pl+Subjv+Pres+Act+NonAdm
të punojnë

xfst[1]: up u punua
punoj+V+3P+Sg+Ind+Aor+Pass+NonAdm
xfst[1]: up u punuan
punoj+V+3P+Pl+Ind+Aor+Pass+NonAdm

xfst[1]: down punoj+V+3P+Sg+Ind+Aor+Pass+NonAdm
u punua
xfst[1]: down punoj+V+3P+Pl+Ind+Aor+Pass+NonAdm
u punuan
```

Die Formen *do të punojë* bzw. *do të kem punuar* werden nicht behandelt, da sie auf der Abstraktionsebene der Syntax besser verarbeitet werden können. Selbst die Möglichkeit der Kombination von *të* mit einem klitischen Pronomen, wie z. B. *ta* (← *të*+*e*) *punojë*, lässt sich besser in der Syntax behandeln. Die Partikeln *të* und *u* sind für die Produktion eine Hilfe, wenn die Grammatik für didaktische Zwecke eingesetzt wird. Zu vorangestellten Partikeln vgl. auch die Modellierung der Adjektive in Abschnitt 5.5. Die Partikeln sind in die eigenständige Datei verb-particles.xfst ausgelagert. Einen Ausschnitt der Datei zeigt Listing 5.7.

Listing 5.7: Ausschnitt: Modellierung der Verbpartikeln

```
...
! "të" vor Konjunktivformen
! read regex [..] (->) {të} " " || .#. _ ?+ "+Subjv" ;
  read regex [..]  ->  {të} " " || .#. _ ?+ "+Subjv" ;

! "u" vor Aorist Passiv
! read regex [..] (->)  {u} " " || .#. _ ?+ "+Aor" "+Pass" ;
  read regex [..]  ->   {u} " " || .#. _ ?+ "+Aor" "+Pass" ;
...
```

Die Partikel *të* wird nur dann eingefügt, wenn auf der Oberseite das Symbol +Subjv vorhanden ist. Genauso wird die Partikel *u* eingesetzt, nämlich wenn die nötigen Eigenschaften +Aor und +Pass vorhanden sind. Im Falle der Übernahme der Zeilen 3 und 7 bei gleichzeitiger Auskommentierung der Zeilen 4 und 8 wären die jeweiligen Partikeln fakultativ, vgl. insbesondere den Pfeil nach rechts (->), der einmal ohne runde Klammern und einmal mit ihnen vorkommt. Ohne die Einbindung dieser Teilregeln würden die Konjunktivformen auch ohne die jeweiligen Partikeln analysiert und produziert.

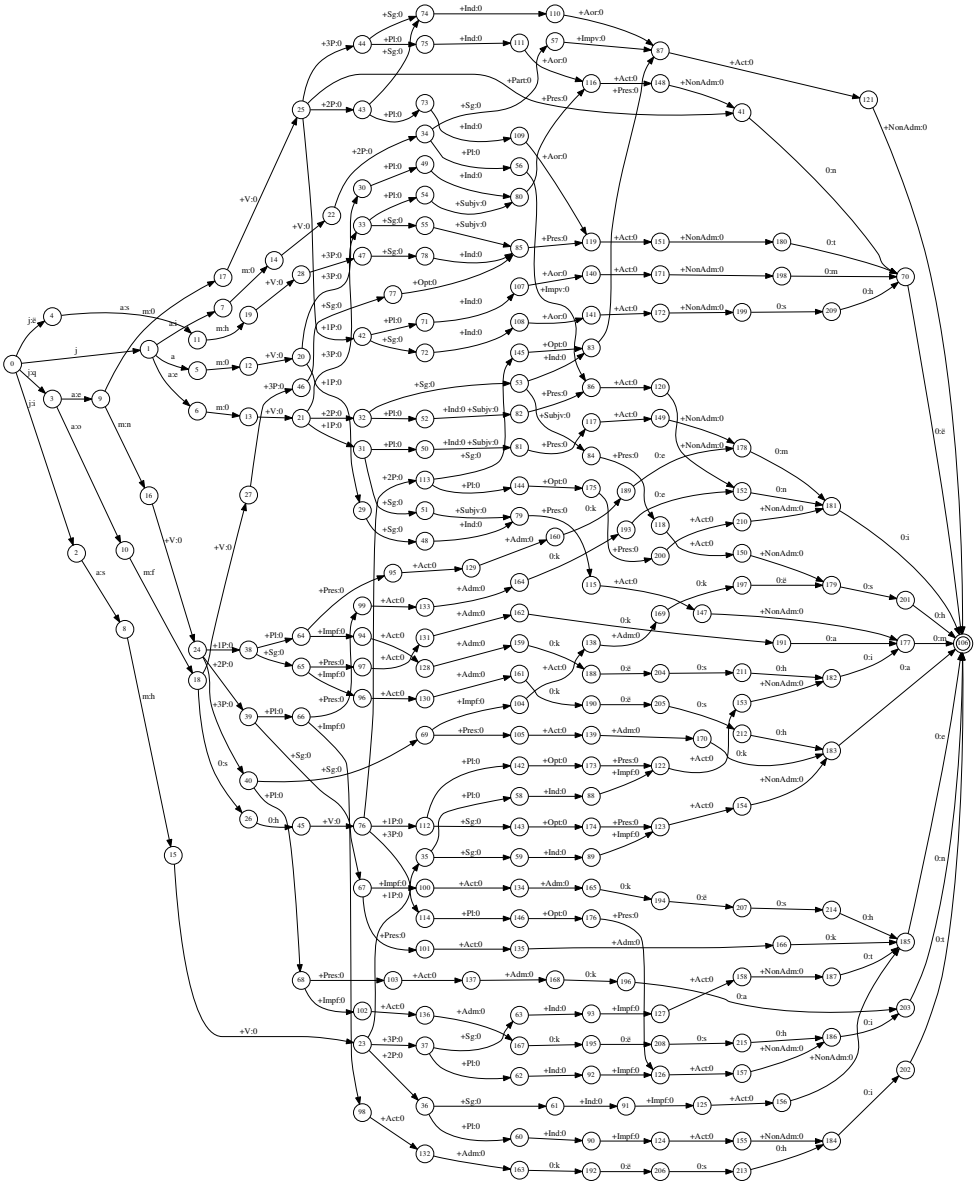
Für das Hilfsverb *jam* sieht das Wortnetz wie in Listing 5.8 aus.

Listing 5.8: Der Automat für das Hilfsverb *jam* in Text-Form.

```
j:ëa:sm:h+V:0+3P:0+Sg:0+Ind:0+Pres:0+Act:0+NonAdm:00:t0:ë
j:qa:om:f0:s0:h+V:0+1P:0+Sg:0+Opt:0+Pres:0+Act:0+NonAdm:00:a
j:qa:om:f0:s0:h+V:0+1P:0+Pl:0+Opt:0+Pres:0+Act:0+NonAdm:00:i0:m
j:qa:om:f0:s0:h+V:0+2P:0+Sg:0+Opt:0+Pres:0+Act:0+NonAdm:0
j:qa:om:f0:s0:h+V:0+2P:0+Pl:0+Opt:0+Pres:0+Act:0+NonAdm:00:i
j:qa:om:f0:s0:h+V:0+3P:0+Pl:0+Opt:0+Pres:0+Act:0+NonAdm:00:i0:n
j:qa:om:f+V:0+3P:0+Sg:0+Opt:0+Pres:0+Act:0+NonAdm:00:t0:ë
j:qa:em:0+V:0+1P:0+Sg:0+Ind:0+Aor:0+Act:0+NonAdm:00:s0:h0:ë
j:qa:em:0+V:0+1P:0+Pl:0+Ind:0+Aor:0+Act:0+NonAdm:00:m0:ë
j:qa:em:0+V:0+2P:0+Sg:0+Ind:0+Aor:0+Act:0+NonAdm:0
j:qa:em:0+V:0+2P:0+Pl:0+Ind:0+Aor:0+Act:0+NonAdm:00:t0:ë
j:qa:em:0+V:0+3P:0+Sg:0+Ind:0+Aor:0+Act:0+NonAdm:0
j:qa:em:0+V:0+3P:0+Pl:0+Ind:0+Aor:0+Act:0+NonAdm:00:n0:ë
j:qa:em:0+V:0+Part:00:n0:ë
j:qa:em:n+V:0+1P:0+Sg:0+Pres:0+Act:0+Adm:00:k0:a0:m
j:qa:em:n+V:0+1P:0+Sg:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h0:a
j:qa:em:n+V:0+1P:0+Pl:0+Pres:0+Act:0+Adm:00:k0:e0:m0:i
j:qa:em:n+V:0+1P:0+Pl:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h0:i0:m
j:qa:em:n+V:0+2P:0+Sg:0+Pres:0+Act:0+Adm:00:k0:e
j:qa:em:n+V:0+2P:0+Sg:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h0:e
j:qa:em:n+V:0+2P:0+Pl:0+Pres:0+Act:0+Adm:00:k0:e0:n0:i
j:qa:em:n+V:0+2P:0+Pl:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h0:i0:t
j:qa:em:n+V:0+3P:0+Sg:0+Pres:0+Act:0+Adm:00:k0:a
j:qa:em:n+V:0+3P:0+Sg:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h
j:qa:em:n+V:0+3P:0+Pl:0+Pres:0+Act:0+Adm:00:k0:a0:n
j:qa:em:n+V:0+3P:0+Pl:0+Impf:0+Act:0+Adm:00:k0:ë0:s0:h0:i0:n
ja:im:0+V:0+2P:0+Sg:0+Impv:0+Act:0+NonAdm:0
ja:im:0+V:0+2P:0+Pl:0+Impv:0+Act:0+NonAdm:00:n0:i
ja:em:0+V:0+1P:0+Sg:0+Subjv:0+Pres:0+Act:0+NonAdm:00:m
ja:em:0+V:0+1P:0+Pl:0+Subjv:0+Pres:0+Act:0+NonAdm:00:m0:i
ja:em:0+V:0+1P:0+Pl:0+Ind:0+Pres:0+Act:0+NonAdm:00:m0:i
ja:em:0+V:0+2P:0+Sg:0+Ind:0+Pres:0+Act:0+NonAdm:0
ja:em:0+V:0+2P:0+Sg:0+Subjv:0+Pres:0+Act:0+NonAdm:00:s0:h
ja:em:0+V:0+2P:0+Pl:0+Subjv:0+Pres:0+Act:0+NonAdm:00:n0:i
ja:em:0+V:0+2P:0+Pl:0+Ind:0+Pres:0+Act:0+NonAdm:00:n0:i
ja:em:0+V:0+3P:0+Sg:0+Subjv:0+Pres:0+Act:0+NonAdm:00:t0:ë
ja:em:0+V:0+3P:0+Pl:0+Subjv:0+Pres:0+Act:0+NonAdm:00:n0:ë
jam:0+V:0+1P:0+Sg:0+Ind:0+Pres:0+Act:0+NonAdm:00:m
jam:0+V:0+3P:0+Pl:0+Ind:0+Pres:0+Act:0+NonAdm:00:n0:ë
j:ia:sm:h+V:0+1P:0+Sg:0+Ind:0+Impf:0+Act:0+NonAdm:00:a
j:ia:sm:h+V:0+1P:0+Pl:0+Ind:0+Impf:0+Act:0+NonAdm:00:i0:m
j:ia:sm:h+V:0+2P:0+Sg:0+Ind:0+Impf:0+Act:0+NonAdm:00:e
j:ia:sm:h+V:0+2P:0+Pl:0+Ind:0+Impf:0+Act:0+NonAdm:00:i0:t
j:ia:sm:h+V:0+3P:0+Sg:0+Ind:0+Impf:0+Act:0+NonAdm:00:t0:e
j:ia:sm:h+V:0+3P:0+Pl:0+Ind:0+Impf:0+Act:0+NonAdm:00:i0:n
```

Dabei können die Übergänge, wie z. B. in der Zeile 1, `j:a`, `a:s`, `m:h` usw., leicht erkannt werden. Der in Listing 5.8 in textueller Form gezeigte Automat wird in graphischer Form in Abbildung 5.10 dargestellt. Eine Zeile in Listing 5.8 entspricht einem Pfad vom Anfang (Zustand `0`) bis zum Ende (Zustand `106`) in Abbildung 5.10.

Abbildung 5.10: Der Automat für das Hilfsverb *jam*.



5.3.7 Erweiterbarkeit der LEXC- und XFST-Dateien

Sowohl neue Lexikoneinträge aus dem Wortschatz der Normsprache als auch solche aus Dialekten können problemlos in das System übernommen bzw. integriert werden.[210] Ebenso können etwaige neue Formen, z. B. regionale Varianten eines Wortes, ohne Schwierigkeiten eingebaut werden.

5.4 Substantive im Rahmen von XFST

Die Substantive lassen sich in deutlich mehr Klassen unterteilen als die Verben, obwohl ihr Flexionsparadigma kleiner ist, als das der Verben. Wie in Abschnitt 3.3.2 erläutert, sorgt die Kategorie Numerus für eine hohe Zahl an Klassen, ca. 200.

In Listing 5.9 sind die ersten Klassen der Substantive (1. Spalte, S–*nnn*) sowie die Genus- und Flexionsmerkmale angegeben.

Listing 5.9: Nominalklassen

```
S-001   f.|~a|~a|~at
S-002   m.|~i|~ë|~ët
S-003   m.|~i|~e|~et
S-004   f.|~a||
S-005   f.|~ja|~e|~et
S-006   m.|~i||
S-007   m.|~i|~|~it
S-008   f.|~a|~e|~et
S-009   m.|~u|~ë|~ët
S-010   f.|~a|~|~të
S-011   m.|~i|~a|~at
S-012   f.||~|~t
S-013   f.|~ja||
S-014   m.|~mi||
S-015   f.|~ra|~ra|~rat
S-016   f.|~ja|~|~të
S-017   m.|~u||
S-018   m.|~u|~nj|~njtë
S-019   m.|~ku|~qe|~qet
S-020   f.|~a|~ë|~ët
...     ...
```

Substantive im Albanischen haben 20 Formen, wenn auch mögliche Überlappungen gezählt werden. Die größte Alternation hängt mit der Kategorie Numerus zusammen, vgl. hierzu die Abschnitte 3.3.2 und 4.3.

[210] Die Lexikoneinträge können sehr dynamisch aufgebaut werden, eine Standardisierung ist jedoch beabsichtigt, um eine konstante Menge der Symbole bzw. der grammatischen Angaben und ein einheitliches Format in LEXC/XFST zu gewährleisten.

Von technischer Seite her betrachtet, sind die Substantive ähnlich den Verben aufgebaut. Die Teilgrammatik besteht aus lexikalischen Daten (LEXC-Dateien) und Regeln (XFST-Dateien).

5.4.1 Einträge der Substantive in LEXC

Für jede Gruppe der Substantivlemmata wurde eine eigene Datei angelegt. Die einzelnen Dateien fassen sowohl die Einträge selbst zusammen, als auch die Informationen über Stämme und Flexionsmerkmale der Lemmata. Wie bei den Verben, wurden später im Laufe der Grammatikentwicklung auch die XFST-Dateien der Substantive aus praktischen Gründen in eine einzige XFST-Datei umformatiert und entsprechend angepasst.

5.4.2 Regeln für Substantive in XFST

Die Regeldatei verwaltet die lexikalischen Daten (LEXC-Datei), die Variablen (var.xfst) sowie die Allomorphie-Regeln (allo.xfst) und Substantivartikel (substantive-article.xfst), bündelt sie in einer Grammatik und schreibt das Ergebnis nach einer Kompilation in eine Datei.

Das Substantiv *djalë* (dt. *Junge*), Klasse S-174, wird beispielsweise folgendermaßen modelliert:

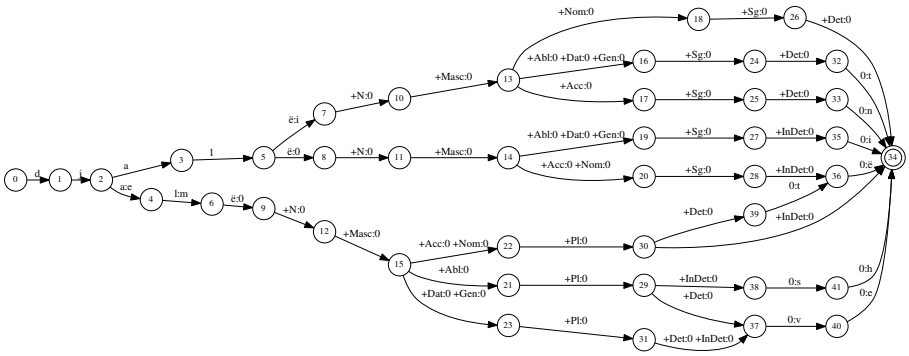
Listing 5.10: Das Lexikon für den Eintrag *djalë, -i* (Ausschnitt)

```
% Erstellung der Flexion (Konkatenativ, als Kaskade)

LEXICON Root

  ! "djalë" N+Masc+Nom+Sg+InDet > "djem" N+Masc+Nom+Pl+InDet
  ! Stämme : sg. "dj{{a}{l}}|+SUFF" > pl. "dj{{e}{m}}|+SUFF"

djalë:dj Stammbildung_dj;

LEXICON Stammbildung_dj

^al  Singular_InDet_Flexion_djal;
^ali Singular_Det_Flexion_djali;
^em  Plural_Flexion_djem;

LEXICON Singular_InDet_Flexion_djal

+N+Masc+Nom+Sg+InDet^ë #;
+N+Masc+Gen+Sg+InDet^i #;
+N+Masc+Dat+Sg+InDet^i #;
+N+Masc+Acc+Sg+InDet^ë #;
+N+Masc+Abl+Sg+InDet^i #;

LEXICON Singular_Det_Flexion_djali

+N+Masc+Nom+Sg+Det^0 #;
+N+Masc+Gen+Sg+Det^t #;
+N+Masc+Dat+Sg+Det^t #;
+N+Masc+Acc+Sg+Det^n #;
```

```
+N+Masc+Abl+Sg+Det^t #;

LEXICON Plural_Flexion_djem

+N+Masc+Nom+Pl+InDet^0 #;
+N+Masc+Gen+Pl+InDet^ve #;
+N+Masc+Dat+Pl+InDet^ve #;
+N+Masc+Acc+Pl+InDet^0 #;
+N+Masc+Abl+Pl+InDet^sh #;

+N+Masc+Nom+Pl+Det^të #;
+N+Masc+Gen+Pl+Det^ve #;
+N+Masc+Dat+Pl+Det^ve #;
+N+Masc+Acc+Pl+Det^të #;
+N+Masc+Abl+Pl+Det^ve #;
...

% Erstellung der Allomorphe/Stämme mit XFST-Regeln
...
read regex   {djal}    -> {djem} || _ ?* "+A_djalë" "+Pl"
 .o.         "{" | "}"  -> 0      || _ ?* "+A_djalë"
 .o.         "+A_djalë" -> 0;


read regex [ {a} -> {e}, {l} -> {m} ] || _ ?* "+A_djalë" "+Pl"
 .o.         "{" | "}"        -> 0     || _ ?* "+A_djalë"
 .o.         "+A_djalë"       -> 0;
...

% Erstellung der Allomorphe/Stämme mit parallelen XFST-Regeln
...
      {ë} -> 0   || _      "+A_djalë" "+N" ?^2 "+Sg" "+InDet"

.o.   {ë} -> {i} || _      "+A_djalë" "+N" ?^2 "+Sg" "+Det"

.o. [ {a} -> {e} || _ {lë} "+A_djalë" "+N" ?^2 "+Pl" ?^1 ,,
      {l} -> {m} || _ {ë}  "+A_djalë" "+N" ?^2 "+Pl" ?^1 ,,
      {ë} -> 0   || _      "+A_djalë" "+N" ?^2 "+Pl" ?^1
    ] ;
...
```

In Listing 5.10 zeigen die Zeilen 1–45 die Pluralbildung des Substantiv *djalë*, sg. nom. mask. *djem*, pl. nom. mask. Durch diese Methode, einer Art Kaskade, werden die Stämme „schritt- oder stufenweise“ gebildet und schließlich die Suffixe angefügt.
Der Eintrag kann auch mittels einer Allo-Regel modelliert werden, und zwar (1) durch direktes Umschreiben des Plural-Stammes, vgl. Zeilen 49–51, oder (2) indem gleichzeitig der Stammvokal *a* in *e* und der Stammkonsonant *l* in *j* umgewandelt werden, wie bspw. sg. $dj|\{\{a\}\{l\}\}|\ddot{e}_{Suff} \rightarrow$ pl. $dj|\{\{e\}\{m\}\}|0_{Suff}$, vgl. hierzu Zeilen 53–55 in Listing 5.10. XFST erlaubt sowohl die Benutzung der einen als auch der anderen Methode, ebenso wie eine Kombination beider Methoden, vgl. [Beesley / Karttunen 2004: 203–278 (§ 4)].
Durch parallele XFST-Regeln, vgl. [Beesley / Karttunen 2004: 142–144 (§ 3.5.3)], ist es möglich mehrere Operationen gleichzeitig und unabhängig voneinander durchzuführen. Bei der Verarbeitung von *djalë*/*djem* in Listing 5.10, vgl. Zeilen 59–68, wurde davon auch Gebrauch gemacht.

Abbildung 5.11: Der Automat für das Substantiv *djalë*.



In Abbildung 5.11 ist zu erkennen, wie die zwei Stämme modelliert werden, vgl. Knoten (Zustände) 0 bis 9. Die Umschreibung von a in e und l in m ist in den Kanten (Übergängen) 2 → 4 und 4 → 6 zu sehen. ë wird entsprechend den bestimmten Formen bzw. Pluralformen in den Kanten 5 → 7, 5 → 8 und 6 → 9 verarbeitet. Bis dahin werden die Formen *djalë*, *djali* und *djem* gebildet. Die weiteren Zustände und Übergänge, bis auf die letzten, dienen der Verarbeitung der grammatischen Information. Zum Schluss, in den letzteren Zuständen und Übergängen, werden die Flexionssuffixe verarbeitet. Bei den Paaren in den Übergängen des Transducers markiert die linke Seite vor dem Doppelpunkt die sogenannte Unterseite, während die rechte Seite die sogenannte Oberseite andeutet.[211] In dieser Form ist auch die Abbildung 5.10, der Automat für das Hilfsverb *jam*, zu interpretieren. Die vielen Zustände und Übergänge ergeben sich aus den Eigenschaften des suppletiven Hilfsverbs *haben*.

In Textform (Wortkette) sieht der in Abbildung 5.11 gezeigte Automat wie in Listing 5.11 aus. Die Zeilen sind wie in Listing 5.8 (Hilfsverb *jam*) zu interpretieren.

[211] Mit dem Begriff *Unterseite* wird in XFST die Generierung bezeichnet, während mit *Oberseite* die Analyse gemeint ist, vgl. [Beesley / Karttunen 2004: 86 f.].

Listing 5.11: Der Automat für das Substantiv *djalë* in Text-Form.

