Maschinelle Sprachverarbeitung für die Klassische Philologie

Sprachverarbeitung ist die wohl wichtigste Grundlage der Philologie. Wir versuchen, Inschriften und mittelalterliche Handschriften zu entziffern, um antike Texte zu rezipieren. Ist uns die Entzifferung einmal gelungen, dann beschäftigen wir uns intensiv mit dem Sprachgebrauch bei bestimmten Gattungen (Cordes 2020, S. 33-43), Personen (Devine & Stephens 2006, S. 452) oder sogar nur einzelnen Werken. Diese starke Ausrichtung auf das Verständnis antiker Sprachen schlägt sich auch in den Lehrplänen nieder (Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen 2019, S. 13). All diese Schritte können maschinell unterstützt werden: Texterkennung, Textstrukturierung, grammatische Analyse und Suche gehören mittlerweile zum Standardrepertoire der einsetzbaren technischen Hilfsmittel für die Lektüre antiker Texte. Die ursprüngliche Motivation zu deren Nutzung ist klar: Je schneller wir auf Texte zugreifen und bestimmte Passagen darin finden können, umso mehr Zeit bleibt uns für die Interpretation, also den Teil, der maschinell bisher kaum unterstützt wird. Aber wie genau kann diese Arbeitsteilung zwischen Mensch und Computer aussehen?

Textgrundlage

Frühneuzeitliche Drucke lateinischer Texte erfordern großen Aufwand, um als digitale Texteditionen einem größeren Publikum zugänglich gemacht zu werden: Sie besitzen oft eine weniger standardisierte Typografie sowie Orthografie und sind mitunter von erheblichem materiellen Verfall gekennzeichnet, der auf den jahrhundertelangen Prozess der Nutzung, Lagerung und Alterung zurückzuführen ist (Springmann & Lüdeling 2017, S. 2). Umso schwieriger wird es bei Handschriften, die in der Regel noch variabler gestaltet und noch älter sind als Drucke (Diem 2010, S. 9). Dennoch gibt es hier erstaunliche Fortschritte, was die automatisierte Erkennung von Schriftzeichen – Optical Character Recognition (OCR) – und deren Übertragung in digitale Formate angeht: Für eine altgriechische Handschrift des Aëtius von Amida wurden mit Hilfe von OCR4all,[1] nach minimaler Vorbereitung, Erkennungsraten von über 95% für die Buchstaben erreicht (Reul et al. 2019, S. 28). Die manuelle Korrektur dieser Vorarbeit nimmt dann noch Einiges an Zeit in Anspruch, allerdings lohnt sich der Einsatz solcher Technologien mitunter schon bei Textpassagen mit nur wenigen Sätzen. Da im Idealfall nicht einmal jeder 20. Buchstabe falsch erkannt wird, beschränkt sich die Korrekturarbeit auf wenige Sekunden pro Satz. Bei monumentalen Editionen wie dem *Corpus Inscriptionum Latinarum*[2] mit über 200.000 Inschriften ist eine solche maschinelle Vorarbeit von unschätzbarem Wert. Sie ermöglicht eine schnellere Erweiterung der existierenden und den Aufbau vieler neuer digitaler Editionen, was angesichts der Millionen von Werken der neulateinischen Literatur (Korenjak 2016, S. 22) ein zentrales Anliegen sein muss, unter anderem zur Erforschung der Rezeptionsgeschichte lateinischer Klassiker. Ein Großteil dieser riesigen Textmenge ist immer noch unerschlossen, obgleich Initiativen wie das *Corpus Corporum*[3] (Roelli 2014) mit seinen über 160 Millionen Wörtern hier Abhilfe zu schaffen versuchen.

Creativ Collection Verlag GmbH

AD ASTRA – Innovationen für den Unterricht

Nachwuchswettbewerb für Latein und Griechisch

Der Deutsche Altphilologenverband (DAV) und der Ernst Klett Verlag schreiben für das Jahr 2021/22 zum zweiten Mal den Nachwuchswettbewerb für Latein und Griechisch aus. Dieser Wettbewerb AD ASTRA richtet sich an junge Lehrkräfte im Referendariat sowie in den ersten fünf Berufsjahren. Eingereicht werden kann eine eigene und in der Praxis selbst erprobte Idee, die ein innovatives Element enthält: eine kluge, clevere und vielleicht auch mutige methodische oder didaktische Neuerung. Diese Idee sollte das Lernen der Schülerinnen und Schüler in den Mittelpunkt stellen, die Freude am Fach wecken und auf andere Lerngruppen übertragbar sein. Die Idee muss schlüssig, überzeugend und nachvollziehbar dargestellt werden.

