Paraphrasenerkennung im Projekt Digital Plato

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

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

Klinker, Fabian
fabian.klinker@tu-dresden.de
Technische Universität Dresden

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

Rücker, Michaela
mruecker1@me.com
Universität Leipzig

Švitek, Mihael
mihael.svitek@tu-dresden.de
Technische Universität Dresden

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

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

Einleitung

Platons Werke wurden seit ihrer Entstehung bis in die heutige Zeit vielfach rezipiert und direkt zitiert, seine enorme Wirkungsmacht ist kaum zu unterschätzen, wie A.N. Whiteheads (1929, 63) berühmter Ausspruch verdeutlicht: “ The safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato “. Mit dem von der VolkswagenStiftung geförderten Projekt *Digital Plato: Tradition and Reception* unter Leitung von Prof. Dr. Paul Molitor, Dr. Jörg Ritter, Prof. Dr. Joachim Scharloth, Prof. Dr. Charlotte Schubert und Prof. Dr. Kurt Sier wird seit April 2016 das Vorhaben verfolgt, diese Rezeption und Nachwirkung Platons bei den griechischen Autoren bis zur Spätantike systematisch zu untersuchen und zwar über das möglichst umfassende Auffinden von Paraphrasen. Der vorliegende Beitrag beschreibt die damit einhergehende grundlegende Problematik am Beispiel und skizziert einen derzeit in Entwicklung befindlichen Ansatz zu deren Lösung.

Die Textgrundlage

Die Werke Platons haben die Zeit weitestgehend überdauert und sind frei digital verfügbar . In den Handschriften sind 43 Werke überliefert, mit Ausnahme der Apologie und der 13 Briefe alle in Dialogform verfasst. Heute werden die Dialoge in neun Gruppen zu je vier Schriften gruppiert (sogenannte Tetralogienordnung). Damit sind wir in der exzeptionellen Lage, alle Werke Platons untersuchen zu können, die in der Antike bekannt waren, auch wenn einige davon ihm fälschlicherweise zugeschrieben wurden oder die Autorenfrage umstritten ist. Diejenigen sieben Werke, die die sogenannte *Appendix Platonica* formen und schon in der Antike für nicht platonisch gehalten wurden, liegen vorerst nicht im Projektfokus (Erler 2006, 27-36).

Die Tetralogien haben einen Umfang von knapp 75.000 Zeilen. Dem gegenüber steht das wesentlich umfangreiche Gesamtwerk der antiken griechischen Autoren, das mit dem *Thesaurus Linguae Graecae* (TLG) in digitaler Form vorliegt und über neun Millionen Textzeilen umfasst.

Problemaufriss Paraphrasenerkennung

Um die Rezeption und Nachwirkung von Platons Werk in der antiken griechischen Literatur untersuchen zu können, sollen Übereinstimmungen zwischen seinen Texten und denen späterer Autoren im TLG gefunden werden. Dies geht bei weitem über das Identifizieren wörtlicher Zitate hinaus, da es das möglichst zuverlässige Auffinden paraphrasiert wiedergegebener Textstellen umfasst. Der Paraphrasenbegriff selbst wird im Rahmen des Projekts derzeit mit dem Arbeitskonzept der ‚Relation’ bestimmt: Wie solche Relationen zwischen dem platonischen Werkkorpus und der übrigen griechischen Literatur der Antike aussehen und welche Aspekte damit erfasst werden können, soll folgendes Beispiel veranschaulichen:

Pl. symp. 206 d 1-2

####μ##### ## #### ## ####### ##### ## ####, ## ## ##### ##μ#####.

Unvereinbar aber ist das Hässliche mit allem Göttlichen, aber das Schöne ist vereinbar.

Plot. enn. III 5, 1, 19-20

μ## ### ####### ######## ### ## ##### ### ## ###.

Denn das Hässliche ist sowohl der Natur als auch dem Gott entgegengesetzt.

Beim Paraphrasieren einer Textstelle kann es zu verschiedenen Phänomenen kommen. Neben fast wortwörtlicher Übernahme einer Textstelle (ggf. mit Auslassungen) können in den Textfluss eingewobene Zitate mit umgestelltem Satzbau auftreten. Das Beispiel geht darüber hinaus: Die wörtliche Übereinstimmung beschränkt sich auf einen geläufigen Ausdruck („das Hässliche“). Zudem wird der Inhalt einerseits nur teilweise wiedergegeben (die Vereinbarkeit vom „Schönen“ und „Göttlichen“ fehlt), andererseits um Neues erweitert (das „Hässliche“ ist nun auch der „Natur“ entgegengesetzt). Mit dem Synonym „ist entgegengesetzt“ statt „ist unvereinbar“ tritt eine lexikalische Varianz in Erscheinung. Zudem wurde das substantivierte Adjektiv „dem Göttlichen“ samt seines attributiven Zusatzes „allem“ durch das Nomen „dem Gott“ ersetzt.

