Semantische Suche in Ausgestorbenen Sprachen: Eine Fallstudie für das Hethitische

Daxenberger, Johannes
daxenberger@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Görke, Susanne
goerkes@uni-mainz.de
Altorientalische Philologie, Institut für Altertumswissenschaften, Johannes Gutenberg-Universität Mainz

Siahdohoni, Darjush
siahdohoni@googlemail.com
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Gurevych, Iryna
gurevych@ukp.informatik.tu-darmstadt.de
Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technische Universität Darmstadt

Prechel, Doris
prechel@uni-mainz.de
Altorientalische Philologie, Institut für Altertumswissenschaften, Johannes Gutenberg-Universität Mainz

Einleitung

Mit dem Auftreten der Keilschrift am Ende des 4. Jt. v. Chr. bis zur Zeitenwende sind zahlreiche Sprachen des Vorderen Orients aufgezeichnet, deren Kenntnis sich heute allein dem Erhalt der Schriftträger dankt: Eine nicht mehr überschaubare Anzahl von Tontafeln stellt das wesentliche Medium zur Rekonstruktion einer alle menschlichen Lebensbereiche umfassenden dreitausendjährigen Geschichte der heutigen Staaten Syrien, Libanon, Türkei, Irak und Iran dar. Zu den besser bezeugten Sprachen gehört neben dem semitischen Akkadischen das isolierte Sumerisch und das indoeuropäische Hethitisch. Auch wenn sich inzwischen diverse Projekte mit der Digitalisierung des keilinschriftlichen Kulturschatzes befassen, z.B. Cohen et al. (2004) und Tyndall (2012), ist der Zugang zu den kulturell, historisch und linguistisch hochbedeutsamen Textcorpora, die zu großen Teilen noch unpubliziert in den Museen der Welt lagern, meist auf Fachwissenschaftler begrenzt. Um eine adäquate Verwendung der durch Grabungen stetig wachsenden Anzahl von Texten auch in fernerliegenden Arbeitsbereichen zu ermöglichen, ist ein umfassendes Angebot von Übersetzungen in moderne Sprachen höchst wünschenswert.

Das hier skizzierte Projekt zielt insbesondere auf den Umstand, dass selbst die (wenigen) vorhandenen Übersetzungen aufgrund der Durchdringung mit autochtonen Termini es oft an Verständlichkeit vermissen lassen. Das Ziel unserer Pilotstudie ist eine digitale Annäherung an Keilschriftsprachen. Wir stellen eine erweiterte Suchfunktion vor, die es auch fachfremden Benutzern erlaubt, intelligente Suchanfragen in den hethitischen und akkadischen Textcorpora zu stellen. Dazu verwenden wir moderne Natural Language Processing (NLP) Methodologie, die automatisiert lexikalisch-semantische Informationen in mehrsprachigen Übersetzungen von aktuell gut 500 Keilschriftdokumenten extrahiert. Durch den Einsatz vollautomatischer Methoden ist das Hinzufügen neuer Übersetzungen jederzeit möglich – es gibt alleine für das Akkadische über eine halbe Million (noch) nicht digitalisierter Quelltexte. Das Ergebnis unserer Studie ist in Form eines webbasierten Tools verfügbar und wurde in einer Benutzerstudie evaluiert. Die primären Anforderungen an das Tool sind a) die Rückgabe von Suchergebnissen, die neben exakten oder fast exakten Treffern auch solche enthalten, die aufgrund semantischer Ähnlichkeit zustande kommen, sowie b) eine intuitive Bedienung durch Nutzer, die weder mit der Sprache noch mit sonstigen kulturellen Gegebenheiten vertraut sind.

Vorarbeiten

Bereits seit Längerem wird an der digitalen Methodik zur Verarbeitung von Sprachen des Alten Orients geforscht. Dabei spielte insbesondere die automatisierte morphologische Verarbeitung eine Rolle, siehe bspw. Barthélemy (1998) und Kataja (1988). Neuere Arbeiten setzen größtenteils auf statistische Verfahren anstelle von regelbasierten Ansätzen. Darunter fallen bspw. Liu et al. (2015) mit einer Studie zur Lemmatisierung für Sumerisch sowie Homburg und Chiarcos (2016) zur Wort-Segmentierung im Akkadischen. Im Rahmen des ORACC Projekts werden Tools zur Annotation der Morphologie in Keilschriftsprachen entwickelt, überwiegend für Akkadisch und Sumerisch. Zur semantischen Analyse von Keilschrifttexten existieren hingegen kaum Arbeiten. Lediglich Jaworski (2008) entwickelte eine Ontologie für sumerische ökonomische Aktivitäten, die mit einer semantischen Grammatik dargestellt werden können. Einen Überblick über die lexikalisch-semantischen Analyseverfahren, die in dieser Arbeit zum Einsatz kommen, gibt bspw. Gurevych et al. (2016). Soweit uns bekannt ist, gab es bislang keine Studien, die untersuchen, inwiefern Keilschrifttexte bzw. deren Übersetzungen

mittels semantischer-lexikalischer Verfahren für ein breiteres Publikum zugänglich gemacht werden können.

