Modellierung eines maschinell lesbaren Lexikons für das Korpus der altäthiopischen Literatur

Alessandro Bausi, Andreas Ellwardt, Cristina Vertan
Universität Hamburg

1. Einführung

Die Entwicklung und ständige Erweiterung des Unicode-Kodierungssytems Unicode[1] sowie der Mark-up-Sprachen XML[2], TEI[3] haben in den letzten Jahren u.a. die digitale textuelle Repräsentation von historischen Dokumenten, die mit unterschiedlichen Alphabeten geschrieben wurden, ermöglicht.

Diese textuelle Repräsentation eröffnet wiederum, im Kontrast zur reinen Speicherung von Bild-Digitalisaten, die Möglichkeit, computergestützte linguistische sowie philologische Untersuchungen auf großen Textmengen durchzuführen. Durch solche Methoden lässt sich beispielsweise eine diachrone Analyse der Sprache gleichzeitig auf mehreren Ebenen (morphologisch, syntaktisch, semantisch) realisieren, vorausgesetzt, die elektronischen Ressourcen wie Lexika oder annotierte Korpora sowie die sprachtechnologischen Prozesse (morphologische Analysierer, Wortart-Tagger, Parser) sind vorhanden.

Während die sprachtechnologischen Ressourcen und Werkzeuge für moderne Sprachen sehr weit entwickelt sind, gelten viele historische Sprachen als stark „under-ressourced“. Laut Krauwer (2003) gibt es ein minimales Set von Ressourcen, die für eine computergestützte Sprachanalyse unabdingbar sind. Dessen Weiterentwicklung stellt die Wissenschaft vor neue Forschungsprobleme, da sich häufig Modelle, die für moderne Sprachen entwickelt wurden, nicht 1:1 auf historische Sprachen übertragen lassen (VertanEtAl.2014)

In diesem Beitrag werden wir die Modellierung und Entwicklung von sprachtechnologischen Ressourcen für das Altäthiopische (Geʿez) erläutern. Die Besonderheiten des Geʿez (s. Sektion 2), bedingen die Entwicklung von neuen Modellen, z.B. im Bereich der Lexika. In Sektion 3 werden wir exemplarisch die Entwicklung eines Lexikon-Modells für Geʿez darstellen, während wir in Sektion 4 die Einbindung des Lexikons in einer Architektur für die diachrone Analyse des Geʿez diskutieren werden.

2. Kurze Darstellung des Altäthiopischen (Geʿez)

Das südsemitische Geʿez ist die Sprache des Königreichs Aksum in der heutigen nordäthiopischen Provinz Tigray, von wo aus die im 4. Jahrhundert beginnende Christianisierung Äthiopiens ihren Anfang nahm. Die in der Folge entstehende reiche Literatur ist in großem Umfang geprägt von Übersetzungen aus dem Griechischen und später, ab dem 13. Jahrhundert, aus dem Arabischen, was durch grammatische Interferenzphänomene reflektiert wird. Während seine Verdrängung als gesprochene Sprache bereits im 9./10. Jahrhundert beginnt, bleibt es als Schriftsprache sehr viel länger erhalten und ist bis in die Gegenwart hinein Liturgiesprache des äthiopischen und eriträischen Klerus.

Das Altäthiopische hat aus einer südsemitischen Schrift ein eigenes Silbenalphabet entwickelt, das bis heute in mehreren modernen Sprachen Äthiopiens und Eritreas Verwendung findet. Innerhalb der semitischen Sprachen fällt es durch die verwendete Rechtsläufigkeit auf, außerdem werden die Vokale vollständig geschrieben. Beides unterscheidet das Geʿez von den ihm nächst verwandten Sprachen Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch. Des weiteren sind Grapheme, die ursprünglich distinkten Phonemen zugeordnet waren, schon früh in identischer phonetischer Realisierung zusammengefallen, was sich konkret bereits in den ältesten überlieferten Handschriftzeugnissen (aber noch nicht in den aksumitischen Inschriften) niederschlägt, wo eine beliebige Austauschbarkeit der Laryngale und Sibilanten jeweils untereinander zu konstatieren ist.
Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Hierbei muss das einzelne Lexem als Kombination von zwei Elementen beschrieben werden, nämlich der Wurzel und dem Schema: Die konsonantische Wurzel gibt veränderliche Positionen zwischen

[1] http://www.unicode.org/
[2] http://www.w3.org/XML/
[3] http://www.tei-c.org/index.xml

ihren, zumeist drei, Wurzelkonsonanten vor, die durch die Vokale des Schemas aufgefüllt werden, häufig, jedoch nicht zwingend, ergänzt um (vokalische oder konsonantische) Affixe.

