Das zoroastrische Mittelpersische
Digitales Corpus und Wörterbuch (MPCD)

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln

Mondaca, Francisco
f.mondaca@uni-koeln.de
Universität zu Köln

Eide, Øyvind
oeide@uni-koeln.de
Universität zu Köln

Colditz, Iris
Iris.Colditz@ruhr-uni-bochum.de
Ruhr-Universität Bochum

Jügel, Thomas
Thomas.Juegel@ruhr-uni-bochum.de
Ruhr-Universität Bochum

Rezania, Kianoosh
kianoosh.rezania@rub.de
Ruhr-Universität Bochum

Zeini, Arash
a.zeini@gmail.com
Freie Universität Berlin

Cantera, Alberto
alberto.cantera@fu-berlin.de
Freie Universität Berlin

Emanuel, Chagai
chagai17@gmail.com
Hebrew University Jerusalem

Shaked, Shaul
msshaul@mscc.huji.ac.il
Hebrew University Jerusalem

Einleitung

Mit diesem Beitrag möchten wir das Projekt “Das zoroastrische Mittelpersische - digitales Corpus und Wörterbuch (Middle Persian Corpus and Dictionary, MPCD)” vorstellen, das im April 2021 seine Arbeit aufgenommen hat. Das MPCD-Projekt wird von der DFG als Langfristvorhaben mit einer geplanten Laufzeit von insgesamt neun Jahren gefördert.[1] Das Vorhaben wird als Kooperationsprojekt der Universitäten Bochum, Berlin, Köln und Jerusalem durchgeführt. Während an den Standorten Bochum, Berlin und Jerusalem der Schwerpunkt auf der philologischen Erschließung des Korpus sowie des darauf aufbauenden Wörterbuchs liegt, ist auf Kölner Seite das Cologne Center for eHumanities (CCeH) für die technische Umsetzung einer kollaborativen Recherche- und Arbeitsumgebung zuständig, deren technischen Entwurf wir in diesem Poster-Beitrag thematisieren und zur Diskussion stellen wollen.

Gegenstand und Ziele des Projekts

Das Mittelpersische war als Amts- und Verkehrssprache insbesondere des Sasanidenreiches (3. – 7. Jhd.) von überkultureller und -religiöser Bedeutung. Von der Spätantike bis zur frühislamischen Zeit verbindet es sprachlich und kulturell die differenten Räume des iranischen Ostens und Westens. Das umfangreiche Korpus der mittelpersischen Texte ist bis heute jedoch nur partiell erschlossen. Ziel des MPCD-Projektes ist deshalb die Erstellung eines Korpus zoroastrisch-mittelpersischer Texte in Pahlavi-Schrift. Dieses mit Abstand größte mittelpersische Korpus (ca. 54 Texte, etwa 687.000 Wörter) wird in Transliteration und Transkription (vgl. dazu Rezania 2020) sowie in Handschriftenphotographien der 15 ältesten Codices, die zum Teil aus dem CAB-Projekt von Alberto Cantera (Corpus Avesticum Berolinense)[2]bezogen werden können, zugänglich sein. Die Texte werden morphosyntaktisch und lexikographisch annotiert und in TEI kodiert. Die morphosyntaktische Annotation der Texte folgt dem Standard *Universal Dependencies*[3], der für die Annotation des Mittelpersischen angepasst wurde, indem das Subset der für die Annotation des Mittelpersischen notwendigen Tags bestimmt und die notwendigen pahlavi-spezifischen Tags hinzugefügt wurden.

Auf Grundlage des Korpus wird anschließend ein digitales Mittelpersisch-Englisch-Lexikon mit ca. 7000 Lemmata erstellt. Digitales Korpus und digitales Wörterbuch stellen im Projekt zwei eng verzahnte Analyseinstrumente dar, die mit unterschiedlichen Schwerpunkten – Syntax und Semantik – ineinandergreifen und auch im Arbeitsprozess eng miteinander verbunden sind. Hierbei kommt eine webbasierte Arbeitsumgebung zum Einsatz, die zum einen die kollaborative Bearbeitung von Korpus und Wörterbuch ermöglicht, zum anderen als Nutzer-Interface für Recherchen und Analysen der aufbereiteten Ressourcen dient.

Mit der Einarbeitung der 15 ältesten Codices in ein digitales Korpus mit darauf aufbauendem Wörterbuch schafft das Projekt einen methodisch neuen Zugang zum gesamten zoroastrisch-mittelpersischen Sprachmaterial, der die Voraussetzung für umfassende linguistische und begriffsgeschichtliche Fragestellungen eröffnet. Mit der engen Verzahnung von Text und Wörterbuch ergänzt das Vorhaben bestehende Textsammlungen zum Mittelpersischen wie bspw. TITUS[4] (Thesaurus Indogermanischer Text- und Sprachmaterialien) und bildet eine wertvolle Aktualisierung gegenüber vorliegenden Wörterbüchern wie MacKenzie (1971) oder Nyberg (1964, 1974). Das Projekt bietet eine Grundlage da-

für, das komplexe Gewebe der Texte der zoroastrisch-mittelpersischen Literatur in seinen internen und externen Bezügen zu identifizieren und damit einer (weithin ausstehenden) kultur- und religionshistorisch differenzierten Beschreibung zuzuführen. Es zielt damit darauf, die ‚horizontalen' (d.h. genrespezifischen) und die ‚vertikalen' (d.h. historischen) Differenzen der Texte und ihres Wortschatzes sichtbar werden zu lassen.