Methodik

Um semantische Suche in Keilschrifttexten zu ermöglichen, haben wir zunächst die transliterierten und übersetzten Texte vorverarbeitet und für die Suche indexiert. Danach werden sie in einer Datenbank abgelegt, in der mittels einer webbasierten Oberfläche gesucht werden kann.

Daten

Die Texte, die im Rahmen dieser Studie verarbeitet wurden, sind überwiegend hethitische, in Keilschrift verfasste Dokumente (Wilhelm 2008). Die Transliterationen und Übersetzungen (auf Deutsch, Englisch, Italienisch und Französisch) wurden an der Johannes Gutenberg-Universität Mainz sowie von Partnern an weiteren Forschungseinrichtungen im In- und Ausland erstellt. Die Originaltexte stammen aus Anatolien (heutige Türkei) und datieren in die zweite Hälfte des 2. Jt. v. Chr. Inhaltlich handelt es sich vornehmlich um religiöse Texte wie bspw. Gebete oder Rituale. Die Dokumente sind auf Satz- oder Teilsatzebene übersetzt und mit den Transliterationen abgeglichen, so dass einfach Bezüge zwischen den Übersetzungen und den Transliterationen hergestellt werden können. Für jedes Dokument existiert ein Einleitungstext, sowie jeweils eine (kommentierte) Übersetzung und eine Transliteration, siehe Abbildung 1. Die Texte sind unabhängig von dieser Arbeit online zugänglich.

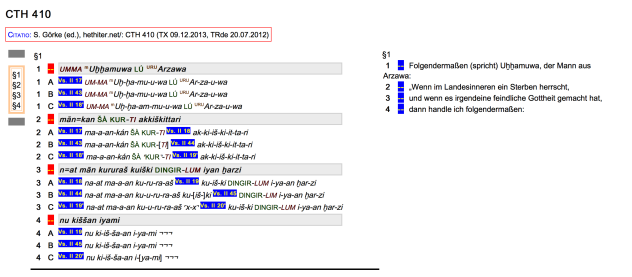

Abbildung 1: Eine manuell erstellte Transliteration (links) und normalisierte Übersetzung (rechts). Quelle: http://www.hethport.uni-wuerzburg.de

NLP Pipeline zur Vorverarbeitung der Texte

Die Übersetzungen und Transliterationen werden direkt aus einem Textformat in eine Pipeline eingelesen, die die weitere linguistische Vorverarbeitung übernimmt. Diese Pipeline erkennt die Struktur der Eingabedokumente, bspw. Dokument-Titel, Sätze, Absätze oder Fußnoten. Außerdem werden die zusammengehörigen Übersetzungen und Transliterationen auf (Teil-)Satzebene gekoppelt. Anschließend werden die mehrsprachigen Übersetzungen mit Hilfe des NLP Frameworks DKPro Core (Eckart de Castilho und Gurevych 2014) analysiert. DKPro Core vereint die Verwendung verschiedener NLP Werkzeuge zur linguistischen Verarbeitung. So ist es möglich, den Inhalt der Dokumente in vier Sprachen zu segmentieren, zu lemmatisieren und nach Wortarten auszuzeichnen. Im nächsten Schritt werden unter Zuhilfenahme des Lesk Algorithmus (Lesk 1986) mehrdeutige Lemmata anhand ihres Kontexts disambiguiert. Dieser Schritt ist die Voraussetzung für die anschließende Zuweisung von sogenannten semantischen Labeln, die einzelne Lemmata mit abstrakteren Konzepten anreichert. Bspw. werden Verben, die eine Bewegung anzeigen, mit einem Label „Bewegung“ gekennzeichnet. Als Ergänzung zu diesen vollautomatischen Verfahren erlaubt es die Pipeline, manuell erstellte Listen für alternative Schreibweisen und Hyperonyme anzuwenden. Darin enthalten sind bspw. geographische Einheiten oder Namen von hethitischen Königen oder Gottheiten, die in den lexikalisch-semantischen Ressourcen, die im Schritt zuvor eingesetzt werden, nicht oder nur teilweise enthalten sind. Bspw. werden verschiedene Namen des Wettergottes (u.a. Taru, Teššup) als solche gelistet. Das Endresultat der Pipeline wird in einem Zwischenformat gespeichert, so dass es anschließend in eine Datenbank importiert werden kann.