3. Arbeitsschritte zu einer computergestützten Analyse des Altäthiopischen

Wie bereits in Sektion 2 erwähnt, sind Geʿez-Dokumente für die gesamte Geschichte des christlichen Orients extrem wertvoll. Manche Überlieferungen von alten griechischen Texten sind in der Originalsprache verloren und nur im Altäthiopischen erhalten. In der Zeit digitaler Bibliotheken erscheint also die Entwicklung von computergestützten Tools für die Geʿez-Sprache umso dringender. Das primäre Ziel des Projekts TraCES[4] ist die Entwicklung eines digitalen Korpus der Geʿez-Sprache, zusammen mit Annotationen auf morphologischer, syntaktischer und semantischer Ebene. Dieses annotierte Korpus soll einerseits eine diachrone Analyse des Altäthiopischen ermöglichen, anderseits soll es selbst als Ressource für weitere computergestützte Prozesse dienen. Langfristig soll eine vergleichende digitale Analyse von altäthiopischen und griechischen (z.B. die in der digitalen PERSEUS Sammlung[5] verfügbaren) oder arabischen sowie anderen christlich-orientalischen Dokumenten möglich sein.

Mit Ausnahme von einigen wenigen Texten gibt es zur Zeit keine verfügbare elektronische Ressource für das Altäthiopische. Daher haben wir uns als erstes der Entwicklung eines maschinell lesbaren Lexikons des Geʿez gewidmet. Dessen Modellierung wird in der nächsten Sektion erklärt.

4. Ein Lexikon-Modell für Geʿez

Die in Sektion 2 erwähnte Austauschbarkeit der Laryngalen und Sibilanten untereinander stellt uns vor eine erste Modellierungsanforderung. Für einen Lexikon-Eintrag muss nicht nur die Grundform, sondern es müssen auch alle möglichen graphischen Varianten gespeichert werden, wobei wohlgemerkt diese graphische Variationen auch in einigen Fällen als selbständige Lexikon-Einträge mit ganz anderer Bedeutung existieren können.

Das Lexikonmodell muss daher eine starke Modularisierung und Verlinkung zwischen den einzelnen Modulen unterstützen. Wir haben uns für das Lemon-Modell (McCraeEtAl.2012) entschieden. Unserer Kenntnis nach, ist dies der erste Versuch, eine semitische Sprache mit dem Lemon-Modell zu beschreiben.
Die Grundkomponenten eines Lemon-Lexikon-Modells für Geʿez wurden wie folgt angepasst.

Die Zitierform eines Wortes in klassischen Lexika semitischer Sprachen ist in der Regel eine verbale Repräsentation der Wurzel in der 3. Person Perfekt Singular maskulin. Diese Form wird in unserem Lemon-Modell als „Lexical Entry“ gespeichert.

Ein „Lexical Entry“ ist mit den folgenden weiteren Modulen verknüpft:

- Das Lexical Form-Modul beinhaltet alle möglichen graphischen Varianten des Lemmas. Jede graphische Variante wird zusammen mit ihrer Transkription gespeichert.
- Das Morphologie-Modul beinhaltet eine Subkomponente für den lexikalischen Eintrag, die das Paradigma, Ausnahmen der morphologischen Realisierung (z.B. Sonderformen im Imperfekt oder Plural) sowie die jeweiligen anderen morphologischen Kategorien für das Lemma umfasst. Das Semantik-Modul setzt sich aus einer Übersetzungs-, einer Korpusevidenz- und einer semantische-Merkmale-Komponente zusammen. Unter Korpusevidenz verstehen wir Beispiele aus Korpora für dieses Lemma oder eine seiner morphologischen Realisierungen. Die Übersetzungen sind unterteilt in eine Übersetzung ins Englische und semantische Äquivalente in anderen Sprachen wie (falls vorhanden) Arabisch, Hebräisch, Syrisch, Koptisch, Griechisch oder sogar Sanskrit.
- Das Syntax-Modul beinhaltet syntaktische Funktion des Lemmas, zusammen mit Beispielen von syntaktischen Bäumen. Dieses Modul wird in einer späteren Projektphase entwickelt.