```
dja:el:më:0+N:0+Masc:0+Nom:0+Pl:0+Det:00:t0:ë
dja:el:më:0+N:0+Masc:0+Nom:0+Pl:0+InDet:0
dja:el:më:0+N:0+Masc:0+Dat:0+Pl:0+InDet:00:v0:e
dja:el:më:0+N:0+Masc:0+Dat:0+Pl:0+Det:00:v0:e
dja:el:më:0+N:0+Masc:0+Gen:0+Pl:0+InDet:00:v0:e
dja:el:më:0+N:0+Masc:0+Gen:0+Pl:0+Det:00:v0:e
dja:el:më:0+N:0+Masc:0+Acc:0+Pl:0+Det:00:t0:ë
dja:el:më:0+N:0+Masc:0+Acc:0+Pl:0+InDet:0
dja:el:më:0+N:0+Masc:0+Abl:0+Pl:0+Det:00:v0:e
dja:el:më:0+N:0+Masc:0+Abl:0+Pl:0+InDet:00:s0:h
djalë:0+N:0+Masc:0+Nom:0+Sg:0+InDet:00:ë
djalë:0+N:0+Masc:0+Dat:0+Sg:0+InDet:00:i
djalë:0+N:0+Masc:0+Gen:0+Sg:0+InDet:00:i
djalë:0+N:0+Masc:0+Acc:0+Sg:0+InDet:00:ë
djalë:0+N:0+Masc:0+Abl:0+Sg:0+InDet:00:i
djalë:i+N:0+Masc:0+Nom:0+Sg:0+Det:0
djalë:i+N:0+Masc:0+Dat:0+Sg:0+Det:00:t
djalë:i+N:0+Masc:0+Gen:0+Sg:0+Det:00:t
djalë:i+N:0+Masc:0+Acc:0+Sg:0+Det:00:n
djalë:i+N:0+Masc:0+Abl:0+Sg:0+Det:00:t
```

Die Zeilen 1–10 stellen die Modellierung der Plural-Formen dar, die Zeilen 11–20 die Formen des Singulars.

Listing 5.12 zeigt Abfragen der besonderen Formen bei Verben und Substantiven, welche in Abschnitt 3.3.4 besprochen wurden.

Listing 5.12: *ç'* und *s'* – *wer /was* und *Negation* bei Verben sowie *wer /was* bei Substantiven.

```
xfst[1]: up ç'punonte
<+What>+punoj+V+3P+Sg+Ind+Impf+Act+NonAdm

xfst[1]: up s'punonte
<+Neg>+punoj+V+3P+Sg+Ind+Impf+Act+NonAdm

xfst[1]: up ç'punë
<+Which/What>+punë+S+Fem+Acc+Pl+InDet
<+Which/What>+punë+S+Fem+Nom+Pl+InDet
<+Which/What>+punë+S+Fem+Nom+Sg+InDet
<+Which/What>+punë+S+Fem+Acc+Sg+InDet

xfst[1]: up s'punë
???
```

Viele Verben bzw. Substantive können mit beiden Formen vorkommen, d. h. es handelt sich um einen produktiven Wortbildungstyp. In geschriebener Sprache kommen sie schätzungsweise seltener vor als in gesprochener Sprache, doch häufig genug, dass sie im vorliegenden System berücksichtigt werden sollten.

5.5 Adjektive im Rahmen von XFST

Listing 5.13 zeigt die ersten 10 Flexionssuffixe der Adjektivklassen. Dabei ist zu erkennen, dass diese im Vergleich zu den Klassen der Verben und Substantive etwas heterogener sind.

Listing 5.13: Adjektivklassen

```
1  Adj-001 ||~e|
2  Adj-002 |i||e
3  Adj-003 |||
4  Adj-004 |i|~me|e
5  Adj-005 |i|~e|e
6  Adj-006 ||~ézë|
7  Adj-007 ||-|-
8  Adj-008 -|-||
9  Adj-009 ||~ë|
10 Adj-010 ||~me|
11 ...     ...
```

Adjektive werden in ihre jeweiligen Haupttypen unterteilt: Adjektive mit vorangestelltem Artikel (Zeilen 2, 4, 5), Adjektive ohne vorangestelltem Artikel (Zeilen 1, 6, 9, 10) und Adjektive, die nicht dekliniert werden (Zeilen 3, 7, 8) . Diese Heterogenität der Adjektive spiegelt sich bei der Implementierung wider.

Listing 5.14 zeigt einige Abfragen an die Grammatik der Adjektive unter xfst, wobei die heterogenen Eigenschaften gut zu erkennen sind.

Listing 5.14: Adjektive unter xfst.

```
1
2  % Adjektive ohne vorangesetzten Artikel
3  xfst[1]: up besnik
4  besnik+A+Nom+Sg+m
5  besnik+A+Gen+Sg+m
6  besnik+A+Dat+Sg+m
7  besnik+A+Acc+Sg+m
8  besnik+A+Abl+Sg+m
9
10 xfst[1]: up besnikë
11 besnik+A+Nom+Pl+m
12 besnik+A+Gen+Pl+m
13 besnik+A+Dat+Pl+m
14 besnik+A+Acc+Pl+m
15 besnik+A+Abl+Pl+m
16
17 xfst[1]: up besnike
18 besnik+A+Nom+Pl+f
19 besnik+A+Nom+Sg+f
20 besnik+A+Gen+Pl+f
21 besnik+A+Gen+Sg+f
22 besnik+A+Dat+Pl+f
23 besnik+A+Dat+Sg+f
```

```
besnik+A+Acc+Pl+f
besnik+A+Acc+Sg+f
besnik+A+Abl+Pl+f
besnik+A+Abl+Sg+f
...

% Zufällige Ausgabe der Adjektivformen
xfst[1]: random-lower
e paasimilueshëm
e shpërthurur
i zbrazët
të popullzuar
shumëkëndore
njëmilionëshe
simetrikë
i stërlodhur
i provueshëm
të jashtëshkruara
e përvëluar
pseudodemokrat
të gllabëruar
i papërshpirtshëm
të përlindur
...

% Kongruenz zischen Adjektiv und vorangestelltem Artikel
xfst[1]: up e afërme
e+Art+Nom+Sg+f+InDet<Whitespace>afërm+A+Fem+Sg+InDet

xfst[1]: up të afërme
e+Art+Nom+Pl+f+InDet<Whitespace>afërm+A+Fem+Pl+InDet

xfst[1]: up i afërme
???

xfst[1]: up i afërm
i+Art+Nom+Sg+m+InDet<Whitespace>afërm+A+Masc+Sg+InDet
...

xfst[1]: up afërme
afërm+A+Fem+Pl+InDet
afërm+A+Fem+Sg+InDet

xfst[1]: up afërm
afërm+A+Neut+Sg+InDet
afërm+A+Neut+Pl+InDet
afërm+A+Fem+Sg+InDet
afërm+A+Masc+Sg+InDet
afërm+A+Masc+Pl+InDe
...
% Komplexe Adjektive (Derivation)
joushqimore <jo+Pref>ushqimor+A+Fem+Sg+InDet
joushqimore <jo+Pref>ushqimor+A+Fem+Pl+InDet

jovegjetarianë <jo+Pref>vegjetarian+A+Neut+Sg+InDet
jovegjetarianë <jo+Pref>vegjetarian+A+Neut+Pl+InDet
jovegjetarianë <jo+Pref>vegjetarian+A+Masc+Pl+InDet

...
```

Die Regeln der Adjektive binden die verschiedenen Typen in eine Grammatik ein. Die erste Klasse flektiert bspw. und hat keinen vorangestellten Artikel. Dabei wird aus der Grundform (Maskulin) eine weitere Form (Feminin) durch das Anhängen eines Suffixes produziert, vgl. *besnik*, m. → *besnik*|*e*, f.

Die zweite Klasse besitzt einen vorangestellten Artikel, der mit anderen Regeln verarbeitet wird, u. a. für die Behandlung seiner Kongruenz mit dem zugehörigen Adjektiv.

5.6 Numeralia im Rahmen von XFST

Die Implementierung der Numeralia im Rahmen von XFST ist im Vergleich zu den anderen Wortarten etwas Besonderes aufgrund ihrer Art der Kombinatorik, sowie ihrer Schreibkonventionen im Albanischen. Als Grundlage diente die Formalisierung der Numeralia, vgl. das Listing 5.15.

Listing 5.15: Formalisierung der Numeralia im Albanischen (Ausschnitt)

```
A    0;
     zëro
1    X


B    1, 2, 3, 4, 5, 6, 7, 8, 9;
     një, dy, trE/trI, katër, pesë, gjashtë, shtatë, tetë, nëntë
9    NJ


C    10;
     dhjetë (NOT një [eins], dy [zwei])
1    DH


D    11--19; [1"mbë"10, ... 9"mbë"10] {x|"MBËDHJETË"}
     një|mbë|dhjetë, dy|mbë|dhjetë, trE|mbë|dhjetë, ...,
     {{B>"mbë">C}};
9    MBË
     {1-9{"mbë">C}};


E    1:20, 2:20; [1"zet", 2"zet"] {"njëZET", "dyZET"};
     {{"e">B}}; dy|zet|e|dy=42
2    1,2 ZET
     {1--2ZET{"e">B}};


F    3:10, (4:10, =2:20) 5:10, 6:10, 7:10, 8:10, 9:10;
     {{"e">B}};
7    3--9 DH => 30, ... 90;
     trI|dhejtë, ... nëntë|dhejtë
     {3-(4)-9DH{"e">B}}; => 30e1, ... 30e9;
     trI|dhejtë|"e"|dy=32
~    dhjetra


...
```

Die Formalisierung der Zahlen stellt aufgrund der Natur der Zahlen einen wichtigen und notwendigen Schritt vor der Implementierung dar. Sie können theoretisch unbegrenzt sein. Aus praktischen Gründen möchte man nicht alle Zahlen im Lexikon eintragen, z. B. alle Kardinalzahlen von 0 bis 1.000.000. Die Ordinalzahlen wären noch aufwändiger, da sie zusätzlich

zu den Kardinalzahlen noch flektiert werden und einen vorangestellten Artikel besitzen.

In Listing 5.16 sind einige Abfragen der Numeralia, sowohl Ordinalia als auch Kardinalia, an die xfst-Numeraliagrammatik dargestellt.

Listing 5.16: Numeralia im Rahmen von xfst

```
 ...      ...
xfst[1]: up një
1+Card

xfst[1]: up njëmbëdhjetë
11+Card

xfst[1]: up njëzetepesë
25+Card

xfst[1]: up njëqind e njëzet e pesë
125+Card
 ...      ...
xfst[1]: down 1+Ord+Nom+Sg+Det+Fem
e para

xfst[1]: down 11+Ord+Nom+Sg+Det+Masc
i njëmbëdhjetë

xfst[1]: down 11+Ord+Nom+Sg+Det+Fem
e njëmbëdhjeta
 ...      ...

xfst[1]: up e njëzetepestë
25+Ord+Nom+Sg+InDet+Fem

xfst[1]: up të njëzetepestin
25+Ord+Acc+Sg+Det+Masc
 ...      ...
```

Die Zeilen 1–12 in Listing 5.16 zeigen einige Abfragen der Kardinalzahlen an die Grammatik der Numeralia. In den Zeilen 15–29 sind Abfragen der Ordinalzahlen dargestellt.

5.7 Pronomina im Rahmen von XFST

Der Großteil der Pronomina wurde direkt als Vollform eingetragen, da die einzelnen Formen so unterschiedlich sind, dass ein gemeinsamer Stamm bzw. ein gemeinsames Flexionsparadigma nicht gegeben ist, vgl. hierzu die Personalpronomina (3.10 und 4.2).[212] Der andere Teil wird in ähnlicher Form wie die Substantive, Numeralia und Adjektive dekliniert. Die Typen wie *secili* (dt. *jeder*), werden wie die Substantive flektiert, vgl. *secili*, *secilit*, *secilit*,

[212] Für die theoretischen Grundlagen der Pronomina, vgl. Abschnitt 3.3.4, für die lexikalischen Einträge, vgl. Abschnitt 4.5.

secilin, *secilit*, usw. Der Typ *cilido* (dt. *jeder, der*), wobei der Teil *do* nicht flektiert wird, sondern nur der Teil *cili*, kann auch im Rahmen von xfst ohne Schwierigkeiten modelliert und implementiert werden.

Listing 5.17: Pronomina im Rahmen von xfst

```
xfst[1]: up unë
unë+PersPron+Nom+Sg+1P

xfst[1]: up ti
ti+PersPron+Nom+Sg+2P

xfst[1]: up ne
ne+PersPron+Acc+Pl+1P+pC_na
ne+PersPron+Nom+Pl+1P

xfst[1]: up ju
ju+u+u+Acc+Pl+3P+Refl
ju+ju+u+Acc+Pl+2P+Refl
ju+pronClitic+Acc+Pl+2P+ju
ju+pronClitic+Dat+Pl+2P+juve
ju+PersPron+Acc+Pl+2P+pC_ju
ju+PersPron+Nom+Pl+2P

 ...     ...

xfst[1]: up secilin
secilin+InDet+Acc+Sg+Masc

xfst[1]: up cilindo
cilindo+InDet+Acc+Masc

 ...     ...

xfst[1]: up imi
imi+PosPron+Used_without_N+Poss_in_Sg+Nom+Sg+1P+Masc
%imi+PosPron+Used_with_N+Poss_in_Sg+Sg+1P+Masc

xfst[1]: up imja
imja+PosPron+Used_without_N+Poss_in_Sg+Nom+Sg+1P+Fem
%imja+PosPron+Used_without_N+Poss_in_Sg+Sg+1P+Fem

xfst[1]: up juaji
juaji+PosPron+Used_without_N+Poss_in_Sg+Nom+Pl+2P+Masc
%juaji+PosPron+Used_without_N+Poss_in_Sg+Pl+2P+Masc

xfst[1]: up juaja
juaja+PosPron+Used_without_N+Poss_in_Sg+Nom+Pl+2P+Fem
%juaja+PosPron+Used_without_N+Poss_in_Sg+Pl+1P+Masc
```

Die komplexen und unterschiedlichen grammatischen Kategorien der Pronomina sieht man auch in Listing 5.17. Ein Vergleich der Zeilen 1–17 mit den Zeilen 21–25 und den Zeilen 29 bis zum Ende zeigt diese Unterschiede.

5.8 Adverbien im Rahmen von XFST

Die Modellierung und Implementierung der Adverbien im Rahmen von XFST entspricht der Komplexität von Pronomina. Da sie nicht flektiert werden, könnten sie einfach ins Format konvertiert und ins Lexikon (lexc) übernommen werden. Einige Formen, wie *së voni* (dt. *letztens*, *spät*, u. ä.), die aus zwei Teilen bestehen, können wie die Adjektive mit vorangestellten Artikel modelliert und implementiert werden.
Bei Adverbien, die einen vorangestellten Artikel besitzen, bestehen grundsätzlich zwei Möglichkeiten der Implementierung: (1) Adverb mit der vorangestellten Partikel *së*, (2) Adverb ohne die vorangestellte Partikel *së*. Im Rahmen der vorliegenden Arbeit wurden beide Varianten implementiert. Bei einem etwaigen Einsatz der Grammatik für didaktische Zwecke wäre die erste Variante besser. Bei einem Einsatz der Morphologie beim Parsing oder einer syntaktischen Analyse wäre unter Umständen die zweite Variante passender. Beide Varianten können sowohl einzeln als auch gleichzeitig eingesetzt werden, indem sie in der Grammatik entsprechend auskommentiert werden.

Listing 5.18: Adverbien im Rahmen von xfst

```
xfst[1]: up keq
keq+N+Masc+Nom+Pl+InDet+Art
keq+N+Masc+Acc+Pl+InDet+Art
keq+Adv

xfst[1]: up lart
lart+Prep
lart+Adv

xfst[1]: up sot
sot+Adv

xfst[1]: up kalimthi
kalimthi+Adv

xfst[1]: up shpejti
shpejti+N+Fem+Acc+Sg+InDet
shpejti+N+Fem+Nom+Sg+InDet
shpejti+Adv+së

xfst[1]: up së shpejti
shpejti+Adv+së

xfst[1]: up voni
voni+Adv+së

xfst[1]: up së voni
voni+Adv+së

xfst[1]: up vonë
vonë+Adv
```

Listing 5.18 zeigt Anfragen an die Grammatik. Die Zeilen 1–14 zeigen Adverbien, die keine vorangestellte Partikel besitzen. Die Zeilen 16–19 und 24–25 zeigen die Variante (1), d. h. die Analyse der Adverbien, die eine vorangestellte Partikel besitzen, die aber bei der Analyse und Produktion nicht berücksichtigt wird. Die Zeilen 21–22 und 27–28 zeigen die Variante (2), wobei die vorangestellte Partikel berücksichtigt wird. Adverbien wie *vonë* (dt. *spät*), können regelhaft in den Typ *së␣voni* umgewandelt werden. Die Zahl dieser Adverbien ist jedoch beschränkt. Sie ist überschaubar und lässt sich auch direkt im Lexikon kodieren.

5.9 Indeklinabilia im Rahmen von XFST

Unter den Begriff Indeklinabilia fallen die indeklinierten Wortarten Interjektionen, Konjunktionen, Partikeln und Präpositionen. Einige von ihnen bestehen aus zwei oder mehr Teilen, sowohl in Kontaktposition, d. h. direkt nacheinander, als auch in Distanzstellung, d. h. durch andere Wortarten unterbrochen.
Die Implementierung der Indeklinabilia im Rahmen von XFST ist gleich der der Lexikoneinträge. Die jeweiligen Einträge, nach Wortart und Typen gruppiert, werden in ihren Ober- und Unterseiten aufgelistet, das heißt in einer lexc-Datei, und von einer xfst-Datei gelesen. Eine Abfrage an die Grammatik der Partikeln unter XFST ist in Listing (5.19) angegeben:

Listing 5.19: Partikel im Rahmen von xfst

```
xfst[1]: up a
a+Par

xfst[1]: up ani
ani+Par

xfst[1]: up ja
ja+Par

xfst[1]: up ndoshta
ndoshta+Par

xfst[1]: up pra
pra+Par

xfst[1]: up sikur
sikur+Par

xfst[1]: up gjer
gjer+Par

xfst[1]: up para
para+Par
para+Prep+Abl
...
xfst[1]: up nga
nga+Conj
```

```
nga+Prep+Nom
...
xfst[1]: up prej
prej+Conj
prej+Prep+Abl

xfst[1]: up me
me+Par
me+Prep+Acc
...
xfst[1]: up pranë
pranë+Prep+Abl
pranë+Adv
...
```

Wie in Listing 5.19 zu erkennen ist, werden keine begleitenden Informationen zu den Partikeln angegeben. Sollten diese Informationen, z. B. semantischer Natur, benötigt werden, lassen sie sich jedoch leicht ergänzen, indem sie direkt im Lexikon eingetragen werden. In der Tat werden bei Indeklinabilia im Rahmen von XFST nur die jeweiligen Lexika (lexc) gelesen.

5.10 Zusätzliche Erweiterungen

Neben den besprochenen Wortarten kommen in Texten auch andere Wörter vor, wie z. B. Namen, sowohl Personenamen als auch Namen anderer Typen, wie Ortsnamen oder Namen der Einwohner dieser Orte.

5.10.1 Namen im Rahmen von XFST

Die Namen sind im Rahmen der vorliegenden Arbeit ähnlich wie die Substantive (5.4) implementiert. Den Unterschied machen einige Beschränkungen bzw. Erweiterungen bei Nomina, die ihren spezifischen Eigenschaften entsprechen, z. B. die Einschränkung des Plurals bei Städtenamen. Eine Abfrage der Ortsnamen ist in Listing 5.20 angegeben:

Listing 5.20: Ortsnamen im Rahmen von xfst

```
xfst[1]: up Pejë
Pejë+N+Fem+Acc+Sg+InDet
Pejë+N+Fem+Nom+Sg+InDet

xfst[1]: up Peja
Pejë+N+Fem+Nom+Sg+Det

xfst[1]: up Pejën
Pejë+N+Fem+Acc+Sg+Det

xfst[1]: up Peje
Pejë+N+Fem+Abl+Sg+InDet+OR
```

Die Bezeichnung `+OR` steht für die Herkunft, dt. etwa *aus Pejë*, z. B. *Birrë Peje*, dt. *Pejaer Bier*.
Da die Ortsnamen wie die übrigen Namen und Substantive dekliniert werden und eine offene Wortklasse bilden, ist ihre Implementierung genauso aufwändig, wie die der übrigen Substantive. Wie im Abschnitt 3.3.2 erklärt, werden die Ortsnamen nach verschiedenen Mustern dekliniert. Noch komplexer wird die Bezeichnung der Einwohner, da noch die grammatische Kategorie Plural hinzukommt.
Personennamen wurden nur prototypisch implementiert, da die Erstellung eines Lexikons dieser Namen sehr viel Aufwand bedeuten und den Rahmen der vorliegenden Arbeit sprengen würde. Die Grammatik der Personennamen kann unter der Voraussetzung, dass die Personennamen in Klassen eingeteilt sind, ohne Schwierigkeiten erweitert werden, indem das Lexikon mit Nomina befüllt wird.

5.10.2 Interpunktion im Rahmen von XFST

Listing 5.21 zeigt einige Anfragen der Interpunktionszeichen an die Grammatik. Dabei werden für Taggingzwecke die Interpunktionszeichen verarbeitet, d. h. übersetzt, wie z. B. „`?`“ in „`pikëpyetje+Intp`“. Die Gegenrichtung, die Produktion, ist in der zweiten Hälfte des Listings gezeigt.

Listing 5.21: Interpunktion im Rahmen von xfst

