Gegenwärtige dialektspezifische Daten und deren Anwendung in der Dialektometrie

Zhekova, Desislava
desi@cis.uni-muenchen.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Krefeld, Thomas
thomas.krefeld@lmu.de
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Herteis, Simeon
simeon.herteis@gmail.com
Centrum für Informations- und Sprachverarbeitung (CIS), LMU, München

Einleitung

Die Datenverarbeitung innerhalb der Geisteswissenschaften ist sehr eng mit den gegenwärtigen technologischen Entwicklungen verbunden und dementsprechend auch stark davon abhängig. Ein sehr gutes Beispiel dafür ist das Gebiet der Dialektologie / Dialektometrie. Klassische Dialektometrie ist eine Forschungsrichtung innerhalb der Linguistik, die sich mit der Erforschung möglichst hochrangiger Ordnungsstrukturen in sprachgeographischen Netzen beschäftigt. Diese Aufgabe wurde bislang hauptsächlich durch die Analyse gesprochener Sprache (z. B. akustische Aufnahmen) oder der sogenannten Fragebögen (z. B. gezielt abgefragte, schriftliche Daten) bewältigt. Ein Nachteil dieser ist allerdings, dass die erhobenen Daten stark beeinflusst oder nicht schriftlich sind. Durch die gegenwärtigen Entwicklungen in der Informationstechnologie sind Sammlungen von neuartigen Dialektdaten erreichbar (die ohne äußeren Einfluss, gesammelt wurden und darüber hinaus in schriftlicher Form als Datensatz vorhanden sind), womit in der Dialektometrie neue Wege gegangen werden können. Ein Beispiel dafür sind neue Medien, wie z. B. Wikipedia, Twitter, digitale Zeitschriften, etc., in denen außerdem Veränderungen in der Gesellschaft schnell abgebildet werden.

Allein in Wikipedia ist eine große Anzahl an Dialekten vertreten, wie zum Beispiel die italienischen Dialekte Lombardisch (31.986 Artikel) , Sizilianisch (25.273 Artikel), Neapolitanisch (14.346 Artikel) etc., die fortlaufend mit neuen Artikeln erweitert werden, die nicht nur von einem, sondern von mehreren Autoren editiert werden. Aus diesen Artikeln kann eine bisher nicht vorhandene Art Korpus erstellt werden, dessen Untersuchung die Beantwortung völlig neuer Fragestellungen möglich werden lässt.

Die Größe dieser neuen Korpora ermöglicht nicht nur neuartige Fragestellungen in der Dialektometrie, sondern auch einen zeitgenössischen und automatisierten Vergleich für die Analyse von Dialekten und ihren linguistischen Eigenschaften (basiert auf statistische Ansätze). Für solche Verfahren ist allerdings nicht nur die vorhandene Datenmenge wichtig, sondern auch die leichte Erreichbarkeit von qualitativen Annotationen und Analysetools. Diese wurden bislang hauptsächlich für die Standardsprachen entwickelt, für Dialekte existieren diese bis jetzt nur in wenigen Ausnahmefällen.

Ein solches Analysetool für die Standardsprache Italienisch ist AnIta (Tamburini / Melandri 2012), ein morphologisches Finite-State-Analysetool, welches bisher nur für das Italienische verwendet werden kann. In AnIta können aber auch viele empirische Belege für Dialekte integriert werden, sodass die maschinelle Bearbeitung vieler italienischer Dialekte möglich wird. Die neuen Dialektwikipedias ermöglichen auch einen halb automatisierten Ansatz dafür.

SiMoN

Überblick

In unserer Softwaredemonstration möchten wir eine vorläufige Erweiterung von AnIta vorstellen, die mit vielen regelmäßigen Verbparadigmen des sizilianischen Dialekts erweitert wurde - SiMoN (Sizilianische Morphologie für NLP-Anwendungen). Die Version der Softwaredemonstration ist schon online erreichbar. Aus Einträgen der sizilianischen Wikipedia wurden Verblemmata (368 sizilianische Lemmata) für das Lexikon von AnIta automatisch extrahiert anhand von dem Auftreten regulären sizilianischen Verbendungen und einer Liste von Verben im Italienischen. Da sich die Verben des Sizilianischen in nur zwei Typen aufteilen (statt wie im Italienischen in drei), sind nur Verbeinträge mit Endungen auf *-ari* und auf *-iri* vorhanden. Die gesamte Zahl, der durch Flexionsparadigmen erfassten Verbformen beläuft sich auf ca. 24.700. Damit bietet SiMoN einen ersten Grundstock für die Entwicklung einer computergestützten, sizilianischen Morphologie.

