Nachhaltige Entwicklung digitaler Ressourcen und Werkzeuge für wenig erforschte historische Sprachen

Feige, Tillmann
tillmann.feige@uni-hamburg.de
Universität Hamburg, Deutschland

González, Alicia
alicia.gonzalez@uni-hamburg.de
Universität Hamburg, Deutschland

Prager, Christian
cprager@gmx.net
NRW Akademie der Wissenschaften und der Künste

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

Werwick, Heiko
heiko.werwick@web.de
Universität Jena, Deutschland

Das Panel wird durch vier Kurzvorträge in Probleme der Nutzung digitaler Werkzeuge für nicht-indoeuropäische Sprachen einführen. Die Vorträge basieren auf Erfahrungen aus langfristig angelegten Projekten und haben jeweils individuelle Lösungen für die spezifischen Anforderungen gefunden, die hauptsächlich durch die Sprache formuliert werden.

Allen Projekten gemein sind Probleme der Nutzung von digitalen Werkzeugen, die insbesondere bei der Verwendung von Quellen historischer oder wenig erforschter Sprachen auftreten.

Aktuelle Content Management Systeme und Annotationstools wurden selten im Hinblick auf Anforderungen aus Orchideenfächern entwickelt. Dies betrifft beispielsweise einige Sprachen mit nichtkonkatenativer Morphologie oder komplexen Schriftsysteme. Daher müssen für erwähnte Sprachen entweder existierende Anwendungen angepasst oder erschaffen werden.

Bei der Adaption können während der Modellierung wichtige Eigenschaften nicht berücksichtigt werden oder bleiben nur als Kommentar erhalten, was eine weitere maschinelle Bearbeitung erschwert. Bezüglich der Datenkodierung ergibt sich das Problem der Ineffizienz. So wurden morphologische Tagsets primär für die indo-europäische Sprachfamilie entwickelt. Für eine tiefe linguistische Annotation müssen aber diese Standards beispielweise für einige semitische Sprachen angepasst werden.

Nicht selten ist die Alternative die Eigenentwicklung projektbezogener Lösungen, die aber aufgrund der Anforderungen mit eigenen Datenformaten arbeiten, und so nicht mehr den geltenden Standards folgen und den Austausch erschweren. Hinzu kommt der immense Zeit- und Ressourcenaufwand bei der Implementierung.

Allerdings sind gerade im deutschsprachigen Raum viele langfristige Projekte auf digitale Tools angewiesen.

Durch eine Vernetzung solcher Projekte können gemeinsame Anforderungen an, und Begrenzungen von aktuellen Lösungen besprochen und Initiativen zur Entwicklung digitaler Tools und Ressourcen koordiniert werden. Daher ist das Ziel dieses Panels eine erste Zusammenführung langfristig ausgerichteter Projekte im deutschen Sprachraum, die mit historischen nicht-indo-europäischen Sprachen im digitalen Kontext arbeiten. Dabei sollen die Probleme der Nachhaltigkeit entwickelter Werkzeuge und Ressourcen, sowie der bearbeiteten Daten besprochen werden. Anschließend werden die vielfältigen Herangehensweisen mit einem Fokus auf drei große Punkte diskutiert:

- Nachhaltigkeit von Repositorien
- Welche Frameworks werden für welche Datentypen benötigt?
- Wie können Informationen über unpräzise Daten gespeichert werden?
- Wie gehen verfügbare Systeme mit Multilingualität um?
- Nachhaltigkeit von (Annotations-)Werkzeugen
- Analyse historischer Daten impliziert die Annotation von Textmaterialien in Sprachen, die aus verschiedenen Gründen zu Problemen führen.
- Welche Annotationstools können genutzt werden? Mit welchen Limitierungen?
- Was bedeutet es, ein neues Tool zu entwerfen?
- Häufige Anforderungen durch strukturell komplexe Sprachen: Multilevel-Annotation, Textkorrektur während der Annotationsphase, Multilevel-Segmentierung
- Nachhaltigkeit des annotierten Materials (Standards)
- Während der Standard TEI-XML als Schnittstellenformat sehr nützlich ist, ergeben sich dennoch Probleme wie:
  - für interne Verarbeitung kann dessen Verwendung hinderlich sein. Daher müssen projekt-spezifische

Lösungen mit standardisiertem Export entwickelt werden.
- Können diese Daten von Dritten in TEI-XML verarbeitet werden?
- Welche anderen Formate können genutzt werden (z.B. JSON)?
- Sind existierende Tagset-Formate ausreichend spezifiziert, um auch nicht-europäische Sprachen taggen zu können?

Herausforderungen in der Nutzung vorhandener Tools für arabische Daten

Alicia González, Tillmann Feige
Universität Hamburg
ERC- Projekt COBHUNI (https://www.cobhuni.uni-hamburg.de/)
Email: ;

Wir beschreiben den Ansatz, einen Korpus der neben modernem auch klassisches Arabisch (siehe Romanov, 2016) enthält, mit computerlinguistischen und semantischen Verfahren analysierbar zu machen. Wir setzen auf bereits vorhandene Software für die Hauptpunkte Annotation und Analyse. Dazu wurde ein Pflichtenheft erstellt, dass mit vorhandenen Tools abgeglichen wurde.