Bitte reichen Sie zur Teilnahme am Wettbewerb folgende Unterlagen ein:

- Deckblatt (Name und Anschrift der Schule / Thema / Jahrgangsstufe(n) / Postanschrift, Telefonnummer und E-Mail-Adresse der Bewerberin/des Bewerbers),
- Darstellung der Idee und ihrer Umsetzung unter Benennung des innovativen Elements, max. 3 Seiten DIN A4 (PDF),
- Unterrichtsmaterialien (PDF, PPT, MPEG, MP3, MP4 etc.) als Anhang unter Angabe der verwendeten Quellen und Literatur, insgesamt max. 15 MB,
- Bestätigung des Bewerbers/der Bewerberin, dass es sich um eine eigene und selbst erprobte Idee handelt,
- Kurzvita (im Schuldienst seit …).

Teilnahmebedingungen:
Referendarinnen und Referendare können prüfungsrelevante Lerneinheiten aus ihren schriftlichen Arbeiten und Lehrproben vor dem Abschluss der Ausbildung weder in Teilen noch als Ganzes einreichen. Eine Jury aus Fachleuten des DAV und des Ernst Klett Verlages trifft eine Auswahl aus den Einsendungen und befindet über die Zuerkennung der Preise. Das Preisgeld wird vom Ernst Klett Verlag gestiftet. Für Platz eins werden 750 €, für Platz zwei 500 € und für Platz drei 250 € ausgelobt. Die Verleihung der Preise findet im Rahmen des DAV-Kongresses in Würzburg im April 2022 statt. Im Falle der Platzierung werden die Teilnehmer zum Kongress eingeladen, um ihre Idee vorzustellen. Ferner wird die Veröffentlichung der prämierten Ideen angestrebt.
Der Beitrag ist einzureichen per E-Mail an: adastra@altphilologenverband.de. Einsendeschluss ist der 31.10.2021
Der Rechtsweg ist ausgeschlossen.

DEUTSCHER ALTPHILOLOGENVERBAND

Klett

Vernetzte Sprachdaten

Doch selbst wenn uns alle erhaltenen altsprachlichen Texte digital zur Verfügung stünden, könnten wir noch nicht ohne Weiteres damit arbeiten. Um unsere Beobachtungen und Forschungen, unsere Interpretationen und Hypothesen mit anderen zu teilen, müssen wir klar und eindeutig kommunizieren, auf welchen Text wir uns beziehen. Informationen wie Autor, Werk, Textpassage und Textausgabe unterliegen dabei einem Standardisierungsprozess, wie er sich in der Abkürzungsliste des Neuen Pauly für antike Textreferenzen niederschlägt. Eine ähnliche, kostenlos zugängliche Form der Kanonisierung ging aus der Textsammlung *PHI Latin Texts* hervor und mündete in der Zuweisung von einzigartigen Identifikatoren für jede beliebige altsprachliche Textstelle in den *Canonical Text Services* (Tiepmar et al. 2014). Über eine entsprechende Schnittstelle kann dann also nicht nur auf die Texte verwiesen, sondern auch ihr Wortlaut direkt abgerufen und durch etwaige Zusatzmaterialien (Übersetzungen, Kommentare etc.) ergänzt werden, wie es in *Alpheios*[4], in der *Perseus Digital Library*[5] und im *Scaife Viewer*[6] umgesetzt wurde.