[4] European Union Seventh Framework Programme IDEAS (FP7/2007-2013), European Research Council, grant agreement no. 338756, project “TraCES – From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages”, http://www1.uni-hamburg.de/ethiostudies/traces.html
[5] http://www.perseus.tufts.edu/hopper/

4.1. Wurzel-Modellierung

Da die Wurzel eine zentrale Stellung in der semitischen Morphologie hat, haben wir als ersten Schritt ein Wurzel-Sublexikon erstellt. Dieses entspricht dem Wurzel-Submodul im morphologischen Modul.

Die Erstellung des Wurzel-Lexikons wurde vollständig automatisiert. Aus einer digitalen Version des trotz seiner Abfassung im Jahre 1865 unverändert als Standardwerk geltenden „Lexicon linguae aethiopicae“ von August Dillmann (Dillmann1865) (im Unicode-Format) wurden zirka 4000 Wurzel-Einträge mit Hilfe von String-basierten Regeln extrahiert.

Für jede Wurzel wurden:

- die vollständige Transkription
- die auf das konsonantische Gerüst zurückgeführte Transkription
- das konsonantischen Wortbildungsschema
- alle graphischen Varianten zusammen mit deren Transkriptionen

durch regel-basierte Verfahren extrahiert. Die Automatisierung ermöglicht zum ersten Mal die Sammlung aller graphischen Varianten für alle 4000 Wurzeln (wobei hervorgehoben werden muss, dass manche Wurzeln bis zu 50 graphische Varianten haben).

Jede Wurzel wird automatisch mit ihren Homophonen (Einträge mit identischer graphischer Form, aber unterschiedlicher Bedeutung) verknüpft. Erfasst werden durch automatische Prozesse auch alle Lexikoneinträge von graphischen Varianten (falls vorhanden).

Das Wurzel-Lexikon wird im XML-Format gespeichert. Dafür wurde ein eigenes XML-Schema entworfen. Eine Java-basierte graphische Oberfläche wurde implementiert. Diese Oberfläche ermöglicht nicht nur die Visualisierung von den Einzeleinträgen und die Navigation durch das Wurzel-Lexikon, sondern auch manuelle Korrekturen, das Löschen oder das Einfügen von neuen Einträgen.

Nach Korrekturen wird das Wurzel-Lexikon:

- als eine „Authority List“ für das Geʿez-Lexikon und
- als Generierungsquelle für Lexikoneinträge

benutzt.

5. Zusammenfassung und weitere Arbeit

In diesem Beitrag haben wir die Modelle für ein Wurzel- und ein Lemma-Lexikon für die Geʿez-Sprache erklärt. Die Wurzel und Lemma-Akquisition werden weitgehend durch computergestützte Prozesse realisiert. Die erstellte Software wird bei der Präsentation des Beitrags vorgeführt.

Das Projekt TraCES wurde im März 2014 begonnen und hat eine Laufzeit von fünf Jahren. Die Erstellung des Lexikons der Geʿez Sprache ist zurzeit die zentrale Arbeit im Projekt, wobei derzeit die Erstellung von Generierungsparadigmen im Vordergrund steht. Mit deren Hilfe werden durch Computerverfahren Lexikoneinträge generiert.
Ein erster Test hat mehr als 13 000 Einträge generiert. Dies zeigt, dass die Automatisierung eine erhebliche Zeitersparnis für die Lexikon-Akquisition ermöglicht.

Literatur

(Dillmann1865) Dillmann, August, *Lexicon linguæ Æthiopicæ cum indice Latino*, Lipsiae 1865.

(Krauwer2003) Krauwer, Steven, „*The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources*“, http://www.elsnet.org/dox/krauwer-specom2003.pdf (09.11.2014)

(McCraeEtAl2012). McCrae, John und Aguado-de-Cea, Guadalupe und Buitelaar, Paul und, Cimiano, Philipp und Declerck, Thierry und Gómez Pérez, Asunción und Gracia, Jorge und Hollink, Laura und Montiel-Ponsoda, Elena und Spohr, Dennis und Wunner, Tobias, *The Lemon Cookbook*, http://lemon-model.net/lemon-cookbook.pdf (09.11.2014)

(VertanET.AL.2014) Vertan, Cristina und Zervanou, Kalliopi und van den Bosch, Antal und Sporeleder, Caroline (Hrsg.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Götheborg, Sweden, 2014, http://www.aclweb.org/anthology/W14-06 (09.11.2014)