Da wir mit arabischen Daten arbeiten, ist eine große Herausforderung die Schrift. Es ist eine linksläufige verbundene Schrift, die durch Konsonanten und lange Vokale repräsentiert wird. Kurze Vokale sind Diakritika, die optional gesetzt werden und gerade bei Referenzen auf religiöse Quellen im Textkorpus vorkommen. Dabei ist vollständige UTF-8 Unterstützung und die saubere Darstellung der Schrift unabdingbar. Dies reduziert die Auswahl erheblich. Hinzu kommt, dass wir auf flexible Import- und Exportmöglichkeiten angewiesen sind. Ähnliche Probleme führen Peralta und Verkinderen auf (Peralta / Verkinderen 2016). Durch unsere Herangehensweise gibt es weitere Einschränkungen wie Mehrebenen-, Multitoken- aber auch Subtoken-Annotation.

Die Auswahl für die semantische Annotation fiel auf WebAnno, dass durch sein spezielles Datenmodell die erforderliche Datenaufbereitung und Kontrolle gestattet.

Als Visualisierungstool haben wir ANNIS ausgewählt, dass ebenfalls Arabisch unterstützt, einen konfigurierbaren Converter mitbringt und Mehrebenenkorpora erlaubt, so dass auch hier die Hauptkriterien erfüllt wurden. Zusätzlich lassen sich potentielle Probleme in der Darstellung durch eine anpassbare HTML-Visualisierung umgehen. Durch Zusammenarbeit mit den Entwicklern beider Programme wurde die Unterstützung für Arabisch stetig ausgebaut.

Im Beitrag werden wir die einzelnen Punkte erläutern und darstellen, warum wir uns für die angeführten Programme und gegen eine Eigenentwicklung entschieden haben, sowie welche Implikationen diese Entscheidung für die Nachhaltigkeit des Projekts, der Daten und der genutzten Tools hat.

Tiefe Mehrebenen-Annotation für semitische Sprachen: der Fall von Ge'ez

Cristina Vertan
Universität Hamburg
ERC-Projekt TraCES (https://www.traces.uni-hamburg.de/)
Email:

Das südsemitische G###z 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, was durch grammatische Interferenzphänomene reflektiert wird. 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 G###z von verwandten Sprachen wie Altsüdarabisch, Arabisch, Hebräisch und Syro-Aramäisch Mit den genannten eng verwandten semitischen Sprachen teilt das Altäthiopische die nichtkonkatenative Morphologie. Durch das äthiopische Silbenalphabet sind Morphemgrenzen in der Schrift nicht darstellbar, so dass beispielsweise ein einzelner Vokal als Bestandteil einer Silbe eine eigenständige Wortart darstellt und tokenisiert werden muss.

Die Komplexität des Annotationstools wird sehr vielfältige linguistische Anfragen und detaillierte Analysen der Sprache ermöglichen, aber auch eine vollautomatische Annotation verhindern. Ein alle morphologischen Merkmale abdeckendes Vektorraum-Modell (das für maschinelle Lernverfahren benutzt werden muss) 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.

Die Besonderheit der entwickelten Lösung (Vertan/ Ellwardt/Hummel 2016) sind:

- automatische Transkription
- manuelle Korrektur der Transkription während des Annotationsprozesses
- semi-automatische Verfahren: automatische Verläufe werden farbig markiert und sind automatisch zur manuellen Korrektur hinterlegt

- Mehrebenenannotation: Linguistik, Edition, Textstruktur
- Anpassungen an unterschiedliche Schriftsysteme und Transkriptionsregeln

Nutzungs- und Nachhaltigkeitsstategien im Projekt "Textdatenbank und Wörterbuch des Klassischen Maya"

Christian M. Prager
NRW Akademie der Wissenschaften und der Künste
http://mayawoerterbuch.de/
Email:

Die Mayaschrift ist das einzig lesbare Schriftsystem der vorspanischen Amerikas. Die über 10.000 Texte sind in einer logographisch-syllabischen Hieroglyphenschrift verfasst und von den rund 800 Zeichen sind erst 60% sicher entziffert. Die Texte enthalten taggenaue Kalenderangaben, die es uns ermöglichen die rund 2000jährige Sprach- und Schriftgeschichte genau zu dokumentieren. Das Projekt (Prager 2015) wird sämtliche Inschriften einschließlich Metadaten in einer Datenbank einzupflegen und darauf basierend ein digitales Wörterbuch des Klassischen Maya zu kompilieren. Herausforderung dabei ist, dass die Schrift noch nicht vollständig entziffert ist und bei der Modellierung zu berücksichtigen ist. Unser Projekt verfolgt den Ansatz, wonach die Bedeutung von Wörtern ihre Verwendung ist - Texte nehmen Bezug auf den Textträger und den Verwendungskontext und nur die exakte Dokumentation sogenannter nicht-textueller Informationen erlaubt es, textuelle und nicht-textuelle Informationsbereiche zueinander in Beziehung zu setzen und bei der Entzifferung von Zeichen und Textstellen zu berücksichtigen. Zum Zweck der Nachhaltigkeit und Nachnutzung greift das Projekt bei der Beschreibung der Artefakte und der relevanten objektgeschichtlichen Ereignisse auf CIDOC CRM zurück, das eine erweiterbare Ontologie für Begriffe und Informationen im Bereich des kulturellen Erbes anbietet. Das entstandene Anwendungsprofil wird durch Elemente aus weiteren Standards und Schemata angereichert und wird damit auch für vergleichbare Projekt nachnutzbar. Die Schemata und erstellten Metadaten werden in einer Linked (Open) Data-Struktur (LOD) abgebildet. Durch die Repräsentation im XML-Format, sowie die Nutzung von HTTP-URIs wird eine einfache Austauschbarkeit und Zitierbarkeit der Daten ermöglicht. Durch diese Umsetzung können Objektmetadaten getrennt vom erfassten Text gespeichert werden und durch die Verwendung der HTTP-URI verlinkt werden. Die Nachnutzung bereits bestehender und fachlich anerkannter Terme trägt darüberhinaus auch zu einer hohen Interoperabilität mit anderen Datenbeständen und Informationssystemen bei. Das ausgestaltete Schema hat eine ontologisch-vernetzte Struktur, die komplexe Beziehungen und Zusammenhänge abbildet.

Interdisziplinäre Digitale Zusammenarbeit für seltene Sprachen und Kulturen

- Eine Fallstudie über jiddische Texte aus der frühen Neuzeit -

Walther v. Hahn (Universität Hamburg), Berndt Strobach (Wolffenbüttel)
Email: , berndt.strobach@freenet.de>

In den Geisteswissenschaften werden häufig die fachlichen Interpretationen und die sprachlichen Erklärungen von verschiedenen Gruppen mit unterschiedlicher Kompetenz bearbeitet. Gute Beispiele sind Studien zu Texten aus semitischen Sprachen, wobei, speziell bei historischen Dokumenten die historische oder geistes- und sozialgeschichtliche Würdigung von Forschern verfasst werden muss, die des Hebräischen, Arabischen, Aramäischen etc. nicht mächtig sind, die sprachwissenschaftlichen Forscher dagegen bei der Interpretation gelegentlich weniger engagiert bleiben. Extremfälle wie Studien über das Sephardische in Spanien (Ladino, Djudezmo) machen etwa solide Kenntnisse zumindest des Spanischen, Hebräischen, Türkischen, Griechischen und Italienischen zur Voraussetzung für seriöse hermeneutische Forschungsergebnisse.Wir berichten über Studien zu jiddischen Texten aus dem Wolffenbüttel des 18. Jahrhunderts, in denen die Rolle der "Hofjuden" und ihres kultur- und sozialgeschichtlichen Hintergrundes diskutiert wird.

Die Herausforderung einer interdisziplinären Zusammenarbeit zwischen Historikern, Sprachwissenschaftlern und Informatikern besteht darin,

1. die Lesbarkeit der Originalquellen für alle Gruppenmitglieder sicher zu stellen (Invertierte Transkriptionen, Vokalisierung, Visualisierung), sowie
2. in der Gruppe eine gemeinsame Behandlung von Vagheit, Unsicherheit und Unbekanntem zu definieren, so dass die Unklarheiten in den einzelnen Forschungsstufen erhalten und im Endergebis sichtbar bleiben (Vagheits-Annotationen und vage Inferenzen). Heute werden derartige Unsicherheiten meist bereits in den Annotationen unterschlagen (von Hahn, 2016).

Bibliographie

Hahn, Walther von (2016): „Humanities meet Computer Science – Digital Humanities between Expectations and Reality“, zu erscheinen in: von Hahn, Walter / Papadima, Liviu / Vertan, Cristina (eds.): *Humanities2020, New*

Trends in Education and Research. Bukarest: University of Bucharest Publishing House.

Peralta, José Haro / Verkinderen, Peter (2016): „‚Find for me!': Building a Context-Based Search Tool Using Python", in: Muhanna, Elias (ed.): *The Digital Humanities and Islamic & Middle East Studies*. Berlin: Walter de Gruyter GmbH 199–231.

Prager, Christian M. (2015): „Das Textdatenbank- und Wörterbuchprojekt des Klassischen Maya: Möglichkeiten und Herausforderungen digitaler Epigraphik", in: Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (eds.): *TextGrid: Von der Community - für die Community: Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*. Glückstadt: Werner Holsbusch 105–124 https://www.academia.edu/17957108/Das_Textdatenbank-_und_W%C3%B6rterbuchprojekt_des_Klassischen_Maya_M%C3%B6glichkeiten_und_Herausforderungen_digitaler_Epigraphik .

Romanov, Maxim (2016): *Creating Frequency-Based Readers for Classical Arabic* http://maximromanov.github.io/2016/05-30.html [letzter Zugriff 1. Dezember 2016].

Vertan, Cristina / Ellwardt, Andreas / Hummerl, Susanne (2016): „Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* http://www.dhd2016.de/abstracts/vorträge-061.html .