Dieser Gedanke der expliziten Vernetzung vorhandener digitaler Ressourcen ist das Kernstück des Prinzips von *Linked Open Data* (Cayless 2019). Dabei geht es darum, der zunehmenden Fragmentierung von Forschung entgegenzuwirken, die aus der Nutzung unterschiedlicher Datenmodelle und -formate hervorgeht. Beispiele dafür sind die Verwendung unterschiedlicher grammatischer Begriffe zur Erklärung von antiker Syntax oder die Speicherung von Texteditionen als Word-, XML- sowie PDF-Dokumente. Ein gängiger Ansatz zur Vernetzung, der sich von der lokalen bis auf die globale Ebene erstreckt, ist anschaulich zu beobachten in der Infrastruktur des LiLa-Projekts: Dort werden wissenschaftlich aufbereitete Textsammlungen wie PROIEL[7] mit kontrollierten Vokabularen wie *Ontolex*[8] verknüpft (Mambrini et al. 2020). Als Vokabular ist in solchen Fällen nicht der Wortschatz eines antiken Werks zu verstehen, sondern – etwas abstrakter – eine einheitliche sprachliche Form zur Beschreibung von Wissen. In diesem Fall sind damit oft Identifikatoren in Form von URLs gemeint (z. B. http://www.w3.org/ns/lemon/ontolex#MultiwordExpression), die als zentrale Anlaufstelle für alle Forschenden dienen, die in ihren Texten eine Information hinzufügen möchten. So dient beispielsweise das Vokabular *Ontolex* dazu, die konkrete sprachliche Umsetzung von kommunikativen Inhalten zu markieren. Die gegebene Beispiel-URL repräsentiert die Information „Hierbei handelt es sich um einen Mehrwortausdruck". Wenn nun also Forschende in einem antiken Text auf eine Phrase wie *cursus honorum* stoßen, können sie die URL zu der Textstelle hinzufügen und beziehen sich dabei nicht auf ihre eigene, subjektive Definition von Mehrwortausdruck, sondern auf eine zentrale, mit anderen Gleichgesinnten ausgehandelte Definition von Mehrwortausdruck. Gegenüber einem analogen oder intuitiven Zugang ergeben sich hier Vorteile wie eine explizite Definition der gesuchten sprachlichen Information (Mehrwortausdruck) sowie die Nachnutzbarkeit der Forschungsdaten durch andere Forschende. Letzteres ist angesichts der oben beschriebenen überwältigenden Menge unerforschter Literatur von besonderer Bedeutung.

Fortgeschrittene sprachliche Analysen durch Künstliche Intelligenz

Wo solche hilfreichen Informationen noch nicht professionell erarbeitet wurden, können

sie durch Verfahren der Künstlichen Intelligenz ergänzt werden. So liefern verschiedene Werkzeuge zunehmend verlässlichere sprachliche Analysen für antike Texte: Die Morphologie und Grundform von Wörtern kann mithilfe von *LemLat*[9] oder *LatMor*[10] bestimmt werden. Häufige Kombinationen mehrerer Wörter, auch im direkten Vergleich mehrerer Textstellen, lassen sich in *Tesserae*[11] ausfindig machen. Kompliziertere syntaktische Analysen, z. B. verschiedene Formen der Reflexivität in der *oratio obliqua*, werden zumindest ansatzweise durch *UDPipe*[12] geliefert und lassen sich dann übersichtlich in *Arethusa*[13] darstellen. Allerdings bezieht sich die automatische Verarbeitung antiker Texte bisher überwiegend auf die sprachwissenschaftlichen Grundlagen. Für die eigentliche literaturwissenschaftliche Interpretation liegen bisher kaum überzeugende Hilfsmittel vor.

Erste vielversprechende Ansätze in die Richtung der Semantik und Hermeneutik sind jedoch in den letzten Jahren zunehmend auf dem Vormarsch. Hierzu zählt insbesondere die Anwendung von fortgeschrittenen Methoden des Maschinellen Lernens auf antike Texte. Sprugnoli et al. 2020 und Bamman & Burns 2020 zeigen überzeugend, wie mit neueren Technologien der Künstlichen Intelligenz antike Texte inhaltlich analysiert werden können. Sei es nun die Abgrenzung des Gebrauchs eines speziellen Worts zwischen zwei Textsammlungen (z. B. *sacer* in paganer und in christlicher Literatur), die nuancierte Unterscheidung verschiedener Bedeutungen desselben Worts innerhalb eines Textes (z. B. *in* als Präposition bei Teilungsprozessen) oder die Bestimmung von Paralleltexten für eine bestimmte Zielpassage (z. B. die Proömien von Vergils *Aeneis* und Ovids *Amores*): Die genannten Forschenden haben zweifelsfrei demonstriert, dass moderne Sprachtechnologie auch für die Bearbeitung komplexer philologischer Fragestellungen eingesetzt werden kann (vgl. auch Pöckelmann et al. 2019, S. 60, zur automatischen Erkennung von Paraphrasen).

