Bachelorarbeit

im Studiengang Computerlinguistik

an der Ludwig- Maximilians- Universität München

Fakultät für Sprach- und Literaturwissenschaften

Department 2

Dialektspezifische Morphologieparadigmen des Sizilianischen für AnIta

vorgelegt von
Simeon Herteis

Betreuer:	Dr. Desislava Zhekova
Aufgabensteller:	Dr. Desislava Zhekova
Bearbeitungszeitraum:	17/21.03.2014 - 26.05.2014

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 26. Mai 2014

.....
Simeon Herteis

Zusammenfassung

Diese Arbeit beschreibt die regelmäßigen Verbparadigmen des sizilianischen Dialekts und formuliert sie als Erweiterung für die italienische *Finite State Morphologie AnIta*. Die Funktionalität, die AnIta mir dieser Erweiterung zur Verfügung stellt, ist interessant für Anwendungen in der italienischen Dialektologie. Die Grundproblematik dialektologischer Forschung wird mit Blick auf *morphologisches Parsing* erläutert. Die Grundlagen von *Finite State Transducern* und der darauf gründenden Anwendungen werden behandelt und ein Überblick über den Aufbau von AnIta gegeben. Mit einer sizilianischen Grammatik als Referenz werden aus Einträgen der sizilianischen Wikipedia nach einem einfachen, automatisierten Ansatz Verblemmata für das Lexikon von AnIta erstellt. Insgesamt wird AnIta um 368 sizilianische Lemmata erweitert. Die *Word Error Rate (WER)* für einen simplen Analysetest mit einer manuell erstellten Verbliste beträgt 70% (entspr. einer Abdeckung von 30%). Die gesamte Zahl der durch Flexionsparadigmen erfassten Verbformen beläuft sich auf ca. 24.700. Damit bietet diese Arbeit einen ersten Grundstock für die Entwicklung einer computergestützten, sizilianischen Morphologie.

Danksagungen

Zunächst möchte ich mich bei Frau Doktor Zhekova herzlich für die Anregungen und Unterstützung bei der Wahl des Themas dieser Arbeit und die professionelle Beratung zu wissenschaftlichen Fragen bedanken. Mein Dank gilt außerdem Herrn Prof. Dr. Thomas Krefeld und Frau Elisa Gallo vom Institut für Romanische Philologie der Universität München für die sehr hilfreichen Hinweise zu adäquaten sprachwissenschaftlichen Quellen zum sizilianischen Dialekt. Ebenso möchte ich Herrn Francis Tyers von der Arctic University of Norway für die technische Beratung und Expertise zu HFST sowie Herrn Fabio Tamburini von der Univesität von Bolognia für die Bereitstellung der Arbeitsmaterialien zu AnIta danken. Zu guter Letzt bedanke ich mich bei Herrn J.K. Bonner für seine exzellente grammatische Einführung zum Sizilianischen.

Inhaltsverzeichnis

1. Dialekte - Von der Feldforschung zur Wikipedia

Den ersten der beiden folgenden Sätze werden viele Italiener möglicherweise nicht verstehen, den zweiten aber schon:

(1) (a) „Studio fa dottu, no maistru"

(b) „Lo studio rende colti, ma non maestri" (Cùnsolo, 1977, S. 79)

Es handelt sich um ein Sprichwort in sizilianischem Dialekt (1 a) und seine italienische Entsprechung (1 b). Die deutsche Übersetzung lautet ungefähr: *„Das Studium macht gelehrt, nicht Meister"*

(2) (a) „Ned gschimpft is globt gnua"

(b) „Nicht geschimpft ist Lob genug"

Dieses bayrische Sprichwort (2 a) wiederum werden viele Deutsche sowie fast alle Italiener nicht verstehen. Die entsprechende deutsche Form (2 b) ist besser verständlich.

Jeder Mensch spricht mindestens eine Sprache, die eigene Landessprache und einen Dialekt, die Mundart des Heimatorts. Die Landessprache, auch als *Standardsprache* bezeichnet, versteht man überall innerhalb des gleichen Sprachraums, etwa Deutsch in Österreich, Südtirol, der Nordschweiz und Deutschland oder Italienisch in der Südschweiz und Italien. Für die Landessprachen gibt es Konventionen zu Rechtschreibung und Grammatik, die festhalten, welche Satz- und Wortformen gültig sind. Die standardsprachlichen Versionen der beiden obigen Sprichwörter (1 b u. 2 b) sind besser verständlich, wegen der einheitlichen Schreibweise und den verwendeten Begriffen, die allgemein bekannt sind.

Einen Dialekt hingegen versteht man nicht unbedingt überall. Die Aussprache ist variabel und verwendete Begriffe sind oft gar nicht geläufig. Grammatik und Wortschatz können sich dabei über kurze räumliche Distanzen stark verändern. Konventionen zur Schreibweise gibt es keine. Die Bewohner zweier Orte, die in verschiedenen bayrischen Regionen liegen oder zweier Städte auf gegensätzlichen Seiten der sizilianischen Insel würden die obigen Sätze wahrscheinlich jeweils ein wenig anders formulieren. Und sich dabei entweder einer anderen Aussprache bedienen, was sich in der Schreibweise niederschlägt, oder eine andere Satzgrammatik benutzen. Sie gebräuchten möglicherweise auch ganz andere Begriffe für das Bezeichnete. Wie stark die Abweichungen ausfallen, hängt ab vom sprachlichen Abstand zur Landessprache oder zum Nachbardialekt. An den obigen Beispielsätzen (1 a u. 2 a) kann man dieses Verhältnis zwischen Dialekt und Landessprache nicht direkt ablesen. Man stellt aber fest, dass man sie gut oder schlecht verstehen kann.

Mit diesen Problemen befasst sich die Dialektologie (s. Abschnitt 2.3), die Lehre von den Dialekten. Sie untersucht das komplexe Zusammenspiel von regionalen Sprechweisen, Mundarten und Landessprache. Das tut sie unter anderem in aufwändiger Feldforschung. Welche praktischen Auswirkungen die beschriebenen Zusammenhänge neben Schriftform und Verständlichkeit noch haben, lässt ein Blick auf alltägliche Kommunikationsmedien erahnen.

Die moderne Informationstechnologie ist seit einiger Zeit sehr gut in der Lage, geschriebene und gesprochene Sprache zu verarbeiten und zu verstehen. Die Anwendungen reichen von Textprogrammen mit automatischer Rechtschreibkorrektur und Diktierfunktion über Sprachsteuerung von elektronischen Geräten, die gesprochene Befehle ausführen, bis hin zu Dialogsystemen,

die einfache Unterhaltungen führen können. Dazu benötigt man spezielle Technologien, die fähig sind die Struktur natürlicher Sprachen abzubilden und für Computeranwendungen nutzbar zu machen. Man spricht hier generell von der *Verarbeitung natürlicher Sprache*, englisch *Natural Language Processing* (NLP). Die Anforderungen an diese Disziplin werden durch die Fortschritte der beteiligten Wissenschaften wie der Computerlinguistik in den letzten Jahren immer besser erfüllt.

Alle hier genannten und die weiteren Verwendungszwecke von NLP-Programmen benötigen jedoch eine Voraussetzung, um gut zu funktionieren: Die Eingabe muss in einer Standardsprache erfolgen. Dialekte werden schlecht bis überhaupt nicht unterstützt. Möchte man das ändern, muss man die vorhandene Technologie an die dafür geltenden Gesetzmäßigkeiten anpassen und erweitern. Anders gesagt: Man braucht ein funktionsfähiges Analysewerkzeug für Dialekte.

Für die Analyse von Sprachen gibt es spezielle Programmlösungen (siehe Abschnitt 3.2). Würden diese Programme zu Dialekten brauchbare Daten liefern, könnten Sprachprogramme „verstehen", wenn man in einem Dialekt mit ihnen spricht. Man könnte dann beispielsweise ein - in Taiwan hergestelltes - Smart-Phone auf bayrisch nach dem nächstgelegenen Biergarten fragen. Und dann eine Nachricht an den „Spezl" senden, dass der Stammtisch „füa heit umziagt, weil'd Maß do weniga deier is", ohne dass die automatische Rechtschreibkorrektur Fehler meldet. Ähnliche Anwendungsfälle sind auch für das Sizilianische oder andere Dialekte denkbar. Suchergebnisse zu guten Biergärten dürften allerdings auf das Bayrische beschränkt bleiben. Der konkrete Anwendungsfall, der den Hintergrund zu dieser Arbeit bildet, ist vor allem die textbasierte Spracherkennung. Die Erkennung gesprochener Sprache wird hier nicht behandelt und dient im oben stehenden Beispiel nur zur Veranschaulichung.

Das benötigte Wissen für die Verarbeitung von Texten in einem Dialekt liefert die Dialektologie zu einem Teil bereits. Es gibt zahlreiche Grammatikwerke, Wörterbücher und Sprachatlanten für die verschiedensten Dialekte. Diese sind aber nicht vollständig, besser gesagt, beschreiben sie immer nur einen, mehr oder weniger räumlich beschränkten, Teil eines Sprachraums. Sie geben keinen Gesamtüberblick zu größeren Sprachräumen, wie das Grammatiken für Standardsprachen bieten. Die vorhandenen Informationen müssen also so zusammengefasst und ergänzt werden, dass möglichst viele Dialektvariationen abgedeckt werden.

Neben den Erfordernissen für die sprachlichen Informationen stellt sich außerdem das Problem der Qualität: Um sicherzustellen, dass Softwarelösungen für eine maschinelle Sprachanalyse fehlerfrei arbeiten, müssen sie getestet werden. Man benötigt Testdaten, mit denen man die Programme prüfen und etwaige Sonderregeln ableiten kann. Seit mehreren Jahren sind größere sprachwissenschaftliche Textsammlungen, sogenannte Korpora, frei im Netz verfügbar. Unter den Prominentesten sind etwa das Brown-Korpus[1] oder das Gutenberg Projekt[2]. Diese Korpora werden in der NLP auch seit Jahren für maschinelles Lernen, statistische Untersuchungen und Bewertung sowie Verbesserung von Software verwendet. Allerdings handelt es sich dabei um Texte in einer Standardsprache. Korpora für Dialekte gibt es kaum.

Seit einiger Zeit gibt es jedoch spezielle Wikipediaseiten, auf denen mehrere tausend Artikel in diversen Dialekten gesammelt sind[3]. Deren Autoren haben zumeist zwar keine linguistische

[1] Handbuch zum Korpus: http://khnt.aksis.uib.no/icame/manuals/brown
[2] http://www.gutenberg.org
[3] Bayerische Wikipedia: https://bar.wikipedia.org
Sizilianische Wikipedia: https//scn.wikipedia.org

Ausbildung, grammatische Inkonsistenzen sind deshalb häufiger. Die Texte könnten aber nach der Überprüfung und Bearbeitung durch Wissenschaftler als Korpora zu Studienzwecken und Tests verwendet werden. Man müsste sich nicht mehr ausschließlich auf die spärlichen und verstreuten Informationen verlassen, die in der Dialektologie bisher zur Verfügung stehen. Je mehr Daten daher in Wikipedia und vergleichbaren Medien zugänglich werden, umso interessanter werden sie für die Forschung und die Entwicklung von Analyseprogrammen für Dialekte.