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 für Sizilianisch zu schaffen. 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 und Cipolla (2001) dokumentieren für die regelmäßigen Verben einiger Zeiten und Modi alternative Formen, die wir verfolgen. 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 vorläufige Analyse des gewonnenen Wikipedia-Korpus zeigte ebenfalls, dass die verschiedenen Varianten der Verben in der Praxis verwendet werden. Stammveränderungen in der sizilianischen Verbgrammatik existieren ebenfalls, diese Fälle werden allerdings mit SiMoN im Moment noch nicht abgedeckt.

	Indikativ: Präsens				Indikativ: Imperfekt			
	parrari		*battiri*		*parrari*		*battiri*	
	Konj.	Var.	Konj.	Var.	Konj.	Var.	Konj.	Var.
1s	parru	-	batti	-	parrava	-avu	battia	-iu -eva -evu -iva -ivu
2s	parri	-	batti	-	parravi	-	battivi	-evi
3s	parra	-	batta	-	parrava	-	battia	-eva -iva
1p	parramu	-	battemu	-	parràvamu	-	batta	-evamu ivamu
2p	parrati	-	battiti	-	parràvavu	-	battiavu	-evavu -ivavu
3p	parranu	-unu	battinu	-unu	parràvanu	-àvunu	battianu	-evanu -ivanu

	Indikativ: Partizip Perfekt				Indikativ: Präteritum				Imperativ	
	aviri + parrari		*aviri + battiri*		*parrari*		*battiri*		*parrari*	*battiri*
	Hilfsverb	Part.	Hilfsverb	Part.	Konj.	Var.	Konj.	Var.	Konj.	Konj.
1s	aiu	parratu	aiu	battutu	parrai	-avi -aiu -avu	battivi	-ìi -iu -ivu	-	-
2s	ai		ai		parrast	-	battisti	-	parra	batti
3s	avi		avi		parrau	-ò	battìu	-	parrassi	battissi
1p	avemu		avemu		parrammu	-amu	battemmu	-emu	parramu	battemu
2p	aviti		aviti		parrastivu	-astu	battistivu	-astu -istu	parrati	battiti
3p	annu		annu		parrarunu	-aru	batterunu	-eru	parrassiru	battissiru

	Gerundium				Imperfekt				Futur	
	stari + parrari		*stari + battiri*		*parrari*		*battiri*		*parrari*	*battiri*
	Hilfsverb	Ger.	Hilfsverb	Ger.	Konj.	Var.	Konj.	Var.	Konj.	Konj.
1s	staiu	parrennu	staiu	battennu	parrassi	-	battissi	-	parrirò	battirò
2s	stai		stai		parrassi	-	battissi	-	parrirai	battirai
3s	sta		sta		parrassi	-	battissi	-	parrirà	battirà
1p	stamu		stamu		parrassimu	-	battissimu	-	parriremu	battiremu
2p	stai		stai		parrassivu	-	battissivu	-	parririti	battiriti
3p	stannu		stannu		parrassiru	-assinu	battissiru	-issinu	parrirannu	battirannu

***Tabelle 1**: Die regelmäßigen Konjugationsformen, die in SiMoN integriert wurden.*

In Tabelle 1 sind die regelmäßigen Konjugationsformen (die in SiMoN vorhanden sind) am Beispiel der sizilianischen Verben *parrari* (Deutsch - reden) und *battiri* (Deutsch - schlagen) aufgeführt. Die Formen beider Verbtypen in den Flexionskategorien Indikativ, Imperativ und Subjunktiv, sowie Konditional und Gerundium sind jeweils vorhanden. Die Paradigmen der unregelmäßigen Hilfsverben *essiri* (Deutsch - sein) und *aviri* (Deutsch - haben) sowie das sehr häufig verwendete *fari* (Deutsch - machen) wurden ebenfalls in SiMoN in die Liste der Lemmata aufgenommen, um Partizipkonstruktionen u. ä. zu erkennen.

Ausblick

Unserer Ziel ist vorerst anhand den Texten der Wikipedia für Standard Italienisch und alle andere Dialektwikipedias weiterhin automatisch dialektspezifische Verben zu extrahieren und damit SiMoN zu erweitern. Damit können zusätzliche Dialekte auch behandelt und entwickelt werden. SiMoN würde dann eine automatisierte morphologische Analyse für reguläre italienische Dialektparadigmen ermöglichen, was wir bis jetzt nur für Sizilianisch anbieten können. Weiterhin ist es geplant auch irreguläre Dialektparadigmen manuell zu integrieren.

Fußnoten

1. Die Zahlen sind von Wikipedia entnommen worden (Stand: August 2015).

Bibliographie

Bonner, J. K. "Kirk" / Cipolla, Gaetano (2001): *Introduction to Sicilian Grammar*. Brooklyn, NY: Legas.

Tamburini, Fabio / Melandri, Matias (2012): „AnIta: A Powerful Morphological Analyser for Italian“, in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey 941-947.