Zu schön um wahr zu sein? Es gibt einen Haken: Die beschriebenen Innovationen wurden bisher hauptsächlich von technisch versierten Angehörigen der *Digital Humanities* vorangetrieben. Für solche Methoden gibt es in der Klassischen Philologie noch keine *Community of Practice*, also keine Gruppe von Forschenden, die regelmäßig entsprechende Werkzeuge nachnutzt, ohne sie selbst entwickelt zu haben. Darum sind viele Probleme und Unwägbarkeiten dieser Sprachmodelle noch nicht so weit erforscht und beseitigt, dass von einem hohen Reifegrad und reibungsloser Einsatzfähigkeit gesprochen werden könnte. Was hier fehlt, ist einerseits eine Verbreitung des notwendigen Wissens in den existierenden Gemeinschaften, um solche Technologien anwenden zu können. Damit einher ginge dann andererseits eine umfangreiche Erhebung der konkreten Anforderungen und eine fachlich begleitete Pilotierung der jeweiligen Werkzeuge. Die transparente, offene Zugänglichkeit des entsprechenden Quellcodes und der zugehörigen wissenschaftlichen Publikation sind der erste essentielle Schritt in eine Richtung, die es uns zukünftig ermöglichen wird, methodische Innovationen schneller und nachhaltiger in der Forschungslandschaft zu verankern.

Schlussfolgerungen

Zusammenfassend lässt sich also festhalten, dass elementare sprachliche Analysen mittlerweile hervorragend maschinell unterstützt werden können. Dazu gehören optische Zeichenerkennung, die Erstellung von Texteditionen sowie die

musterbasierte Suche und Referenzierung von Textpassagen. Etwas kompliziertere Techniken wie *Linked Open Data* oder die Bestimmung von Wortarten und syntaktischen Funktionen genießen momentan großes Interesse, sind aber bisweilen fehlerbehaftet und benötigen darum etwas mehr Aufwand zur Korrektur der Ergebnisse. Sie verzeichnen allerdings auch große Fortschritte in der Weiterentwicklung, weshalb hier von einer zunehmenden Einsatzreife in den nächsten Jahren ausgegangen werden muss. Als vielversprechendster Neuankömmling im Bereich der maschinellen Sprachverarbeitung gilt momentan das *Natural Language Understanding* (Beyer et al. 2021), also die Erschließung von Textinhalten durch Künstliche Intelligenz. Mit seinen ungleich komplexeren Sprachmodellen zeigt es hervorragende Ansätze zur Aufarbeitung komplizierter philologischer Fragen, die bisher als technisch unlösbar galten. Dazu zählt etwa die detaillierte Untersuchung von Wortbedeutungen bis hinunter auf die Ebene einzelner Sätze und unter Berücksichtigung des jeweiligen Kontextes. Je nach Bedarf können dann auch relevante Parallelstellen identifiziert und für die weitere Interpretation hinzugezogen werden, wobei die Parallele nicht, wie früher oft üblich, nur in zitierten Wortgruppen, sondern auch in vagen Anspielungen gefunden werden kann. Dadurch werden z. B. wichtige Forschungsfragen zur Intertextualität in der altsprachlichen Literatur unterstützt.

Links:

1) http://www.ocr4all.org/de/home.php
2) https://cil.bbaw.de/hauptnavigation/das-cil/geschichte-des-cil
3) http://www.mlat.uzh.ch/MLS/
4) https://alpheios.net/
5) http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman
6) https://scaife.perseus.org/
7) https://proiel.github.io/
8) https://www.w3.org/2016/05/ontolex/
9) http://www.lemlat3.eu/
10) https://www.cis.uni-muenchen.de/~schmid/tools/LatMor/
11) https://tesserae.caset.buffalo.edu/
12) https://lindat.mff.cuni.cz/services/udpipe/
13) https://www.perseids.org/tools/arethusa/app/#/

Literatur:

Bamman, D., & Burns, P. J. (2020): Latin BERT: A Contextual Language Model for Classical Philology. ArXiv Preprint ArXiv:2009.10053, S. 1-14.

Beyer, A., Schulz, K., & Cordes, L. (2021): BridgeClassics. Künstliche Intelligenz für die Klassische Philologie. https://doi.org/10.5281/zenodo.4745781.

Cayless, H.A. (2019): Sustaining Linked Ancient World Data, in: M. Berti (Hrsg.), Digital classical philology: Ancient Greek and Latin in the digital revolution (Vol. 10, S. 35-50), Berlin/Boston.

Cordes, L. (2020): Wenn Fiktionen Fakten schaffen. Faktuales und fiktionales Erzählen in den spätantiken Panegyrici Latini, in: D. Breitenwischer, H.-M. Häger, & J. Menninger (Hrsg.), Faktuales und fiktionales Erzählen II. Geschichte – Medien – Praktiken (S. 31–56), Baden-Baden. https://doi.org/10.5771/9783956505126-31.