Ferner sind bspw. die Verwendung von Antonymen oder Metaphern denkbar, die die Erkennung einer Rezeption zusätzlich erschweren.

Vorarbeiten

Da sich die aufzufindenden paraphrasierten Textstellen nicht auf einen beliebigen Autor, sondern auf Platon beziehen, ergeben sich einige Vorteile für die Suche. Der überschaubare Umfang der Texte ermöglicht die manuelle bzw. teil-automatisierte Extraktion von Informationen aus den Werken Platons. Dazu gehören Listen mit den vorkommenden Substantiven, Verben, Eigennamen oder Stoppwörtern. Aber auch die Auflistung der zentralen Konzepte der platonischen Philosophie ist für die spätere Paraphrasenerkennung hilfreich. Zudem liegen verschiedene Übersetzungen in elektronischer Form vor, die in das Projekt einfließen .

Einen großen Gewinn stellt auch die Vorarbeit des an der Universität Leipzig durchgeführten Projektes eAQUA dar, welches ein Werkzeug zur Zitationsanalyse entwickelt hat und damit die vorkommenden Zitate im Korpus bereitstellt.

Für die Bewertung und Extraktion von Paraphrasen aus einem Textkorpus gibt es bereits verschiedene Ansätze, wie Androutsopoulos und Malakasiotis (2010) in einem Übersichtsartikel zusammengetragen haben. Diese basieren häufig auf einer Kontextanalyse und der Annahme, dass Worte in einem ähnlichen Kontext auch eine ähnliche Bedeutung haben. So können für jedes Textsegment einer festen Länge (n-Gramme, meist mit n # 5) die Kontexte aller Vorkommen betrachtet und als ein Vektor repräsentiert werden. Ähnlichkeitsmaße auf Vektoren erlauben nun den Vergleich zweier Textsegmente. Für das Auffinden von Paraphrasen müssen auf diesem Weg alle Textsegmente miteinander verglichen werden. Allerdings sind die so extrahierten Paraphrasen bzw. -fragmente sehr kurz, im Gegensatz zu den teils umfangreichen Rezeptionen, die im Rahmen des Projekts gefunden werden sollen. Zielführender sind Vorgehen, die zunächst Anker, d.h. eine Gemeinsamkeit zwischen zwei Textstellen, suchen und in einem zweiten Schritt die Fundstellen ausweiten. Solche Anker können bspw. einzelne Wörter, die oben beschriebenen n-Gramme sowie syntaktische oder semantische Repräsentationen einer Textstelle sein. Naheliegend ist, die Fundstellen im zweiten Schritt auf den umliegenden Satz auszuweiten. Stattdessen kann auch die Sinneinheit über semantische Informationen rekonstruiert werden, wie ein erfolgreich auf englischsprachige Korpora angewandtes Verfahren von Regneri, Wang und Pinkal (2014) aufzeigt.

Viele der bestehenden Verfahren zum Extrahieren von Paraphrasen erlangen ihre Effektivität mittels umfangreicher Annotationen der zu Grunde liegenden Texte, welche durch Parser mit einem gewissen Wirkungsgrad automatisch bestimmt werden können. Dieser Wirkungsgrad ist wiederum stark von zu Grunde liegenden Trainingskorpora und damit der Sprache der betrachteten Texte abhängig. So ist die Entwicklung für moderne Sprachen sehr weit vorangeschritten. Die Anwendung auf Altgriechisch ist hingegen deutlich seltener und auf weniger umfangreiche Korpora beschränkt (Mambrini und Passarotti 2012). Einer der ersten Vertreter ist das regelbasierte Analysewerkzeug Morpheus, das unter anderem Lemmata bestimmen kann (Crane 1991). In einer aktuellen Studie von Celano, Crane und Majidi (2016) wurden fünf aktuelle POS-Tagger mit Hilfe der Ancient Greek Dependency Treebank (Bamman und Crane 2011) trainiert und auf ihre Wirksamkeit getestet, wobei der Mate-Tagger mit einer Genauigkeit von 88% am besten abschnitt. Das entsprechende Modell wurde dem Projekt zur Verfügung gestellt. Auch wenn die Parser sich stets weiterentwickeln, bleiben insbesondere die lange Zeitspanne und die vielfältigen Genres in dem von uns betrachteten Korpus problematisch, sodass die Parser und damit auch die darauf aufbauenden Verfahren zur Paraphrasenerkennung qualitativ schlechtere Ergebnisse produzieren als für moderne Sprachen (Dik und Whaling 2008).

