Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte

Vertan, Cristina
cristina.vertan@uni-hamburg.de
Universität Hamburg, Deutschland

Ellwardt, Andreas
andreas.ellwardt@uni-hamburg.de
Universität Hamburg, Deutschland

Hummerl, Susanne
sanne.hummel@gmx.de
Universität Hamburg, Deutschland

Kurze Darstellung des Altäthiopischen (Ge#ez)

Das südsemitische Ge#ez 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 Ge#ez von verwandten Sprachen wie 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 ihren, zumeist drei, Wurzelkonsonanten vor, die durch die Vokale des Schemas aufgefüllt werden, häufig, jedoch nicht zwingend, ergänzt um (vokalische oder konsonantische) Affixe. Das äthiopische Silbenalphabet bringt dabei mit sich, dass Morphemgrenzen in der Schrift nicht darstellbar sind, sodass beispielsweise ein einzelner Vokal als Bestandteil einer Silbe eine eigenständige Wortart darstellen kann und tokenisiert werden muss; z. B. ist im zweisilbigen Wort ### /be·tu/ das /u/ als pronominales Suffix zu bet-*u* (*sein* Haus) zu segmentieren.

Ein Tagest für die morphologische Annotation des Ge#ez

Zur Analyse morphologischer Merkmale des Altäthiopischen wurde erstmals ein feingliedriges Tagset von 30 Wortarten entwickelt. Wir unterscheiden vier Klassen von Wortarten: Nomina, Verben, Existentiale, Partikel. Diese Klassen untergliedern sich weiter in die folgenden Wortarten:

- Nomina: Nomen (Eigennamen und Substantive), Pronomen (mit 10 Wortarten) und Zahlen (Kardinal- und Ordinalzahlen)
- Verben
- Existentiale (affirmativ und negativ)
- Partikel (14 Wortarten, z. B. Konjunktionen, Präpositionen, Adverbien)

Den Wortarten wurden entsprechende grammatische Kategorien zugewiesen (Genus, Numerus, Kasus, Person usw.). Ein Sonderfall ist die Bestimmung des Genus für das Nomen. Hier ist das Altäthiopische häufig uneindeutig sowohl in der morphologischen Markierung, als auch in der syntaktischen Kongruenz. Bei eindeutiger Kennzeichnung des grammatischen Genus wird daher weiter dahingehend spezifiziert, wodurch das jeweilige Genus bestimmt ist: durch das morphologische Schema, durch die Syntax und/ oder aufgrund des natürlichen Geschlechts (z. B. Mutter). Daher kann ein Nomen im selben Satz im Genus mehrfach bestimmt sein (z. B. syntaktisch maskulin und dem Schema nach feminin). Ist das grammatische Genus nicht eindeutig zu bestimmen, wird es als „unmarkiert“ annotiert.

Das Annotationstool

Die Komplexität des in Sektion 2 dargestellten Annotationstools wird zwar zum einen sehr vielfältige linguistische Anfragen und eine detaillierte Analyse der Sprache ermöglichen, andererseits jedoch verhindert ebendiese Komplexität eine automatische Annotation. Ein Vectorspace-Modell (das für maschinelle Lernverfahren benutzt werden muss) das alle morphologischen Merkmale abdecken würde, wäre zu groß. Vorstellbar ist lediglich

eine flache automatische Annotation (z. B. der Wortarten); jedoch wird auch für eine solche zunächst eine relativ große Menge an Trainingsdaten benötigt. Daher ist die Entwicklung eines Werkzeugs für die manuelle Annotation ein obligatorischer Schritt.

Das eigens für das Altäthiopische entwickelte Annotationstool berücksichtigt die spezifischen Besonderheiten der Sprache, von denen einige oben in Sektion 1 kurz skizziert wurden. Aufgrund der dargestellten Eigenheiten sowohl des Silbenalphabets als auch der semitischen Morphologie kann der Text nicht unmittelbar in der Originalschrift annotiert werden. Eine Annotation kann daher ausschließlich in der Transliteration erfolgen (siehe obiges Beispiel *bet-u*).