Devine, A. M., & Stephens, L. D. (2006): Latin Word Order: Structured Meaning and Information, Oxford.

Diem, M., & Sablatnig, R. (2010): Recognizing Characters of Ancient Manuscripts. Proc. SPIE 7531, Computer Vision and Image Analysis of Art, 7531, S. 1-12. https://doi.org/10.1117/12.843532.

Korenjak, M. (2016): Geschichte der neulateinischen Literatur: Vom Humanismus bis zur Gegenwart, München.

Mambrini, F., Cecchini, F. M., Franzini, G., Litta, E., Passarotti, M. C., & Ruffolo, P. (2020): LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web. Umanistica Digitale, 4.8, S. 63-78.

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (Hrsg.) (2019): Kernlehrplan für die Sekundarstufe I Gymnasium in Nordrhein-Westfalen. Latein. https://www.

schulentwicklung.nrw.de/lehrplaene/lehrplan/206/g9_l_klp_3402_2019_06_23.pdf
Pöckelmann, M., Ritter, J., & Molitor, P. (2019): Word Mover's Distance angewendet auf die Paraphrasenextraktion im Altgriechischen, in C. Schubert, P. Molitor, J. Ritter, K. Sier, & J. Scharloth (Hrsg.), Platon Digital. Tradition und Rezeption (S. 45-60). Propylaeum Heidelberg. https://books.ub.uni-heidelberg.de/propylaeum/reader/download/451/451-30-84795-1-10-20190507.pdf.
Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., & Puppe, F. (2019): OCR4all – An open-source tool providing a (semi-) automatic OCR workflow for historical printings. Applied Sciences, 9.22, S. 1-30. https://doi.org/10.3390/app9224853.
Roelli, P. (2014): The Corpus Corporum, a new open Latin text repository and tool. Archivum Latinitatis Medii Aevi-Bulletin Du Cange (ALMA).
Springmann, U., & Lüdeling, A. (2017). OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. Digital Humanities Quarterly, 11.2, Article 2.
Sprugnoli, R., Moretti, G., & Passarotti, M. (2020): Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. IJCoL. Italian Journal of Computational Linguistics, 6 (6-1), S. 29-45. https://doi.org/10.4000/ijcol.624.
Tiepmar, J., Teichmann, C., Heyer, G., Berti, M., & Crane, G. (2014): A new implementation for canonical text services. Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), S. 1-8. https://www.aclweb.org/anthology/W14-0601.

Konstantin Schulz

Ovid, Vater Rumäniens

Der Titel dieses Aufsatzes – Theodor Haeckers „Vergil, Vater des Abendlandes" nachempfunden – ist gewiss eine plakative Kurzformel, aber er soll die besondere Beziehung der Rumänen zu Ovid auf den Punkt bringen.

Ovidiu

Wer ‚Ovidiu' in eine Internet-Suchmaschine eingibt, wird feststellen, dass sich die große Masse der Fundstellen nicht auf den Schöpfer der *Metamorphosen*, sondern auf unzählige rumänische ‚Namensvettern' bezieht, so beliebt ist Ovidiu als männlicher Vorname. Nicht nur das. Man wird bei dieser Recherche auf Stadt und Insel bei Constanța stoßen, die – eine nicht nur in Europa unübliche Form der Ehrung – beide den Namen des Dichters tragen. Jedes rumänische Geschichtsbuch, ob für Erwachsene oder für Jugendliche, enthält einen ausführlichen Hinweis auf Ovid, angereichert durch Zitate aus seinem Werk – eben nicht aus den *Metamorphosen*, sondern aus den *Tristia* und den *Epistulae Ex Ponto*.

Diese besondere Verbundenheit Rumäniens und der Rumänen mit dem Dichter hat unterschiedliche Gründe. Zunächst: Der Verbannte von Tomi – nicht Herodot, Strabon oder Vergil im Skythenexkurs der *Georgica* (3,349-383) – ist derjenige antike Autor, der das ausführlichste und anschaulichste Bild vom realen Leben in der Dobrudscha, der Keimzelle des romanisierten Rumänien, geliefert hat, in kräftigen, wenn auch düsteren Farben.

Und: Ovid gilt den Rumänen als Begründer ihrer Nationalliteratur. Traditionelle spanische Literaturgeschichten beginnen mit Seneca (Córdoba), Martial (Calatayud), Lucan (Córdoba), Columella (Cádiz); die Verbindung Ovids mit Rumänien aber ist weitaus enger. Er war ja der Erste, der auf dem Boden des ‚Römerlandes'