Für den bayrischen Dialekt etwa sind bisher knapp 10.000 Artikel[4] in der Wikipedia vorhanden, ein relativ geringer Bestand. Für das Sizilianische allerdings beträgt die Anzahl der Artikel mehr als das Doppelte, circa 24.000[5]. Italienische Dialekte sind klarer zum italienischen Standard abgegrenzt als die deutschen Dialekte. Die Grammatik ist eigenständiger und weiter ausgebildet. Grammatikbeschreibungen sind daher meist verlässlicher.

Es wird deshalb der sizilianische Dialekt als Studienobjekt dieser Arbeit gewählt. Aus einer Einführung in die Grammatik des Sizilianischen (Bonner, 2001) wird eine Grundlage für die maschinelle Untersuchung dieses italienischen Dialekts geschaffen. Das Ziel ist eine Erweiterung von AnIta (Tamburini u. Melandri, 2012), einem Analyseprogramm für das Italienische, sodass sizilianische Begriffe parallel zum Italienischen erkannt werden. Dabei dienen Einträge aus dem sizilianischen Wiktionary[6] als Infomationsquelle für eine automatische Erfassung und Kategorisierung sizilianischer Wortformen dienen. Mit Hilfe der Texte aus der Wikipedia wird die Erweiterung anschließend geprüft. Außerdem wird beschrieben, wie zusätzliche Informationen hinzugefügt werden können und welche Anwendungsmöglichkeiten das entwickelte Programm bietet. Im folgenden werden nun zunächst die sprachwissenschaftlichen und technischen Grundlagen erläutert, die für diese Arbeit relevant sind.

[4] knapp 10.000 Artikel (Stand: 24.5.2014), siehe: https://bar.wikipedia.org/wiki/Spezial:Statistik
[5] knapp 24.000 Artikel (Stand: 24.5.2014), siehe: http://scn.wikipedia.org/wiki/Spiciali:Statistiche
[6] http://scn.wiktionary.org/wiki/P%C3%A0ggina_principali

2. Wortbildung, Wortanalyse & Mundartforschung

2.1. Morphologisches Parsing

In allen Sprachen entstehen durch bestimme Prozesse laufend neue Wörter. Nach bestimmten Regeln lassen sich flexibel neue Wörter erzeugen, die ein gültiger Teil der Sprache sind. Man spricht in diesem Zusammenhang von Wortbildung und ihrer Produktivität. Mit den Bestandteilen von Wörtern einer Sprache und den Regeln, aus denen sie gebildet werden, befasst sich die sogenannte Morphologie. Sie ist das Studium der Formen von Wörtern und der Wortbildung (Glück, 2010)[7]. Die Bestandteile von Wörtern nennt man Morpheme. Ein Morphem verkörpert die kleinste bedeutungstragende Einheit einer Sprache. Die Wortbildung wird in verschiedene, teilweise von der Wortart abhängige, Prozesse untergliedert. Eine Auswahl ist wie folgt:

Deklination: Anpassung von Substantiven an Kasus und Numerus

Konjugation: Beugung von Verben

Komparation: Steigerung von Adjektiven und Adverbien

Flexion: Wortbeugung, Überbegriff für Deklination, Konjugation und Komparation

Komposition: Zusammensetzung von bestehenden zu neuen Wortformen

Konversion: Änderung der Wortart bei gleicher Wortform (z.B. Nominalisierung)

Derivation: Ableitung einer neuen Wortform aus einer bestehenden

Klitisierung: Verkürzung oder Verschmelzung von Wörtern durch Auslassung von Buchstaben

In der Computerlinguistik nennt man maschinelle Analysewerkzeuge, die die oben aufgeführten Bildungsprozesse erkennen können, *morphologische Parser* [8]. Parser lesen Sätze ein, zerlegen sie in einzelne Bestandteile und geben diese, mit Anmerkungen versehen, wieder aus. Bei der Untersuchung von Sprache auf der Wortebene spricht man von *Morphologischer Analyse* oder *Morphological Parsing* (vgl. Jurafsky u. Martin, 2009). Beim morphologischen Parsing wird ein Wort in seine Morpheme zerlegt und diese werden analysiert. Tabelle 2.1 gibt eine Übersicht von Anwendungen, denen morphologisches Parsing zu Grunde liegt.

Die Produktivität bestimmter Wortbildungsprozesse ist im Vergleich zu anderen sehr hoch. Mittels morphologischem Parsing kann man neue, noch unbekannte Wörter erkennen. Dabei werden die speziellen Wortbildungsphänomene und -prozesse einer Sprache auf Basis der bestehenden Regeln abgebildet. Neue Wortformen werden analysiert und anhand ihrer Morphemzusammensetzung erkannt. Verfügt ein unbekanntes Wort etwa über ein Morphem eines schon bekannten Verbs und zusätzlich noch über eine Adjektivendung, ist es höchstwahrscheinlich ein neues Adjektiv, das von einem bestehenden Verb abgeleitet wurde. Dieses Prinzip des „Wortbaukastens" verschafft gegenüber einem Vollformenlexikon zudem den Vorteil, dass die Zahl der nötigen Einträge sehr viel geringer ist und schneller verarbeitet werden kann. Die vorhanden

[7] Sämtliche folgenden Definitionen zu Begriffen der Sprachwissenschaft sind, soweit nicht anders gekennzeichnet, ebenfalls diesem Werk entnommen.

[8] von *parsing,* dt. etwa Satz-/Syntaxaanalyse, (siehe Glück, 2010)

Tabelle 2.1: Anwendungsbereiche für morphologisches Parsing

Gebiet	Funktion	Beispiel	Ergebnis
Internetsuche	Suchbegriffe grammat. erweitern (Konjugation, Derivation)	„Treiber installieren"	+ „Treiber installiert" + „Treiber Installation" + „Treiber Deinstallation" ...
POS-Tagging[9]	Wortkategorien erkennen	„Käsekuchen schmeckt gut"	Käsekuchen/Nomen schmeckt/Verb.-Präs. gut/Adjektiv
Rechtschreib-prüfung	red. Lexika + neue Wörter „erraten"	Verbstamm + Endung „genieß-bar" → Adj. neu: „ungenießbar"	genießen, genießbar: Verb → Adj.; gültig Un-genieß-bar: möglw. Adj.; gültig
masch. Übersetzung	Wortform u. Bedeutung erkennen + Disambiguierung[10]	„säe"	- säen, V., 1. Pers., Präs. oder - säen, V., 1./3. Pers, Konj. I

Wörter einer Sprache können nach den vorgegebenen Regeln aus einer reduzierten „Stammliste" erzeugt werden, die nur jeweils die Grundform und den Wortstamm enthält.

Ein Algorithmus, nach dem diese Art von Parsern effizient programmiert werden können, ist der endliche Transduktor (ET) oder in der englischen Entsprechung der *Finite State Transducer*. Bei der Umsetzung der Morphologie einer Sprache in ein ET-System spricht man von einer *Finite State Morphology* (FSM). Mithilfe von einigen ET-Programmen, die im folgenden Abschnitt vorgestellt werden, können FSM für die morphologische Wortanalyse entwickelt werden.

2.2. Parser für Standardsprachen

Das zum Gegenstand der Bearbeitung genommene Analyseprogramm AnIta(Tamburini u. Melandri, 2012) verwendet für das morphologische Parsing eine Technologie die auf die Automatentheorie(vgl. Jurafsky u. Martin, 2009; Pfister u. Kaufmann, 2008) aufbaut. Automaten, genauer gesagt die Transduktorautomaten, sind der Schlüsselalgorithmus für effiziente Programmierung von Anwendungen wie sie in Tabelle 2.1 beschrieben sind. Neben AnIta mit Italienisch wurden unter anderem auch Morphologien für Türkisch (Çağrı Çöltekin, 2010), Deutsch (Schmid u. a., 2004) und Finnisch (Pirinen, 2008) auf Basis dieses Algorithmus entwickelt. Das Prinzip der Automaten ist vereinfacht in den Abbildungen 2.1 und 2.2 auf der nächsten Seite dargestellt. ET sind im Prinzip eine Art von Übersetzungsapparat: Ein eingegebenes Wort, d.h. die einzel-

[9] *Part Of Speech Tagging*, Kategorisierung d. Elemente eines Satzes in jew. Wortart
[9] Zweideutigkeit auflösen

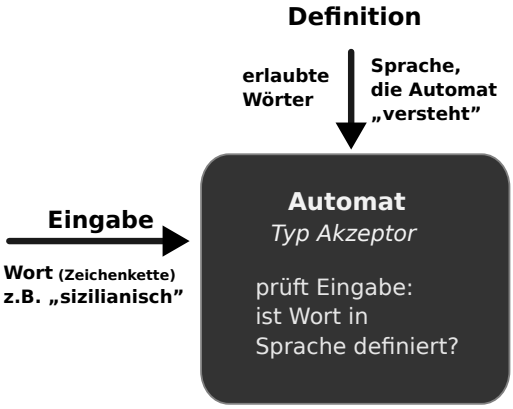

Abbildung 2.1: Akzeptorautomat

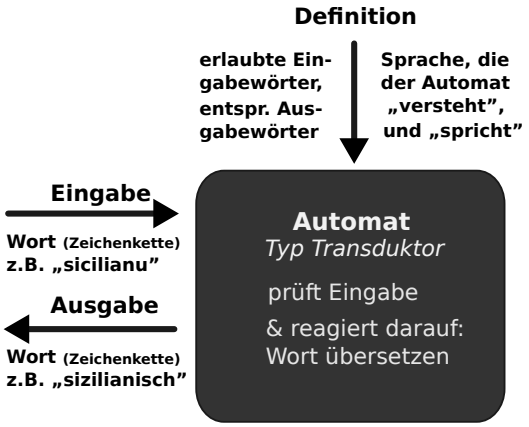

Abbildung 2.2: Transduktorautomat

nen Zeichen darin, werden nach den definierten Regeln des ET (Details siehe Abschnitt 3.1 auf Seite 12), in eine entsprechende Ausgabe umgewandelt. Der ET-Algorithmus eignet sich für die Bearbeitung folgender Aufgaben in der maschinellen Sprachverarbeitung(Jurafsky u. Martin, 2009):

- Erkennung:

  z.B. von Wörtern einer Sprache

- Generierung:

  z.B. von Wörtern und Sätzen einer Sprache

- Übersetzung:

  z.B. von Wörtern, die in Morpheme zerlegt und entsprechend ihrer Zusammensetzung oder Kategorie in der am ehesten passenden Bedeutung in eine anderen Sprache übersetzt werden

- Relationsberechnung:

  z.B. Ausgabe der möglichen Bedeutungen zu einem Wort mit bestimmten Morphemen (Disambiguierung)

Die weiteren Begriffe, Funktionsweisen und mathematische Definitionen von Automaten sind in Abschnitt 3.1 auf Seite 12 erläutert.