```
xfst[1]: up +
plus+Intp

xfst[1]: up ?
pikëpyetje+Intp

xfst[1]: up :
dypika+Intp

xfst[1]: up ,
presje+Intp

xfst[1]: up /
thyesë+Intp
shenjë thyese+Intp

xfst[1]: up \
shenjë thyese e praptuar+Intp

xfst[1]: up )
kllapë mbyllëse+Intp

xfst[1]: up [
kllapë katërore hapëse+Intp

xfst[1]: up !
pikëçuditje+Intp

xfst[1]: up ~
përafërsisht+Intp
```

```
31|
32| xfst[1]: up >
33|
34| xfst[1]: up ">"
35| më e madhe se+Intp
36|
37| xfst[1]: down minus+Intp
38| -
39|
40| xfst[1]: down plus+Intp
41| +
42|
43| xfst[1]: down dypika+Intp
44| :
```

5.10.3 Abkürzungen im Rahmen von XFST

Ähnlich wie die Interpunktionszeichen sind auch die Abkürzungen aufgebaut und implementiert.

Listing 5.22 zeigt einen Ausschnitt der Anfragen an die Grammatik für Abkürzungen.

Listing 5.22: Abkürzungen im Rahmen von xfst

```
1| ...     ...
2|
3| xfst[1]: up kg
4| kilogram+Abbr
5|
6| xfst[1]: up l
7| litër+Abbr
8|
9| xfst[1]: up l.
10| litër+Abbr
11|
12| xfst[1]: up km.
13| kilometër+Abbr
14|
15| xfst[1]: up km
16| kilometër+Abbr
17|
18| xfst[1]: up VL
19| Vlorë+Abbr
20|
21| xfst[1]: up nr.
22| numër+Abbr
23|
24| xfst[1]: up vjet.
25| i,e vjetëruar+Abbr
26|
27| xfst[1]: up hist.
28| i,e kohës së kaluar+Abbr
29|
30| xfst[1]: up b.f.
31| bie fjala+Abbr
32|
33| xfst[1]: up dmth.
34| ???
35|
36| xfst[1]: up d.m.th.
37| do më thënë+Abbr
```

```
38
39 xfst[1]: up AQSh
40 Arkivi Qëndror i Shtetit+Abbr
41
42 xfst[1]: up CD
43 Corps diplomatique+Abbr
44 compact disc+Abbr
45
46 ...      ...
```

Die Zeilen 3–4 zeigen Maßeinheiten, die als Abkürzungen betrachtet werden und mit der gleichen Grammatik implementiert worden sind. Die Zeilen 36–37 zeigen eine Abfrage für „`d.m.th.`“, eine der häufigsten Abkürzungen. Sie sind mit der Abkürzung `Abr` (engl. *abbreviation*) versehen.

5.11 Wortbildung im Rahmen von xfst

Beim Testen der Morphologiekomponente der Flexion stellt sich heraus, dass einige Wortformen nicht analysiert werden können, da sie komplexe Wortformen sind, deren Lemmata nicht im Lexikon enthalten sind – obwohl das zugrundeliegende Lexikon im Vergleich zu den Lexika des Albanischen eine überdurchschnittliche Zahl an Lemmata enthält. Eine Sammlung dieser fehlenden Wortformen und entsprechende Lemmaeinträge im Lexikon wäre die erste Idee, um sie beim nächsten Testlauf behandeln zu können. Doch einige Neuschöpfungen sind nur für einen speziellen Kontext geschaffen (okkasionelle Verwendungen), sodass ihre lemmatisierte Form im Wörterbuch nicht gerechtfertigt werden kann, wie z. B. [*anije*] *bashkëpeshkuese*, dt. ein Schiff, das (nur) zusammen mit einem anderen Schiff oder mehreren anderen Schiffen in Zusammenarbeit fischen geht. Den Eintrag *bashkëpeshkues/-e* (Adj.) findet man nicht in einem Wörterbuch. Dennoch kann die Bedeutung des Wortes entschlüsselt werden und die Wortbildung ist für den Hörer/Leser transparent.
Sofern diese Wortformen nach bekannten Strukturen gebaut sind, könnten sie abgefangen werden, indem Wortbildungsstrukturen modelliert und implementiert werden.[213] Einige dieser Strukturen decken Wortarten oder bestimmte Eigenschaften einer Wortart ab, die eine große Zahl der Lemmata im Lexikon ausmachen, weshalb sich ihr Einsatz lohnt.

[213] Vgl. hierzu [Celex 1994: German] und [Motsch 2004] zu Möglichkeiten der Modellierung komplexer morphologischer Strukturen der Wortformen.

Wie in Abschnitt 3.4, insbesondere in den Unterabschnitten 3.4.2 und 3.4.4 erläutert, wurde im Rahmen der vorliegenden Arbeit ein Versuch unternommen einige Typen der Wortbildung zu modellieren und zu implementieren, die eine hohe Abdeckung ermöglichen. Einige andere Typen, die nicht produktiv genug sind, wurden nicht behandelt.

5.11.1 Derivation im Rahmen von XFST

Für die Derivation werden getrennte Lexika verwendet, wobei die der Verben, Substantive, Adjektive und Adverbien sogenannte Stammlexika sind. Das bedeutet, dass ein Eintrag wie *punoj*, Verb (dt. *arbeiten*) und *punë*, Substantiv (dt. *Arbeit*), usw. in der Form *pun|* steht. Sie können automatisch aus den Grundlexika erstellt werden.
Aus diesen Lexika werden die abgeleiteten Formen gebildet, indem die jeweiligen Mittel der Wortbildung in systematischer Weise konkateniert werden. Listing 5.23 zeigt einen Ausschnitt der Derivationsgrammatik, wobei Substantive vom Stammtyp 1, d. h. Substantive der Klasse 1, zu einem Verb der Klasse 3 umgeschrieben werden, vgl. Zeile 2. Auch die umgekehrte Richtung, die Umwandlung eines Verbstamms in ein Substantiv, ist in gleicher Weise modelliert, vgl. die Zeile 6. Dadurch wird ein Verbstamm vom Typ 3 in ein Substantiv vom Typ 1 umgewandelt.

Listing 5.23: Derivationsregeln im Rahmen von xfst

```
1  ...     ...
2
3  define DERIVsv001  "+V":0 ( Pref ) "+":0 SSTEM001  {ë}:0 {|} "+sV-
                      Deriv":0 {|} VFLEX003  ;
4
5
6  define DERIVsv001z "+V":0 ( Pref ) "+":0 SSTEM001z       {|} "+sV-
                      Deriv":0 {|} VFLEX003  ;
7
8
9  ...     ...
10
11
12 define DERIVvs001  "+S":0 ( Pref ) "+":0 VSTEM003 {oj}:0 {|} "+vS-
                      Deriv":0 {|} SFLEX001  ;
13
14
15 define DERIVvs001z "+S":0 ( Pref ) "+":0 VSTEM003 {oj}:0 {|} "+vS-
                      Deriv":0 {|} SFLEX001z ;
16
17 ...     ...
```

Einige Anfragen an die Derivationsgrammatik sind in Listing 5.24 zu sehen.

Listing 5.24: Derivation im Rahmen von xfst

```
 ...   ...

xfst[1]: up garazhoj
garazh+sV-Deriv+1P+Sg+Ind+Pres+Act+NonAdm+R-DERIVs003v004
garazh+sV-Deriv+1P+Sg+Subjv+Pres+Act+NonAdm+R-DERIVs003v004

 ...   ...

xfst[1]: up akullzohet
akullzim+sV-Deriv+3P+Sg+Ind+Pres+Pass+NonAdm+R-DERIVsv004

 ...   ...

xfst[1]: up akullzuar
akullzim+sV-Deriv+Part+R-DERIVsv004

 ...   ...

xfst[1]: up daktilograf
daktilografoj   +vS-Deriv+S+f+Nom+Sg+InDet+R-DERIVvs001z

 ...   ...

xfst[1]: up deduksionon
deduksion+sV-Deriv+2P+Sg+Ind+Pres+Act+NonAdm+R-DERIVs003v004
deduksion+sV-Deriv+3P+Sg+Ind+Pres+Act+NonAdm+R-DERIVs003v004

 ...   ...

xfst[1]: up degazime
degazoj+vS-Deriv+S+Masc+Dat+Pl+InDet+R-DERIVvs004003
degazoj+vS-Deriv+S+Masc+Dat+Pl+InDet+R-DERIVvs003
degazoj+vS-Deriv+S+Masc+Gen+Pl+InDet+R-DERIVvs004003
degazoj+vS-Deriv+S+Masc+Gen+Pl+InDet+R-DERIVvs003
degazoj+vS-Deriv+S+Masc+Acc+Pl+InDet+R-DERIVvs004003
degazoj+vS-Deriv+S+Masc+Acc+Pl+InDet+R-DERIVvs003
degazoj+vS-Deriv+S+Masc+Nom+Pl+InDet+R-DERIVvs004003
degazoj+vS-Deriv+S+Masc+Nom+Pl+InDet+R-DERIVvs003

 ...   ...

xfst[1]: up dëgjimesh
dëgjoj+vS-Deriv+S+Masc+Abl+Pl+InDet+R-DERIVvs003

xfst[1]: up dëgjimeve
dëgjoj+vS-Deriv+S+Masc+Dat+Pl+Det+R-DERIVvs003
dëgjoj+vS-Deriv+S+Masc+Gen+Pl+Det+R-DERIVvs003
dëgjoj+vS-Deriv+S+Masc+Abl+Pl+Det+R-DERIVvs003

 ...   ...

xfst[1]: up konfliktuan
konflikt+sV-Deriv+3P+Pl+Ind+Aor+Pass+NonAdm+R-DERIVs003v004
konflikt+sV-Deriv+3P+Pl+Ind+Aor+Act+NonAdm+R-DERIVs003v004

 ...   ...
```

Die Daten für die Behandlung der Derivation (und Komposition, s. u. 5.11.2) sind so organisiert, dass das linguistische Wissen ohne Schwierigkeiten ziemlich direkt (fast *deklarativ*) als xfst–Regeln umgesetzt werden kann.

5.11.2 Komposition im Rahmen von XFST

Die Komposition kann bis zu einem gewissen Grad wie die Derivation aufgebaut und implementiert werden. Im Unterschied zur Derivation können bei der Komposition mehrere Wortbildungselemente gleichzeitig an einer Wortform beteiligt sein. Dies macht die Komposition im Vergleich zur Derivation deutlich komplexer. Ein Ausschnitt der Grammatik der Komposition ist in Listing 5.25 gegeben.

Listing 5.25: Kompositionsregeln im Rahmen von xfst

```
1  ...     ...
2
3  define COMP1 SSTEM001       {|}:0 VSTEM003 {oj}:0 VFLEX ;
4
5  define COMP2 SSTEM001 0:{ë} {|}:0 VSTEM003 {oj}:0 VFLEX ;
6
7  ...     ...
```

Zeile 3 des Listing 5.25 zeigt eine Regel der Komposition, die einen Substantivstamm vom Typ 1 mit einem Verbstamm vom Typ 3 und anschließend mit dem Flexionsparadigma des Verbs konkateniert.

Listing 5.26 zeigt einige Anfragen an die Grammatik.

Listing 5.26: Derivation und Komposition im Rahmen von xfst

```
1
2   ...   ...
3
4  xfst[1]: up deindustrializojnë
5  de|industrializoj+V+3P+Pl+Ind+Pres+Act+NonAdm+xFreqV004-COMP
6  de|industrializoj+V+3P+Pl+Subjv+Pres+Act+NonAdm+xFreqV004-COMP
7
8   ...   ...
9
10 xfst[1]: up shumëkomentuar
11 shumë|koment+sV-Deriv+Part+R-DERIVs003v004
12 shumë|komentim+sV-Deriv+Part+R-DERIVsv004
13 shumë|komentoj+V+Part+xFreqV004-COMP
14
15  ...   ...
16
17 xfst[1]: up ripozicionuan
18 ri|pozicion+sV-Deriv+3P+Pl+Ind+Aor+Act+NonAdm+R-DERIVs003v004
19 ri|pozicion+sV-Deriv+3P+Pl+Ind+Aor+Pass+NonAdm+R-DERIVs003v004
20
21  ...   ...
22
23 xfst[1]: up superprivilegjuar
24 super|privilegj+sV-Deriv+Part+R-DERIVs003v004
25
26  ...   ...
27
28 xfst[1]: up vetëkontrolluar
29 vetë|kontroll+sV-Deriv+Part+R-DERIVs003v004
30
```

```
 ...   ...

xfst[1]: up vetëfitojë
vetë|fitim+sV-Deriv+3P+Sg+Subjv+Pres+Act+NonAdm+R-DERIVsv004
vetë|fitoj+V+3P+Sg+Subjv+Pres+Act+NonAdm+xFreqV004-COMP

 ...   ...

xfst[1]: up vetëcilësohej
vetë|cilësim+sV-Deriv+3P+Sg+Ind+Impf+Pass+NonAdm+R-DERIVsv004

 ...   ...

xfst[1]: up vetëfrymëzuar
vetë|frymëzim+sV-Deriv+Part+R-DERIVsv004

 ...   ...

xfst[1]: up vetëjustifikohet
vetë|justifikim+sV-Deriv+3P+Sg+Ind+Pres+Pass+NonAdm+R-DERIVsv004

 ...   ...

xfst[1]: up punëkrijues
punë|krijues+SS-COMP+Masc+Acc+Sg+InDet+R-SS-COMPssS020004
punë|krijues+SS-COMP+Masc+Nom+Sg+InDet+R-SS-COMPssS020004
punë|krijues+SvA-COMP+Masc+Acc+Sg+InDet+R-SV-COMPsvaS020004
punë|krijues+SvA-COMP+Masc+Nom+Sg+InDet+R-SV-COMPsvaS020004

 ...   ...
```

In Listing 5.27 wird der Typ der Komposita <Häufiges Wort>+<Wortart/Klasse> vorgestellt. Es geht darum, produktive Lemmata wie *shumë* (dt. *viel*), *jo* (dt. *nicht-*, *un-*) usw. mit den Lemmata verschiedener Wortarten und Klassen zu konkatenieren.

Listing 5.27: Komposition im Rahmen von xfst (2)

```
 ...     ...

xfst[1]: up agrokulture
agro|kulturë+S+Fem+Dat+Sg+InDet+xFreqS001-COMP
agro|kulturë+S+Fem+Gen+Sg+InDet+xFreqS001-COMP
agro|kulturë+S+Fem+Abl+Sg+Det+xFreqS001-COMP
agro|kulturë+S+Fem+Abl+Sg+InDet+xFreqS001-COMP

 ...     ...

xfst[1]: up antidhunë
anti|dhunë+S+Fem+Nom+Sg+InDet+xFreqS001-COMP
anti|dhunë+S+Fem+Acc+Sg+InDet+xFreqS001-COMP

 ...     ...

xfst[1]: up bashkëfitojnë
bashkë|fitoj+V+3P+Pl+Ind+Pres+Act+NonAdm+xFreqV004-COMP
bashkë|fitoj+V+3P+Pl+Subjv+Pres+Act+NonAdm+xFreqV004-COMP

 ...     ...

xfst[1]: up joligësinë
jo|ligësinë+S+Fem+Nom+Sg+InDet+xFreqS001-COMP
jo|ligësinë+S+Fem+Acc+Sg+InDet+xFreqS001-COMP
```

```
 ...     ...

xfst[1]: up mosdëshira
mos|dëshirë+S+Fem+Nom+Sg+Det+xFreqS001-COMP
mos|dëshirë+S+Fem+Acc+Pl+InDet+xFreqS001-COMP
mos|dëshirë+S+Fem+Nom+Pl+InDet+xFreqS001-COMP

 ...     ...
```

Aufgrund fehlender Daten, insbesondere einer empirischen Klassifikation des Wortschatzes und der systematischen Darstellung der Wortbildungsregeln, die eine ausführliche Implementierung ermöglichen, kann das Thema Komposition leider nur begrenzt behandelt werden. Selbst grundlegende Daten, wie bspw. Flexionsparadigmen sind in einigen Fällen, bei einigen Klassen gar vollständig, unklar. Sie schwanken bzw. sind nicht klar definiert. Trotz dieser Umstände wurden einige Typen der Komposita berücksichtigt, die häufig vorkommen. Ein anderer Teil, genauer die lexikalisierten Komposita, sind direkt ins Lexikon aufgenommen. Sie findet man auch in den herkömmlichen Wörterbüchern, wie in einem Rechtschreibwörterbuch oder Definitionswörterbuch.
Bei der Erstellung der Regeln in den Teilgrammatiken für Derivation und für Komposition konnten leider keine diachronen Angaben über die Lemmata berücksichtigt werden. Konkrete Beispiele wären die bereits erwähnten Wörter *punë*, dt. *Arbeit* (S) und *punoj*, dt. *arbeiten* (V). Der Stamm *pun|* kann sowohl für die Ableitung des Verbs vom Substantiv, als auch für die Ableitung des Substantivs vom Verb verwendet werden. Wenn die Etymologie bekannt ist, können leicht die entsprechenden Einträge im Lexikon ausgeschlossen werden, sodass nur der richtige Eintrag übrig bleibt.
Bei der Komposition kommt es vor, dass einige Kontakte der Grapheme entstehen, für welche die Rechtschreibregeln eine Umschreibung vorsehen. So z. B. bei der Komposition *zëëmbël*, dt. [*Stimme+süß*], d. h. *angenehme Stimme*, (←*zë+ëmbël*) werden beide *ës* beibehalten, also nicht umgeschrieben, während bei *gojëmbël*, dt. [*Mund+süß*], d. h. *jemand, der schön und freundlich mit jemandem redet*, (←*gojë+ëmbël*) das vorgehende *ë* wegfällt, die Rechtschreibregeln also eine Umschreibung vorsehen.[214] Es gibt auch Wörter, die mehrsilbig sind und eine Endbetonung vorweisen, wie z. B. *shullë́*, zu dt. *eine Stelle, an der die Sonne scheint*, was bedeutet, dass bei einer Komposition dieses Wortes als erste Komponente mit einem Wort, das mit einem Vokal anfängt, *ë* (bei *shullë*) nicht wegfallen würde.[215]

[214] Vgl. [Drejtshkrimi 1974: 20–22 (§ 5 b, Shën. 2)].
[215] Vgl. [Drejtshkrimi 1974: 18 f. (§ 3)].

Listing 5.28 zeigt einen Ausschnitt der Regel, die es erlaubt bei einsilbigen Wörtern, welche auf *ë* enden, im Kontakt mit einem Wort, das mit einem Vokal anfängt, das vorangehende *ë* zu tilgen.

Listing 5.28: Vokalkontakte bei Komposita

```
 ... ...
{ëa} -> {a} || Vokal ?* _ .o.
{ëe} -> {e} || Vokal ?* _ .o.
{ëë} -> {ë} || Vokal ?* _ .o.
{ëi} -> {i} || Vokal ?* _ .o.
{ëo} -> {o} || Vokal ?* _ .o.
{ëu} -> {u} || Vokal ?* _ .o.
{ëy} -> {y} || Vokal ?* _ .o.
 ... ...
```

Die Ausnahmen, wie z. B. ein mögliches Wort *shullë|ëndërr*, zu dt. *ein Traum, in einer Sonnenstelle [zu sein]*, die eine kleine Zahl ausmachen, werden vorher gesondert behandelt, damit sie nicht von diesen Umschreibung betroffen werden.

5.12 Die Hauptgrammatik und ihre Bestandteile

Die Hauptgrammatik besteht aus folgenden Teilgrammatiken:

- Verben: Die Teilgrammatik besteht aus den Dateien verb-lexicon.lexc, verb-particles.xfst, verb-rules.xfst, verb-vars.xfst und verb-allos.xfst. Die Hauptdatei ist verb-rules.xfst. Das Lexikon enthält ca. 6 950 Einträge.[216]
- Substantive: Die Teilgrammatik besteht aus den Dateien substantive-lexicon.lexc, substantive-rules.xfst, substantive-vars.xfst, substantive-articles.xfst und substantive-allos.xfst. Die Zahl der Lexikoneinträge beträgt ca. 42 850.
- Adjektive: Die Teilgrammatik besteht aus den Dateien adjective-lexicon.lexc, adjective-rules.xfst, adjective-vars.xfst, adjective-articles.xfst und adjective-allos.xfst. Die Zahl der Lexikoneinträge beträgt ca. 16 250.
- Numeralia: Die Zahl der Einträge beträgt 28. Die Grammatik generiert und analysiert Zahlen bis 999 999 999, sowohl Kardinal- und Ordinalzahlen. Mit einfachen Erweiterungen der Regeln nach dem Muster für

[216] Die Zahlen (für alle Teillexika) sind gerundet angegeben.

Tausend oder Million kann die Grammatik theoretisch unendlich viele Zahlen parsen.

- Pronomina: Die Teilgrammatik besteht aus den Dateien pronoun-lexicon.lexc, pronoun-vars.xfst und pronoun-grammar.xfst. Die Zahl der Einträge im Teillexikon beträgt ca. 800. Es handelt sich um Vollformen. Eine Allo-Datei ist hier nicht nötig.
- Adverbien: Die Teilgrammatik besteht aus den Dateien adverb-lexicon.lexc, adverb-vars.xfst und adverb-grammar.xfst. Es sind ca. 3 050 Einträge im Lexikon enthalten.

Die Teilgrammatiken der folgenden Wortarten bestehen aus den Dateien <name>-lexicon.lexc, <name>-vars.xfst und <name>-grammar.xfst. Die Zahl der Einträge liegt insgesamt unter 1 000.

- Präpositionen
- Konjunktionen
- Interjektionen
- Partikeln
- Artikel
- Onomatopoetika

Die folgenden Teilgrammatiken können verschiedene Mengen von Lexikoneinträgen enthalten. Während es sich bei den Interpunktionszeichen nur um wenige Einträge handelt, beträgt die Anzahl der Einträge bei Abkürzungen maximal einige hunderte und bei den Teilgrammatiken der Namen kann sie theoretisch unbeschränkt sein, in der Praxis variiert sie je nach Bedarf, jedoch bei umfangreichen Grammatiken von ca. 10 000 bis ca. 50 000.

- Abkürzungen
- Interpunktionszeichen
- Namen (Personennamen), ca. 1150
- Namen (Ortsnamen), ca. 2250
- Herkunftsnamen der Personen (Ortsnamen), ca. 450

Wortbildung (Derivation und Komposition): Die Zahl der Einträge hier entspricht der der Substantive, Adjektive, Verben und Adverbien sowie der der Wortbildungsmittel, d. h. der Präfixe, Suffixe usw. zusammen. Die Letzteren machen zusammen ca. 500 Einträge aus. Zusammen stehen diese den Wortbildungsregeln zur Verfügung. Die Zahl der möglichen Produktionen ist deutlich höher und entspricht der der möglichen Kombinationen der Regeln mit Lexikoneinträgen.

Hypothesenregeln annotieren Wortformen, die durch die Grammatik nicht erkannt werden, morphologisch. Sie werden in der Regel nicht eingesetzt, da einige Flexionssuffixe für sehr viele grammatische Eigenschaften stehen. Die Hypothesenregeln können eingebunden werden oder als separate Grammatik für die nicht erkannten Wortformen aus der Hauptgrammatik eingesetzt werden. Die Zahl der Einträge entspricht der Summe der Zahlen der Substantiv-, Adjektiv- und Verbflexionssuffixe. Ein Beispiel ist im Abschnitt 6.5.3, Listing 6.5, gegeben.

Listing 5.29 zeigt die Stelle, in der die Teilgrammatiken in die Hauptgrammatik eingebunden werden.

Listing 5.29: Hauptgrammatik und ihre Einzelkomponenten