Umsetzung im Projekt

Das verbreitete Vorgehen zur Extraktion von Paraphrasen über die Suche von Ankern wird für dieses Projekt durch die in Abschnitt 4 beschriebenen Vorarbeiten und die Entwicklung einer interaktiven Arbeitsumgebung für Suche und Auswertung von Paraphrasen praktikabel. Zwei Anwendungsszenarien sind dabei zu unterscheiden: die Suche ausgehend von einem Textstück und das Auffinden möglichst aller Rezeptionen Platons im Korpus. Der Fokus der aktuellen Arbeiten liegt zunächst in der ersten Aufgabe, ist ihre Bewältigung doch Grundlage für die zweite. Im Folgenden wird ein erster Ansatz beschrieben, der derzeit umgesetzt wird.

Ausgehend von einem Textstück, wie einem Satz von Platon, werden geeignete Anker für die Suche gewählt. Statt dabei alle Wörter zu berücksichtigen, kann die Auswahl auf Basis der angefertigten Listen auf bestimmte Wortarten oder auf die Begriffe der platonischen

Philosophie beschränkt werden. Diese erste Vorfilterung reduziert die Anzahl der Suchwörter auf möglichst aussichtsreiche Kandidaten, um die anschließende Auswertung handhabbar zu halten. Dennoch ist eine gewisse Unschärfe, d.h. die Erweiterung eines Suchwortes zu einer Menge verwandter Worte, sinnvoll, um eine ganze Reihe von möglichen Rezeptionen abzudecken. Das Suchwort wird dazu durch die Verknüpfung von Wortrelationen erweitert, bspw. um seine Synonyme sowie verschiedene Übersetzungen samt deren Synonyme, die wiederum ins Altgriechische zurückübersetzt werden (siehe Abbildung 1).

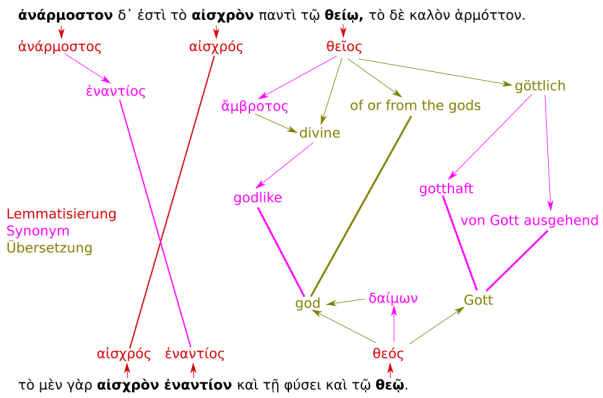

Abb. 1: Durch die Verknüpfung von Lemmatisierung, Synonymen und Übersetzungen kann das Suchwort #### ('dem Göttlichen') um ### ('der Gott') erweitert und so ein drittes Ankerpaar für das Beispiel gefunden werden.

Diese Erweiterung von Wortrelationen wird erfolgreich für die Verschlagwortung von Themen in Briefkorpora (Hildenbrandt et al. 2015) und in ähnlicher Form zum Aufbau eines *WordNet* für Altgriechisch (Bizzoni et al. 2014) genutzt. Ergänzt um zusätzliche Relationen, wie bspw. Hyperonym- respektive Hyponymbeziehungen, wird die Suche robuster gegenüber verschiedenen Formen der Paraphrasierung. Dabei gilt: Je mehr möglichst kurze Verbindungen zwischen zwei Wörtern liegen, desto größer ist die Aussagekraft dieses Paares. Wahrscheinliche Kandidaten für Rezeptionen sind dann Textstellen, in denen sehr viele aussagekräftige Anker nahe beieinander wiedergefunden werden.