Mehrere Programme bzw. Programmiersprachen setzen den ET-Algorithmus in Werkzeuge um, die für eine maschinelle Sprachverarbeitung im Allgemeinen und Morphologieuntersuchungen im Speziellen verwendet werden können. Hier ein kurzer Überblick über die wichtigsten Bibliotheken:

XFST Das *Xerox*-Paket (Beesley u. Karttunen, 2003) ist eines der älteren Werkzeuge, besser gesagt eine Werkzeugsammlung. Sie wurde mit dem Ziel entwickelt, basierend auf ET, rechnergestützte Morphologien einfach formulierbar zu machen. Das Format XFST (*Xerox Finite State Technology*) ist eine Art Quasi-Standard. Zunächst unter kommerzieller Lizenz am *Xerox Palo Alto Research Center*[11] entwickelt, wurde es später für nicht-kommerzielle Nutzung freigegeben. Das *XFST*-Format wird von vielen Morphologieprojekten verwendet.

foma Das foma-Progamm (Hulden, 2009) ist eine eigenständige, quelloffene Implementierung für XFST. Es kann XFST-Quellcode verarbeiten und ausführen.

OpenFst Das OpenFst-Projekt (Allauzen u. a., 2007), ebenfalls frei verfügbar, verwendet sein eigenes Format für Transduktoren. Interessantes Merkmal dieser Implementierung ist, dass sie eine spezielle Form von Transduktoren verwendet, die eine zusätzliche Gewichtung der möglichen Übergänge erlaubt. Damit kann zum Beispiel vermerkt werden, wie produktiv ein bestimmter Wortbildungsprozess ist oder welche Bedeutung von zwei möglichen Übersetzungen eher zutrifft.

SFST Mit dem freien Projekt *Stuttgart Finite State Transducer* (Schmid, 2006) gibt es eine weitere Programmiersprache (neben XFST), mit der Transduktoren definiert werden können.

HFST Das Programm *Helsinki Finite State Technology* ist eine quelloffene Entwicklungsumgebung für Transduktoren, ähnlich wie Xerox-XFST. Sie wurde für das AnIta-Projekt verwendet. Eine kurze Einführung gibt der Abschnitt 3.2.

2.3. Dialektologie

Wie zur Einleitung bereits angemerkt, ist die Dialektologie die Lehre von den Dialekten. Sie wird auch Dialekt- oder Mundartforschung genannt. Sie versucht den Wissensstand zu regionalen Mundarten einer Landessprache und deren Vokabular und Grammatik zu erhalten sowie zu vergrößern. Man teilt die Dialektologie in die Dialektdokumentation und Dialektgeographie auf. (Löffler, 2003; Niebaum u. Macha, 2006; Benincà, 1996). Inwiefern ihre Ziele mit dieser Arbeit zusammenhängen, soll ein kurzer Blick auf die Geschichte der Dialektforschung verdeutlichen.

Wissenschaftler und Sprachbegeisterte begannen schon sehr früh damit, regionale Mundarten zu dokumentieren. Erste Vokabularsammlungen entstanden bereits zwischen dem 16. und 17. Jahrhundert. Für die Etablierung der modernen Dialektologie im Europa der neueren Zeit und den verwandten Disziplinen wie der Dialektgeographie und Dialektsoziologie sorgte aber vor allem das Aufkommen von Sprachatlanten. Die ersten wegweisenden Kartensammlungen

[11] www.parc.com

stellten Gilliéron u. Edmont (1902) und Wenker u. Wrede (1926) jeweils für den deutschen und französischen Sprachraum zusammen. Die Methoden der Datenerhebung für die Karten, die von diesen Autoren begründet wurden, stützen sich auf direkte Feldforschung. Zum Ende des 20. Jahrhunderts befuhren speziell geschulte Feldforscher die verschiedenen Sprachregionen und befragten ausgewählte Informanten mithilfe umfangreicher Kataloge. Ohne einheitliche Schriftform mussten sämtliche erfassten Dialekte in einer genormten Lautschrift festgehalten werden. Die Erhebungszeiträume erstreckten sich über Jahre. Auch die Arbeiten zu aktuelleren Projekten wie etwa der *Sprachatlas für Bayern* (Hinderling u. König, 2005) oder der *Sprach- und Sachatlas Italiens und der Südschweiz* (Jaberg u. Jud, 1928) in Italien laufen oder liefen über ein Jahrzehnt und sind noch längst nicht vollständig.

Zu diesem rein technischen Aufwand gesellt sich noch ein anderes Problem. Die Antworten der für die Erhebungen persönlich oder mit Fragebögen befragten Personen sind nicht hundertprozentig authentisch. Die Befragten befinden sich in einer Art „Prüfungssituation", die nicht den alltäglichen Bedingungen entspricht und passen dabei ihre Wortwahl oder Aussprache dem Gegenüber bzw. der Situation und vermeintlichen Erwartungen an. Diesen Effekt kann man bei Texten von dialektspezifischen Wikipediaseiten vernachlässigen (nahezu, siehe Ende d. nächsten Absatz). Da die Autoren die Artikel nicht auf Nachfrage verfassen, verwenden sie ihren Dialekt so, wie sie es im Alltag würden.

Digitale Ausgaben von Sprachatlanten wie der *Digitale Wenkeratlas*[12] (auf Basis von Wenker u. Wrede, 1926) und der *AdIS*[13](auf Basis von Jaberg u. Jud, 1928) erleichtern die Verwaltung und Verarbeitung dialektbezogener Forschungsmaterialien, stützen sich aber, ebenso wie ihre Originalfassungen, auf manuell zusammengetragene Daten. Im Vergleich dazu erleichtern die neuen Wikipediaseiten für Dialekte potentiell den Zugang zu Forschungsmaterial enorm. Die Sprecher der Dialekte müssen nicht mehr direkt befragt werden, da sie selbständig Texte in ihrer Mundart verfassen. Allerdings gehen dabei vor allem Informationen zur Aussprache verloren, die Orthographie eines Wortes kann nämlich nur einen kleinen Teil der Aussprache wiedergeben. Außerdem passt man sich beim Schreiben auch immer ein Stück weit der Umgebung an. Im Fall der dialektspezifischen Wikipedia ist das die Ausdrucksform, die andere Autoren in Artikeln verwenden. Aus diesen Gründen tendieren die Texte aus der Wikipedia zu einem „inoffiziellen" Standard[14]. Dennoch eröffnen sich durch diese neuen Quellen in Kombination mit den morphologischen Parsern neue Ansätze für die Dialektologie, vor allem auch für die vergleichende Forschung zu verschiedenen Dialekten des selben Sprachraums.

2.4. Morphologisches Parsing in der Dialektologie: Neuland

Die zuvor aufgestellte Behauptung, dass Dialekte von Parsern nicht unterstützt werden (S. 3) ist nicht ganz korrekt. Tatsächlich gibt es keine Parsing-Programme, die speziell auf einen Dialekt zugeschnitten sind. Attila Novák hat jedoch eine Morphologie für sechs kaum dokumentierte uralische Sprachen, sogenannte Kleinsprachen, zusammengestellt (Novák, 2006). Diese Kleinsprachen der somojedischen und finno-ugrischen Sprachfamilie werden von Bevölkerungsminder-

[12]http://www.diwa.info/titel.aspx
[13]http://www.adis.gwi.uni-muenchen.de/AIS.php
[14]Diese Feststellung wurde bei einer gemeinsamen Betrachtung der fraglichen Wikipediaseiten von Prof. Dr. Thomas Krefeld am Institut für Romanische Philologie der Universität München gemacht

heiten in Russland gesprochen. Auch wenn diese wissenschaftliche Arbeit von ihrer Motivation keine direkt dialektologische Forschung darstellt, kann man sie annähernd als solche betrachten, zumindest was die Ausgangslage und den Ansatz betrifft. Für diese Minderheitensprachen sind diese sehr ähnlich zu denen der Dialektologie. Im Übrigen liegt der Unterschied zwischen Dialekt und der Sprache einer Minderheit gewissermaßen nur in der politisch-historischen Betrachtungsweise; Sprachen von Bevölkerungsminderheiten sind letztendlich ebenfalls eine lokale Mundart.

Zum Zeitpunkt der Recherche zur dieser Arbeit stellten die Forschungen Nováks die einzige bekannte Studie dar, in der morphologische Analysewerkzeuge für lokale Sprachvarianten entwickelt wurden. Novák strebte die Dokumentation der uralischen Kleinsprachen an, die teilweise sogar vom Aussterben bedroht sind. Sie wurden bisher entweder gar nicht oder nur unzureichend untersucht. Einen Teil der Ergebnisse seiner Forschungen bereitete seine Forschungsgruppe in einer Software auf, die ursprünglich für finno-ugrische Sprachforschung entwickelt wurde. Den anderen Teil, die somojedischen Kleinsprachen hat man im XFST-Format formuliert. Das bereits erwähnte Analysewerkzeug AnIta verwendet ebenfalls XFST für seine Morphologie. Das Ergebnis der Forschungsarbeiten sind funktionsfähige Morphologieanwendungen für fünf der sechs untersuchten Sprachen.

Nachdem also mit den uralischen Kleinsprachen gewissermaßen dialektähnliche Sprachformen erfolgreich in eine transduktorbasierte Morphologie übersetzt wurden, AnIta für Italienisch im selben Format einen Parser bereitstellt und zudem die besprochenen Quellen für Dialekte verbesserte Bedingungen schaffen, scheint der Zeitpunkt günstig, die Entwicklung eines Dialekt-Parsers zu beginnen. Das soll in dieser Arbeit geschehen. Genauer gesagt soll AnIta um die Fähigkeit ergänzt werden, italienischen Dialekt neben Standarditalienisch zu analysieren.

Diese Arbeit erhebt nicht den Anspruch auf Vollständigkeit, als Schwerpunkt für die Erarbeitung - die gesamte Grammatik des Sizilianischen kann hier nicht behandelt werden - ist die Konjugation (Verbbeugung) regelmäßiger Verben gewählt. AnIta soll um die Paradigmen (ein Paradigma enthält alle flektierten Formen eines Wortes) dieser Verben erweitert werden.

Die genutzten Programme, HFST und AnIta, sowie die Grundlagen der Automatentheorie werden im nächsten Abschnitt besprochen. Die dann folgenden Abschnitte beschreiben den Dokumentationsvorgang, der sich am Ansatz von AnIta orientiert und die Erweiterung für dessen Datensätze. Dabei werden aus einem Grammatikwerk für das Sizilianische (Bonner, 2001) die benötigten grammatischen Informationen zu verbalen Flexionsparadigmen zusammengestellt und zu AnIta hinzugefügt. Zusätzlich wird aus den Artikeln der sizilianischen Wikipedia und der zugehörigen Wiktionaryseite jeweils ein Korpus erstellt und aus diesen Lemmata (Wörterbucheinträge) für sizilianische Verben beziehungsweise eine Testliste für die Erweiterung extrahiert. Die von Bonner zusammengestellten regelmäßigen Verben werden ebenfalls aufgenommen.