Die Transliteration wiederum ist nur bedingt durch automatische Regeln beschreibbar. Phänomena wie Konsonantenverdoppelung oder Kontraktion von zwei Silben können nicht durch klare Regeln beschrieben werden. Einige solcher Phänomena ließen sich zwar mittels weiterer Ressourcen automatisch regeln (z. B. Konsonantenverdoppelung bei bestimmten Verbklassen), jedoch müssten derartige Informationen aus einem (bisher noch nicht vorhandenen) digitalen Lexikon extrahiert werden. Auch über die genannten Schwierigkeiten hinaus wäre die Entwicklung einer (semi-)automatischen Transliteration ein äußerst zeitaufwendiger Prozess. Daher haben wir uns für die folgenden Arbeitsschritte entschieden:

- Die Texte werden mittels eines automatischen regelbasierten Prozesses transkribiert.
- Die Transkription wird manuell korrigiert (entspricht der Transliteration des Textes)

Aus den oben genannten Gründen ist der in der Arbeit mit anderen Sprachen gängige Arbeitsablauf – zuerst Textkorrektur, dann Annotation – hier nicht möglich. Ein solcher ist jedoch bei bereits existierenden Tools Bedingung, wenn eine Mehrebenen-Annotation angestrebt wird. Das CorA-Tool (vg. Bollmann et al.) ermöglicht zwar Korrekturen synchron mit der Annotation, jedoch sind nicht mehr als zwei Annotationsebenen möglich; auch eine Mehrwortannotation ist nicht erlaubt. Für die Annotation muss ein XML-Schema des Tagsets vorliegen und es werden alle möglichen Kombinationen von morphologischen Merkmalen je Wortart generiert. Da sämtliche Kombinationsmöglichkeiten dem Benutzer in Form einer Dropdown-Liste präsentiert werden, ist das Tool in der Anwendung mit dem sehr umfangreichen Tagset für das Altäthiopische ungeeignet. Ein anderes Tool, das relativ häufig angewendet wird, ist WebAnno . Dieses Tool ermöglicht eine Annotation mit mehr als zwei Ebenen, jedoch sind Korrekturen im Text während der Annotation nicht möglich.

Im TraCES Projekt implementieren wir eine neuartige Architektur, die sowohl Änderungen im Text als auch eine Mehrebenen-Annotation ermöglicht.

Wir betrachten als Grundtext den Originaltext in der altäthiopischen Schrift. Die Transliteration bildet die erste und die morphologische Annotation die zweite Ebene, wobei die Transliteration und der Originaltext bei allen Arbeitsschritten synchronisiert bleiben. Im folgenden Abschnitt beschreiben wir das Datenmodell, das diese Architektur ermöglicht.

Die Basiseinheit in unserem System ist ein Wort, das eine einmalige ID zugewiesen erhält. Ein Wort hat folgende Komponenten:

- Eine Liste der einzelnen Fidal -Objekte, wo ein Fidal-Objekt aus einer ID und einem Label (dem Fidal-Buchstaben) besteht.
- Eine Liste einzelner Silben-Objekte, wo ein Silben-Objekt aus einer ID und einer Liste von einzelnen Buchstaben-Objekten besteht
- Ein Buchstaben-Objekt hat immer eine ID und ein Label (das graphische Symbol)

Die Zusammengehörigkeit aller Komponenten wird durch die ID-Zusammensetzung gesichert:

Eine Wort ID ist aus vier Komponenten zusammengesetzt

> Projekt ID + Dokument ID + W + automatisch generierte Random ID

Ein Fidal-Buchstaben-Objekt wird dann durch

> Wort ID + F + Random-Nummer

identifiziert, während ein Transliterationssilben-Objekt:

> Wort ID + TF + Random-Nummer

als ID hat.