Semantische Suchmaschine: Back- und Frontend

Eine MYSQL Datenbank nimmt die Dokumente inklusive der von der NLP Pipeline generierten zusätzlichen semantischen Informationen auf und legt diese in indexierten Tabellen ab. Suchanfragen über das Webinterface werden in entsprechende Abfragen auf die Tabellen übersetzt. Die Anordnung der Suchergebnisse wird über eine Priorisierung der verschiedenen zusätzlichen Informationen geregelt. Wörtliche Treffer werden entsprechend höher gerankt als solche, die durch Übereinstimmung mit semantischen Labeln oder alternativen Schreibweisen zustande kommen.

Das Frontend der Suchmaschine besteht aus dem Eingabefeld für einen oder mehrere Suchbegriffe. Die Suchergebnisse werden pro Dokument gebündelt und angeordnet nach der Güte der Übereinstimmung mit dem Suchbegriff. Abbildung 2 zeigt die Benutzeroberfläche nach einer Suchanfrage. Ein Klick auf ein Suchergebnis öffnet ein Fester, das den Inhalt des gesamten Dokuments jeweils als Übersetzung und Transliteration zeigt.

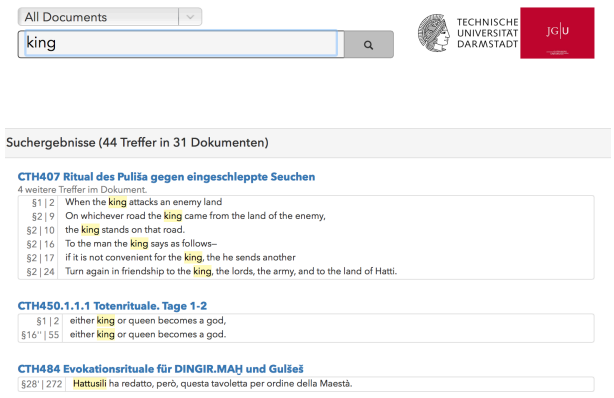

Abbildung 2: Das Frontend mit Ergebnissen zu einer Suchanfrage.

Evaluation

Um zu überprüfen, ob die Suchmaschine die eingangs gestellten Anforderungen erfüllt, haben wir eine anonyme Online-Benutzerstudie mit 23 Fragen unter 27 Teilnehmern durchgeführt. Die Mehrheit der Teilnehmer waren Studierende an deutschen Universitäten. Etwa die Hälfte hatte einen geisteswissenschaftlichen Studienhintergrund, die andere Hälfte einen technischen. Inhaltlich bestand die Benutzerstudie aus einer kurzen Einleitung sowie drei Teilen mit Fragen. Der erste Teil beinhaltete einfache Fragen, die das allgemeine Verständnis der Suchabfragen überprüfen sollten (bspw. Suche nach einem Begriff in einem bestimmten Dokument). Der zweite Teil zielt explizit auf den semantischen Teil der Suchfunktion ab (bspw. Suche nach dem Namen einer Gottheit). Im dritten Teil wurde die allgemeine Bedienbarkeit und Nützlichkeit des Tools erfragt.

Mit wenigen Ausnahmen wurden die Aufgaben aus dem ersten Teil der Benutzerstudie von allen Teilnehmern korrekt gelöst. Im zweiten Teil mussten diverse hethitische Gottheiten, Könige oder Städte namentlich benannt werden, diese Aufgabe konnten sämtliche Teilnehmer korrekt lösen. Eine Frage, in der die (nicht vorhandene) Beziehung zwischen zwei Gottheiten anhand von Suchergebnissen bestimmt werden sollte, wurde nur von etwa zwei Dritteln der Teilnehmer korrekt gelöst. Tabelle 1 fasst die Abfragen und Ergebnisse aus dem dritten Teil der Benutzerstudie zusammen.

Kriterium	Durchschnittswert (1-5)
Sortierung der Ergebnisse basierend auf einer konkreten Suchanfrage	4.15
Benutzerfreundlichkeit der Weboberfläche	4.19
Allgemeine Qualität der Suchergebnisse	4.26
Nützlichkeit für Fachfremde	4.3

Tabelle 1: Kriterien und Bewertungen (Auswahlmöglichkeiten zwischen 1 = sehr schlecht und 5 = sehr gut) des dritten Teils der Benutzerstudie.