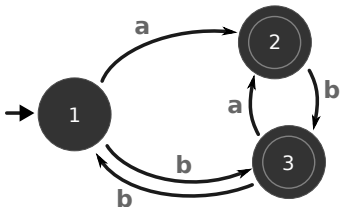

Abbildung 3.1: Einfacher Akzeptor

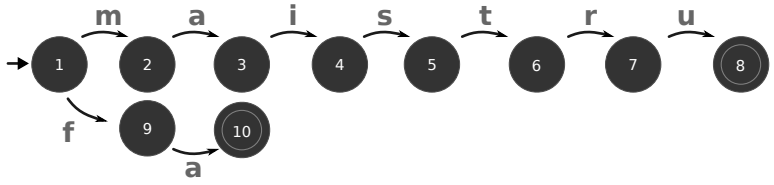

Abbildung 3.2: Akzeptor für ganze Wörter

3. Transduktoren & Sprachanalyse auf Italienisch

Die zuvor in Abschnitt 2.2 eingeführten Transduktoren werden in den folgenden Abschnitten eingehender betrachtet. Im Abschnitt 3.3 wird dargelegt, wie AnIta bzw. die HFST-Umgebung die Eigenschaften der Transduktoren konkret einsetzen, um morphologische Analysen von Eingaben zu erzeugen.

3.1. Automaten

Ein Automat ist ein Zustandsmodell. Jedem seiner Zustände sind ein oder mehrere „Pfade" zugeordnet, mit denen der Zustand von anderen aus erreicht werden kann. Diese Pfade oder Übergänge erhalten jeweils ein Symbol, das ein beliebiges Zeichen oder eine Kette von Zeichen sein kann. Es gibt zwei Arten von Automaten: Akzeptoren und Transduktoren.

3.1.1. Der endliche Akzeptor

Der Automat vom Typ *Akzeptor* wird oft nur als Automat bezeichnet. Automaten lassen sich mathematisch beschreiben und implementieren. Ein endlicher Automat (EA) verfügt per Definition über eine endliche Zahl von Zuständen. Für die praktische Anwendung in Rechnerprogrammen werden EA genutzt. Automaten können außerdem deterministisch oder nichtdeterministisch sein. Der Unterschied liegt bei den möglichen Übergängen.

Bei nichtdeterministischen Automaten (NDEA) gibt es mehrere Übergänge, die mit dem selben Symbol gewählt werden können. Die Anzahl der möglichen Pfade ist nicht mehr linear sondern proportional zur Zustandsmenge. Man kann sich die Zustände dabei analog zu Weggabelungen oder Kreuzungen vorstellen. Jedes mal, wenn man vor einer Abzweigung steht, muss man sich für einen von mindestens zwei Wegen entscheiden. Je mehr Kreuzungen es gibt, umso mehr Möglichkeiten gibt es für die Route, die man wählt. Deterministische Automaten (DEA) bieten in jedem Zustand für ein gelesenes Symbol nur einen möglichen Übergang zu einem nächsten Zustand. Die Menge der möglichen Symbolfolgen ist linear zur Zustandsmenge. In der Analogie der Abzweigungen kämen an einer deterministischen Kreuzung zwei entgegengesetzte

Einbahnstraßen zusammen. Man kann in einer Richtung nur jeweils einen Weg wählen. DEA können dadurch mit Rechnern effizienter verarbeitet werden.

Ein einfacher, deterministischer endlicher Automat kann ähnlich wie in Abbildung 3.1 auf der vorherigen Seite aussehen. Die Elemente eines Automaten sind Mengen und Produkte von Mengenoperationen. Sie werden für Abbildung 3.1 folgendermaßen definiert:

Eingabealphabet Menge der im Automat verwendeten Symbole: $\Sigma = \{a, b\}$

Zustände des Automaten: $Q = \{1, 2, 3\}$

Funktion der möglichen Übergänge: $\delta : \{< 1 \to a \to 2 >, < 2 \to b \to 3 >, < 1 \to b \to 3 >, \ldots\}$[15]

Startzustände in denen die Eingabeerkennung beginnt, hier nur: $s = \{1\}$

Finalzustände in denen der Symbolfolge enden kann: $F = \{2, 3\}$, $F \subseteq Q$

Dieser Automat bildet abstrakte Symbolfolgen der Buchstaben a und b ab. Mögliche Eingabefolgen die der Automat erkennt, wären zum Beispiel: *a, abab, ba, bba, usw.* Die Liste lässt sich weiter fortführen. Hier zeigt sich der Vorteil dieses Algorithmus: Schon bei einer übersichtlichen Definition von Symbolen und Übergängen lassen sich eine Vielzahl von möglichen Zeichenketten erkennen. Man spricht dabei von der Sprache $L(A)$, die ein Automat A definiert bzw. erkennt. Sie ist die Menge aller erlaubten Kombinationen der Elemente des Alphabets. Ein Programm, dass einen Automaten implementiert, kann erkennen, ob eine Eingabe Teil seiner Sprache ist, oder nicht.

Mit Automaten können natürlich nicht nur Sprachen aus abstrakten Zeichenketten erkannt werden. Auch natürliche Wörter können abgebildet werden. Der sehr einfache DEA in Abbildung 3.2 auf der vorherigen Seite erkennt etwa die Wörter *fa* und *maistru* (s. Bsp. 1 a auf Seite 2). Um alle Wörter im Sprichwort zu erkennen, müsste die Menge der Zustände und Symbole sowie die Übergänge nur entsprechend erweitert werden. Um ganze Sätze und Sprachen zu einzulesen, muss in der Definition zusätzlich zum Alphabet noch die Grammatik formal als Übergangsregeln hinterlegt werden. Kombinierte Informationen zu Wortstruktur und Satzsyntax (Morphosyntax) können in Form von entsprechenden Symbolen in einen Automaten eingefügt werden. Damit lassen sich aus Automaten komplexe Grammatikprüfer konstruieren, die zum Beispiel die Zusammensetzung von Wörtern anzeigen oder kontrollieren.

Dem Akzeptorautomaten fehlt jedoch eine entscheidende Komponente: Die Ausgabe. Wie er bisher beschrieben wurde, kann er lediglich Zeichenfolgen als gültig oder ungültig nach seiner Definition erkennen und mit dem Durchlaufen der angesprochenen Zustände reagieren. Er kann keine Ausgabe erzeugen und zum Beispiel zu einem gelesenen Wort die Wortklasse oder morphologische Zusammensetzung ausgeben, die als Übergangssymbole vorliegen.

3.1.2. Der endliche Transduktor

Diese Fähigkeit hat der endliche Transduktor (ET). Ein ET ist ein endlicher Automat mit zwei zusätzlichen Elementen, einem Alphabet, das die Menge der Ausgabesymbole festlegt und einer zweiten Funktion, die zusammen mit der ersten aus dem EA jedem Übergang ein zum

[15] lies: von Zustand 1 wird a gelesen und Zustand 2 erreicht, von Zustand 2 b gelesen und Zustand 3 erreicht

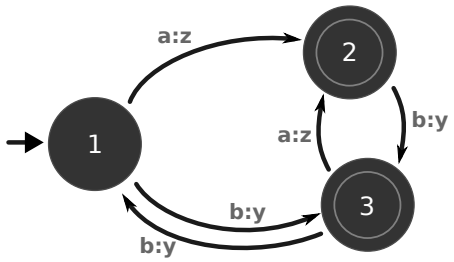

Abbildung 3.3: Einfacher Transduktor

Eingabesymbol gehörendes Gegenstück aus den Ausgabesymbolen zuordnet. Ausgaben können alternativ auch Zuständen zugeordnet sein. Abbildung 3.3 zeigt den zuvor definierten Automat mit zusätzlichen Ausgabesymbolen. Ein Transduktor stellt folglich die Symbole seiner beiden Alphabete miteinander in Relation. Die erweiterte Definition für den Transduktor lautet dann:

Eingabealphabet $\Sigma = \{a, b\}$

Ausgabealphabet $\Delta = \{y, z\}$

Zustände *wie oben*

Übergangsfunktion *wie oben*

Ausgabefunktion $\sigma = \{< 1 \rightarrow z \rightarrow 2 >, < 2 \rightarrow y \rightarrow 3 >, < 1 \rightarrow y \rightarrow 3 >, \ldots\}$

Startzustände *wie oben*

Finalzustände *wie oben*

Das obige Beispiel von Abbildung 3.3 ist wie der EA aus Abbildung 3.1 deterministisch. Die Determinisierung für komplexere ET geschieht mit sogenannten sequentiellen und subsequentiellen ET (Jurafsky u. Martin, 2009). Der zu determinisierendene ET wird dabei in einzelne Abschnitte unterteilt, sodass diese immer deterministisch sind. Die Definition wird dann so angepasst, dass ein Eingabesymbol genau einen Übergang mit genau einer Ausgabe vorgibt. Subsequentielle Transduktoren haben noch eine nützliche Eigenschaft: Sie können umgekehrt werden und damit in der Implementierung für Grammatikanalysen einmal regulär als Analysator, einmal als Generator eingesetzt werden, der mögliche Wörter seiner Sprache ausgibt.

Die interne Verarbeitung von morphosyntaktischen Informationen zu Wörtern auf der einen und die Ausgabe der konkreten, nicht mit Informationen versehenen Formen dieser Wörter auf der anderen Seite funktioniert über eine sogenannte Zweiebenenmorphologie (*Two Level Morphology*, Pfister u. Kaufmann, 2008; Glück, 2010). In der ersten und unteren der beiden Ebenen wird die interne morphosyntaktische Struktur eines Wortes festgehalten. Die hinterlegten Kennzeichnungssymbole sind mittels der definierten Übergänge nur in einer Richtung sichtbar, bei der Analyse eines Wortes. Die zweite, obere Ebene einhält nur die konkrete Wortform, bzw.

Tabelle 3.1: HFST - Unterstütze Formate

Name	**Art**	**Syntax**	**Format/API**
SFST	Eigenständige Sprache zu ET	SFST-PL	sfst/text
XeroX	XFST-Kompilierer & -Bibliothek	XFST	lexc/text, twolc/text
foma	freier XFST-Kompilierer & -Bibliothek	”	”
OpenFst	Programmbibliothek für gewichtete ET	FST	fst/text, C++-Objekt

die Oberflächenform. Bei der Generierung eines Wortes, werden Symbole die morphologische Eigenschaften darstellen mit einer leeren Ausgabe ersetzt. Dieses Verhalten von Ersetzungen steuern dabei kontextabhängige Übergangsregeln, die im ET formuliert sind.

3.2. Helsinki-FST

Der Programmverbund des HFST-Projekts bietet eine Entwicklungsumgebung für den Transduktoralgorithmus und darauf basierende Morphologien. Die erste Veröffentlichung stammt aus dem Jahr 2009 (Lindén u. a., 2009). Eine Neuauflage gibt es seit 2011 (Lindén u. a., 2011). Das Programm und seine Quelldaten können von der Internetseite des Projekts[16] bezogen werden.

