Entwicklungsstand im Projekt 'Digital Plato'

Kath, Roxana
roxana.kath@me.com
Universität Leipzig, Deutschland

Keilholz, Franz
franz.keilholz@tu-dresden.de
Technische Universität Dresden, Deutschland

Pöckelmann, Marcus
marcus.poeckelmann@informatik.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Rücker, Michaela
mruecker1@me.com
Universität Leipzig, Deutschland

Wöckener-Gade, Eva
woeckener-gade@uni-leipzig.de
Universität Leipzig, Deutschland

Yu, Xiaozhou
xiaozhou.yu@tu-dresden.de
Technische Universität Dresden, Deutschland

Einleitung

Das interdisziplinäre Forschungsprojekt *Digital Plato*[1] untersucht die Rezeption und Nachwirkung des platonischen Werkes in der griechischen Literatur bis in die Spätantike mit einem Fokus auf nicht-wörtlichen Referenzen. Der folgende Beitrag und das dazugehörige Poster geben einen Überblick über die wichtigsten Zwischenergebnisse, die während der zurückliegenden ersten Hälfte der Projektlaufzeit erzielt wurden. Neben der Organisation des Korpus sowie dessen linguistischer Anreicherung, dem Aufbau eines Wortnetzes und die Einbettung in einen Vektorraum, zählen dazu die Erfassung und Kategorisierung bekannter Referenzstellen, die Adaption der CTS zur wortgenauen Referenzierung, die theoretische Annäherung und schließlich die semi-automatische Suche nach neuen Paraphrasen.

Teilbereiche

Textgrundlage

Die Textgrundlage des Projekts ist der Thesaurus Linguae Graecae (TLG), der zunächst in ein XML-Format nach TEI-Standard überführtwurde. Da das Textkorpus als statisch anzusehen ist, wir im Laufe des Projektes aber mehrfach zusätzliche Annotationen hinterlegen und aktualisieren möchten, haben wir in einem zweiten Schritt Text und Annotationen nach dem Single-Source Prinzip voneinander getrennt. Dazu wurde der eigentliche Inhalt des Werks als unveränderliche Quelle (Single-Source) in Form einer einfachen fortlaufenden Textdatei angelegt, auf welche wiederum Dateien mit zusätzlichen Informationen (Standoff-Markup) referenzieren.Eine Referenz gibt dabei an, am wievielten Zeichen der Single-Source die Annotation startet und endet. Auch Texthervorhebungen und strukturelle Auszeichnungen des ursprünglichen TLG-Formats wurden so erfasst.

Für jede Form der Annotation wird eine eigene Datei mit Standoff-Markup angelegt, sodass eine Bearbeitung dieser keinen Einfluss auf die übrigen Auszeichnungen hat und sogar überlappende Annotationen ermöglicht werden.

Annotationen

Der TLG enthält hauptsächlich strukturelle Annotationen. Über eine Kombination verschiedener bestehender linguistischer Werkzeuge, wie Morpheus (Crane 1991) und Mate Tagger (Bohnet und Nivre 2012), werden dem Korpus nachträglich Lemmata und morphologische Informationen in Form von Standoff-Markup hinzugefügt. Ferner sollen auf dieser Basis auch die Nominalphrasen automatisch erkannt und ausgezeichnet werden. Solche Angaben helfen bei der Suche nach für das Projekt relevanten Textstellen.

Helleninet

Über die Verknüpfung diverser Wörterbücher wurde im Rahmen des Projekts ein Wortnetz für das Altgriechischegeneriert. Die Helleninet getaufte Struktur stellt ein weiteres wichtiges Standbein für die Einordnung der Relationen zwischen Wörtern und Textstellen dar.

Worteinbettung

Die Wörter eines Korpus können mit Hilfe statistischer Verfahren in einen Vektorraum eingebettet werden, sodass dieser semantische Beziehungen zwischen den Wörtern abbildet. Vorteilhaft hierbei ist, dass lediglich ein hinreichend großes Korpus benötigt wird, da das Verfahren auf den Kontexten der Wörter und nicht auf Vorwissen zur Sprache aufbaut. Das genutzte Verfahren word2vec (Mikolov et al. 2013) erlaubte uns dank seiner Performanz die Durchführung einer umfangreichen Evaluation, um eine für das Projekt möglichst optimale Einbettung zu finden.

Semi-automatische Rezeptionserkennung

Um weitere Referenzen auf Platon im Korpus aufzuspüren, verfolgen wir verschiedene (semi-)automatische Ansätze, die über die aus der Literatur bekannten Ansätze hinausgehen. Beim auf der DHd 2017 vorgestellten 'Rütteln' (Kath et al. 2017) handelt es sich um ein exploratives Verfahren, bei dem interaktiv von einer Textstelle Platons ausgehend einzelne Worte mit sinnverwandten Wörtern (bspw. Synonyme oder Übersetzungen) ersetzt werden und anschießend nach der modifizierten Textstelle im Korpus gesucht wird.