```
#!/home/bk/bin/xfst -l

load stack verbs/verbs.fst;
define Verbs;

load stack nouns/nouns.fst;
define Nouns;

load stack adjectives/adjectives.fst;
define Adjectives;

load stack numbers/numbers.fst;
define Numbers;

load stack pronouns/pronouns.fst;
define Pronouns;

load stack adverbs/adverbs.fst;
define Adverbs;

load stack articles/articles.fst;
define Articles;

load stack conjunctions/conjunctions.fst;
define Conjunctions;

load stack particles/particles.fst;
define Particles;

load stack prepositions/prepositions.fst;
define Prepositions;

load stack interjections/interjections.fst;
define Interjections;

```

```
load stack onomatopoetica/onomatopoetica.fst;
define Onomatopoetica;

load stack abbreviations/abbreviations.fst;
define Abbreviations;

load stack interpunctuations/interpunctuations.fst;
define Interpunctuations;

load stack place-names/names.fst;
define PlaceNames;

load stack inhabitant-names/names2.fst;
define InhabitantNames;

load stack composition/composition.fst;
define Composition;

! load stack hypothesis/hypothesis.fst;
! define Hypothesis;

read regex
 Verbs | Nouns | Adjectives | Numbers | Pronouns
| Articles | Adverbs | Conjunctions | Particles
| Prepositions | Interjections | Onomatopoetica
| Interpunctuations | Abbreviations | PlaceNames
| InhabitantNames | Composition
!| Flexion
!| Derivation
!| Hypothesis
;

turn stack
compose net
save stack AlbanianMorphology.fst
```

Die einzelnen Grammatiken werden unabhängig voneinander organisiert und implementiert. Die Hauptgrammatik bindet die Kompilate der einzelnen Teilgrammatiken in einer Datei ein und ermöglicht somit ihren Einsatz als ein komplettes Werkzeug.

5.13 Eigenschaften des Morphologie-Systems

Das Morphologie-System ist in Modulen organisiert, welche der Klassifikation des Wortschatzes in Wortarten folgen. Die internen Eigenschaften einer Wortart, wie z. B. die Konjugationsklassen der Verben oder Deklinationsklassen bei Substantiven und Adjektiven wurden soweit wie möglich einheitlich behandelt. Bei Variation des Stammes im Flexionsparadigma eines Lexikoneintrages wurden Allomorphieregeln benutzt, um die Stammallomorphe zu generieren. Diese werden mit den passenden Suffixparadigmen, welche zuvor in Klassen organisiert wurden, kombiniert. Es handelt sich um eine Kombination Klasse von Stamm bzw. Stammallomorph + Klasse von Suffixparadigma. Alternativ wäre es möglich, die Allomorphe der Stämme

direkt in ein Lexikon einzutragen, ohne sie durch Regeln für Stammallomorphe zu erstellen. Die Allomorphmethode hat demgegenüber den Vorteil, dass Allomorphe systematisiert und somit auch besser gewartet werden können. Auch eine etwaige Erweiterung oder Änderung des Wortschatzes wäre leichter zu handhaben.
Für die Modellierung und Implementierung der Wortbildung zeigt die gesammelte Erfahrung, dass abhängig von den Eigenschaften der Sprache verschiedene Strategien und Modelle gewählt werden müssen. Mit einem gut klassifizierten lexikalischen Wissen und einer guten morphologischen Beschreibung wäre die Behandlung der Flexion deutlich leichter.
Die Komponente der Wortbildung, die sowohl Derivation als auch Komposition beinhaltet, nutzt Stämme, die in dieser Komponente erstellt bzw. erweitert werden, und zusätzliches morphologisches Wissen, bspw. über Fugenelemente, Präfixe und Suffixe, um neue Wortformen zu modellieren. Obwohl diese Komponente als Prototyp zu sehen ist, deckt sie die regulären Flexionsklassen und die produktiven Typen der Wortbildung, die bekannterweise einen großen Teil des Wortschatzes ausmachen, ab.

5.14 Zusammenfassung des 5. Kapitels und Schlussbemerkungen

Die Ausschnitte aus den jeweiligen Wortarten und aus dem zusätzlichen sprachlichen Material, sowie eine Übersicht über die Komponenten der Grammatik geben einen Überblick über das erstellte Morphologiesystem im Rahmen von XFST für das Albanische.
Einige Teilkomponenten, wie z. B. die der Implementierung der Namen, insbesondere der Personennamen, konnten nur prototypisch erstellt werden. Auch die Komponente der Wortbildung deckt nicht alle Phänomene ab, da die zur Verfügung stehenden Daten und Literatur leider beschränkt sind.
Trotz dieser Einschränkungen kann das Morphologiesystem für verschiedene Zwecke im Rahmen der maschinellen Sprachverarbeitung eingesetzt werden. Sie deckt die Flexion des Albanischen sowie die Indeklinabilia ab, zuzüglich der Haupttypen der Derivation und der häufigsten Phänomene der Komposition.
Einige Wortformen sind sogar unter den präskriptiven Werken nicht einheitlich/einstimmig. So sind zum Beispiel die Formen des Verbs *shes* (dt. *verkaufen*) bei [Munishi 1998] *shitja*, *shitje*, *shiste* (vgl. Muster 59), während sie bei [Bega/Bega 2007] (vgl. Muster 67) *shisja*, *shisje*, *shiste* sind.

6 Testressourcen und Evaluierung der Arbeit

Nach der Erstellung eines Produktes ist es üblich, dieses zu testen. Dadurch wird festgestellt, was für eine Qualität erreicht wurde. Der Test gewinnt an Bedeutung, wenn das Programm bzw. das Werkzeug in Bereichen eingesetzt wird, die eine grundlegende Bedeutung haben, wie es bei der maschinellen morphologischen Verarbeitung der Fall ist. Sie kann u. a. in einer Verarbeitungskette (Pipeline) eingesetzt werden, wobei ihre Ausgabe als Eingabe einer anderen Komponente verwendet wird. Falsche oder unklare Ausgaben mindern bzw. verfälschen die Qualität der Komponenten, welche diese Ausgaben verwenden.
Zunächst werden in Abschnitt 6.1 die Testformen besprochen. Die vorhandenen Korpora des Albanischen, behandelt in Abschnitt 2.4.1, werden in Abschnitt 6.2 erneut erwähnt, und zwar diesmal als mögliche Testressourcen. In Abschnitt 6.3 geht es um die Erstellung eines kleinen Korpus zu Testzwecken. Wichtige Eigenschaften und einige morphologische und morphosyntaktische Besonderheiten der Texte werden in Abschnitt 6.4 vorgestellt. Im nächsten Punkt (6.5) werden die Methoden, die zum Testen verwendet wurden, sowie Statistiken und weitere Daten zu den Erkennungsraten der erstellten Morphologie geliefert. Schließlich werden in Abschnitt 6.7 eine Zusammenfassung des Kapitels sowie Schlussbemerkungen gegeben.

6.1 Testformen

Um das erstellte Werkzeug zu testen, reicht eine kleine Stichprobe nicht aus. Wenn die Morphologie-Komponente mehrere hunderttausend (bis über die Millionengrenze hinaus) Wortformen erkennen kann, ist es wünschenswert, dass einige tausend Wortformen, bspw. 100 000, abgefragt werden. Zu den verbreitetsten Testformen zählen folgende:

- Testen durch Stichprobenabfragen
- Testen durch ein Vollformlexikon
- Testen durch Wortformlisten

Ein Test durch Stichprobenabfragen, wie z. B. direkt an einem xfst-Laufzeit-Prompt, wäre unpraktisch, nicht ausreichend und außerdem unübersichtlich.
Ein Vollformlexikon, wie in Abschnitt 4.14, Listings 4.16 und 4.17, zu sehen, kann für solche Testzwecke verwendet werden. Ein solches fehlt jedoch wie viele andere Ressourcen für das Albanische in diesem Bereich.
Parallel oder alternativ dazu können Wortformlisten, die aus einem balancierten Textkorpus oder mehreren Textkorpora extrahiert wurden zum Testen verwendet werden. Der Vorteil gegenüber einzelnen Abfragen liegt hierbei darin, dass die Listen automatisch erstellt werden können und in der Regel ausreichend lang sind (bei Korpora mit über einer Million laufender Wortformen). Sie enthalten deklinierte und konjugierte Wortformen, höchstwahrscheinlich auch Neologismen und okkasionelle Wortformen, die nicht in Wörterbüchern enthalten sind bzw. für gewöhnlich erst nach mehrfacher Belegung und ihrer Lexikalisierung dort aufgenommen werden.
Diese Wortformlisten wären in einer Form zu organisieren, die es ermöglicht, systematisch die Morphologiekomponente zu testen, wiederholt und mit veränderbaren Parametern. Eine Sammlung dieser Testwerkzeuge, auch Test-Suite genannt, erleichtert oft das Testen und ermöglicht auch einen Vergleich der verschiedenen Ergebnisse miteinander – nicht zuletzt geben sie auch ein Feedback für die Verbesserung der Software.

6.2 Reichen die vorhandenen Ressourcen zum Testen aus?

Ein Vollformlexikon ist bis jetzt nicht vorhanden. Es wäre auch mit einem Wortschatz mittlerer Größe sehr nützlich gewesen, denn dadurch hätte man die Wortformen des Grundwortschatzes am besten testen können. [ECI/MCI–SQ_AL 1994] enthält ein Lexikon, das die Lemmata von [FDSH 1976] enthält, dargestellt als eine Art Wortliste, die in Abschnitt 2.5 (Listings 2.23 und 2.24) gezeigt wurde. Es bietet außer den in Kapitel 3 und in Kapitel 4 beschriebenen Nennformen leider bei Weitem nicht die Paradigmen der jeweiligen Klassen verschiedener Wortarten.
Vom Fehlen eines Korpus für das Albanische wurde bereits berichtet. In [ECI/MCI–SQ_AL 1994] ist ein nennenswerter Text enthalten. Doch von einem Korpus kann noch lange nicht die Rede sein. Seit 2010 werden im Rahmen des Projektes Europarl, vgl. [SE-TIMES 2010] Texte in albanischer Sprache gesammelt. Diese Texte könnten als Korpus dienen, doch bis jetzt ist nur eine kleine Zahl von Texten vorhanden. Dennoch sieht

das Projekt vielversprechend aus. Umso mehr, da es sich um ein Parallelkorpus mit mehreren Sprachen handelt. Die aktuelle Verwendung der albanischen Sprache und die heutigen Tendenzen bezüglich der Verwendung und Entwicklung des Wortschatzes im Albanischen sind leider nicht ausreichend abgedeckt.
Aus den genannten Gründen und aufgrund der Tatsache, dass es immer noch kein Korpus für das Albanische gibt, das für die maschinelle Sprachverarbeitung geeignet wäre, wird im Rahmen der vorliegenden Arbeit ein Korpus mit dem Ziel erstellt, Wortformlisten, sowie Bi- und Trigramme daraus zu extrahieren, die für Testzwecke der Morphologie-Komponente dienen können.

Die Wortliste sollte:

- in erster Linie aus einem balancierten Korpus extrahiert worden sein,
- ausreichend groß sein, und
- den aktuellen Sprachgebrauch widerspiegeln.

Alternativ könnte ein (Universal-)Wörterbuch bzw. Definitionswörterbuch als eine Art balanciertes Kleinkorpus dienen, um daraus Testlisten mit Wortformen zu erstellen. Ein solches deckt jedoch oft den neuesten Wortschatz erst einige Zeit bzw. Jahre später ab.
Die Erstellung eines Korpus ist nicht Teilaufgabe der vorliegenden Arbeit, sondern dient nur als Zwischenschritt, um die genannte Wortformliste und die N-Gramme zu erstellen. Die Letzteren sind notwendig, da viele Wörter einiger Wortarten wie z. B. Adjektive, analytisch gebaut werden, d. h. aus zwei bzw. mehreren Teilen/Wörtern bestehen. So können Informationen über die Sequenz, Kongruenz, Varietät und über weitere Eigenschaften dieser Wörter in Form von Belegen gesammelt werden, um die Morphologie-Komponente dadurch zu überprüfen.

6.3 Wortformlisten zum Testen

Nachdem die Option eines Vollformlexikons und die der einzelnen Stichproben nicht möglich bzw. unpraktisch sind, bleiben Wortformlisten übrig. Wie sehen die Wortformlisten aus? Sie sind ganz einfache Auflistungen von Wortformen, d. h. eine Wortform pro Zeile. Es werden zwei Typen unterschieden: eine positive und eine negative Testliste.[217] Die positiven Testlisten

[217] Im Abschnitt 6.5.1 wird näher auf den Einsatz der Testlisten eingegangen.

bestehen aus Wortformen, die korrekt geschrieben (und gebildet) sind, d. h. einer Rechtschreibregelung konform sind. Die Formen in einer negativen Testliste sind absichtlich falsch geschrieben, um zu sehen, ob sie auch als inkorrekte Formen vom Programm erkannt werden.
Diese Listen (1) können aus Wortformen bestehen, die zufällig aus allen Wortarten und möglichst aus allen Klassen der Wortarten entnommen sind. Eine fast vollkommene Methode (2) wäre es, die Paradigmen der Deklinations- bzw. Konjugationsklassen zu nehmen. So hätte man alle Elemente eines Paradigmas prüfen können. Eine andere Methode (3) wäre, eine Liste der Wortformen aus einem balancierten Korpus zu extrahieren und damit die erstellte Grammatik zu testen.
Die Erstellung dieser Listen, genauer die Sammlung der beinhalteten Wortformen, macht die Qualität des Tests aus. Je besser die Liste den entsprechenden Wortschatz abdeckt, desto umfangreicher kann das Tool durch die Wortformen der Sprache überprüft werden.
Für Testzwecke stellt XFST das Tool lookup zur Verfügung. Im Abschnitt 6.5.1 wird näher darauf eingegangen. Es nimmt eine Liste mit Wortformen und testet sie, indem einerseits in der Ausgabe die erkannten Wortformen mit den entsprechenden Kategorien versehen werden, andererseits die nicht-erkannten Wortformen mit einem Fragezeichen (?) markiert werden, vgl. z. B. Listing 6.1.

6.4 Texte und einige ihrer Besonderheiten

Die Texte, insbesondere diejenigen, die als Fachtexte bezeichnet werden, beinhalten nicht nur gewöhnliche Textzeichen, sondern sind mit vielen zusätzlichen Textelementen „angereichert“. Oft braucht man nicht so weit zu gehen, um „ungewöhnliche“ Texte zu finden. Schon einige Abkürzungen, insbesondere in der Presse, wie etwa *zv/ministri* ← zëvendës/ministri ← zëvendësministri (dt. *stellvertretender Minister*) oder *p/ligji* ← projekt/ligji ← projektligji (dt. *Gesetzentwurf*), bereiten die ersten Schwierigkeiten bei der Verarbeitung der Daten aus Korpora. Einige Textwörter können ins Lexikon eingetragen werden, andere müssen gesondert behandelt werden, damit sie keine Schwierigkeiten bereiten.
Auch mehrteilige Wörter (engl. multi-word units) bereiten Schwierigkeiten und können nur als einzelne Wortformen behandelt werden, etwa Namen oder Datumsangaben, bspw. *Malësia e Madhe* (Name dt. etwa. „Große Berglandschaft“) [*Malësi e Madhe*, *Malësisë së Madhe*, (në) *Malësi të Madhe*, ...] (flektierte Formen), *28 Nëntor 1912* (28. November 1912), *nga 12 deri 14 prill të*

vitit 2012 (dt. *vom 12. bis 14. April des Jahres 2012*), *Rruga A 2* (dt. *Straße A 2*), *nr. 12 A* usw., vgl. hierzu auch [van Halteren 1999], u. a. [Grefenstette 1999].
Einige Texte können selbst in der Tagespresse komplexe Kombinationen beinhalten, wie z. B. „55 lekë/m^3“ Es folgt ein Beispiel, in dem m^3 ausformuliert wurde, vgl. *metër kub*:

> *Kështu, çdo metër kub ujë të harxhuar, konsumatorët familjarë të Lushnjes do të paguajnë me 44 lekë, bizneset me nga 110 lekë, ndërsa institucionet buxhetore me 100 lekë.* [218]

Ein weiteres Beispiel wäre die folgende Textpassage:

> *Duke qenë se dy perimetrat janë tek 4 metra dhe se P=2πR, ku π=3.14 dhe 2R=diametrin, rezulton një diametër diku tek 1.30-1.40 metra. Pesha nuk mund të përcaktohet pa e peshuar ose pa ditur çfarë materiali është që të shumëzohet vëllimi me dendësinë specifike, po nga rrezja me formulën V=(4/3)πR3 nxjerrim se vëllimi duhet të jetë diku tek 1.15 metër kub.*[219]

Fälle wie *41 mijë e 51 banorë të qytetit*, wobei Numeralia in gemischter Form (Wort/Zahl) geschrieben werden, müssen gesondert behandelt werden, nämlich als einzelne Wortformen, welche in einem Folgeschritt auf der morphosyntaktischen Abstraktionsebene verarbeitet werden müssten.

6.4.1 Neologismen und okkasionelle Verwendungen

Bei der maschinellen morphologischen Verarbeitung der Texte begegnet man Wortformen, deren Grundformen nicht im Lexikon enthalten sind, z. B. aufgrund seines geringen Umfanges, und dementsprechend von Systemen, die strikt regelbasiert arbeiten, nicht analysiert werden können. In diesen

[218] Übersetzung: *So, jeden Kubikmeter verbrauchtes Wasser werden die Familien von Lushnje mit 44 Lek (d. i. Währung in Albanien), Geschäfte mit 110 Lek und die Behörden mit 100 Lek bezahlen.*

[219] Shekulli 02/03/2012: <http://www.shekulli.com.al/shekulli/2012/03/02/sfera-e-↪cuditshme-e-gjetur-ne-rrugen-rreshen-kalimash/> (*Sfera e çuditshme e gjetur në rrugën Rrëshen-Kalimash*). Übersetzung: *Da die beiden Umfänge bei 4 Metern sind und P=2πR, wobei π=3.14 und der Umfang gleich 2R ist, ergibt sich ein Durchmesser zwischen 1.30-1.40 Metern. Das Gewicht kann nicht bestimmt werden, ohne ihn zu wiegen oder ohne das Material zu kennen, damit das Volumen mit der spezifischen Dichte multipliziert werden könnte, aber vom Radius her, durch die Formel V=(4/3)πR3 berechnet, müsste das Volumen ca. 1.15 Kubikmeter sein.*

Fällen wird oft das Lexikon um die neuen unbekannten Wörter/Wortformen erweitert, um sie bei einer späteren Verarbeitung berücksichtigen zu können. Auch in einem noch so umfangreichen Lexikon können manche Wortformen nicht erkannt werden, z. B. Entlehnungen oder Neologismen wie im folgenden Satz:

> *Italianët e dinin fort mirë se atje ai mund të bëhej pre e ndonjë atentati të fanolistëve, sikurse ndodhi në të vërtetë.*[220]

Das Wort bedeutet *ein Anhänger von Fan Noli* (Thoefan Stilian Noli).[221] Diese Wortbildung findet immer wieder Verwendung, so dass sie als Lemma (Substantiv) in [Dhrimo/Memushaj 2011] enthalten ist. In früheren Standardwörterbüchern findet man das Wort nicht.
Die Wörter *kryeqeverisësit* („kryeqeverisës") (dt. *der Hauptbeamte*) (‚Haupt' + ‚Regierender'), *mëtejshmëria* („mëtejshmëri") (dt. *die Weiterführung*) und *përshoqërimit* („përshoqërim") (dt. *die Mitbegleitung*), sind im Vergleich zu *fanolistëve* Wörter, genauer Wortformen, die aus dem bestehenden Wortschatz kreiert wurden. Als Neologismen können die letzteren leichter oder problemlos erkannt werden, indem die Segmentierung bzw. die Konkatenation der Wortbildungselemente durch entsprechende Regeln erkannt wird, z. B. kann *përshoqërim* als eine Derivation von *shoqërim* erkannt werden.
Ein weiteres Beispiel für eine Verwendung, die auf das Adjektiv *i␣, e␣përkushtuar* (dt. *mit Zuwendung* u. ä.) zurückgeht: „*një punë e përkushtueshme disavjeçare …*" (dt. ≈ *sich einer Arbeit mit Hingabe jahrelang widmen …*). Die Form *e␣përkushtueshme* findet man nicht in allen Wörterbüchern, sie kann jedoch im gegebenen Satz ohne Schwierigkeiten verstanden werden. Die übliche Form wäre *e␣përkushtuar*. Statt der Ableitung aus der Partizipialform des Verbs *përkushtoj* wurde in diesem Fall die Bildung der Adjektive mit dem Suffix *-shëm/-shme* verwendet.

6.4.2 Ambiguität

Bei der automatischen morphologischen Analyse eines Textes/Korpus kommt es vor, dass bei einem Textwort mehrere Analysen möglich sind. Dieses als Ambiguität bekannte Phänomen bereitet Schwierigkeiten bei

[220] Ausschnitt aus dem Artikel „Servilizmi politik e historik", Ausgabe vom 27.6.2012 (on-line: <http://gazeta-shqip.com/lajme/2012/06/27/servilizmi-politik-e-historik/>) der unabhängigen Tageszeitung *Gazeta Shqip*. Übersetzung: *Die Italiener wussten sehr gut, dass er dort Opfer eines Attentats von Fanolisten werden könnte, wie es in der Tat auch geschah.*

[221] 1882–1965, Orthodoxer Priester, Übersetzer, Schriftsteller, Historiker und Politiker.

der Analyse, die als Tags den Wörtern hinzugefügt wird. Einige Beispiele, welche als Folge von Homonymie bzw. Homomorphie eine Mehrdeutigkeit aufweisen:

- alban. *thot*, dt. *sagen* / alban. *thahet*, dt. *austrocknen*
  - *ai tha se ...* [→ *sagen*];
    er sagte, dass ... ;
    ai u tha atyre ... [→ *sagen*];
    er sagte denen, dass ... ;
  - *U tha në mbledhje se ...* [→ *sagen*];
    In der Versammlung wurde gesagt, dass ... ;
  - *pema u tha në mungesë uji ...* [→ *austrocknen*];
    Der Baum war mangels Wasser ausgetrocknet.
- *para*, dt. *Geld* vs. *para*, dt. *früher*;
  - *shumë para*, dt. *viel Geld* vs.
  - *shumë para*, dt. *viel früher, vorher*, räumlich *para*, dt. *vor*;
- *shoh*, dt. *sehen* vs. *pe*, dt. *Faden*
  - *pe*: $shoh_{[aor.2sg.]}$ vs.
  - *pe*: $pe_{[nom.unbest.sg.]}$;

Einen hohen Grad an Ambiguität zeigen die „kurzen Wörter“, die Partikeln, die Präpositionen und die Artikel. Einige Beispiele wären *e*, *i*, *të*, *së*, die verschiedene syntaktische und semantische Rollen erfüllen können.

6.5 Evaluierung der Morphologie-Komponente

Für die Evaluierung der Morphologie-Komponente wird das von XFST bereitgestellte Werkzeug lookup verwendet. Damit ist es möglich, einen einfachen und überschaubaren Test von Wortformen durchzuführen.

6.5.1 Testen der Morphologie mit XFST

Die Datei grammar-test.lup listet die Teilautomaten auf und ermöglicht in einem Schritt deren Durchsuchung mit Hilfe von lookup (Parameter -f, für *one or more strategies - lists of cascaded fst-s*). Für die Durchsuchung wird eine vom xfst-Tool kompilierte Grammatik gebraucht, wie z. B. cat test.lst | lookup -f regeln/grammatik-binary.fst.
Die sogenannten *positiven* und *negativen Listen* mit Wortformen sind ein Standardweg, die erstellten Grammatiken zu testen. In eine positive Liste werden korrekte Wortformen, üblicherweise in großer Zahl, eingetragen und von der Grammatik geprüft. Sie müssten korrekte Ergebnisse liefern, d. h., die Grammatik müsste sie erkennen und mit richtigen Analyse- bzw. Produktionsangaben versehen. Dabei ist nicht auszuschließen, dass Wortformen irrtümlich „erkannt“ werden und mit entsprechenden Informationen versehen werden. Das würde bedeuten, dass die Grammatik einer Verbesserung bedarf. Die Fälle der „irrtümlichen Erkennung“ sollen insbesondere vermieden werden, um die statistischen Daten (Erkennungsraten) der Grammatik nicht zu verfälschen.

Ein Beispiel für den Test einer positiven Liste oder mehrerer positiver Testlisten: cat *.test-list-positive.txt | lookup -f grammar-test.lup

Listing 6.1 zeigt einen Ausschnitt des Tests einiger Wortformen mittels des Programms lookup.

Listing 6.1: Look-Up-Test im Rahmen von XFST