Die in Abschnitt 2.2 vorgestellten, freien Projekte *foma*, *OpenFst* und *SFST* vereint das HFST-Projekt in einem einzigen Paket. Die Autoren haben Schnittstellen zwischen die einzelnen Bestandteile eingefügt und so eine volle Unterstützung für die verschiedenen Formate geschaffen. Dadurch können Transduktoren kombiniert werden, die ursprünglich in unterschiedlichen Formaten vorliegen. Nähere technische Details zum HFST-Projekt finden sich bei Lindén u. a. (2011) und auf der Wikiseite[17] des Projekts. Außerdem haben die Autoren die Architektur von HFST so gestaltet, dass sie die nachträgliche Einbindung weiterer Bibliotheken für Transduktorentwicklung erlaubt. Eine Übersicht der unterstützten Formate gibt Tabelle 3.1. Mit dem HFST-Paket können formale Morphologien und andere automatenbasierte Anwendungen aus der Sprachverarbeitung bequem in einer einheitlichen Umgebung erstellt und verglichen werden. Neben AnIta, das im nächsten Abschnitt besprochen wird, haben seit der ersten Veröffentlichung von 2009 bereits eine Reihe von Projekten Morphologien auf Basis des HFST-Pakets erstellt, zum Beispiel OMorFi (Pirinen, 2008) für das Finnische.

[16] http://hfst.sourceforge.net
[17] https://kitwiki.csc.fi/twiki/bin/view/KitWiki/HfstHome

3.3. AnIta

AnIta wurde von Tamburini u. Melandri (2012) der Universität von Bologna entwickelt. Das Analyseprogramm für italienische Morphologie umfasst ca. 110.000 Lemmata des italienischen Wortschatzes. Unter den diversen morphologischen Prozessen des Italienischen, die mit AnIta untersucht werden können, findet man die Konjugation und Klitisierung von Verben, die Deklination, Derivation und Komposition von Nomen und Steigerungsformen von Adjektiven. Außerdem sind spezielle, ausspracheabhängige Wortveränderungen, sogenannte morphophonetische Phänomene wie Abschwächung von Konsonanten, Auslassungen und Verschmelzung als eigene Regeln formuliert. Zwei Komponenten bilden den AnIta genannten morphologischen Parser.

Das Lexikon

Das Lexikon einhält einzelne Worteinträge in einem Wurzellexikon und Sublexika, sogenannte Kontinuationsklassen (KKL). Im Wurzellexikon stehen die aus Grundform und Wortstamm bestehenden Lemmata. Diese enthalten außerdem grammatische Informationen wie die Wortklasse, das Genus oder ein besondere Ausspracheeigenschaften (phonetische Merkmale). Jedem Lemma wird eine Kontinuationsklasse zugewiesen. Diese Unterlexika bestimmen diejenigen Morpheme, die an den jeweilig zugeordneten Stamm angehängt werden. Daraus ergibt sich eine Baumstruktur, über die die Oberflächenformen der morphologischen Paradigmen hierarchisch aus der Grundform des Lemmas dargestellt werden können.
Zu Verben wird nach der Grundform und dem Stamm ein Verweis auf die entsprechende Konjugationsklasse hinzugefügt. Die Klasse enthält dann die jeweiligen Konjugationsmuster für die verschiedenen Flexionskategorien Modus, Tempus, Person und Numerus. Die regelmäßigen Beugungsformen der Verbendungen ARE, ERE und IRE stehen in den entsprechenden KKL *SufVerbAre*, *SufVerbEre* und *SufVerbIre*. Für Verben, die nur für bestimmte Kategorien unregelmäßige und ansonsten regelmäßige Formen annehmen, stehen die betroffenen Endungen in eigenen Lexika. Ebenso wird mit Verben verfahren, die für manche Kategorien den Stamm verändern, für andere aber nicht. Weiter gibt es Spezialklassen, die Pronomen und Artikel enthalten, die sich mit Verben zu klitisierten Wörtern zusammenfügen. Vollständig unregelmäßige Verben sind mit ihren Paradigmen direkt in das Wurzellexikon geschrieben. Nach diesem Kontinuationsklassen-Prinzip erkennt das Programm dann jene Wortformen, die aus den Lemmaeinträgen im Wurzellexikon generierbar sind.

Die Regeldatei

Die Regeldatei enthält die erlaubten Kombinationen der im Lexikon definierten Wortarten und beschreibt spezielle morphophonetische Prozesse mithilfe von kontextbasierten Regeln. Nach ihnen werden bei der Übersetzung in einen ET die möglichen Übergänge definiert. Außerdem werden die Ein-/Ausgabepaare deklariert, die jedem gelesenen Symbol ein auszugebendes Symbol zuordnen. Damit wird die Angabe von morphologischen Eigenschaften, wie die Wortart, der Modus oder das Tempus, je nach „Richtung" des ETs an- oder ausgeschaltet. Entweder werden die Eigenschaftssymbole für die Anzeige der Oberflächenform eines Wortes entfernt aus oder für die Analyse eines Wortes hinzugefügt. AnItas Lexikon und seine Regeldatei werden mit HFST zu einen ET im Binärformat zusammengefügt. Dieser ET fungiert ist ein Generator für ita-

lienische Wortformen, invertiert an ihn, erhält man einen Analysator, der für eine konkreten flektierten Wortform aus dem Lemma die Grundform inklusive der zugeordneten Symbole für morphosyntaktische Eigenschaften ermittelt.

4. Sizilianische Erweiterung (SiMoN)

Die Beschreibung des sizilianischen Dialektes mittels spezifischer Paradigmen stützt sich auf die Grammatik von Bonner (2001). Die dort beschriebene sizilianische Verbgrammatik wird in der Struktur von AnIta umgesetzt und in einen ET konvertiert. Der resultierende Generator und sein Gegenstück, der Analysator werden jeweils mit denen von AnIta kombiniert. Das Ergebnis ist ein Transduktorpaar, das gleichzeitig italienische und sizilianische Formen untersuchen und erzeugen kann. Im Folgenden erhält diese *si*zilianische *Mo*rphologie für *N*LP-Anwendungen die Bezeichnung *SiMoN.*

4.1. Lexikonaufbau

Die Verbparadigmen werden analog zu AnIta ihrer Endung entsprechend in die KKL gruppiert. Unregelmäßige Verben und Verben mit regelmäßigen als auch unregelmäßigen Konjugationsmustern (s. S. 16) sind gegenwärtig noch nicht berücksichtigt. Die Verben des Sizilianischen teilen sich, statt wie im Italienischen in drei, in nur zwei Typen auf, in Verben mit Endungen auf Ari und auf Iri. SiMoN enhält für die beiden regelmäßigen Verbtypen jeweils in die Lexika SufVerbAri und SufVerbIri.

Das erstellte Lexikon befindet sich in Ausschnitten im Anhang. Es liegt als eigenständige Datei vor und kann mit den entsprechenden HFST-Werkzeugen mit dem Lexikon AnItas zusammengeführt werden. Für diese Arbeit wurde die das Original der Regeldatei von AnIta für die Verwendung mit neueren HFST-Versionen und die sizilianischen Verbtypen angepasst. Hinweise zur Kopplung von AnIta und SiMoN finden sich im Anhang (A.1 auf Seite 28)

4.2. Dokumentierte Paradigmen

Der Fokus der zu untersuchenden Paradigmen liegt in dieser Arbeit auf den Konjugationsmustern regelmäßiger Verben. Das vorderste Ziel ist es hier, eine Grundlage für die Verbanalyse zu schaffen. Die regelmäßigen Verbparadigmen des sizilianischen zu dokumentieren ist problematischer als es zunächst den Anschein macht. Man muss sich des zu Beginn der Arbeit (s. Seite 2) erläuterten Variantenreichtums bewusst sein. Im Gegensatz zum Italienischen gibt es für einige Verben eine große Zahl an Wahlmöglichkeiten für Endungen konjugierter Formen, die regional unterschiedlich verbreitet und gleichermaßen gültig sind. Bonner (2001) dokumentiert für die regelmäßigen Verben einiger Zeiten und Modi alternative Formen. Diese Alternativformen gehören alle zum selben Paradigma. Daher gibt es im jeweiligen Lexikon der beiden Verbtypen in SiMoN teilweise mehrfache Einträge zur Konjugation der ersten, zweiten oder dritten Person. Eine vergleichende Analyse des gewonnenen Wikipedia-Korpus zeigt ebenfalls, dass die verschiedenen Varianten der Verben in der Praxis verwendet werden. Stammveränderungen (16) in der sizilianischen Verbgrammatik existieren ebenfalls, diese Fälle werden allerdings mit SiMoN im Moment noch nicht abgedeckt.

Für die erstellten Paradigmen wurden hauptsächlich die Verbtabellen von Bonner (2001) als Referenz verwendet und mit den Grammatikinformationen in der sizilianischen Wikipedia[18]

[18] http://scn.wikipedia.org/wiki/Wikipedia:Gramm%C3%A0tica

verglichen. Die vollständige regelmäßige Konjugation ist in SiMoN abgebildet, mit Ergänzungen aus der Wikipedia, falls dort weitere Formen zu finden waren.
In den Tabellen 4.1 auf der nächsten Seite bis 4.9 auf Seite 23 auf den folgenden Seiten sind die regelmäßigen Konjugationsformen am Beispiel der sizilianischen Verben parrari (reden) und battiri (schlagen) aufgeführt. Die einzelnen Tabellen geben jeweils die Formen beider Verbtypen in den Flexionskategorien Indikativ (Tab. 4.1 auf der nächsten Seite - 4.5 auf Seite 22), Imperativ und Subjunktiv (Tab. 4.6 und 4.7 auf Seite 22), sowie Konditional und Gerundium (Tab. 4.8 und 4.9 auf Seite 23). Der Teil *a* enthält jeweils Verben mit Endungen Ari in der Grundform, der Teil *b* die Formen für Verben mit Iri.

Zusätzlich zur diesen Flexionskategorien wurden die von Bonner (2001) in dessen Beispielen verwendeten regelmäßigen Verben als Lemmata zu SiMoN hinzugefügt und ihrer jeweiligen Endungsklasse zugeordnet. Die Paradigmen der unregelmäßigen Hilfsverben *essiri* (sein) und *aviri* (haben) sowie das sehr häufig verwendete *fari* (machen) wurden ebenfalls in die Liste der Lemmata aufgenommen, um Partizipkonstruktionen u.ä. zu erkennen. Diese grundlegende Basis für die Verberkennung stellt die angestrebte Erweiterung für AnIta dar.

