Das Latin Text Archive (LTA) – Digitale Historische Semantik von der Projektentwicklung an der Universität zur Institutionalisierung an der Akademie

Geelhaar, Tim
geelhaar@em.uni-frankfurt.de
Goethe Universität Frankfurt am Main, Deutschland

Einleitung

Die Arbeit zur digitalen historischen Semantik in Frankfurt am Main erreicht mit der neuen webbasierten Plattform und Datenbank „Latin Text Archive“ (LTA)[1] im Rahmen der Dateninfrastrukturen der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) ein neues Level. Nach über zehn Jahren Entwicklungsarbeit an den Datenbanken „Historical Semantics Corpus Management (HSCM)“[2], „Frankfurt Latin Lexicon (FLL)“ sowie an der Webseite „www.comphistsem.org“[3] ergaben sich drei grundlegende Herausforderungen: (1) Nachhaltigkeit und Verfügbarkeit der geleisteten Arbeit mussten gesichert, (2) Benutzerfreundlichkeit und Funktionalität verbessert sowie (3) Akzeptanz und Verankerung innerhalb der Fachwissenschaft gesteigert werden. Das LTA antwortet auf diese Herausforderungen und führt mit historischen Referenzkorpora neue Arbeitsinstrumente ein. Das LTA ist auch das Ergebnis einer neuen strategischen Partnerschaft, um die bisherige Arbeit aus der stets zeitlich begrenzten Projektentwicklung an der Universität in einen dauerhaften Betrieb an einer Akademie zu überführen.

Zur Standortbestimmung der Frankfurter Arbeit lässt sich Michael Piotrowskis Definition von *digital humanities* (Piotrowski 2018: 2) anwenden: Danach ist sie den *applied digital humanities* zuzuorten, da ein konkretes geschichtswissenschaftliches Forschungsziel mit digitalen Techniken verfolgt wird. Der Historiker Bernhard Jussen hat zur Umsetzung seines Postulates einer „kulturellen Semantik“ nach Wegen gesucht, wie computerlinguistische Methoden helfen können, Bedingungen, Mittel und Formen multimedialer Sinnproduktion in vergangenen Gesellschaften zu erforschen (Jussen 2000: 24ff.). Es geht um die kontrollierte Analyse von semantischem Wandel innerhalb der lateinischen Textproduktion in poströmischer Zeit. Ein Anwendungsgebiet ist die politische Geschichte. So lässt sich mittels der computergestützten Semantik danach fragen, welche impliziten politischen Ordnungsmodelle in Texten sichtbar werden, bevor mit der Wiederentdeckung der Politik des Aristoteles der Begriff des Politischen aufkommt? Außerdem kann die Begriffs- und Ideengeschichte Verwendungszusammenhänge von zentralen Vokabeln untersuchen (Geelhaar 2015; Schwandt 2018). Am Ende sollen aber nicht nur Forschungsergebnisse und -methoden, sondern auch Arbeitsinstrumente und Forschungsdaten einem breiten fachwissenschaftlichen Publikum ohne Programmierkenntnissen zur Weiterverwendung bereitstehen. Hierzu hat Bernhard Jussen den Computerlinguisten Alexander Mehler und dessen Text Technology Laboratory[4] für eine Kooperation gewonnen, in der die texttechnologische und die geschichtswissenschaftliche Seite ihre jeweiligen Agenden verfolgen und vom interdisziplinären Austausch profitieren. Aus dieser Konstellation ergibt sich, dass die Vorstellung des LTA aus geschichtswissenschaftlicher Perspektive den Fokus nicht auf technische bzw. technologische, sondern auf programmatische und anwenderbezogene Aspekte legt.