Systementwurf

Der Schwerpunkt des Posters liegt auf der Präsentation des Systementwurfs der kollaborativen Recherche- und Arbeitsumgebung sowie der dort eingesetzten Technologien. Dies umfasst zum einen eine Beschreibung der funktionalen Elemente der Nutzerschnittstelle, die dem Anwender als Forschungsumgebung dient, indem sie verschiedene Werkzeuge bereitstellt (z.B. Suche, Verknüpfung von Korpus und Wörterbuch, Export in TEI-Format). Zum anderen werden die Systemarchitektur und die für deren Umsetzung verwendeten Technologien thematisiert.

Die Daten werden mithilfe das Python-Webframework Django[5] modelliert und in PostgreSQL persistiert. Für die Suche in den Daten werden wir Elasticsearch[6] einsetzen. Suche und CRUD-Operationen werden über eine REST-API verfügbar sein, die sich konzeptuell an Vorarbeiten aus dem Projekt VedaWeb[7] orientiert (vgl. dazu Mondaca et al. 2019a, 2019b). Für das Frontend werden wir eine Single-Page-Applikation mit React.js[8] entwickeln.

Wesentliche Funktionen des Frontend sind zum einen die Darstellung der digitalisierten Handschriften und der transliterierten und transkribierten Texte sowie die Möglichkeit zur differenzierten Suche nach linguistischen Parametern, die unter Verwendung von Konzepten aus dem VedaWeb-Projekts implementiert wird (vgl. Kiss et al. 2019). Zum anderen soll das Frontend im Sinne eines Redaktionssystems einen separaten Bereich für die Bearbeitung anbieten, der durch die Nutzerverwaltung nur angemeldeten Benutzern zugänglich ist. Ein solcher Bereich für Korpus und Wörterbuch kann unter Verwendung einer Nutzerverwaltung jeweils als separater View umgesetzt werden, in dem Korrekturen und (Neu-)Eingaben vorgenommen werden.

Mit der Fokussierung auf den Systementwurf möchten wir mit dem Poster vor allem einen kompakten Überblick über die Nutzungsmöglichkeiten sowie über die technische Architektur der geplanten Arbeitsumgebung geben.

Fußnoten

1. https://gepris.dfg.de/gepris/projekt/452473565
2. https://www.geschkult.fu-berlin.de/e/iranistik/forschung/CAB/index.html
3. https://universaldependencies.org/
4. http://titus.uni-frankfurt.de/indexe.htm
5. https://www.djangoproject.com/
6. https://www.elastic.co/
7. https://vedaweb.uni-koeln.de/
8. https://reactjs.org/

Bibliographie

Kiss, Börge / Kölligan, Daniel / Mondaca, Francisco / Neuefeind, Claes / Reinöhl, Uta / Sahle, Patrick (2019): "It Takes a Village: Co-developing VedaWeb, a Digital Research Platform for Old Indo-Aryan Texts." In: Steven Krauwer und Darja Fišer (Hg.), *TwinTalks at DHN 2019 – Understanding Collaboration in Digital Humanities*. Kopenhagen, 2019.

MacKenzie, David N. (1971): *A Concise Pahlavi Dictionary*. London: Oxford University Press.

Mondaca, Francisco / Rau, Felix / Neuefeind, Claes / Kiss, Börge / Kölligan, Daniel / Reinöhl, Uta / Sahle, Patrick (2019a): "C-SALT APIs - Connecting and Exposing Heterogeneous Language Resources." In: *Book of Abstracts of the Digital Humanities Conference 2019 (DH2019)* 09.07-12.07.2019. Utrecht, Netherlands.

Mondaca, Francisco / Schildkamp, Philip / Rau, Felix (2019b): "Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data". In: *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference, Sintra, Portugal*. Brno: Lexical Computing CZ, s.r.o., 907–921.

Nyberg, Henrik S. (1964): *A Manual of Pahlavi. Part I: Texts, Alphabets, Index, Paradigms, Notes and an Introduction.* Wiesbaden: Harrassowitz.

Nyberg, Henrik S. (1974): *A Manual of Pahlavi. Part II: Ideograms, Glossary, Abbreviations, Index, Grammatical Survey, Corrigenda to Part I*. Wiesbaden: Harrassowitz.

Rezania, Kianoosh (2020): "A Suggestion for the Transliteration of Middle Persian Texts in Zoroastrian Middle Persian: Digital Corpus and Dictionary (MPCD): A Three Layered Transliteration System". In: *Estudios Iranios y Turanios* 4: 153–73.