Ein Transliterationsbuchstabe wird durch:

> Wort ID + Transliterationssilben ID + L + Random-Nummer

identifiziert.

Durch dieses System ist es zu jeder Zeit möglich, jeden einzelnen Buchstaben zu identifizieren und zu referenzieren. Da wir jeden Buchstaben als Objekt betrachten, trennen wir die graphische Realisierung von der linguistischen Annotation; so sind etwa die Annotation und die graphische Repräsentation eines Buchstabens in der Transliteration Labels für ein und dasselbe Objekt.

In der Abbildung 1 wird dieses Modell für die Wortgruppe ######„ *und vor allem nämlich*“ (eigentlich # - ## - ## - #, aber graphisch quasi ein „Wort“) dargestellt:

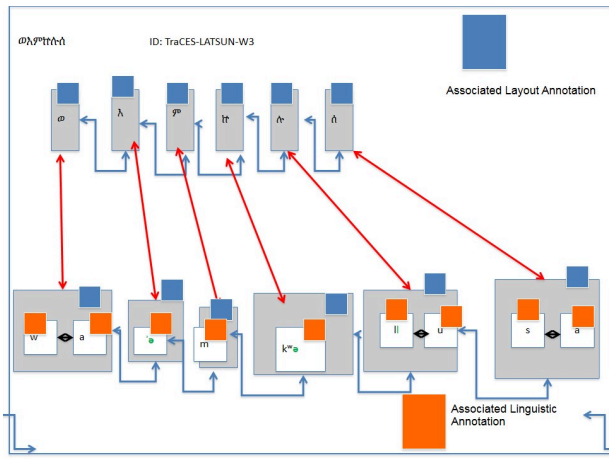

***Abb. 1**: Annotations-Modell für ######*

Zusammenfassung

In diesem Artikel beschreiben wir ein neuartiges Modell für ein Annotationstool, das sowohl eine Annotation mit gleichzeitiger Korrektur, als auch eine Mehrebenen-Annotation ermöglicht. Wir begründen, warum die Entwicklung eines speziellen Modells für die Annotation des Altäthiopischen notwendig war. Möglicherweise könnte das Modell mit wenigen Änderungen auch für andere Sprachen benutzt werden. Eine Demonstration des ersten Prototyps wird auch möglich sein.

Acknowledgements

Das Projekt TraCES wird durch einen Grant der European Science Foundation unterstützt (Grant Agreement 338756).

Die in diesem Artikel beschriebenen Ergebnisse sind das Resultat der Arbeit des gesamten TraCES-Teams: Alessandro Bausi (Projektleiter), Wolfgang Dickhut, Daria Elagina, Andreas Ellwardt, Susanne Hummel, Vitagrazia Pissani, Eugenia Sokolinski, Cristina Vertan.

Fußnoten

1. „Fidal“ ist der Terminus technicus für das äthiopische Silbenalphabet.

Bibliographie

Bollmann, Marcel / Petran, Florian / Dipper, Stefanie / Krasselt, Julia (2014): "CorA: A web-based annotation tool for historical and other non-standard language data", in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Gothenburg, Sweden 86-90 https://aclweb.org/anthology/W/W14/W14-0612.pdf [letzter Zugriff 30. Dezember 2015].

Castilho, Richard Eckart de / Biemann, Chris / Gurevych, Iryna / Yimam, Seid Muhie (2014): "WebAnno: a flexible, web-based annotation tool for CLARIN", in: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands http://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf [letzter Zugriff 10. Februar 2016].

Marquez, Lluis / Rodriguez, Horacio / Carmona, Josep / Montolio, Josep (1999): "Improving POS Tagging Using Machine-Learning Techniques", in: *Proceedings of the 1999 Joint SIGDAT conference on Empirical Methods in Natural Language Processing and very large corpora* 53-62 http://www.aclweb.org/anthology/W99-0608 [letzter Zugriff 30. Dezember 2015].