Ein ähnliches, aber systematisches Vorgehen stellt die n-Gramm-Suche dar. Nach einer umfangreichen Normalisierungwerden die n-Gramme verschiedener Längen für das gesamte Korpus indiziert. Anschließend können alle übereinstimmenden n-Gramme effizient ermittelt werden.

Ein drittes Verfahren basiert auf der Word Mover's Distance (Kusner et al. 2015), einem Distanzmaß für zwei Wortgruppen auf Grundlage einer Worteinbettung. Ausgehend von einer Textstelle wird das Korpus hierbei nach Textstellen mit möglichst geringer Distanz durchsucht(Pöckelmann et al. 2017).Die systematische Evaluation an Hand des im Projekt erstellten Goldstandard zeigt, dass dieses Verfahren zu sehr guten Ergebnissen führt.

Referenzierungssystem

Zur wortgenauen Referenzierung von Textstellen im Korpus wurden CTS-URNs adaptiert, d.h. *Uniform Resource Names* nach der Notation der *Canonical Text Services* (Blackwell und Smith 2014). Im Unterschied zum Standard werden die Wörter einer Zeile ebenfalls durchnummeriert, sodass in einer Subreferenz nicht das Wort selbst, sondern dessen Position genutzt werden kann. Um Probleme mit durch Zeilenumbruch getrennten Wörtern zu vermeiden, werden beide Teile in ihrer jeweiligen Zeile mitgezählt. Ein entsprechender Konverter bildet die CTS-URN auf Positionen in den Single-Sources sowie umgekehrt ab.

Goldstandard und Referenzannotierer

Mit Hilfe des im Projekt entwickelten, graphischen Werkzeugs # des Referenzannotierers # wurde eine zuvor erstellte Sammlung bekannter Rezeptionen mit Annotationen verschiedener Kategorien versehen. Der Goldstandard erlaubt durch diese umfassende Kategorisierung eine statistische Auswertung und hilft bei der Begriffsbildung. Zudem bildet er die Grundlage zur systematischen Evaluation der automatischen Suchverfahren und für einen umfassenden Thesaurus.

Begriffsbildung

Das Projekt arbeitet an einer theoretischen Ausdifferenzierung des Paraphrasenbegriffs: als konstitutiv werden hierbei Ähnlichkeiten von Textstellen zueinander angesehen, die sich auf der Wortebene abbilden. Hierbei nehmen wir wie anderen Ansätze die Notwendigkeit eines ‘Dritten’ an, um die Relation von Texten zueinander zu charakterisieren, nur verorten wir dies weniger stark im Bereich der Semantik, der schwer operationalisierbar ist. Vielmehr zielen wir auf eine fruchtbare Synthese dieser Theorie, die Paraphrasen ohne Annahme von Autorenintentionen oder der Bestimmung von Abhängigkeitsverhältnissen zwischen Texten beschreibbar macht, mit bestehenden Ansätzen zur Bestimmung von Ähnlichkeit zwischen Texten aus den DH.

Fußnoten

1. Gefördert durch die VolkswagenStiftung. Weitere Informationen auf der Projektseite unter: https://digital-plato.org/

Bibliographie

Blackwell, Christopher / Smith, Neel (2014): "The Canonical Text Service (CTS)" http://cite-architecture.github.io/cts/ [letzter Zugriff 18. September 2017].

Bohnet, Bernd / Nivre, Joakim (2012): "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing" in: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*: 1455-1465.

Crane, Gregory (1991): "Generating and parsing classical greek" in: *Literary and Linguistic Computing* 6(4):243–245.

Kath, Roxana / Keilholz, Franz / Klinker, Fabian / Pöckelmann, Marcus / Rücker, Michaela / Švitek, Mihael / Wöckener-Gade, Eva / Yu, Xiaozhou (2017): "Paraphrasenerkennung im Projekt Digital Plato" in: *Tagungsband der 4. Jahrestagung der Digital Humanities im deutschsprachigen Raum*: 266-270.

Kusner, Matt J. / Sun, Yu / Kolkin, Nicholas I. / Weinberger, Kilian Q. (2015): "From Word Embeddings To Document Distances" in: *Proceedings of the 32. International Conference on Machine Learning*: 957–966.

Mikolov, Tomas / Sutskever, Ilya / Chen, Kai / Corrado, Greg S. / Dean, Jeff (2013): "Distributed representations of words and phrases and their compositionality" in: *Advances in Neural Information Processing Systems* 26: 3111–3119.

Pöckelmann, Marcus / Ritter, Jörg / Wöckener-Gade, Eva / Schubert, Charlotte (2017): "Paraphrasensuche mittels word2vec und der Word Mover’s Distance im

Altgriechischen"in: Digital Classics Online, Band 3, Ausgabe 3, S. 24-36.