```
1
2  # ./start-test-grammar.sh
3  Reading script from "grammar-test.lup"
4  0%>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>100%
5  0%>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>100%
6
7    *****  LEXICON LOOK-UP  *****
8
9  ndrittë ndrit   +V+3P+Sg+Opt+Pres+Act+NonAdm
10
11 ndritte ndritte +?
12
13 gjettë  gjej    +V+3P+Sg+Opt+Pres+Act+NonAdm
14
15 ishte   jam     +V+3P+Sg+Ind+Impf+Act+NonAdm
16
17 ka      kam     +V+3P+Sg+Ind+Pres+Act+NonAdm
18
19 jam     jam     +V+1P+Sg+Ind+Pres+Act+NonAdm
20
21 punonte punoj   +V+3P+Sg+Ind+Impf+Act+NonAdm
22
23 punote  punote  +?
24
25 ishin   jam     +V+3P+Pl+Ind+Impf+Act+NonAdm
```

```
26|
27| ...
28|
29| LOOKUP STATISTICS (success with different strategies):
30| strategy 0:     0 times          (0.00 %)
31| strategy 1:     0 times          (0.00 %)
32| strategy 2:     0 times          (0.00 %)
33| strategy 3:     0 times          (0.00 %)
34| strategy 4:     0 times          (0.00 %)
35| strategy 5:     0 times          (0.00 %)
36| strategy 6:     10 times         (7.63 %)
37| strategy 7:     50 times         (38.17 %)
38| strategy 8:     0 times          (0.00 %)
39| strategy 9:     0 times          (0.00 %)
40| not found:      71 times         (54.20 %)
41|
42| corpus size:    131 words
43| execution time: 0 sec
44| speed:          131 words/sec
45|
46|   *****  END OF LEXICON LOOK-UP  *****
```

Es wurden nur 131 (zufällig gewählte) Wortformen getestet, vgl. Zeile 42 in Listing 6.1. Davon wurden 60 erkannt und 71 nicht. Die „Strategien“ (engl. strategies) in den Zeilen 29–39 stehen für verschiedene Teilgrammatiken. Die erkannten Wortformen wurden von den Teilgrammatiken 6 und 7 verarbeitet. In den Zeilen 9 bis 25 sind in der ersten Spalte jeweils die Wortformen angegeben, in der zweiten die Grundformen der jeweiligen Wortformen, gefolgt von ihren entsprechenden grammatischen Eigenschaften in der dritten Spalte. Die nicht erkannten Wortformen sind in der dritten Spalte mit +? markiert.

6.5.2 Morphologische Annotation mit einem Vollformlexikon

In Abschnitt 4.14 wurde ein Vollformlexikon vorgestellt, das zu Testzwecken erstellt wurde. Listing 6.2 zeigt einen Ausschnitt eines Tests. Dabei handelt es sich um den Einsatz eines Perl-Skriptes, das zum Taggen benutzt wird und das Vollformlexikon als Ressource verwendet. Dies ermöglicht, wenn auch wegen der bekannten Beschränkungen des Vollformlexikons nicht das komplette Taggen der Wortformen eines Texts, so doch zumindest teilweise einige seiner Segmente zu verarbeiten und einzelne Stichproben durchzuführen.[222]

Listing 6.2 zeigt einen Ausschnitt einer Annotation am Beispiel des Personalpronomen *atyre*, dt. *denen/deren*.

[222] Außerdem beinhaltet es nicht alle Wortarten, sondern nur die deklinierbaren, nämlich Substantive, Adjektive, Verben und Pronomina, und davon nur den Grundwortschatz.

Listing 6.2: Tagging

```
atyre
{
  'Pron' => [
              {
               'LE' => 'ata',
               'Tags' => [
                           'Pron-001_Pers.AbP3m',
                           'Pron-001_Pers.DP3m',
                           'Pron-001_Pers.GP3m'
                         ]
              },
              {
               'LE' => 'ato',
               'Tags' => [
                           'Pron-001_Pers.AbP3f',
                           'Pron-001_Pers.DP3f',
                           'Pron-001_Pers.GP3f'
                         ]
              }
             ]
};
```

Um zu dieser Ausgabe zu kommen, werden mit Hilfe von Perl-Skripten eine Wortliste, gewonnen aus einem Text, und das Vollformlexikon eingelesen und miteinander verglichen. Im Falle einer Übereinstimmung einer Wortform aus der Liste mit einer im Vollformlexikon werden der Wortform in der Liste die Eigenschaften aus dem Lexikon hinzugefügt. Falls zu einer Wortform in der Liste mehrere Einträge im Vollformlexikon vorhanden sind, werden sie alle je nach Eigenschaften (wie Wortart, Lemma usw.) sortiert hinzugefügt. Der Wortform *atyre* entsprechen sowohl das Lemma *ata* als auch *ato*, was dementsprechend separat eingetragen wird.
Die Eigenschaften, die die Wortformen besitzen, sind mit `'Tags'` markiert. Es sind pro Eintrag drei. Ein weiteres Beispiel, bei dem die Eigenschaften nicht ambig sind, zeigt Listing 6.3.

Listing 6.3: tagging-pron

```

atë {
        'Pron' => [
                    {
                      'LE' => 'ajo',
                      'Tags' => [
                                  'Pron-001_Pers.AcS3f'
                                ]
                    },
                    {
                      'LE' => 'ai',
                      'Tags' => [
                                  'Pron-001_Pers.AcS3m'
                                ]
                    }
                  ]
      };
```

Während es in den Listings 6.2 und 6.3 um eine Übereinstimmung der Wortformen zweier Lemmata innerhalb einer Wortart geht, handelt es sich in Listing 6.4 um eine Übereinstimmung der Formen der Lemmata aus unterschiedlichen Wortarten, nämlich Substantiv und Verb. Dies ist entsprechend mit `'S'` und `'V'` markiert. Das Tag `'LE'` steht für Lemma.

Listing 6.4: Tagging-2

```
mund
{
 'S' => [
          {
           'LE' => 'mund',
           'Tags' => [
                       'S-006_Acc.Sg.Indef;',
                       'S-006_Nom.Sg.Indef;'
                      ]
          }
         ],
 'V' => [
          {
           'LE' => 'mund',
           'Tags' => [
                       'V-003_1.Pers.Sg.Indv.Pres.Act.Adm-;',
                       'V-003_1.Pers.Sg.Sbjv.Pres.Act.Adm-;',
                       'V-003_2.Pers.Sg.Indv.Pres.Act.Adm-;',
                       'V-003_2.Pers.Sg.Impv.Pres.Act.Adm-;',
                       'V-003_3.Pers.Sg.Indv.Pres.Act.Adm-;',
                       'V-003_3.Pers.Sg.Indv.Pres.Pas.Adm-;'
                      ]
           }
          ]
};
```

Die Ausgabe aus dieser Annotation wurde als eine Ressource verwendet, um in einzelnen Fällen die Ausgaben der Verarbeitung mit XFST zu überprüfen, indem beide Ausgaben miteinander verglichen wurden.
Die Nachteile dieser Annotation gegenüber einer Verarbeitung mit XFST sind, dass sie umständlicher ist, mit mehr Wartungsaufwand verbunden ist und dass die Möglichkeiten (Features), die XFST bietet, implementiert werden müssen. Dies war nicht das Ziel der vorliegenden Arbeit.

6.5.3 Neologismen und Hypothesen

In Abschnitt 6.4.1 wurde auf das Thema *Neologismen und okkasionelle Verwendungen* eingegangen. Dabei wurde ein weiterer Fall nicht behandelt, bei dem sowohl eine Lexikonerweiterung als auch eine umfangreiche Grammatik einige Textwörter erkennen können.

Um zumindest die grammatische Funktion der Textwörter zu erkennen, bietet es sich aufgrund der Flexionsmarkierung einiger Wortarten des Albanischen an, einige Regeln im Rahmen von XFST zu schreiben, um Textwörter, die nicht erkannt wurden, gesondert zu überprüfen. Dabei werden nur potentielle grammatische Merkmale annotiert, den Wortformen können jedoch keine Grundformen zugeordnet werden.
Einen Ausschnitt zeigt Listing 6.5. Es handelt sich um erfundene Wortformen, die jedoch auf Flexionsendungen des Albanischen enden:

Listing 6.5: Hypothesen ...