Ein Vorteil, den diese Arbeit - dank Wikipedia - im Vergleich zu der von Tamburini u. Melandri genießt, ist die Erfassung von Flexionsparadigmen auf digitalem Weg. Diese müssen nicht von Hand aus einem Lexikon in ein digitales Format übertragen werden, sondern können nach entsprechender Verarbeitung über den Korpus gewonnen werden. Mittels eines für diese Arbeit erstellten, sehr einfachen Filterprogramms[19] wurde der Korpus automatisiert nach Verbformen mit regelmäßigen Endungen durchsucht. Dabei diente als Referenz eine aus dem Wiktionary extrahierte Liste von Verben. Insgesamt wurden 368 Lemmata auf diese Weise gesammelt. Bekannte unregelmäßige Verben wurden, falls nötig, entfernt. Dieses Verfahren ist noch ausbaufähig. Denn es bleibt ein großes Problem bei der Lemmaerstellung: Die Unterscheidung von regel- und unregelmäßigen Verben ist schwierig. Bei Bonner (2001) finden sich zwar abschnittsweise Listen mit eingeführten Verben, mit Informationen zur Konjugationsart, jedoch sind diese Listen nicht sehr umfangreich (insgesamt etwa 60 Verben). Auch anhand des Wikipediakorpus können viele Verbformen nicht eindeutig bestimmt werden, da nicht alle Konjugationsformen in den Artikeln vorkommen. Erweiterte, anspruchsvollere Filter zur Endungsüberprüfung sind möglicherweise ein Mittel, diese Kategorisierungsprobleme zu reduzieren.

[19] Pythonskript auf beigefügter CD enthalten(*paradigm-collector-0.2.py)*

Tabelle 4.1: Regelmäßige Konjugation im Indikativ: *Präsens*

(a) Verbendung Ari

	Konjugationsform	Variationen
Inf	**parr*ari***	
1s	parr*u*	-
2s	parr*i*	-
3s	parr*a*	-
1p	parr*amu*	-
2p	parr*ati*	-
3p	parr*anu*	*-unu*

(b) Verbendung Iri

	Konjugationsform	Variationen
Inf	**batt*iri***	
1s	batt*i*	-
2s	batt*i*	-
3s	batt*a*	-
1p	batte*mu*	-
2p	batt*iti*	-
3p	batt*inu*	*-unu*

Tabelle 4.2: Regelmäßige Konjugation im Indikativ: *Imperfekt*

(a) Verbendung Ari

	Konjugationsform	Variationen
Inf	**parr*ari***	
1s	parr*ava*	*-avu*
2s	parr*avi*	-
3s	parr*ava*	-
1p	parr*àvamu*	-
2p	parr*àvavu*	-
3p	parr*àvanu*	*-àvunu*

(b) Verbendung Iri

	Konjugationsform	Variationen				
Inf	**batt*iri***					
1s	batt*ia*	*-iu*	*-eva*	*-evu*	*-iva*	*-ivu*
2s	batt*ivi*	-evi	-	-	-	-
3s	batt*ia*	*-eva*	-iva			
1p	batt*a*	*-evamu*	*ivamu*			
2p	batt*iavu*	*-evavu*	*-ivavu*			
3p	batt*ianu*	*-evanu*	*-ivanu*			

Legende:
1s, 1p stehen jew. für *1. Person Singular* und *Plural* (2s, 2p, usw. analog),
in der Zeile Inf steht die Verbgrundform

Tabelle 4.3: Regelmäßige Konjugation im Indikativ: *Partizip Perfekt*

(a) Verbendung Ari

	Hilsverb	**Partizip**
Inf	aviri +	**parr*ari***
1s	*aiu*	
2s	*ai*	
3s	av*i*	parr*atu*
1p	av*emu*	
2p	av*iti*	
3p	*annu*	

(b) Verbendung Iri

	Hilsverb	**Partizip**
Inf	aviri +	**batt*iri***
1s	*aiu*	
2s	*ai*	
3s	av*i*	batt*utu*
1p	av*emu*	
2p	av*iti*	
3p	*annu*	

Legende: siehe Tabelle 4.1
Anmerkung: Das (unregelm.) Hilfsverb für das Partizip des Perfekts ist *aviri* (haben)

Tabelle 4.4: Regelmäßige Konjugation im Indikativ: *Präteritum*

(a) Verbendung Ari

	Konjugationsform	**Variationen**		
Inf	**parr*ari***			
1s	parr*ai*	*-avi*	*-aiu*	*-avu*
2s	parr*ast*	-	-	-
3s	parr*au*	*-ò*	-	-
1p	parr*ammu*	*-amu*	-	-
2p	parr*astivu*	*-astu*	-	-
3p	parr*arunu*	*-aru*	-	-

(b) Verbendung Iri

	Konjugationsform	**Variationen**		
Inf	**batt*iri***			
1s	batt*ivi*	*-ìi*	*-iu*	*-ivu*
2s	batt*isti*	-	-	-
3s	batt*ìu*	-	-	-
1p	batt*emmu*	*-emu*	-	-
2p	batt*istivu*	*-astu*	*-istu*	-
3p	batt*erunu*	*-eru*	-	-

Legende:
1s, 1p stehen jew. für *1. Person Singular* und *Plural* (2s, 2p, usw. analog),
in der Zeile Inf steht die Verbgrundform

Tabelle 4.5: Regelmäßige Konjugation im Indikativ: *Futur*

(a) Verbendung Ari

	Konjugationsform
Inf	**parr*ari***
1s	parr*irò*
2s	parr*irai*
3s	parr*irà*
1p	parr*iremu*
2p	parr*iriti*
3p	parr*irannu*

(b) Verbendung Iri

	Konjugationsform
Inf	**batt*iri***
1s	batt*irò*
2s	batt*irai*
3s	batt*irà*
1p	batt*iremu*
2p	batt*iriti*
3p	batt*irannu*

Legende: siehe Tabelle 4.1

Tabelle 4.6: Regelmäßige Konjugation: *Imperativ*

(a) Verbendung Ari

	Konjugationsform
Inf	**parr*ari***
2s	parr*a*
3s	parr*assi*
1p	parr*amu*
2p	parr*ati*
3p	parr*assiru*

(b) Verbendung Iri

	Konjugationsform
Inf	**batt*iri***
2s	batt*i*
3s	batt*issi*
1p	batt*emu*
2p	batt*iti*
3p	batt*issiru*

Anmerkung: Der Imperativ im Sizilianischen wird nur für den Präsens verwendet

Tabelle 4.7: Regelmäßige Konjugation im Subjunktiv: *Imperfekt*

(a) Verbendung Ari

	Konjugationsform	**Variationen**
Ind	**parr*ari***	
1s	parr*assi*	-
2s	parr*assi*	-
3s	parr*assi*	-
1p	parr*assimu*	-
2p	parr*assivu*	-
3p	parr*assiru*	-*assinu*

(b) Verbendung Iri

	Konjugationsform	**Variationen**
Inf	**batt*iri***	
1s	batt*issi*	-
2s	batt*issi*	-
3s	batt*issi*	-
1p	batt*issimu*	-
2p	batt*issivu*	-
3p	batt*issiru*	-*issinu*

Anmerkung: Die *Präsenskonjugation* im Subjunktiv ist im Sizilianischen nicht mehr in Gebrauch, stattdessen wird das Präsens des Indikativ verwendet.

Tabelle 4.8: Regelmäßige Konjugation im Konditional: *Präsens*

(a) Verbendung Ari

	Konjugationsform
Inf	**parr*ari***
1s	parr*iria*
2s	parr*airissi*
3s	parr*airia*
1p	parr*iriamu*
2p	parr*iriavu*
3p	parr*irianu*

(b) Verbendung Iri

	Konjugationsform
Inf	**batt*iri***
1s	batt*iria*
2s	batt*irissi*
3s	batt*iria*
1p	batt*iriamu*
2p	batt*iriavu*
3p	batt*irianu*

Tabelle 4.9: Regelmäßige Konjugation der Verlaufsform (*Gerundium*)

(a) Verbendung Ari

	Hilfsverb	**Gerundform**
Inf	stari +	**parr*ari***
1s	sta*iu*	parr*ennu*
2s	*stai*	
3s	st*a*	
1p	st*amu*	
2p	st*ati*	
3p	st*annu*	

(b) Verbendung Iri

	Hilfsverb	**Gerundform**
Inf	stari +	**batt*iri***
1s	sta*iu*	batt*ennu*
2s	*stai*	
3s	st*a*	
1p	st*amu*	
2p	st*ati*	
3p	st*annu*	

Anmerkung: Das (unregelm.) Hilfsverb für das Partizip des Perfekts ist *stari* (sein)

Legende:
1s, 1p stehen jew. für *1. Person Singular* und *Plural* (2s, 2p, usw. analog),
in der Zeile Inf steht die Verbgrundform

5. Bewertung von SiMoN

SiMoN kann mit HFST zum einen als Generator, zum anderen als Analysator ausgeführt werden. Der Generator erzeugt insgesamt rund 24.700 Verbformen aus den gegebenen Lemmata und den zugehörigen Kontinuationsklassen. Für den Analysator wurde aus dem Wikipedia-Korpus eine Testliste von 180 zufällig ausgewählter Verbformen mit regelmäßigen Endungen zur Bewertung des Erfassungsvermögens von SiMoN erstellt[20]. Von diesen Verbformen wurden 54 als Teil eines Flexionsparadigmas erkannt. Das entspricht einer Genauigkeit von 30% bzw. einer Fehlerrate (*WER* , vgl. Tamburini u. Melandri, 2012) von 70%.

Eine derartige Evaluation ist natürlich sehr rudimentär. Sie ermöglicht aber zumindest eine erste Einschätzung der bisherigen Abdeckung, die SiMoN erreicht. Standardisierte Werte zu Genauigkeit (Precision) und Trefferquote (Recall) können zu diesem Zeitpunkt noch nicht gebildet werden. Tamburini u. Melandri nutzen für die Evaluation von AnIta u.a. Referenzwerte eines Taggers für italienische Wortfrequenzen und vergleichen ihre Ergebnisse mit anderen Morphologieimplementationen des Italienischen. Für das Sizilianische sind vergleichbare Daten noch nicht verfügbar. Denkbar wäre für zukünftige Messungen eine einfache Variante, Verzerrungen zu messen, die als Precision und Recall angegeben werden können: Das HFST-Paket bietet die Möglichkeit, auf Basis der vorhandenen Regeln und Lemmata neue Wörter durch Mutmaßung einer Wortkategorie zuzuweisen. Dieses Verfahren wurde in Tabelle 2.1 auf Seite 7 bereits angedeutet (Beispiel Wörter „erraten"). Damit könnte die irrtümliche Zuordnung eines Wortes in die falsche Klasse protokolliert und für die tatsächlich korrekt erkannten Wörter (Precision) im Verhältnis zu fälschlich nicht erkannten und fälschlich erkannten Wörtern festgehalten werden (Precision). Mit diesen Werten sollten bessere Referenzgrößen zur Berechnung von Precision und Recall geliefert werden. Im Moment ist diese Möglichkeit durch die fehlende Implementierung weiterer Wortklassen noch nicht gegeben.

In Anbetracht des noch relativ kleinen Lexikons mit nur ca. 300 Lemmata ist dieses Ergebnis nicht überraschend. Durch die verwendete Wörterbuchstruktur wird ein Wort nur dann erkannt, wenn es in den Lemmata enthalten ist. Dementsprechend verbessert eine Erweiterung des Lexikons, vor allem um unregelmäßige Verben, am ehesten die Ergebnisse.