Das Latin Text Archive: Teil des DTA-Markenstamms

Das LTA ist eine frei zugängliche, webbasierte Plattform zu Korpusaufbau und Korpusanalyse sowie auch eine Datenbank. Technisch baut es auf den am Deutschen Textarchiv (DTA, DFG-gefördert zwischen 2007 und 2016, siehe Geyken et al. 2018) entwickelten Komponenten zur Textpräsentation auf und erweitert somit den Markenstamm des DTA um eine lateinische Textkomponente. Diese umfasst die Textproduktion im lateinischsprachigen Europa von (zunächst) 400 bis 1500. Die versammelten Texte basieren auf kritischen und somit für die Geschichtswissenschaft validen Editionen, soweit sie in Open Access verfügbar sind. Sie werden zum Zweck des Text Mining um den kritischen Apparat gekürzt[5], im TEI-P5-Format nach dem für HSCM entwickelten Datenmodell aufbereitet, vollständig lemmatisiert und mit einer aufwändigen Metadaten-Annotierung angereicht, die zugleich die Vernetzung zu anderen digitalen Ressourcen herstellt – nicht zuletzt zu den textgebenden Institutionen selbst.[6] Hierbei handelt es sich u. a. um die *Monumenta Germaniae Historica* (MGH). Dieses wichtige deutsche Editionsunternehmen für mittelalterliche Texte stellt seine Editionen im „openMGH“-Projekt unter Creative-Commons-Lizenz in TEI-konformen Versionen zur Verfügung.[7] Doch erst durch die Datenintegration ins LTA können auch reine Anwender vom openMGH-Projekt profitieren, da die Editionen nun erst wortstatistisch vergleichend analysiert werden können. Des Weiteren ist das LTA auf kontrollierte

Datenerweiterung durch die gezielte Aufbereitung und Übernahme aus projektexternen Quellen ausgelegt, um in Zukunft ein repräsentatives Korpus historischer, lateinischer Textproduktion analysierbar zu machen. Hierzu können die Daten entweder als Gesamtkorpus und nach freier Auswahl durch den Anwender an Analysemodule weitergereicht werden, von denen die Voyant Tools bereits verfügbar sind.[8] Die in den Vorgängerprojekten entwickelten Analysetools werden als weiteres, externes Modul zugänglich gemacht. Dabei handelt es sich um Konkordanz- und Kookkurrenzanalysen sowie die Berechnung semantischer Netzwerke.

Zusammenhang zwischen LTA und HSCM

Das LTA unterscheidet sich in mehrfacher Hinsicht von seinen Vorgängeranwendungen. Die Überführung des Datenbestandes aus HSCM in das LTA als Teil der von der BBAW betreuten Dateninfrastruktur dient dem Zweck der nachhaltigen Verfügbarkeit. Zudem wird das LTA als explizites Parallelangebot zum DTA vom Renommee der BBAW profitieren, die durch eigene Datenprojekte nicht nur eine ausgezeichnete Expertise in den DH vorweisen kann, sondern auch bereits hohe Anerkennung in den Geisteswissenschaften genießt. Zudem gewinnt das LTA durch die Anlehnung an das DTA, da es einen Wiedererkennungseffekt in der Benutzerführung gibt, der das Arbeiten mit dem LTA erleichtert. Die wesentlichen Neuerungen gegenüber HSCM bestehen aber nicht nur in der verbesserten Benutzerführung; wichtiger noch ist die Trennung der Primärdatenaufbereitung von der Datenverwaltung, indem die Daten in HSCM kuratiert und im LTA zur Verfügung gestellt werden. Das Preprocessing neuer Texte wird weiterhin über HSCM als Teil des eHumanities- Desktops[9] laufen und über den eigens vom Text Technology Lab entwickelten TT Lab Tagger (vor der Brück/Mehler 2016; Eger/Gleim/Mehler 2016) für die automatische Lemmatisierung lateinischer Texte, über das dafür nötige morphologische Lexikon („Frankfurt Latin Lexicon“) sowie über Editoren zur kontrollierenden, manuellen Nachlemmatisierung und zur nachträglichen Korrektur des TEI-Codes. Die vollständig bearbeiteten Texte werden anschließend in das LTA überführt, wo es nicht mehr möglich sein wird, in den jeweiligen Source-Code des Textes einzugreifen. Dies erlaubt eine feste Indexierung (auch der Lemmatisierungsinformationen), wodurch die Schnelligkeit bei der Verarbeitung von Suchanfragen bedeutend gesteigert wird. Darüber hinaus ist durch den Datentransfer die Analyse des Materials nicht mehr auf die in HSCM vorhandenen Tools beschränkt, sondern können im Grunde von allen denkbaren Toolkits wie eben den Voyant Tools oder Diacollo[10] weiterverwendet werden.