```
xfst[1]: up zzzzonte
zzzzonte+S+Fem+Gen+Sg+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Gen+Pl+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Dat+Sg+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Dat+Pl+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Abl+Sg+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Abl+Sg+Det<UNCONFIRMED>
zzzzonte+S+Fem+Nom+Pl+InDet<UNCONFIRMED>
zzzzonte+S+Fem+Acc+Pl+InDet<UNCONFIRMED>
zzzzonte+V+3P+Sg+Ind+Impf+Act+NonAdm<UNCONFIRMED>
xfst[1]:
```

Für eine Annotierung dieser Wortformen werden die Endungen verwendet, d. h., die rechte Seite (Ende der Wortform) ist entscheidend, während die linke (Anfang des Wortes) nicht berücksichtigt wird. In einem weiteren Schritt könnte noch Gebrauch von einem Minimal-Edit-Distance-Algorithmus gemacht werden, um die restlichen nicht erkannten Wortformen in Kombination mit Hypothesenregeln zu verarbeiten.

6.5.4 Erweiterung des Lexikons und der Morphologie

Das ursprüngliche Lexikon wurde anfangs um die wichtigsten Ortsnamen des albanischsprachigen Raumes sowie um Namen, die in der Sprache, Kultur, Geschichte, Wirtschaft und Wissenschaft häufiger vorkommen, erweitert. Ebenso wurden Ländernamen und Einwohnerbezeichnungen ins Lexikon aufgenommen.
Darüber hinaus wurde das Lexikon während der Entwicklung der Morphologie ständig getestet. Nach jedem Testlauf wurden die erkannten Fehler seitens des Lexikons und der Grammatik kontinuierlich beseitigt, indem entweder ein Lexikoneintrag korrigiert oder ein fehlender hinzugefügt wurde. Zuvor wurden solche Einträge richtig kategorisiert und systematisiert. Gegebenenfalls wurden auch die Morphologiekomponenten, wie z. B. die Flexion, oder bestimmte Regeln überarbeitet.

Die Wörter bzw. Wortformen wurden aus verschiedenen Texten extrahiert, vor allem aus Berichten der Ministerien, etwa der Wirtschaft, und der Pressemitteilungen der Regierungen in Tirana[223], Prishtina[224] und *Agjencia Telegrafike Shqiptare* (dt. *Albanische Nachrichtenagentur*)[225] ab dem Jahr 2011 (bis Februar 2014).
Mit jedem hinzugefügten neuen Text wurden die Wortformen aus dem neuen Text der vorherigen Liste der Wortformen hinzugefügt, falls sie dort nicht enthalten waren. So wurde die Liste der Wortformen länger. Diese Einträge wurden in einer abstrahierten Form (d. h. als Lemmata) in das Lexikon aufgenommen. Konkret: Falls die Wortformen *adoleshenti* und *adoleshentit* (dt. *Adoleszent*) nicht im Lexikon enthalten waren, wurden sie in einem Text gefunden und als *neu* markiert. Sie wurden als Lemma, d. h. *adoleshent* (ohne Flexionsmarkierung) ins Lexikon aufgenommen und dabei der entsprechenden Klasse zugeordnet, damit der Eintrag mit dem passenden Flexionsparadigma und damit den passenden Regeln der morphologischen Verarbeitung verbunden wird.

6.5.5 Der Test der Morphologie

Um die Morphologiekomponente zu prüfen, wurde einerseits der Text des Romans *Koncert* verwendet und andererseits selbsterstellte Testlisten. Die Letzteren sind wiederum entweder automatisch erstellte und auf Rechtschreibung geprüfte Wortformlisten oder selbst erstellte Wortformlisten, die niedrigfrequente Wortformen enthalten.
Zum Testen der Morphologiekomponente wurden drei verschiedene Tests durchgeführt, um grundlegende Aspekte der Wortformen zu berücksichtigen: (1) Testen anhand von wortartübergreifenden Frequenzklassen, (2) Testen nach Wortarten und (3) Testen anhand von Frequenzklassen innerhalb der Wortarten, vgl. 6.5.6, 6.5.7 und 6.5.8.
Durch den ersten Test werden Wortformen, die in *Koncert* vorkommen, nach Frequenzklassen[226] kategorisiert und als solche geprüft. Bevor die Statistiken vorgestellt werden, bedarf es noch folgender Hinweise, die den Text von *Koncert* betreffen:

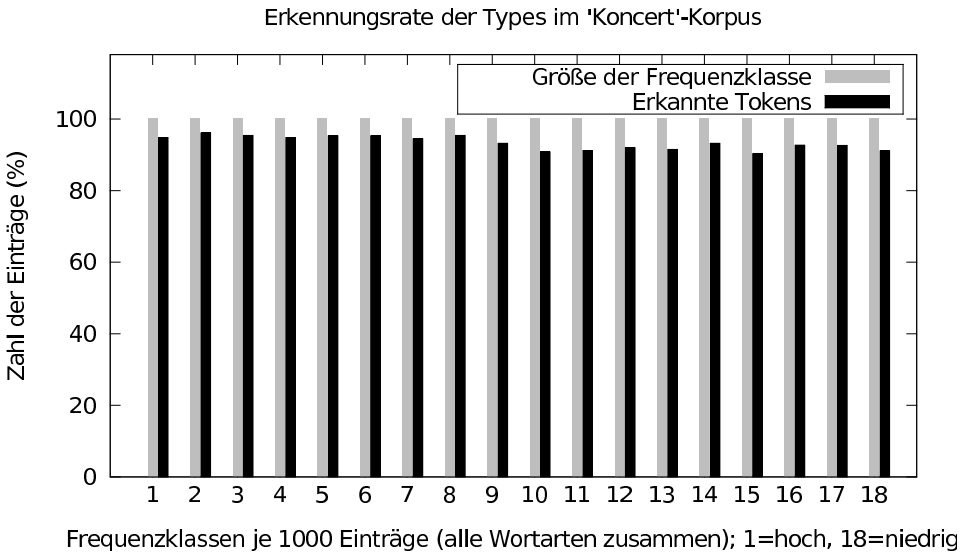
[223] <http://kryeministria.al/> , 21.2.2014..
[224] <http://www.rks-gov.net/> , 21.2.2014.
[225] <http://www.ata.gov.al/> , 21.2.2014..
[226] Vgl. [Baayen 2001: 1-24] zum Thema Wortfrequenzen.

Einige Wortformen, die im Korpus vorkommen, wie z. B. *m'e* [227] entsprechen nicht der Rechtschreibung. Sie müssen bei der Erhebung der Statistiken und dementsprechend bei der Durchführung der Tests berücksichtigt werden.
Da der Korpustext wahrscheinlich durch einen Scanprozess gewonnen wurde, kommen nicht selten „Tippfehler“ vor, wie *kishteqenëaq* statt *kishte␣qenë␣aq* oder *klshin* statt *kishin*, *ngushëliime* statt *ngushëllime* und *qerarnikës* statt *qeramikës*. Diese Fälle müssen genauso berücksichtigt werden.

6.5.6 Wortartübergreifende Frequenzklassen

Abbildung 6.1 zeigt in graphischer Form die Erkennungsrate der Wortformen, gruppiert nach Frequenzklassen. Gruppe 1 beinhaltet die ersten 1000 häufigsten Wortformen unabhängig von ihrer Wortart. Gruppe 2 beinhaltet die nächsten 1000 Wortformen usw. Mit der steigenden Zahl der Gruppe nimmt die Frequenz der Wortformen ab. So bleiben bei den letzten Gruppen nur Hapax Legomena, d. h. Wortformen mir der absoluten Häufigkeit 1, übrig. Gruppe 1 fasst Wortformen von der Frequenz 13693 bis hin zur Frequenz 18 zusammen. Die letzten 9213 Wortformen, aufgeteilt in den letzten 10 Gruppen, kommen im Korpus nur einmal vor.

Abbildung 6.1: Erkennungsrate nach Frequenzklassen der Types im 'Koncert'.



[227] Ismail Kadare, *Koncert në fund të dimrit*, S. 291, Tiranë: Onufri, 2011.

Erkennungsraten in Zahlen sind in Tabelle 6.1 angegeben:

Tabelle 6.1: Testergebnisse in Zahlen.

Frequenzklasse (Types)	Erkennungsrate (%)	Frequenz(bereich)
Freq. 1	94,80 %	13693–18
Freq. 2	95,60 %	18–8
Freq. 3	95,30 %	8–5
Freq. 4	94,40 %	5–4
Freq. 5	95,30 %	4–3
Freq. 6	95,00 %	3–2
Freq. 7	94,50 %	2–2
Freq. 8	95,20 %	2–2
Freq. 9	93,10 %	2–1
Freq. 10	90,50 %	1
Freq. 11	91,10 %	1
Freq. 12	92,00 %	1
Freq. 13	91,40 %	1
Freq. 14	92,70 %	1
Freq. 15	90,30 %	1
Freq. 16	92,40 %	1
Freq. 17	92,60 %	1
Freq. 18	90,64 %	1

Die Erkennungsraten sind hier dank vieler Namen und Fremdwörter etwas niedriger als erwartet. Der Durchschnittswert (Abbildung 6.1) ist 93.23 %.

6.5.7 Testen nach Wortarten

Durch den zweiten Test werden ebenso die Wortformen von *Koncert*, gruppiert nach Wortarten, geprüft. Tabelle 6.2 zeigt die Ergebnisse des Tests der Morphologie mit dem Text des Romans *Koncert*.[228] Es geht darum, zu zeigen, wie die Erkennungsrate der Morphologie ist, da die einzelnen Wortarten unterschiedliche Eigenschaften besitzen und auch eine unterschiedliche Anzahl an Formen aufweisen. In Tabelle 6.2 sind nur Types angegeben, da zum Testen der Morphologiekomponente die Types relevant sind. Diese Werte sind in Abbildung 6.2 (Typen) und 6.3 (Tokens) graphisch dargestellt.

[228] Vgl. hierzu [Koncert 1994].

Tabelle 6.2: Testergebnisse des Koncert-Korpus.

Analyse des Romans *Koncert* (nach Wortarten)	
Wortarten	Types
Substantive	98,12 %
Verben	97,17 %
Adjektive	96,43 %
Adverbien	98,16 %
Pronomina	98,00 %
Numerale	95,00 %
Präpositionen	98,00 %
Konjunktionen	99,00 %
Interjektionen	97,00 %
Partikel	99,00 %

Abbildung 6.2: Erkennungsrate K. (Types)

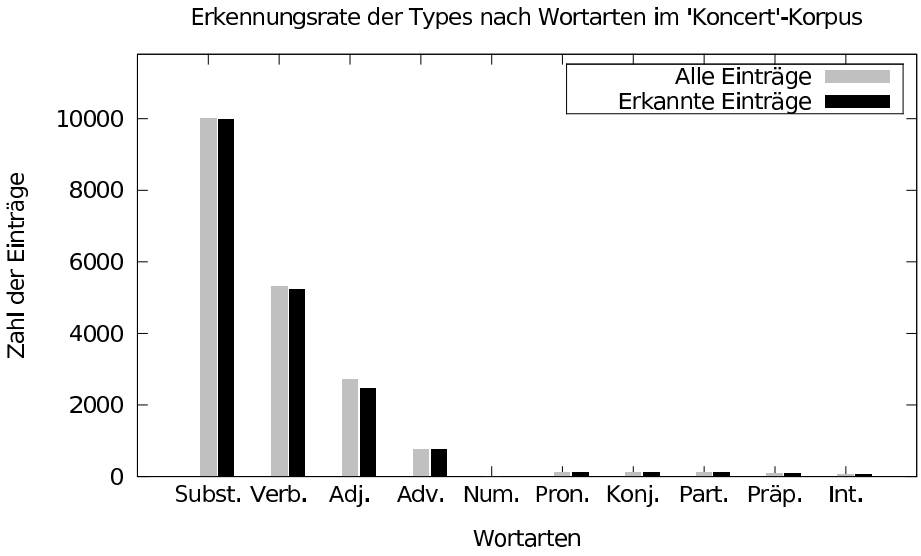


Abbildung 6.3 soll nur einen Eindruck geben, in welchem Maße die Tokens in einem Text im Vergleich zu Types verteilt sein können – konkret, welche Verteilung sie im Text des Romans *Koncert* haben. Insbesondere fällt auf, dass eine kleine Zahl der Types bei den geschlossenen Wortarten, wie Pronomina, Präpositionen, Konjunktionen, Partikel usw., auf eine relativ große Zahl Tokens verteilt sind.

Abbildung 6.3: Erkennungsrate K. (Token)

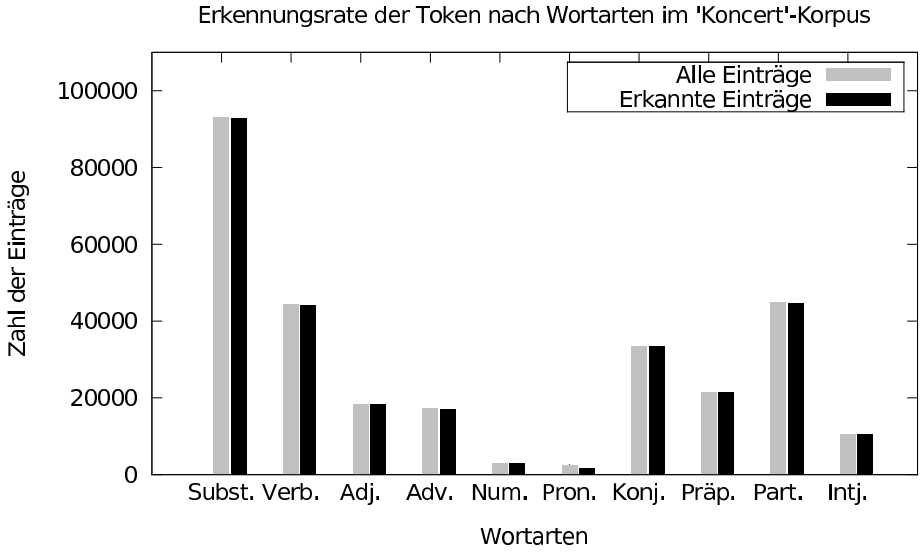


Tabelle 6.3 zeigt die Daten des Tests der Morphologie mit ausgewählten Wortformen. Ziel war es dabei, Wortformen, die selten vorkommen, zu testen. Die Wortformen wurden vorher überprüft, ob sie korrekt im Sinne der Rechtschreibregelung sind. Eine Unterscheidung Type/Token wurde dabei nicht vorgenommen.
Im Unterschied zu Tabelle 6.3 wurden für die Erstellung der Wortformlisten in Tabelle 6.4 Daten genommen, die aus dem Internet extrahiert wurden. In dieser Tabelle sind Testergebnisse der Wortlisten, die einerseits häufig vorkommende Wortformen sowie andererseits selten vorkommende Wortformen, die als Hapax Legomena bekannt sind, d. h. nur einmal vorkommen oder niederfrequent sind, enthalten. Letztere diente dazu, Neologismen abzufangen. Auch diese Wortformen wurden vorher in Wortarten klassifiziert und überprüft, ob sie korrekt im Sinne der Rechtschreibregelung sind. Eine Untersuchung der nicht erkannten Wortformen zeigt, dass die meisten Fälle mit einem im Lexikon nicht vorhandenen Lemma zu erklären sind. Weitere Ursachen sind Fehler in der Kodierung der grammatischen Kategorien im Lexikon oder in der Grammatik sowie falsche Klassifikationen von Lemmata, vgl. hierzu auch Abschnitt 6.6 zur Fehleranalyse.

Tabelle 6.3: Testergebnisse der Test-Wortliste 1.

Testergebnisse der Test-Wortliste 1 (nach Wortarten)	
Wortarten	Types
Substantive 100 000 Wortformen	98,48 %
Verben 39 476 Wortformen	97,22 %
Adjektive 33 874 Wortformen	99,26 %
Adverbien 1 556 Wortformen	95,18 %

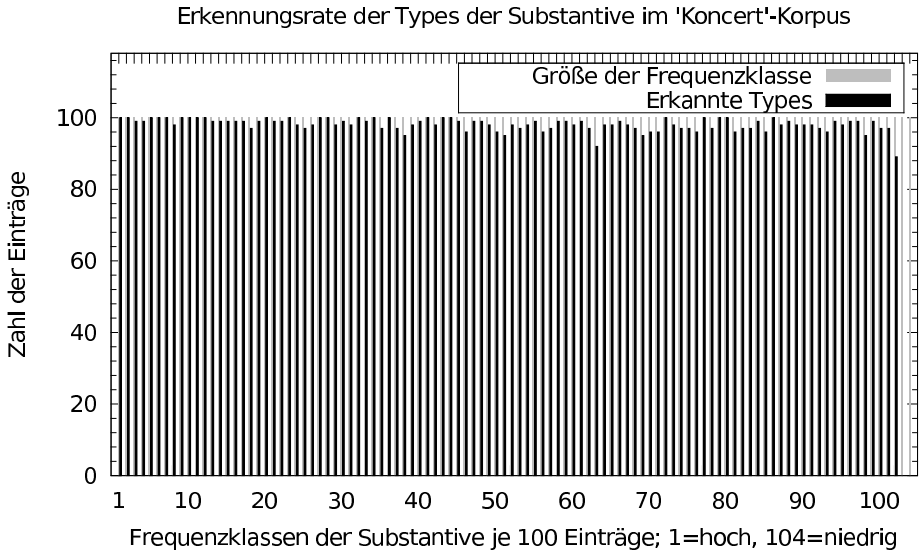
Tabelle 6.4: Testergebnisse der Test-Wortliste 2.

Testergebnisse der Test-Wortliste 2 (nach Wortarten)	
Wortarten	Types
Substantive 40 000 Wortformen	97,93 %
Verben 60 000 Wortformen	97,88 %
Adjektive 10 000 Wortformen	97,52 %
Adverbien 2 341 Wortformen	96,80 %

6.5.8 Frequenzklassen innerhalb der Wortarten

Als Nächstes werden Wortformen der vier häufigsten Wortarten, nämlich der Substantive, Adjektive, Verben und Adverbien, geteilt in Frequenzklassen innerhalb der jeweiligen Wortart, getestet.
Abbildung 6.4 zeigt die Erkennungsraten der Substantive nach Frequenzklassen. Es sind insgesamt 104 Frequenzklassen, wobei eine Frequenzklasse 100 Einträge beinhaltet. Wie in der Abbildung zu erkennen ist, ist die Erkennung der einzelnen Klassen von den hochfrequenten Klassen bis hin zu denen mit der Frequenz 1 relativ gleichmäßig verteilt. Oft wird erwartet, dass die hochfrequenten Klassen gute Erkennungsraten erzielen, während die niedrigfrequenten Klassen proportional zur fallenden Frequenz absinkende Werte vorweisen. Die guten Erkennungsraten der niedrigfrequenten Klassen können neben der Morphologiekomponente, u. a. der Derivation und Komposition, auch dank dem guten Lexikon erzielt werden.

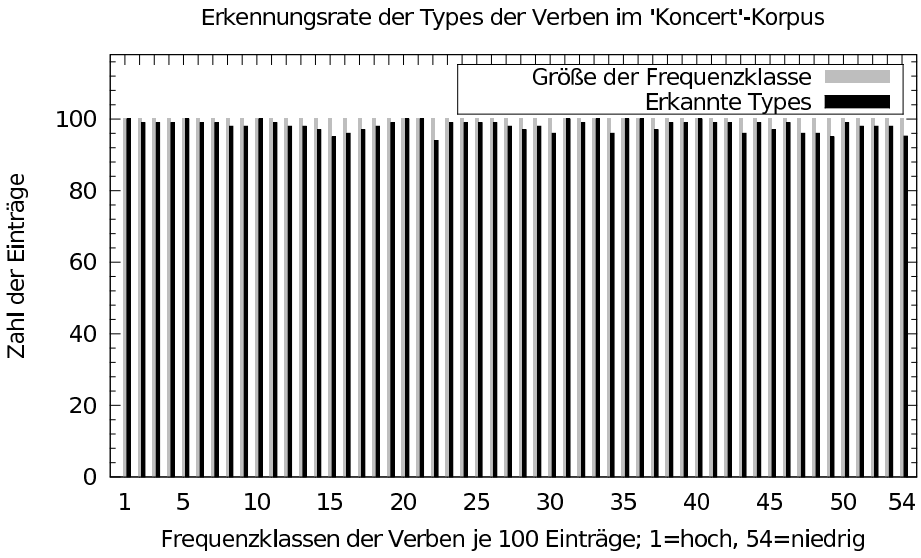
Abbildung 6.4: Erkennungsrate der Substantive nach Frequenzklassen



Die letzten zwei Klassen (103 und 104) in Abbildung 6.4 sind leer, da die Scanfehler im Text beseitigt wurden. Aus zwei Wortformen (eine richtig erkannte und eine nicht erkannte Wortform) wurde eine Wortform. So reduzieren sich die Klassen von 104 auf 102.

In Analogie zu Abbildung 6.4 zeigt Abbildung 6.5 die Erkennungsraten der Verben nach Frequenzklassen. Es sind insgesamt 54 Frequenzklassen mit je 100 Einträgen. Wie es bei Substantiven der Fall ist, zeigt sich auch bei Verben eine Erkennungsrate, die kleine Schwankungen aufweist, jedoch bei allen Frequenzklassen Werte im gleichen Bereich zeigt, obwohl das Paradigma eines Verbs deutlich mehr Formen besitzt als das eines Substantivs. Auf der anderen Seite ist die Zahl der Substantive/Namen in einer Sprache deutlich größer als die der Verben.

Abbildung 6.5: Erkennungsrate der Verben nach Frequenzklassen



Die Erkennungsraten der Adjektive nach Frequenzklassen sind in Abbildung 6.6 dargestellt. Die Adjektive sind in insgesamt 27 Frequenzklassen verteilt. Adverbien bilden eine noch kleinere Gruppe von Wortformen als die Adjektive, die in nur acht Klassen verteilt sind. Die Erkennungsraten der Adverbien nach Frequenzklassen sind in Abbildung 6.7 angegeben.
Die Hapax Legomena bei Substantiven, Verben, Adjektiven und Adverbien, die mehr als eine Klassengröße sind, werden vorher mit Hilfe eines Zufallsgenerators, des Unix-tool shuf, gemischt, damit sie unabhängig von der alphabetischen Ordnung in Klassen (mit Frequenz 1) verteilt werden. Die Erkennungsrate der einzelnen Frequenzklassen kann sich beim Ausführen eines neuen Tests ändern, wobei vorher die Klassen neugebildet werden, ohne dass sich die Erkennungsrate aller Hapax Legomena zusammen ändert.

Abbildung 6.6: Erkennungsrate der Adjektive nach Frequenzklassen

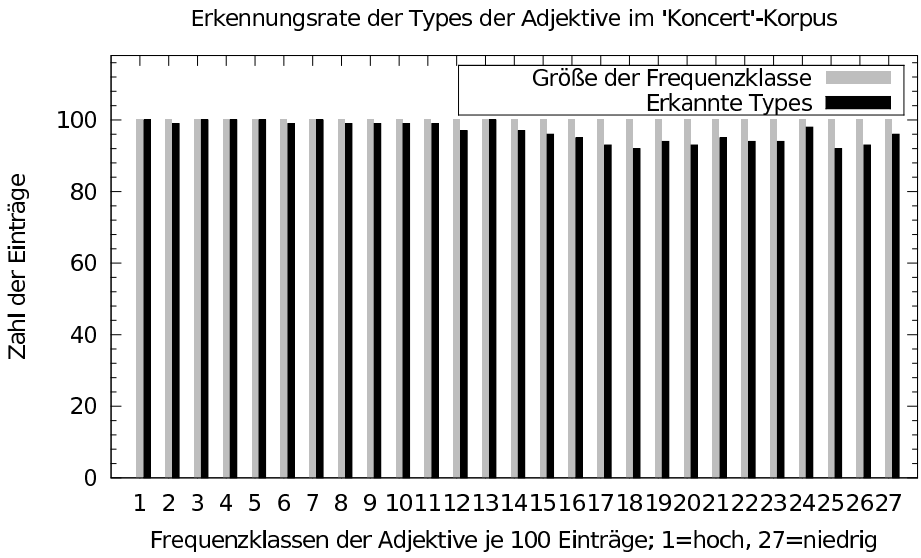
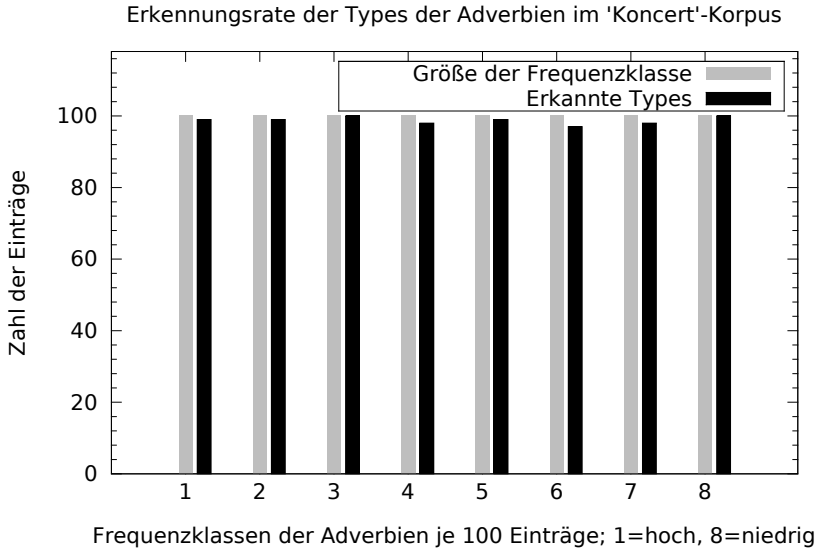


Abbildung 6.7: Erkennungsrate der Adverbien nach Frequenzklassen



Einige Wortformen im *Koncert*-Korpus wie z. B. *Dyfuçiasi*, *mikrofonavënësit*, *mosnënvleftësimin*, *pakmëpërparshmën*, *punëra* (dt. *Arbeiten*, Nicht-Standard) (Substantive), sowie *biografibukura*, *brengëgërryese*, *hënëzbardhur*,

miqësiprishës (Adjektive) sind neue Wortschöpfungen, die die Morphologiekomponente nicht analysiert – sie können jedoch mit der Erweiterung der Wortbildungsregeln ohne großen Aufwand abgedeckt werden. Die Struktur des System erlaubt eine Erweiterung diesbezüglich.

6.6 Fehleranalyse

Eine Beobachtung der fehlgeschlagenen Einträge nach einem Testdurchlauf von Wortlisten lässt erkennen, dass fast alle Fehler auf der Seite des sprachlichen Wissens liegen. Das heißt, am häufigsten fehlt ein Lexikoneintrag (lexc), gefolgt von einer falschen Klassifikation eines Eintrages. Darüber hinaus kommen Fehler in den Flexionsparadigmen und dazugehörigen grammatischen Kodierungen sowie schließlich Fehler im Bereich der Wortbildung vor. Die Regeln, die die Wortbildung behandeln, sind leider nicht vollständig und decken die kleinen Gruppen nicht ab. Es konnten nur die Gruppen behandelt und implementiert werden, die einen hohen Grad an Produktivität und Kombinierbarkeit aufweisen. Dieser Punkt bleibt eine Fehlerquelle. Man könnte dies umgehen, indem man die neugebildeten Wörter aus den Gruppen, die eine kleinere produktive Stärke zeigen, vorläufig direkt ins Lexikon aufnimmt, bis die Wortbildungsregeln dafür formalisiert und implementiert sind.
Auf der technischen Seite ließen sich während der Entwicklung der Morphologiekomponente auch Fehler beobachten, die in diesen Fällen selbstverständlich unverzüglich behoben wurden. Fehler in diesem Bereich fallen schneller als sprachliche Fehler auf. Die Beseitigung von Fehlern auf der lexc- und xfst-Seite sollte anhand der überschaubaren Zusammensetzung der Lexika und der Grammatik leicht zu handhaben sein.

6.6.1 Recall

Die Wortformen, die von der Grammatik bearbeitet werden, können leicht in solche unterteilt werden, die mit grammatischer Information angereichert wurden, und in soche, die nicht mit dieser Information versehen wurden. Die erste Gruppe kann als Recall bezeichnet werden.[229] Die nicht erkannten Wortformen, in XFST mit +? markiert, werden nicht dazu gezählt. Die Recall-Werte der vorgelegten Morphologiekomponente entsprechen den bereits bekannten Tests aus den Abschnitten 6.5.5, 6.5.6, 6.5.7 und 6.5.8.

[229] Man vergleiche hierzu Listing 6.1.

6.6.2 Precision

Die analysierten Wortformen (Recall) können noch Fehler enthalten oder nicht ausreichend mit Information ausgestattet sein. Der Teil der Information, der einer korrekten grammatischen Information entspricht, wird hier mit dem Begriff Precision gleichgestellt. Anhand von Stichproben – 10 Tests mit je 1000 zufällig gewählten Wortformen – ergibt der Wert der Precision einen Durchschnittswert von 98,87 %. Dabei wurde auch die wortartspezifische Verteilung der Wortformen berücksichtigt, d. h., in den Tests überwog die Zahl der Substantive, Verben, Adjektive und Adverbien gegenüber der Zahl der Wortformen anderer Wortarten.

6.7 Zusammenfassung des 6. Kapitels und Schlussbemerkungen

Die statistischen Angaben der Tests, gezeigt in der Tabelle 6.2 (Text des Romans *Koncert*) sowie in den Tabellen 6.3 und 6.4, zeigen Erkennungsraten, die es durchaus erlauben, die entwickelte Morphologiekomponente für Zwecke der maschinellen Sprachverarbeitung einzusetzen. Sicherlich kann die Erkennung verbessert werden, indem zuerst das Lexikon erweitert wird, die Regeln der Flexion (im Rahmen von XFST) verbessert werden und die Komponente für Wortbildung vervollständigt wird.

7 Schlussbemerkungen, Forschungsbeitrag und Ausblick

Mit jedem Testdurchlauf der vorliegenden Morphologiekomponente, der stets mit Aufnahme neuer Texte zu den bisherigen durchgeführt wurde, kamen neue Aufgaben ans Licht. Es liegt in der Natur des Problems, dass die Zahl der Wörter theoretisch unendlich ist. Jedoch erleichtert die im Rahmen der vorliegenden Arbeit erstellte Morphologie viele Prozesse bei der Gewinnung neuer Daten, seien sie lexikalische Daten oder Wissen über Morphologie. So ist zum Beispiel eine nicht erkannte Wortform ein Hinweis auf einen fehlenden lexikalischen Eintrag oder eine nicht vorhandene morphologische Regel, was zur anschließenden Weiterentwicklung der Morphologiekomponente führt.
An dieser Stelle werden die Ergebnisse der vorliegenden Arbeit in Form der Schlussbemerkungen wiedergegeben (7.1). Es geht mehr darum, die Eckpunkte zu nennen, als Inhalte mit Details wiederzugeben. Darauf folgen einige Hinweise zum Forschungsbeitrag (7.2). Mit einem Ausblick über die Möglichkeiten der Verwendung und Erweiterung der Daten und der Morphologiekomponente schließt das Kapitel ab (7.3).

7.1 Schlussbemerkungen

Ausgangspunkt der vorliegenden Arbeit war die Tatsache, dass es immernoch (2010/2011) keine vollständige Morphologie für Zwecke der maschinellen Sprachverabeitung des Albanischen gab. Dieser Zustand machte die Lösung vieler Probleme und das Erledigen bestimmer Aufgaben schwierig bis unmöglich. Schwierig, da man durch Umwege oder partielle Lösungen eingeschränkt war, die Probleme vollständig und korrekt zu lösen. Dabei waren die Ergebnisse entsprechend entweder linguistisch nicht gut genug, z. B. wegen mangelnder Ressourcen, oder nicht genügend effizient im technisch-methodischen Sinne, nicht ausreichend getestet, usw. Für den Verfasser der vorliegenden Arbeit als Autor einer Morphologiekomponente