[20] Die Liste kann auf der CD im Anhang eingesehen werden

6. Potentiale für Parsing in der Dialektologie

Die Betrachtungen zu den sprachwissenschaftlichen Gegebenheiten aus den ersten Abschnitten und der Ergebnisse zu SiMoN zeigen: Maschinelle Analyseprogramme für Dialektmorphologie sind technisch anspruchsvoll und stellen viele Herausforderungen, zuallererst an die Sprachdokumentation selbst. Dank den verfügbaren Programmen und Bibliotheken kann die eigentliche Beschreibung der Morphologie jedoch vergleichsweise einfach durchgeführt werden, sobald der erforderliche Dokumentationsgrad des Zieldialekts gegeben ist.

Trotz der niedrigen Erkennungsrate von SiMoN ist das Ergebnis in der Gesamtheit durchaus befriedigend. Die sizilianischen Wikipedia und die Wiktionaryseite konnten, wenn auch noch nicht auf optimale Weise, einmal als Quelle für das Lexikon und einmal zu dessen Überprüfung genutzt werden[21]. Eine weitere Aufbereitung der erstellten Korpora und eine Verfeinerung der automatischen Paradigmensuche dürften die Erweiterung des Lexikons und damit eine Verbesserung der Leistung erheblich erleichtern. Das gesetzte Ziel, einen Grundstock an sizilianischen Paradigmen in AnIta einzupflegen wurde erreicht. AnIta „spricht" nun ein wenig sizilianisch.

6.1. Anregungen zu Verbesserung und Ergänzung

Wie bereits ausgeführt, ist die bisher erarbeitete Morphologie des sizilianischen Dialekts nur ein erster Grundstein. Bis zur Vervollständigung zu einem robusten Parser ist es noch ein weiter Weg. Die nachstehenden Vorschläge skizzieren einige Ergänzungen, die nach Begutachtung der Ergebnisse vorrangig sind:

1. Vervollständigung der Verbparadigmen um Sonderklassen mit Stammveränderungen und unregelmäßiger Konjugation
2. Aufnahme weiterer Wortklassen, vor allem von Pronomina und Substantiven sowie entsprechender Flexionsparadigmen (Deklination, Komparation, etc.)
3. Abbildung von Derivations- und Konversionsprozessen
4. Vervollständigung von Regeln und Lexika zu Auslassungen bei bestimmten Artikel-Wort-Paaren, Wortverkürzungen und Verschmelzung von Pronomen und Artikeln (Klitisierung, Elimination sowie Assimilation) als auch Konsonantenverdopplung (Gemination)

Zusätzlich zu diesen Punkten wäre eine Prüfung sinnvoll, inwieweit sich die italienischen Morphologieregeln AnItas, die für SiMoN nahezu unverändert eingesetzt werden, als Vorlage für die sizilianische Morphologie eignet und wo sizilianische Spezialfälle Formulierung eigener Regeln erfordern.

6.2. Ausblick

Die Bedeutung des in dieser Arbeit behandelten Themas, der computergestützten Morphologieanalyse von Dialekten, wird mit der weiteren Entwicklung der besprochenen Materialquellen wie

[21] Stand der Datensätze aus Wikipedia: 30. April 2014, Datensätze aus dem Wiktionary: 9. Mai 2014

Wikipedia und digitaler Atlanten mehr Gewichtung bekommen. Für die klassische Dialektforschung wie zum Beispiel die räumlich-sprachliche Differenzanalyse dürften der Einsatz und die Weiterentwicklung von Parsinganwendungen dann nach und nach attraktiver werden.

Natürlich bieten die Parser keine Generallösung, wie sie das Beispiel mit der Sprachsteuerung eines Smartphones von Seite 3 suggeriert. Es wird immer Dialekte oder Teile davon geben, die auch mit den Mitteln des transduktorgestützten morphologischen Parsings nicht umfassend beschrieben werden können. Das zeigt sich vor allem am großen Vorkabularbedarf. Dennoch liegt in der Zusammenführung von Dialektologie und NLP großes Potenzial, das die Bemühungen der beiden Disziplinen näher an ihre jeweiligen Ziele bringen kann.

A. Anleitungen & Code-Quellen

A.1. Verwendung von AnIta und SiMoN

Für die Verwendung von AnIta und seiner Erweiterung ist das HFST-Paket notwendig. Es kann von der Projektseite[1] für die Plattformen MAC/UNIX und Windows heruntergeladen und entsprechend den enthaltenen Anweisungen installiert werden. Die in dieser Arbeit eingesetzte Version ist *hfst 3.7.0*[2].

A.1.1. Kompilierung

Die folgenden Befehle übersetzten die jeweiligen Quelldateien von AnIta und SiMoN in Binärcode und kombinieren sie zu einem Analysetransduktor, der Italienisch und Sizilianisch parallel erkennt.

Lexika

```
$ hfst-lexc -v -o lemmata-it.hfst italiano.lexc
$ hfst-lexc -v -o lemmata-scn.hfst siciliano.lexc
```

Ruft den LexC-Kompilierer mit erweiterten Ausgabeinformationen auf. Die Ausgabedatei[3] wird mit -o festgelegt.

Regelsätze

```
$ hfst-twolc -v -o regeln-it.hfst italiano.twolc
$ hfst-twolc -v -o regeln-scn.hfst siciliano.twolc
```

Gleicher Aufruf für Twol-Kompilierer.

Generatoren

```
$ hfst-compose-intersect -v -o generator-it.hfst lemmata-it.hfst regeln-it.hfst
$ hfst-compose-intersect -v -o generator-scn.hfst lemmata-scn.hfst regeln-scn.hfst
```

Bildet die Schnittmenge der Lexion- und Regelsatz-Transduktoren. Das Ergebnis ist ein Vollformengenerator.

Analysatoren

```
$ hfst-invert -i generator-it.hfst -o analysator-it.hfst
$ hfst-invert -i generator-scn.hfst -o analysator-scn.hfst
```

Invertiert den Generator, der resultierende Transduktor kann Wortformen einlesen.

[1] http://hfst.sourceforge.net/

[2] Ein vorkompiliertes Paket für debianbasierte Systeme mit 64bit-Architektur ist auf dem beigefügten Datenträger enthalten.

[3] Die Namen der Ausgaben in den hier aufgeführten Befehlen sind Beispielnamen, sie können nach belieben verändert werden.

Kombination zu SiMoN Für die Paralellanalyse von Sizilianisch und Italienisch müssen die oben erzeugten Transduktoren kombiniert werden:

```
$ hfst-union -v -o SiMoN-analysator.hfst -1  generator-it.hfst[4]
-2 generator-scn.hfst
```

A.1.2. Anwendung

Worterzeugung

```
$ $hfst-fst2strings -n 5 SiMoN-generator.hfst
l_aviri+V_FIN+IND+PRES+2+SING:ai
l_aviri+V_FIN+SUBJ+PRES+2+SING:ai
l_aviri+V_FIN+IMP+PRES+2+SING:ai
l_aviri+V_FIN+SUBJ+PRES+1+SING:aia
l_aviri+V_FIN+SUBJ+PRES+3+SING:aia
```

Eine Liste aller möglichen Wörter kann mit dem Befehl *hfst-fst2strings* erzeugt werden.

Wortanalyse Die erzeugten Transduktoren werden für die Analyse mit dem Nachschlagewerkzeug von HFST, *hfst-lookup*, aufgerufen:

```
$ hfst-lookup SiMoN-analysator.hfst
> non
non l_non+ADV
> dormivo
dormivo l_dormire+V_FIN+IND+IMPERF+1+SING
> durmivu
durmivu l_durmiri+V_FIN+IND+IMPERF+1+SING
durmivu l_durmiri+V_FIN+IND+PAST+1+SING
```

Testliste Für eine rudimentäre Auswertung wurde der erzeugte Analysator auf die folgende Liste von Verbfomen angewendet.

A.2. SiMoN: Quellcode

A.2.1. Lexikon

Der nachstehende Quellcode enthält Aufgrund des großen Umfangs (über 800 Zeilen) nur den Anfang der Lexikondatei und Auschnitte aus den Lemmaeinträgen für zwei der drei unregelmäßigen Verben. Die vollständigen regelmäßigen Konjugationsmuster sind in Abschnitt 4.2 für die jeweilige Flexionskategorie aufgeführt. Die automatisiert erstellten Lemmata können in der Quelldatei auf der CD eingesehen werden.

[4]Der Zeilenumbruch dient der Formatierung und wird nicht eingegeben

```
1  Multichar_Symbols
2  +V_ARI +V_IRI +V_FIN +V_NOFIN +V_PP
3  +SING +PLUR
4  +INF +IND +SUBJ +COND +IMP +PART +GER
5  +PRES +IMPERF +PAST +FUT
6  +1 +2 +3
7  +GLI +VEL_A +VEL_Y
8  +C_CE +C_CI +C_GLI +C_LA +C_LE +C_LI +C_LO +C_ME +C_MI +C_NE +C_SE +
       C_SI +C_TE +C_TI +C_VE +C_VI
9  LEXICON Root
10 !- Note: some entries have been duplicated without
11 !- diacritics in order to be able to recognize 'misspelled'
12 !- words missing them
13 !- Multiple entries for 1st, 2nd or 3rd person (all tenses/modes)
14 !- record local variations in used forms (palermo,catania,etc.) for
       the same conjugation pattern
15 !------- regular verbs ------------
16 l_abbruciari:abbruci+V_ARI SufVerbAri ;
17 l_durmiri:durm+V_IRI SufVerbIri ;
18 l_finiri:fin+V_IRI SufVerbIri ;
19 l_parrari:parr+V_ARI SufVerbAri ;
20 l_pàrtiri:part+V_IRI SufVerbIri ;
21 l_partiri:part+V_IRI SufVerbIri ;
22 l_parlari:parl+V_ARI SufVerbAri ;
23 l_purtari:purt+V_ARI SufVerbAri ;
24 l_rispunniri:rispun+V_IRI SufVerbIri ;
25 !------- assumed regular verbs -----
26 !- CAUTION! REVISIION NEEDED.
27 !- These forms have been autogenerated from the sicilian Wikipedia,
28 !- guessing their membership of category by the surface forms they
       appear in the Wiki
29 !- Irregular forms and forms of changing stems might be contained as
       well
30 !-## conjugation paradigm for essiri
31 l_essiri+V_IRI: SiriEssiri ;
32 l_siri+V_IRI: SiriEssiri ;
33 !-- Note: since siri is a variation of essiri, the conjugation
       patterns
34 !-- have been moved to an extra lexicon in order to avoid cloning the
        whole paradigm.
35 §
36 !-## conjugation paradigm for aviri
37 !--## indicative
```

```
38 l_aviri+V_FIN+IND+PRES+1+SING:haiu #;
39 l_aviri+V_FIN+IND+PRES+1+SING:aiu #;
40 l_aviri+V_FIN+IND+PRES+2+SING:hai #;
41 l_aviri+V_FIN+IND+PRES+2+SING:ai #;
42 l_aviri+V_FIN+IND+PRES+2+SING:a' #;
43 l_aviri+V_FIN+IND+PRES+3+SING:havi #;
44 l_aviri+V_FIN+IND+PRES+3+SING:avi #;
45 l_aviri+V_FIN+IND+PRES+1+PLUR:havemu #;
46 l_aviri+V_FIN+IND+PRES+1+PLUR:avemu #;
47 l_aviri+V_FIN+IND+PRES+1+PLUR:avimu #;
48 l_aviri+V_FIN+IND+PRES+2+PLUR:haviti #;
49 l_aviri+V_FIN+IND+PRES+2+PLUR:aviti #;
50 l_aviri+V_FIN+IND+PRES+3+PLUR:hannu # ;
51 l_aviri+V_FIN+IND+PRES+3+PLUR:annu #;
52
53 ...
54
55 !-## conjugations for fari
56 !--## indicative
57 '_fari+V_FIN+IND+PRES+1+SING:fazzu #;
58 l_fari+V_FIN+IND+PRES+2+SING:fai #;
59 l_fari+V_FIN+IND+PRES+2+SING:fa' #;
60 l_fari+V_FIN+IND+PRES+3+SING:fa #;
61 l_fari+V_FIN+IND+PRES+3+SING:faci #;
62 l_fari+V_FIN+IND+PRES+1+PLUR:facemu #;
63 l_fari+V_FIN+IND+PRES+2+PLUR:faciti #;
64 l_fari+V_FIN+IND+PRES+3+PLUR:fannu #;
65 l_fari+V_FIN+IND+IMPERF+1+SING:facia #;
66 l_fari+V_FIN+IND+IMPERF+1+SING:faceva #;
67 l_fari+V_FIN+IND+IMPERF+1+SING:faciva #;
68 l_fari+V_FIN+IND+IMPERF+2+SING:facivi #;
69 l_fari+V_FIN+IND+IMPERF+2+SING:facevi #;
70 l_fari+V_FIN+IND+IMPERF+3+SING:facia #;
71 l_fari+V_FIN+IND+IMPERF+3+SING:faceva #;
72 ...
```