Durch die Unschärfe des Verfahrens sind auch viele Kandidaten zu erwarten, die keine Paraphrasen sind und nicht als Rezeption des platonischen Werkes angesehen werden können. Die Ergebnisse sollen daher durch eine Arbeitsumgebung zunächst automatisch bewertet und sortiert werden. Ausgehend von einzelnen Treffern und einer transparenten Visualisierung, wie das System zur Entscheidung gelangte, soll eine interaktive Exploration der Texte die effiziente Recherche ermöglichen. Das beinhaltet das Wichten bzw. Entfernen einzelner Relationen sowie das manuelle Einordnen der gefunden Textstellen. So können Fallbeispiele und Phänomene näher untersucht, aber auch neue entdeckt werden. Die qualifizierte Bewertung auf der Basis der Fachexpertise von Altertumswissenschaftlern hilft wiederum, die Sammlung bereits bekannter Rezeptionen zu erweitern und die zu Grunde liegenden Algorithmen zu verbessern.

Die aus eAQUA bekannten Zitate sind für den Beginn des Projekts eine wichtige Unterstützung. Sie erlauben einen ersten Einblick in den Umfang der Rezeption Platons. Über die Verteilung lassen sich besonders häufig zitierte Passagen ermitteln, was möglicherweise auch Rückschlüsse auf die Fundstellen von Paraphrasen zulässt. Eine naheliegende, aber zu prüfende Hypothese ist, dass häufig zitierte Stellen auch anderweitig übernommen wurden. Das könnte zum zeitnahen Auffinden bisher unentdeckter Paraphrasen führen bzw. eine aufwendige Untersuchung an diesen Stellen rechtfertigen, um an besonders interessante Fallbeispiele zu gelangen.

Fußnoten

1. Siehe bspw. Perseus Digital Library http://www.perseus.tufts.edu/hopper/collection?collection=Perseus%3Acorpus%3Aperseus%2Cauthor%2CPlato
2. Eine entsprechende Liste findet sich bei Gigon und Zimmermann (1974, 301ff.)
3. Siehe bspw. Schleiermacher (Deutsch) oder Ü. Fowler (Englisch)
4. Siehe http://www.eaqua.net/
5. Siehe https://code.google.com/archive/p/mate-tools/

Bibliographie

Androutsopoulos, Ion / Malakasiotis, Prodromos (2010): „A Survey of Paraphrasing and Textual Entailment Methods“, in: *Journal of Artificial Intelligence Research* 38: 135–187.

Bamman, David / Crane, Gregory (2011): „Ancient Greek and Latin dependency treebanks“, in: *Language Technology for Cultural Heritage* 79–98 DOI:10.1007/978-3-642-20227-8_5.

Bizzoni, Yuri / Boschetti, Federico / Diakoff, Harry / Del Gratta, Riccardo / Monachini, Monica / Crane, Gregory (2014): „The Making of Ancient Greek WordNet“, in: *Proceedings of LREC 2010*.

Celano, Giuseppe G. A. / Crane, Gregory / Majidi, Saeed (2016): „Part of Speech Tagging for Ancient Greek“, in: *Open Linguistics* 2 (1), ISSN (Online) 2300–9969 10.1515/opli-2016-0020.

Crane, Gregory (1991): „Generating and Parsing Classical Greek“, in: *Literary and Linguistic Computing* 6 (4): 243–245 10.1093/llc/6.4.243

Dik, Helma / Whaling, Richard (2008): „Bootstrapping Classical Greek Morphology“, in: *DH2016: Book of Abstracts* 105–106.

Erler, Michael (2006): *Platon*. München: C.H.Beck.

Gigon, Olof / Zimmermann, Laila (1974): *Platon. Begriffslexikon*. Zürich: Artemis Verlag.

Hildenbrandt, Vera / Kamzelak, Roland S. / Molitor, Paul / Ritter, Jörg (2015): „im Zentrum eines Netzes [...] geistiger Fäden - Erschließung und Erforschung thematischer Zusammenhänge in heterogenen Briefkorpora“, in: *Datenbank-Spektrum : Zeitschrift für Datenbanktechnologie*: 15 (2015, 1): 49–55.

Mambrini, Francesco / Passarotti, Marco (2012): „Will a parser overtake Achilles? First experiments on parsing the Acient Greek Dependency Treebank“, in: *11th International Workshop on Treebanks and Linguistic Theories*, Lisbon, Portugal.

Regneri, Michaela / Wang, Rui / Pinkal, Manfred (2014): „Aligning predicate-argument structures for paraphrase fragment extraction“, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Whitehead, Alfred North (1929): *Process and Reality: An Essay in Cosmology*. New York.