für Verben des Albanischen[230] lag somit die Erstellung einer vollständigen Morphologie, d. h. einer, die alle Wortarten abdeckt und ohne Einschränkungen einsatzfähig ist, nahe.
Als erster Schritt wurde die genannte Komponente für die automatische Verarbeitung der albanischen Verben komplett neukonzipiert und implementiert, anstatt sie in der ursprünglichen Form in das neue System (XFST) einzubinden. Einige lexikalische Daten wurden aus [Kabashi 2003] übernommen und demensprechend komplett überarbeitet. Zu den übernommenen lexikalischen Daten kamen noch ca. 10 000 Einträge, hauptsächlich Substantive und Adjektive, die von Herrn R. Memushaj (Universität Tirana, Albanien) zur Verfügung gestellt wurden. Schließlich wurde das Lexikon mit Daten, die aus dem in Abschnitt 6.2 vorgestellten Textkorpus gewonnen wurden, erweitert. Der Restteil der lexikalischen Daten, wie z. B. einige Pronomina, wurde aus dem muttersprachlichen Wissen des Verfassers der vorliegenden Arbeit erstellt. Somit war es möglich, ein neues Lexikon zu entwickeln, das eine bessere Abdeckung ermöglicht.
Um das maschinenlesbare Lexikon zu erstellen, musste ein Großteil der lexikalischen und grammatischen Informationen (linguistisches Wissen) zuerst gefunden, gesammelt und anschließend in ein EDV-System eingegeben werden.
Neben der Übernahme der o. g. Daten sowie der Berücksichtigung der Literatur über das Albanische, wurde zusätzlich ein datengestützter Ansatz verwendet, bei dem die Unterklassen der Wortarten, insbesondere der Substantive, der Namen und der Adjektive teils automatisch bestimmt wurden, wie in den Abschnitten 3.3.2, 3.3.3, 4.3 und 4.4 vorgeführt wurde. So wurden ca. 50 000 Einträge aufwändig linguistisch kategorisiert.
Die Erstellung der Morphologie war mit der Organisation einer enorm großen Menge Daten verbunden. Insbesondere die morphologische Kategorisierung der lexikalischen Einheiten einzelner Wortarten nahm viel Zeit in Anspruch. Praktisch jeder Fehler in der Kategorisierung der lexikalischen Einheiten resultiert in vielen falschen Analysen bzw. in entsprechend vielen falschen Produktionen. Das Lexikon bzw. die Morphologiekomponente wurde mit verschiedenen Namen, Abkürzungen, Maßeinheiten usw. vervollständigt, die in Texten nicht selten vorkommen, aber in Lexika und Lehrbüchern nur zum Teil (exemplarisch oder partiell) berücksichtigt werden. Zuletzt kam ein unterschätzter Teil, nämlich die Erstellung der Testressourcen. Um die Morphologiekomponente zu testen, mussten Wort(form)listen erstellt werden, die korrekt (positiv), oder entsprechend

[230] Vgl. hierzu [Kabashi 2003].

falsch (negativ) sind. Ebenso ist die Erstellung dieser Listen nach Wortarten, vor einer Morphologiekomponente oder während ihrer Entwicklung, mit sehr viel manueller Arbeit verbunden. Diese Listen könnten nur partiell automatisch aus den vorhandenen Daten und linguistischem Wissen erstellt werden.
Durch die erstellte Morphologie und durch ihre praktische Einsatzfähigkeit wurde das Ziel erreicht, ein System/Werkzeug zu erstellen, das Wortformen aller Wortarten des Albanischen analysieren und produzieren kann.

7.2 Forschungsbeitrag

Die oft unterschätzte Aufgabe der Systematisierung und Strukturierung der lexikalischen Daten, sowie der Aufbau eines Lexikons für die MSV wurde erfolgreich bearbeitet.
Die bei der Erstellung dieser Daten erkannten Probleme können bei der möglichen Erweiterung des Lexikons leicht vermieden werden, insbesondere bei der Kategorisierung des Wortschatzes.
Das Vorhandensein einer maschinellen Morphologiekomponente erleichtert die Verarbeitung verschiedener linguistischer Daten. Ohne eine maschinelle Morphologie wäre die Verarbeitung in einigen Bereichen von niedriger Qualität, sogar fast unmöglich. Somit ist mit der erstellten Morphologie ein grundlegender Baustein gesetzt, um das Albanische in dieser und in darauf aufbauenden Abstraktionskomponenten maschinell zu verarbeiten.

7.2.1 Vergleich mit Tagger und Stemmer

Im Vergleich zum Tagger von [Trommer/Kallulli 2004], vgl. kurze Beschreibung in Abschnitt 2.6.1, ist die vorliegende Morphologie deutlich ausführlicher. Sie bietet u.a. neben der Analyse, der Aufgabe eines Taggers, gleichermaßen auch die Fähigkeit der Produktion (Generierung). Weiterhin bietet die vorliegende Arbeit auch Ansätze der Wortbildung, wie in Kapiteln 5 und 6 beschrieben ist.
Die Arbeiten [Kadriu 2010] und [Hasanaj 2012], vgl. kurze Beschreibung in Abschnitt 2.6.1, unterscheiden sich von der vorliegenden Arbeit. Alle drei Arbeiten behandeln zwar die albanische Sprache, jedoch in unterschiedlichem Umfang, aus verschiedenen Blickwinkeln und verwenden unterschiedliche Verarbeitungsmethoden.
[Kadriu 2010] basiert auf einem maschinellen Lernverfahren, während die vorliegende Arbeit dies nicht tut, sondern regelbasiert funktioniert.

Während [Kadriu 2010] nur zwei Wortarten behandelt, deckt die vorliegende Arbeit alle Wortarten ab und behandelt noch zum Teil Wortbildung, wie in Kapiteln 5 und 6 beschrieben ist.
[Hasanaj 2012] unterscheidet sich von der vorliegenden Arbeit sehr: In der Analyse, was den Umfang angeht, und insbesondere in der Produktion, denn ein Tagger ist nur für die Analyse konzipiert, während die vorgelegte XFST-Morphologiekomponente auch produzieren kann.
Die Arbeit von Sadiku und Biba [2012] unterscheidet sich von der vorliegenden darin, dass sie Stemming, vgl. 2.6.2, also die Extraktion des Stammes aus einer Wortform, als Aufgabe hat und nicht die morphologische Information. Dies liegt im Bereich der Analyse und nicht der Produktion.
Die Arbeiten [Kadriu 2010], [Hasanaj 2012] sowie [Sadiku/Biba 2012] entstanden parallel zur Entwicklung der vorliegenden Arbeit, doch die Aufgaben, die sie erledigen können, sind kleiner bzw. sind andere als die der vorliegenden Arbeit. Zum Beispiel kann eine Morphologiekomponente die Aufgabe eines Stemmers komplett abdecken.

7.2.2 Vergleich mit Annotierung von Korpora

Die Erstellung eines Korpus war nicht das Ziel der vorliegenden Arbeit. Eine Möglichkeit ihrer Verwendung ist jedoch die Annotation/das Taggen eines Korpus. [Arkhangelskij et al. 2012] beabsichtigen auch eine morphologische Annotation. Sie können z. Z.[231] schätzungsweise 1/3 der Textwörter annotieren, bewertet nach einigen (25) Stichproben auf den gezeigten Texten nach den einzelnen Suchvorgängen. Auch in diesem Vergleich ist das im Rahmen der vorliegenden Arbeit entwickelte System mit seiner Analyse-Funktion (Fähigkeit zu annotieren) mit ca. 94–98% Erkennungsrate, vgl. Kapitel 6, der genannten Arbeit [Arkhangelskij et al. 2012] deutlich überlegen – unabhängig davon, wie das Korpus annotiert wird, maschinell, manuell oder in einer gemischten Form.
[Caka/Caka 2012] bauen ein Korpus und beabsichtigen es mit Hilfe von statistischen Methoden, bzw. mit darauf basierenden Werkzeugen, und sogar zum Teil manuell zu annotieren. Getaggt sind, laut Autoren, die ersten 10%, d. h. 100 000 Textwörter der 1 000 000 Wortformen, die das Korpus beinhaltet.[232] Auch in diesem Fall deckt das in der vorliegenden Arbeit vorgestellte System die (Funktion der) Annotation leicht ab.

[231] Vgl. <http://web-corpora.net/AlbanianCorpus/search/index.php>, 24.2.2014.
[232] Vgl. [Caka/Caka 2012: 648] und [op. cit.: 653].

7.2.3 Ein Beitrag für die morphologische Analyse

Nach dem besten Wissen des Autors dieser Zeilen ist das entwickelte System die erste Gesamtbehandlung der albanischen Morphologie in großem Umfang. Die vorliegende Arbeit bietet die Behandlung häufiger Personennamen (Vor- und Nachname) [ca. 5000] und vieler Einwohner- und Ortsnamen [ca. 1000], alle samt ihrer Flexionsformen, sowie der häufigsten bzw. wichtigsten Interpunktionszeichen, Maßeinheiten und Abkürzungen. Zusätzlich kann die vorgestellte Morphologiekomponente die verbreitetsten Wortbildungstypen des Albanischen behandeln, eine Dimension, die bisher im Rahmen der maschinellen Verarbeitung der Morphologie nicht behandelt wurde.
Als wichtige Eigenschaft der vorliegenden Arbeit kann die Aufbaustruktur des praktischen Teils genannt werden, denn nicht enthaltene Wörter (Lemmata) können leicht eingebaut werden. Der Rahmen deckt die Funktionalität der Morphologie (und Wortbildung) vollständig ab, d. h. wenn ein Wort fehlt, müsste es nur an der passenden Stelle, mit passenden Eigenschaften, eingetragen und die dazugehörige Grammatik kompiliert werden. Ebenso kann im Falle einer Erweiterung der Derivations- oder Kompositionsregeln vorgegangen werden. Beide Komponenten können leicht erweitert werden. In Analogie zu vorhandenen Regeln können neue definiert oder die bestehenden angepasst werden. Auch für diejenigen, denen das Programmieren zu abstrakt scheint, sollte dies möglich sein. Der praktische Teil kann mit leichten Anpassungen in das Softwarepaket foma, vgl. Abschnitt 2.3.3, portiert werden.

7.3 Ausblick

Die Erweiterung des Lexikons um weitere lexikalische Einheiten, selten vorkommende Lemmata, ggf. mit regionalen Varianten oder fachspezifische Einträge wäre als erster Schritt denkbar bei einer Weiterentwicklung der vorgestellten Morphologie im Rahmen der vorliegenden Arbeit. Das Hinzufügen von neuen Namen wäre eine wichtige Ergänzung des Lexikons. Sie machen einen nicht zu unterschätzenden Teil einer Wortformliste (gewonnen aus einem Text) aus. Hier ist keine Beschränkung notwendig: Es können alle möglichen Namen aufgenommen werden, seien diese Personennamen, geographische Namen oder bspw. einige Produktnamen.
Eine vertiefte Behandlung des Teilbereichs der Komposition würde die Morphologiekomponente vervollständigen und ihre Erkennungsrate verbessern. Neue Erkenntnisse, insbesondere detaillierte aus den Wortstrukturen

und deren Kombinationen im Rahmen der Derivation und Komposition, wären für die vorliegende Arbeit sehr nützlich sein. Gemeint sind detaillierte korpusbasierte Erkenntnisse der einzelnen Klassen und Subklassen samt ihren Ausnahmen und zusätzlichen besonderen Eigenschaften im Vergleich zu allgemeinem theoretischen Wissen aus der traditionellen Morphologie- und Wortbildungslehre.
Um dem eingangs geschilderten Problem, dass die Zahl der Wörter theoretisch unendlich ist, besser Herr zu werden, läge es nahe, die vorgestellten Hypothesenregeln in Abschnitt 6.5.3 zu optimieren bzw. zu erweitern, oder neue zu erstellen. Damit wäre es möglich anhand von Flexionsmarkierung, Groß- und Kleinschreibung der Wortformen, sowie eines phonologischen Kontextes ein neues Wort bzw. seine Wortform zu erraten bzw. zu erkennen. Im Rahmen der Wortbildung wäre die entsprechende Erweiterung um Regeln, die auf den Kombinationsmöglichkeiten der Wörter (Lexeme, Morpheme usw.) miteinander basieren, nötig.

Literaturverzeichnis

[Arkhangelskij et al. 2012] Timofej Arkhangelskij / Mikhail Daniel / Maria Morozova / Alexandar Rusakov: „Korpusi i gjuhës shqipe: Drejtimet kryesore të punës“. 635–642. [In:] [Ismajli 2012].

[Atkins/Rundell 2008] B. T. S. Atkins, Michael Rundell: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press, 2008.

[Bátori et al. 1989] István Bátori / Winfried Lenders / Wolfgang Puschke: *Computational Linguistics / Computerlinguistik*. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen / An International handbook on Computer Oriented Language Research and Applications [HSK 4]. Berlin / New York: Walter de Gruyter, 1989.

[Baayen 2001] Harald R. Baayen: *Word Frequency Distributions*. Dodrecht / Boston / London: Kluwer Academic Publishers, 2001.

[Beesley/Karttunen 2003] Kenneth R. Beesley / Lauri Karttunen: *Finite State Morphology*. Stanford: CSLI Publications, 2003.

[Bega/Bega 2007] Batjar Bega / Sokol Bega: *Albanian Verbs*. An overview of features and usage of Albanian verbs. Tiranë: Pegi, 2007.

[Beutel 2011] Björn Beutel: *Documentation for Malaga 7.12. User's and Programmer's Manual*. Online: `<http://home.arcor.de/bjoern-beutel/malaga/>`, 28.7.2014.

[BNC–xml 2007] *British National Corpus. Version 3. BNC XML Edition*. Distributed under license by Oxford University Computing Services on behalf of the BNC Consortium, 2007.

[Buchholz/Fiedler 1987] Oda Buchholz / Wilfried Fiedler: *Albanische Grammatik*. Leipzig: VEB, Verlag Enzyklopädie 1987.

[Buchholz et al. 1993] Oda Buchholz / Wilfried Fiedler / Gerda Uhlisch: *Wörterbuch Albanisch–Deutsch*. 6. Auflage. Leipzig & München: VEB, Verlag Enzyklopädie & Langenscheidt 1993.

[Bussmann 2002] Hadumod Bussmann (Hrsg.): *Lexikon der Sprachwissenschaft*. 3. Auflage. Stuttgart: Kröner, 2002.

[Buxheli 2007] Ludmila Buxheli: *Modelet e caktimit rasor në gjuhën e sotme shqipe*. Tiranë: Neraida, 2007.

[Buxheli 2008] Ludmila Buxheli: *Formimi i foljeve në gjuhën e sotme shqipe*. Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, Departamenti i Gramatikës dhe i Kulturës së Gjuhës, 2008.

[Buxheli 2009] Ludmila Buxheli: *Fjalët e përngjitura në gjuhën shqipe*. Tiranë: Akademia e Shkencave e Republikës së Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë; Kumi, 2009.

[Caka/Caka 2012] Nebi Caka / Ali Caka: „Korpusi i gjuhës shqipe". 643–656. [In:] [Ismajli 2012].

[CELEX 1994] R. H. Baayen / R. Piepenbrock / L. Gulikers, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995. Online: <http://celex.mpi.nl/> , 28.7.2014.

[Clark et al. 2010] Alexander Clark / Chris Fox / Shalom Lappin (Eds.): *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley & Blackwell, 2010.

[Claus/Schwill 2001] Volker Claus / Andreas Schwill: *Duden Informatik. Ein Sachlexikon für Studium und Praxis*. 3. Auflage. Herausgegeben von Meyers Lexikonredaktion. Bearbeitet von Prof. Dr. Volker Claus und Dr. Andreas Schwill. Mannheim / Leipzig / Wien / Zürich: Dudenverlag, 2011.

[Çeliku et al. 1998] Mehmet Çeliku / Mustafa Karapinjalli / Ruzhdi Stringa: *Gramatika praktike e gjuhës shqipe*. Tiranë: Toena, 1998.

[Çeliku et al. 2011] Mehmet Çeliku / Mustafa Karapinjalli / Ruzhdi Stringa: *Gramatika praktike e gjuhës shqipe*. Tiranë: ILAR, 2011.

[DATR] *DATR, a language for the lexical knowledge representation*. Online: <http://www.informatics.susx.ac.uk/research/groups/nlp/datr/datr.html> , 28.7.2014.

[Demiraj B. 1990] Bardhyl Demiraj: „Die Bildung der Zahlwörter 11 bis 19 im Albanischen". 185–189. [In:] *Studia Albanica* 27:2. Tiranë, 1990.

[Demiraj Sh. 1994] Shaban Demiraj: *Historische Grammatik der albanischen Sprache*. Wien: Österreichische Akademie der Wissenschaften, 1994.

[Dhrimo et al. 2002] Ali Dhrimo / Edmond Tupja / Eshref Ymeri: *Fjalor sinonimik i gjuhës shqipe*. Tiranë: Toena 2002.

[Dhrimo et al. 2007] Ali Dhrimo / Edmond Tupja / Eshref Ymeri: *Fjalor sinonimik i gjuhës shqipe*. Botimi i dytë. Tiranë: Toena 2007.

[Dhrimo/Memushaj 2011] Ali Dhrimo / Rami Memushaj: *Fjalor drejtshkrimor i gjuhës shqipe*. Tiranë: Infbotues, 2011.

[Dornseiff et al. 2004] Franz Dornseiff / Uwe Quasthoff / Herbert Ernst Wiegand: *Dornseiff – Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearbeitete Auflage. Mit vollständigem alphabetischen Zugriffsregister, lexikographisch-historischer Einführung und ausgewählter Bibliographie. Berlin: Walter de Gruyter, 2004.

[Drejtshkrimi 1974] Androkli Kostallari (Kryetar) / Mahir Domi / Eqrem Çabej / Emil Lafe: Akademia e shkencave e RP të Shqipërisë. Instituti i Gjuhësisë dhe Letërsisë: *Drejtshkrimi i gjuhës shqipe*. Botim i posaqëm. Tiranë: Shtëpia botuese e librit shkollor, 1974.

[Duden-Grammatik 2009] Kathrin Kunkel-Razum / Franziska Münzberg: *Duden – Die Grammatik*. 8., überarbeitete Auflage. Mannheim / Wien / Zürich: Dudenverlag, 2009.

[ECI/MCI 1994] *European Corpus Initiative Multilingual Corpus I (ECI/MCI) CD-ROM*. Utrecht: ELSNET 1994. Online: <http://www.elsnet.org/resources/eciCorpus.html> , 28.7.2014.

[Eggers et al. 1980] Hans Eggers / Heinz-Dirk Luckhardt / Heinz-Dieter Maas / Monika Weissgerber (Hrsg.): *SALEM – Ein Verfahren zur automatischen Lemmatisierung deutscher Texte.* Tübingen: Max Niemeyer Verlag, 1980.

[Engelberg 2009] Stefan Engelberg: *Lexikographie und Wörterbuchbenutzung*. 4., überarb. u. erw. Aufl. Tübingen: Stauffenburg-Verlag, 2009.

[Fdsh 1976] Androkli Kostallari / Mahir Domi / Emil Lafe / Nikoleta Cikuli: *Fjalori drejtshkrimor i gjuhës shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 1976.

[Ferber 2002] Reginald Ferber: „Dokumentensuche und Dokumenterschließung“. 913–934 (§ E 4). [In:] [Rechenberg/Pomberger 2002].

[Fiedler 2005] Wilfried Fiedler: *Die Pluralbildung im Albanischen.* Prishtinë: Akademia e Shkencave dhe e Arteve e Kosovës, Botime të veçanta LXXIX, Libri 34, Seksioni i Gjuhësisë dhe i Letërsisë, 2005.

[Fisseni et al. 2005] Bernhard Fisseni / Hans-Christian Schmitz / Berhard Schröder / Petra Wagner (Hrsg.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn*. Frankfurt am Main / Berlin / Bern / Bruxelles/ New York / Oxford / Wien: Peter Lang, 2005. (Sprache, Sprechen und Computer, Bd. 8).

[Fitschen 2004] Arne Fitschen: *Ein Computerlinguistisches Lexikon als komplexes System*. Dissertation. Universität Stuttgart, 2004.

[Fjalori 1980] Androkli Kostallari (Kryeredaktor) / Jani Thomaj / Xhevat Lloshi / Miço Samara: *Fjalor i gjuhës së sotme shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 1980.

[Fjalori 1984] Androkli Kostallari (Kryeredaktor) / Jani Thomaj / Miço Samara / Josif Kole / Palok Daka / Pavli Haxhillazi / Hajri Shehu / Kornelja Sima / Thanas Feka / Beatriçe Keta / Agim Hidi: *Fjalor i shqipes së sotme*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 1984.