Geschichtswissenschaftliche Referenzkorpora

Die dritte wesentliche Neuerung sind die unter geschichtswissenschaftlichen Aspekten kontrollierten Referenzkorpora[11], um Veränderungen im Sprachgebrauch zeitlich wie genrespezifisch berechnen und visualisieren zu können. Wie die DTA-Referenzkorpora beinhalten diese Referenzkorpora ganze Werke und nicht nur Samples wie linguistische Korpora (z. B. das British National Corpus). Das Clustering von Texten wird nach Vierteljahrhunderten und nicht nach Zehn-Jahres-Schritten wie im DTA geschehen[12], weil eine präzisere zeitliche Zuordnung aufgrund fehlender Datierungen und mitunter komplizierten Überlieferungsgeschichten nicht möglich ist. Dieser Arbeitsschritt erforderte zudem eine erneute klassische Quellenkritik zur Chronologie einzelner Texte. Die Textmengen pro Zeiteinheit sollen quantitativ nicht zu sehr voneinander abweichen, was angesichts der teilweise eklatanten Disparität historischer Schriftproduktion eine große Herausforderung darstellt. Außerdem wird, soweit möglich, der Verbreitungsgrad handschriftlicher Überlieferung berücksichtigt, wenngleich das LTA ansonsten an der Idee des Textes als abstrakter Größe aus konzeptionellen Gründen festhalten muss.[13] Das erste Referenzkorpus besteht aus narrativen Texten, die aus den Scriptores-Reihen der MGH stammen und historiographische wie hagiographische Texte beinhalten. Künftige Korpora werden Briefe bzw. Urkunden und vor allem auch theologische bzw. juristische Traktate umfassen, um somit Sprachgebrauch in verschiedenen Genres vergleichen zu können.

Fußnoten

1. http://lta.bbaw.de
2. Jussen/Mehler/Ernst 2007; Cimino/Geelhaar/Schwandt 2015. HSCM wurde zwischen 2008 und 2014 aus den Mitteln des Gottfried Wilhelm Leibniz-Preises der DFG sowie aus den Mitteln des LOEWE-Schwerpunktes „Digitale Humanities“ finanziert, um im BMBF-Projekt „Computational Historical Semantics“ (2013-2016) weiterentwickelt zu werden.
3. Eine Präsentation der Frankfurter Projekte für das CCeH 2017 findet sich unter: http://www.geschichte.uni-frankfurt.de/43013259/geelhaart (alle folgenden Links wurden eingesehen am 6.10.2018)
4. https://www.texttechnologylab.org/
5. Kritisch hierzu Fischer 2017: S266.
6. Es gibt Verlinkungen, soweit möglich, zum Repertorium Fontium Mediae Aevi (www.geschichtsquellen.de) , zu VIAF (Personen und Werke) und zum Katalog der Staatsbibliothek zu Berlin für bibliographische Angaben.
7. http://www.mgh.de/dmgh/openmgh/