A.2.2. Regelsätze

Die Regeldatei von AnIta wurde nur geringfügig verändert, die Symbole für die drei Verbendungen wurden mit denen der zwei Sizilianischen ersetzt. Untenstehendes Listing enthält nur die Veränderten Zeilen.

```
Alphabet
a à â b c d e è é f g h i ì j k l m n o ò p q r s t u ù v w x y z
...
%+V%_ARI:0 %+V%_IRI:0 %+V%_FIN:0 %+V%_NOFIN:0 %+V%_PP:0
...
Vowel = a â à e è é i ì o ò u ù ;
Definitions
Head = %+NN: | %+NN%_P: | %+V%_ARI: | %+V%_IRI: | %+V%_FIN: | %+V%
    _NOFIN: | %+V%_PP: | %+ADJ: | %+ADJ%_DIM: | %+ADJ%_IND: | %+ADJ%
    _IES: | %+ADJ%_NUM: | %+ADJ%_POS: | %+ADV: | %+PRON: | %+PRON%_P:
    | %+PRON%_DIM: | %+PRON%_IND: | %+PRON%_IES: | %+PRON%_REL: | %+
    PRON%_POS: | %+PRON%_PER: | %+INT: | %+PREP: | %+PREP%_A: | %+CONJ
    %_C: | %+CONJ%_S: | %+ART: ;
...
Verbi = %+V%_ARI: | %+V%_IRI: | %+V%_FIN: | %+V%_NOFIN: | %+V%_PP: ;
Coniug = %+V%_ARI: | %+V%_IRI: ;
...
```

Literatur

[Allauzen u. a. 2007] Allauzen, Cyril ; Riley, Michael ; Schalkwyk, Johan ; Skut, Wojciech ; Mohri, Mehryar: OpenFst: A General and Efficient Weighted Finite-State Transducer Library. Version: 2007. http://dx.doi.org/10.1007/978-3-540-76336-9_3. In: Holub, Jan (Hrsg.) ; Žďárek, Jan (Hrsg.): *Implementation and Application of Automata* Bd. 4783. Springer Berlin Heidelberg, 2007. – DOI 10.1007/978–3–540–76336–9_3. – ISBN 978–3–540–76335–2, S. 11–23 2.2

[Beesley u. Karttunen 2003] Beesley, Kenneth R. ; Karttunen, Lauri: *Finite State Morphology*. 2003 http://www.fsmbook.com 2.2

[Benincà 1996] Benincà, Paola: *Piccola storia ragionata della dialettologia italiana*. [2. ed.]. Padova : Unipress, 1996. – ISBN 88–8098–083–1 2.3

[Bonner 2001] Bonner, J. K. ; Cipolla, Gaetano (Hrsg.): *Introduction to Sicilian grammar*. Brooklyn, NY : Legas, 2001. – xiv, 225 p. : ill., map : 25 cm + 1 CD (12 cm), 1 Beil. (23 S.). – ISBN 1–881901–25–4 1, 2.4, 4, 4.2, 4.2

[Çağrı Çöltekin 2010] Çağri Çöltekin: A Freely Available Morphological Analyzer for Turkish. In: Nicoletta Calzolari (Hrsg.) ; Khalid Choukri (Hrsg.) ; Bente Maegaard (Hrsg.) ; Joseph Mariani (Hrsg.) ; Jan Odijk (Hrsg.) ; Stelios Piperidis (Hrsg.) ; Mike Rosner (Hrsg.) ; Daniel Tapias (Hrsg.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta and Malta : European Language Resources Association (ELRA), 2010. – ISBN 2–9517408–6–7 2.2

[Cùnsolo 1977] Cùnsolo, Felice (Hrsg.): *Proverbi siciliani commentati*. Palermo : Vespro, 1977 1 b

[Gilliéron u. Edmont 1902] Gilliéron, Jules ; Edmont, Edmond: *Atlas linguistique de la France*. Paris : Champion, 1902-1906 2.3

[Glück 2010] Glück, Helmut (Hrsg.): *Metzler-Lexikon Sprache*. 4., aktualisierte und überarb. Aufl. Stuttgart [u.a.] : Metzler, 2010. – XXXIV, 814 S. : Ill., graph. Darst., Kt.. – ISBN 978–3–476–02335–3 2.1, 8, 3.1.2

[Hinderling u. König 2005] Hinderling, Robert (Hrsg.) ; König, Werner (Hrsg.): *Bayerischer Sprachatlas*. Heidelberg : Winter, 2005-2010 2.3

[Hulden 2009] Hulden, Mans: Foma: a Finite-State Compiler and Library. In: *Proceedings of the Demonstrations Session at EACL 2009*. Athens, Greece : Association for Computational Linguistics, April 2009, 29–32 2.2

[Jaberg u. Jud 1928] Jaberg, Karl ; Jud, Jakob: *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen (Schweiz) : H. Champion, 1928-56 2.3

[Jurafsky u. Martin 2009] Jurafsky, Dan ; Martin, James H.: *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 2. ed., internat. ed. Upper Saddle River : Pearson Education International, Prentice Hall, 2009 (Prentice-Hall-series in artificial intelligence). – ISBN 0–13–504196–1 2.1, 2.2, 2.2, 3.1.2

[Lindén u. a. 2011] Lindén, Krister ; Axelson, Erik ; Hardwick, Sam ; Pirinen, TommiA ; Silfverberg, Miikka: HFST—Framework for Compiling and Applying Morphologies. Version: 2011. http://dx.doi.org/10.1007/978-3-642-23138-4_5. In: Mahlow, Cerstin (Hrsg.) ; Piotrowski, Michael (Hrsg.): *Systems and Frameworks for Computational Morphology* Bd. 100. Springer Berlin Heidelberg, 2011. – DOI 10.1007/978–3–642–23138–4_5. – ISBN 978–3–642–23137–7, 67–85 3.2, 3.2

[Lindén u. a. 2009] Lindén, Krister ; Silfverberg, Miikka ; Pirinen, Tommi: HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. Version: 2009. http://dx.doi.org/10.1007/978-3-642-04131-0_3. In: Cerstin Mahlow (Hrsg.) ; Michael Piotrowski (Hrsg.): *State of the Art in Computational Morphology* Bd. 41. Springer Berlin Heidelberg, 2009. – DOI 10.1007/978–3–642–04131–0_3. – ISBN 978–3–642–04130–3, 28–47 3.2

[Löffler 2003] Löffler, Heinrich: *Dialektologie: Eine Einführung.* Tübingen : Narr, 2003 (Narr-Studienbücher). – ISBN 3–8233–4998–8 2.3

[Niebaum u. Macha 2006] Niebaum, Hermann ; Macha, Jürgen: *Germanistische Arbeitshefte.* Bd. 37: *Einführung in die Dialektologie des Deutschen*. 2. Tübingen : Niemeyer, 2006. – ISBN 3–484–26037–8 2.3

[Novák 2006] Novák, Attila: Morphological Tools for Six Small Uralic Languages. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006 2.4

[Pfister u. Kaufmann 2008] Pfister, Beat ; Kaufmann, Tobias: *Sprachverarbeitung: Grundlagen und Methoden der Sprachsynthese und Spracherkennung.* Berlin [u.a.] : Springer, 2008 (Springer-Lehrbuch). – ISBN 978–3–540–75909–6 2.2, 3.1.2

[Pirinen 2008] Pirinen, Tommi: Open Source Morphology for Finnish using Finite-State Methods / Technical Report. Department of Linguistics, University of Helsinki. 2008. – Forschungsbericht 2.2, 3.2

[Schmid 2006] Schmid, Helmut: A Programming Language for Finite State Transducers. Version: 2006. http://dx.doi.org/10.1007/11780885_38. In: Yli-Jyrä, Anssi (Hrsg.) ; Karttunen, Lauri (Hrsg.) ; Karhumäki, Juhani (Hrsg.): *Finite-State Methods and Natural Language Processing* Bd. 4002. Springer Berlin Heidelberg, 2006. – DOI 10.1007/11780885_38. – ISBN 978–3–540–35469–7, 308–309 2.2

[Schmid u. a. 2004] Schmid, Helmut ; Fitschen, Arne ; Heid, Ulrich: SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In: *LREC*, European Language Resources Association, 2004 2.2

[Tamburini u. Melandri 2012] Tamburini, Fabio ; Melandri, Matias: AnIta: a powerful morphological analyser for Italian. In: Calzolari, Nicoletta (Hrsg.) ; Choukri, Khalid (Hrsg.) ; Declerck, Thierry (Hrsg.) ; Uğur Doğan, Mehmet (Hrsg.) ; Maegaard, Bente (Hrsg.) ; Mariani. Joseph (Hrsg.) ; Odijk, Jan (Hrsg.) ; Piperidis, Stelios (Hrsg.): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul and Turkey : European Language Resources Association (ELRA), 2012. – ISBN 978–2–9517408–7–7 1, 2.2, 3.3, 4.2, 5

[Wenker u. Wrede 1926] Wenker, Georg (Hrsg.) ; Wrede, Ferdinand (Hrsg.): *Deutscher Sprachatlas : auf Grund des Sprachatlas des Deutschen Reichs*. Marburg (Lahn) : Elwert, 1926/27 2.3