[GermaNet 2006] : *GermaNet 4.0* (2006). Tübingen: Seminar für Sprachwissenschaft. Abt. Computerlinguistik. Eberhard-Karls-Universität, 2006.

[Geyken / Hanneforth 2006] : Alexander Geyken / Thomas Hanneforth: *TAGH: A Complete Morphology for German Based on Weighted Finite State Automata*. 55–66 [In:] *Finite-State Methods and Natural Language Processing*, *Lecture Notes in Computer Science* 2006, Volume 4002/2006. Berlin / New York: Springer, 2006; Online: <http://www.tagh.de/> , 28.7.2014.

[Gibbon 2010] Dafydd Gibbon: „Lexika für multimodale Systeme“. 515–523. [In:] [Klabunde et al. 2010].

[Gjinari et al. 2007] Jorgji Gjinari (drejtues) / Bahri Beci / Gjovalin Shkurtaj / Xheladin Gosturani: *Atlasi dialektologjik i gjuhës shqipe*. Pjesa e parë. Tiranë: Akademia e Shkencave e Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë, 2007. / Vol. 1. Napoli: Univ. degli Studi di Napoli l'Orientale, Dipartimento di Studi dell'Europa Orientale, 2007.

[Gjinari et al. 2008] Jorgji Gjinari (drejtues) / Bahri Beci / Gjovalin Shkurtaj / Xheladin Gosturani: *Atlasi dialektologjik i gjuhës shqipe*. Pjesa e dytë. Tiranë: Akademia e Shkencave e Shqipërisë. Instituti i Gjuhësisë dhe i Letërsisë, 2008. / Vol. 2. Napoli: Univ. degli Studi di Napoli l'Orientale, Dipartimento di Studi dell'Europa Orientale, 2008.

[Glas 1975] Reinhold Glas: „Das LiMaS-Korpus, ein Textkorpus für die deutsche Gegenwartssprache". 63–66. [In:] *Linguistische Berichte* 40 (1975).

[Glück 2011] Helmut Glück (Hrsg.): *Metzler Lexikon Sprache*. 4. Auflage. Stuttgart / Weimar: Metzler, 2011.

[Goldsmith 2010] John A. Goldsmith: "Segmentation and Morphology". 364–393. (§ 14). [In:] [Clark et al. 2010].

[Görz/Paulus 1988] Günther Görz / Dietrich Paulus: "A finite state approach to German verb morphology". 212–215 [In:] *COLING-88, Proceedings of the 12th conference on Computational linguistics – Volume 1*. Association for Computational Linguistics.

[Görz et al. 2003] Günther Görz / Claus-Reiner Rollinger / Josef Schneeberger (Hrsg.): *Handbuch der Künstlichen Intelligenz*. München: Oldenbourg, 4. Auflage, 2003.

[Grefenstette 1999] Gregory Grefenstette. "Tokenization". 117–133 [In:] [van Halteren 1999].

[van Halteren 1999] Hans van Halteren (Ed.): *Syntactic wordclass tagging*. (Text, speech and language technology 9); Dordrecht [u. a.]: Kluwer Acad. Publ., 1999.

[Hamp 2006] Eric P. Hamp: "The anatomy of survival in Arbëresh numeral forms". 23–29 [In:] *Int`l. J. Soc. Lang. 178 (2006)*.

[Hasanaj 2012] Besmir Hasanaj: *A Part of Speech Tagging Model for Albanian*. Saarbrücken: Lambert Academic Publishing, 2012.

[Hausmann 1985] Franz Josef Hausmann: „Lexikographie“ (Kapitel 13), 367–411. [In:] *Handbuch der Lexikologie*. Hrsg. v. Christoph Schwarze & Dieter Wunderlich. Königstein/Ts.: Athenäum Verlag 1985.

[Hausser 1996] Roland Hausser (Hrsg.): *Linguistische Verifikation*. Tübingen: Niemeyer, 1996.

[Hausser 2001] Roland Hausser: *Foundations of Computational Linguistics.* Human-Computer Communication in Natural Language. 2nd Edition. Berlin / New York: Springer 2001.

[Heid 2003] Ulrich Heid: „Morphologie und Lexikon“. 665–709 (§ 17). [In:] [Görz et al. 2003].

[Helbig 2008] Herrman Helbig: *Wissensverarbeitung und die Semantik der Natürlichen Sprache.* 2. überarbeitetet Auflage. Berlin / New York: Springer, 2008.

[Heringer 2010] Hans-Jürgen Heringer: *Einführung in die Morphologie.* Paderborn: Wilhelm Fink / UTB 3204, 2010.

[Hess et al. 1983] Klaus Hess / Jan Brustkern / Winfried Lenders: *Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration.* Max Niemeyer Verlag, Tübingen 1983.

[Hetzer/Finger 1993] Armin Hetzer / Zuzana Finger: *Lehrbuch der vereinheitlichen albanischen Schriftsprache*. 5. Auflage. Hamburg: Buske 1993.

[Hopcroft et al. 2001] John E. Hopcroft / Rajeev Motwani / Jeffrey D. Ullman: *Introduction to Automata Theory, Languages and Computation*. 2. Ed. Boston: Addison–Wesley, 2001.

[Hopcroft et al. 2002 de] John E. Hopcroft / Rajeev Motwani / Jeffrey D. Ullman: *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie*. 2., überarbeitete Auflage. (Übersetzung der [Hopcroft et al. 2001]). Übersetzung: Sigrid Richter und Ingrid Tokar; Fachlektorat: Manfred Paul. München: Pearson Studium, 2002.

[Hysa 2004] Enver Hysa: *Formimi i emrave me ndajshtesa në gjuhën shqipe*. Tiranë: Akademia e Shkencave e Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 2004.

[Ismajli 2012] Rexhep Ismajli (Ed.): *Shqipja dhe gjuhët e Ballkanit* ([Akten der Konferenz] *Albanian and Balkan Languages*) Prishtinë, 10–11 November 2011. Prishtinë / Tiranë: Academy of Sciences and Arts of the Republic of Kosovo & Academy of Sciences of the Republic of Albania, 2012.

[Kabashi 2003] Besim Kabashi: *Automatische Wortformerkennung für das Albanische*. Magisterarbeit im Fach „Linguistische Informatik", Universität Erlangen–Nürnberg, 2003.

[Kabashi 2005] Besim Kabashi: „Disa propozime për modelimin e informacionit në leksikografinë kompjuterike." 179–184. [In:] *Seminari Ndërkombëtar për Gjuhën, Letërsinë dhe Kulturën Shqiptare, XXIV.* Libri 24/1. Prishtinë: Universiteti i Prishtinës, 2005.

[Kabashi 2007] Besim Kabashi: "Pronominal clitics and valency in Albanian. A computational linguistics prespective and modelling within the LAG-Framework". 339–352. [In:] Thomas Herbst / Katrin Götz-Votteler (Eds.): *Valency. Theoretical, descriptive and cognitive issues*. (Trends in Linguistics. Studies and Monographs, 187). Berlin / New York: Mouton de Gruyter, 2007.

[Kabashi 2009] Besim Kabashi: „Das albanische Alphabet aus sprachtechnologischer Sicht". 175–208. [In:] Bardhyl Demiraj (Hrsg.): *Der Kongress von Manastir. Herausforderung zwischen Tradition und Neuerung in der albanischen Schriftkultur.* Hamburg: Verlag Dr. Kovač, 2009. (Akten der 3. Deutsch-Albanischen Kulturwissenschaftlichen Tagung. Ludwig-Maximilians-Universität München, 7.–8. Oktober 2008.)

[Kabashi 2012] Besim Kabashi: „Korpuse gjuhësore për shqipen". 627–634. [In:] [Ismajli 2012].

[Kadriu 2010] Arbana Kadriu: "Modeling a Two-Level Formalism for Inflection of Nouns and Verbs in Albanian". 301–312. [In:] Shkelzen Cakaj (Ed.): *Modeling Simulation and Optimization – Focus on Applications*. Rijeka (Croatia): InTech, 2010.

[Kennedy 1998] Graeme Kennedy: *An Introduction to Corpus Linguistics*. London / New York: Addison–Wesley Longaman, 1998.

[Klabunde 1998] Ralf Klabunde: *Formale Grundlagen der Linguistik*. Tübingen: Narr, 1998.

[Klabunde 2010] Ralf KLABUNDE: „Automatentheorie und formale Sprachen“. 66–93 (§ 2.2). [In:] [KLABUNDE ET AL. 2010].

[Klabunde et al. 2010] Kai-Uwe CARSTENSEN / Christian EBERT / Cornelia EBERT / Susanne JEKAT / Ralf KLABUNDE / Hagen LANGER (Hrsg): *Computerlinguistik und Sprachtechnologie. Eine Einführung*. 3. Auflage 2010. Heidelberg: Spektrum Akademischer Verlag, 2010.

[Kostallari et al. 1984] Androkli KOSTALLARI / Emil LAFE / Minella TOTONI / Nikoleta CIKULI: *Gjuha letrare shqipe*. Elemente të normës letrare. Botimi i dytë. Prishtinë: ETMM 1984.

[Koncert 1994] Ismail KADARE: „Koncert në fund të dimrit“. Tiranë: „Naim Frashëri“, 1988. *albo1*. [In:] [ECI/MCI 1994].

[Koskenniemi 1993] Kimmo KOSKENNIEMI: *A Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki. Department for General Linguistics: Publications, No. 11., 1983.

[Kunze/Lemnitzer 2007] Claudia KUNZE / Lothar LEMNITZER: *Computerlexikographie.* Tübingen: Gunter Narr Verlag, 2007.

[Lemnitzer/Zinsmeister 2010] Lothar LEMNITZER / Heike ZINSMEISTER: *Korpuslinguistik. Eine Einführung*. 2. Auflage. Tübingen: Narr Verlag, 2010.

[Lobin 2001] Henning LOBIN: *Informationsmodellierung in XML und SGML* Berlin / Heidelberg / New York: Springer, 2001.

[Lobin/Lemmnitzer 2004] Henning LOBIN / Lothar LEMMNITZER (Hrsg.): *Texttechnologie. Perspektiven und Anwendungen.* Tübingen: Stauffenburg, 2004.

[Lüdeling / Kytö 2008] Anke LÜDELING / Merja KYTÖ (Edited by): *Corpus Linguistics*. An International Handbook Vol. 1 (Handbooks of Linguistics and Communication Science [HSK], 29.1). Berlin: Mouton de Gruyter, 2008.

[Lüdeling / Kytö 2009] Anke LÜDELING / Merja KYTÖ (Edited by): *Corpus Linguistics*. An International Handbook Vol. 2 (Handbooks of Linguistics and Communication Science [HSK], 29.2). Berlin: Mouton de Gruyter, 2009.

[Mahlow/Piotrowski 2010] Cerstin Mahlow / Michael Piotrowski (Eds.): *State of the Art in Computational Morphology*. Workshop on Systems and Frameworks for Computational Morphology (SCFM). *Communications in Computer and Information Science* 41. Heidelberg / Berlin / New York [etc.]: Springer, 2010.

[Matthews 1997] Peter-Hugoe Matthews: *Oxford concise dictionary of linguistics*. Oxford University Press, 1997.

[Memushaj 2003] Rami Memushaj: „Për një ndarje tjetër në klasa të foljeve të shqipes". 40–62. [In:] *Studime filologjike* Nr. 1-2/2003, Tiranë: Akademia e Shkencave e Republikës së Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 2003.

[Memushaj 2004] Rami Memushaj: *Shqipja standarde*, Tiranë: Toena, 2004.

[Memushaj 2010] Rami Memushaj: *Fonetika e shqipes standarde*. Botimi i dytë i përmirësuar. Tiranë: Toena, 2010.

[Mitkov 2003] Ruslan Mitkov (Ed.): *The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2003.

[Morfologjia 1995] Fatmir Agalliu et al., Mahir Domi (Kryeredaktor): *Gramatika e gjuhës shqipe*. [Vëllimi I] Fatmir Agalliu et al., Shaban Demiraj (Redaktor): *Morfologjia*. Tiranë: Akademia e Shkencave e Republikës së Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë 1995.

[Motsch 2004] Wolfgang Motsch: *Deutsche Worbildung in Grundzügen*. 2. überarbeitete Auflage. Berlin / New York: Walter de Gruyter, 2004.

[Munishi 1998] Zijadin Munishi: *Zgjedhimi i foljeve*. Prishtinë: Libri Shkollor, 1998.

[Newmark et al. 1982] Leonard Newmark / Philipp Hubbard / Peter Prifti: *Standard Albanian*. A reference grammar for students. Stanford: Stanford University Press 1982.

[Newmark 1999] Leonard Newmark: *Oxford Albanian–English Dictionary*. New York: Oxford University Press, 1999.

[Pala 2005] Karel Pala: "The Balkanet Experience". 353–366. [In:] [Fisseni et al. 2005].

[Piton et al. 2007] Odile Piton / Klara Lagji / Remzi Përnaska: "Electronic Dictionaries and Transducers for Automatic Processing of the Albanian Language". 407–413. [In:] *Lecture Notes in Computer Science*. Volume 4592, Natural Language Processing and Information Systems, Berlin / New York, 2007.

[Rechenberg/Pomberger 2002] Peter Rechenberg / Gustav Pomberger: *Informatik-Handbuch*. 3., aktualisierte Auflage. München / Wien: Carl Hanser Verlag, 2002.

[Rechenberg 2002] Peter Rechenberg: „Formale Sprachen und Automaten". 89–110 (§ A3). [In:] [Rechenberg/Pomberger 2002].

[Ressuli 1985] Namik Ressuli: *Grammatica albanese*. Bologna: Pàtron, 1985.

[Roark/Sproat 2007] Brian Roark / Richard Sproat: *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press, 2007.

[Sadiku/Biba 2012] Jetmir Sadiku / Marenglen Biba: "Automatic Stemming of Albanian through a rule-based approach". 173–190. [In:] *Journal of International Scientific Publications: Language, Individual & Society*. Volume 6, Part 1. Info Invest (Bulgaria), 2012.

[Samara 1985] Miço Samara: *Çështje të antonimisë në gjuhën shqipe*. Tiranë: Akademia e Shkencave e RPS të Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 1985.

[Samara 1999] Miço Samara: *Parafjalët në shqipen e sotme. Vështrim leksiko-semantik*. Tiranë: Panteon 1999.

[Schaeder/Willée 1989] Burkhard Schaeder / Gerd Willée: „Computergestützte Verfahren morphologischer Beschreibung". 188–203 [In:] [Bátori et al. 1989].

[SE-Times 2010] Francis M. Tyers / Murat Alperen: "South-East European Times: A parallel corpus of the Balkan languages". Online: `http://opus.lingfil.uu.se/SETIMES.php` (`http://www.setimes.com`), 28.7.2014.

[Schmid et al. 2004] Helmut Schmid / Arne Fitschen / Ulrich Heid: *SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection*. Proceedings of LREC 2004. Online: `<http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/LREC04/smor.pdf>`, 28.7.2014.

[Schneider 1997] Hans-Jochen Schneider (Hrsg.): *Lexikon Informatik und Datenverarbeitung*. 4. Auflage. München / Wien: Oldenburg, 1997.

[Snoj 1994] Marko Snoj: *Rückläufiges Wörterbuch der albanischen Sprache*. Hamburg: Buske, 1994.

[Sproat 1992] Richard Sproat: *Morphology and Computation*. Cambridge, Massachusetts / London, England: The MIT Press, 1992.

[Sproat 2000] Richard Sproat: "Lexical Analysis". 37–57. [In:] Robert Dale / Hermann Moisl / Harold Somers: *Handbook of Natural Language Processing*. New York / Basel: Marcel Dekker, Inc., 2000.

[Sulejmani 1984] Fadil Sulejmani: *Praktikumi i gjuhës shqipe*. Prishtinë: ETMM, 1984.

[Thomai 2001] Jani Thomai: *Leksiku dialektor e krahinor në gjuhën e sotme shqipe*. Tiranë: Akademia e shkencave e shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 2001.

[Thomai 2004] Jani Thomai: *Veçori leksiko-semantike të ndajfoljeve me prapashtesa në gjuhën shqipe*. Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe Letërsisë, 2004.

[Thomai et al. 2004] Jani Thomai / Miço Samara / Hajri Shehu / Thanas Feka (Redaktor shkencor Jani Thomai): *Fjalor sinonimik i gjuhës shqipe*. Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë, 2004.

[Thomai 2005] Jani Thomai: *Leksikologjia e gjuhës shqipe*. Botimi i tretë. Tiranë: Botimet Dudaj, 2005.

[Thomai 2009] Jani Thomai: *Prejardhja kuptimore në gjuhën shqipe*. (*Semantikë leksikore*). Tiranë: Akademia e Shkencave e Shqipërisë, Instituti i Gjuhësisë dhe Letërsisë & EDFA, 2009.

[Trommer 1997] Jochen Trommer: *Eine Theorie der albanischen Verbflexion in* `mo_lex`. Magisterarbeit. Universität Osnabrück, 1997.

[Trommer 2010] Jochen Trommer: „Morphologie". (§ 3.3) 236–263. [In:] [Klabunde et al. 2010].

[Trommer/Kallulli 2004] Jochen Trommer / Dalina Kallulli: "A Morphological Analyzer for Standard Albanian". 1271–1274. [In:] Proceedings of LREC 2004.

[Turano 2010] Giuseppina Turano: "La duttile morfologia dell'albanese tra derivati e composti". 395–412. [In:] Giovanni Belluscio / Antonio Mendicino [a cura di]: *Scritti in onore di* Eric Pratt Hamp *per il suo 90. compleanno*. Rende: Università della Calabria. Centro Editoriale e Librario, 2010.

[Xhuvani / Çabej 1962] Aleksandër Xhuvani / Eqrem Çabej: *Parashtesat e gjuhës shqipe*. Tiranë: Universiteti Shtetëror i Tiranës. Instituti i Historisë e Gjuhësisë 1962.

[Xhuvani / Çabej 1975] Aleksandër Xhuvani / Eqrem Çabej: *Prapashtesat e gjuhës shqipe*. [In:] Mahir Domi (Redaktor) et al.: *Çështje të gramatikës së shqipes së sotme II*. Tiranë: Akademia e Shkencave e RPS të Shqiperisë. Instituti i Gjuhësisë dhe i Letërsisë, 1975.

[Wahlster 2000] Wolfgang Wahlster (Editor): *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin / New York: Springer, 2000.

[Wahrig 1998] Gerhard Wahrig: *Wahrig – Wörterbuch der deutschen Sprache*. 2. Aufl. der Neuausgabe 1997, neu herausgegeben von Dr. Renate Wahrig-Burfeind, 1998; 1. Auflage 1978, herausgegeben von Prof. Dr. Gerhard Wahrig; München: Deutscher Taschenbuch Verlag [dtv].

[Wahrig 2012] Gerhard Wahrig: *Brockhaus WAHRIG Deutsches Wörterbuch*. 9., neubearb. u. aktualis. Aufl. Gütersloh/München: 2012.

[WALBU 2012] Helmut Schumacher / Jacqueline Kubczak / Renate Schmidt / Vera de Ruiter: *WALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Gunter Nach Verlag, 2004.

[Wynne 2005] Martin Wynne (Ed.): *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 2005.

[Yli-Jyrä et al. 2006] Anssi Yli-Jyrä / Lauri Karttunen / Juhani Karhumäki (Eds.): *Finite-State Methods and Natural Language Processing*. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1–2, 2005. Revised Papers. Berlin / Heidelberg / New York: Springer, 2006. (Lecture Notes in Computer Science, 4002).

[Yli-Jyrä et al. 2010] Anssi Yli-Jyrä / András Kornai / Jacques Sakarovitch / Bruce W. Watson (Eds.): *Finite-State Methods and Natural Language Processing*. 8th International Workshop, FSMNLP 2009. Pretoria, South Africa, July 21-24, 2009, Revised Selected Papers. Berlin / Heidelberg / New York: Springer, 2010. (Lecture Notes in Computer Science, 6062).

Die automatische Sprachverarbeitung hat seit ihren Anfängen deutlich an Bedeutung gewonnen. Sie ist heute in einigen Bereichen wie z.B. bei der Suche im Internet unverzichtbar und nicht mehr wegzudenken. Ein Werkzeug für die automatische Wortformerkennung und -produktion ist ein grundlegender Baustein für viele Anwendungen. Sie kann in vielen Bereichen eingesetzt werden, sowohl als eigenständige Anwendung, z.B. für didaktische Zwecke oder zur morphologischen Annotation von Korpora, als auch als unterstützende Komponente für Anwendungen wie die syntaktische Analyse von Texten.

Das hier vorgestellte System ist ein automatisches Werkzeug für folgende Aufgabengebiete: Analyse der Rechtschreibung, Lemmatisierung, Annotation der Wortarten, vollständige morphologische Analyse von Wortformen. Das System kann auch im umgekehrten Modus verwendet werden, d.h. Wortformen aus einem gegebenen Lemma und seinen morphologischen Eigenschaften generieren.

Das System deckt die Flexion der albanischen Nomina, Verben, Adjektive, Numeralia, Adverbien und Pronomina ab, sowie die nicht flektierenden Wortarten und die häufigsten Typen der Wortbildung. Es wurde mit einer Reihe von Testlisten aus unterschiedlichen Quellen getestet.

Mit diesen Eigenschaften eröffnet sich für das Morphologie-Werkzeug ein breites Spektrum von Anwendungsfällen in der maschinellen Verarbeitung der albanischen Sprache.

ISBN 978-3-944057-40-8
9 783944 057408

FAU UNIVERSITY PRESS 2015