Diskussion

In der Gesamtheit zeigen die Ergebnisse der Benutzerstudie, dass das Tool die eingangs gestellten Anforderungen erfüllt. Neben den zu lösenden Aufgaben gab es auch die Möglichkeit, per Freitextfeld Rückmeldung zu geben. Die so identifizierten Probleme sind zurückzuführen auf a) fehlende Erklärungen zur Formulierung von Suchanfragen, b) Fehlern in den manuell erstellten Listen mit alternativen Schreibweisen und Hyperonymen, und c) irreführender Hervorhebung von Wörtern bei Ergebnissen, die auf semantische Übereinstimmung zurückzuführen sind.

Zu den allgemeinen Herausforderungen bei der Aufbereitung der Daten für die semantische Suche zählt u.a. die Fragmenthaftigkeit diverser Texte. Da solche Phänomene zu Fehlern in der NLP Vorverarbeitung (bspw. bei der Segmentierung) führen, wurde eine Komponente in die Pipeline integriert, die Lücken soweit möglich repariert. Der „Vocabulary Gap“ zwischen den Termini in den lexikalisch-semantischen Ressourcen und dem in den Übersetzungen tatsächlich verwendeten Vokabular hat letztlich dazu geführt, dass zusätzlich manuell erstellte Wortlisten eingesetzt wurden. Diese Listen müssen allerdings nur einmal erstellt werden und haben einen überschaubaren Umfang.

Neben der Behebung der oben genannten Probleme ist als nächster Schritte u.a. vorgesehen, das Backend um eine Funktion zum einfachen Upload neuer, transliterierter und übersetzter Texte in die Datenbank zu erweitern. Wir sind zuversichtlich, dass mit dieser Studie ein erster Schritt hin zu einer einfacheren Erschließung des Inhalts keilschriftlicher Quellen genommen ist.

Fußnoten

1. Zugänglich unter http://semsearch.ukp.informatik.tu-darmstadt.de.
2. http://oracc.museum.upenn.edu
3. http://www.hethiter.net

4. Die gesamte Verarbeitungspipeline wurde hier veröffentlicht: https://github.com/UKPLab/DHd2017-semsearch-cuneiform
5. Die Listen können hier eingesehen werden: https://github.com/UKPLab/DHd2017-semsearch-cuneiform
6. Das Geschlecht wurde nicht erfasst, wir beziehen uns jeweils auf alle Teilnehmerinnen und Teilnehmer.

Bibliographie

Barthélemy, François (1998): „A morphological analyzer for akkadian verbal forms with a model of phonetic transformations“, in: *Proceedings of the Workshop on Computational Approaches to Semitic Languages* 73–81.

Cohen, Jonathan / Duncan, Donald / Snyder, Dean / Cooper, Jerrold / Kumar, Subodh / Hahn, Daniel / Chen, Yuan / Purnomo, Budirijanto / Graettinger, John (2004): „iClay: Digitizing Cuneiform“, in: *Proceedings of the International conference on Virtual Reality, Archaeology and Intelligent Cultural Heritage* 135–143.

Eckart de Castilho, Richard / Gurevych, Iryna (2014): „A broad-coverage collection of portable NLP components for building shareable analysis pipelines“, in: *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT at COLING* 1–11.

Gurevych, Iryna / Eckle-Kohler, Judith / Matuschek, Michael (2016): *Linked Lexical-Semantic Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.

Homburg, Timo / Chiarcos, Christian (2016): „Word Segmentation for Akkadian Cuneiform“, in: *Proceedings of the International Conference on Language Resources and Evaluation*.

Jaworski, Woyciech (2008): „Contents Modelling of Neo-Sumerian Ur III Economic Text Corpus“, in: *Proceedings of the International Conference on Computational Linguistics* 369–376.

Kataja, Laura / Koskenniemi, Kimmo (1988): „Finite-state description of semitic morphology: a case study of Ancient Akkadian“, in: *Proceedings of the Conference on Computational Linguistics* 313–315.

Lesk, Michael (1986): „Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone“, in: *Proceedings of the Annual International Conference on Systems Documentation* 24–26.

Liu, Yudong / Burkhart, Clinton / Hearne, James / Luo, Liang (2015): „Enhancing Sumerian Lemmatization by Unsupervised Named-Entity Recognition“, in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* 1446–1451.

Tyndall, Stephen (2012): „Toward Automatically Assembling Hittite-language Cuneiform Tablet Fragments into Larger Texts“, in: *Proceedings of ACL-2012* 243–247.

Wilhelm, Gernot (2008): „Die Edition der Keilschrifttafeln aus Bo#azköy und das Projekt ‚Hethitische Forschungen‘ der Akademie der Wissenschaften und der Literatur, Mainz“, in: Wilhelm, G. (ed.): *Hattuša - Bo#azköy: Das Hethiterreich im Spannungsfeld des Alten Orients*. Wiesbaden: Harrassowitz 73–86.