8. https://voyant-tools.org/ Zu dessen Anwendung in der Geschichtswissenschaft siehe Schwandt 2018: 125-133.
9. HSCM ist ein Modul der VRE „eHumanities Desktop 2.2“ (www.hudesktop.hucompute.org) des Text Technology Laboratory.
10. https://clarin-d.de/de/kollokationsanalyse-in-diachroner-perspektive
11. Zu den Schwierigkeiten mit historischen Korpora und von Historikern organisierten Korpora siehe Geelhaar 2015: 11f.
12. Unser Kooperationspartner IRHT/CNRS verfolgt im Corpus-Building-Project VELUM (http://www.agence-nationale-recherche.fr/Project-ANR-17-CE27-0015) eine sehr viel gröbere Stratifikation.
13. Eine Korpusanreicherung mittels digital edierter Handschriften ist technisch in HSCM/LTA realisierbar, würde aber zu Inkonsistenzen im Materialbestand führen. Hierzu auch Fischer 2017: S280.

Bibliographie

Vor der Brück, Tim / Mehler, Alexander (2016): *„TLT-CRF: A Lexicon-supported Morphological Tagger for Latin Based on Conditional Random Fields”*, in: “Proceedings of the 10th International Conference on Language Resources and Evaluation”.

Eger, Steffen/ Gleim, Rüdiger / Mehler, A. (2016): *„Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art”*, in: “Proceedings of the 10th International Conference on Language Resources and Evaluation”.

Cimino, Roberta / Geelhaar, Tim / Schwandt, Silke (2015): *“Digital Approaches to Historical Semantics: new research directions at Frankfurt University”*. In: Storicamente 11. http://storicamente.org/historical_semantics [letzter Zugriff 12.10.2018] 7. DOI: 10.12977/stor594

Fischer, Franz (2017): *„Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Critism”*, in: Speculum 92/S1: S266-S287. https://doi.org/10.1086/693823

Geelhaar, Tim (2015): *“Talking About christianitas at the Time of Innocent III (1198–1216): What Does Word Use Contribute to the History of Concepts?”* in: Contributions to the history of concepts 10/2: 7–28. https://doi.org/10.3167/choc.2015.100202

Geyken, Alexander / Boenig, Matthias / Haaf, Susanne / Jurish, Bryan / Thomas, Christian / Wiegand, Frank (2018): *„Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN“*, in: **Lobin, Henning/ Schneider, Roman / Witt, Andreas (Hgg.)**: *Digitale Infrastrukturen für die germanistische Forschung* (= Germanistische Sprachwissenschaft um 2020, Bd. 6). Berlin/Boston: De Gruyter, 219–248. https://doi.org/10.1515/9783110538663-011

Jussen, Bernhard, Mehler, Alexander / Ernst, Alexandra (2007): *„A Corpus Management System for Historical Semantics. Sprache und Datenverarbeitung“*, in: International Journal for Language Data Processing 31/2: 81-87.

Jussen, Bernhard (2000): *Der Name der Witwe. Erkundungen zur Semantik der mittelalterlichen Bußkultur*. (VMPIG, Bd. 158). Göttingen.

Piotrowski, Michael (2018): *„Digital Humanities – An explication”*, in: **Burghardt, Manuel, Müller-Birn, Christian (eds.)**: *INF-DH 2018 – Workshopband*, 25. Sept. 2018, Berlin https://doi.org/10.18420/infdh2018-07

Schwandt, Silke (2018): *„Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora“*, in: Geschichte und Gesellschaft 44: 107-134. https://doi.org/10.13109/gege.2018.44.1.107

Mehler, Alexander / vor der Brück, Tim / Gleim, Rüdiger / Geelhaar, Tim (2015): *„Towards a Network Model of the Coreness of Texts: An Experiment in Classifying Latin Texts using the TTLab Latin Tagger”*, in Text Mining: From Ontology Learning to Automated text Processing Applications, C. Biemann and A. Mehler, Eds., Berlin/New York: Springer, 2015, pp. 87-112.

Mehler, Alexander / Schwandt, Silke / Gleim, Rüdiger / Jussen, Bernhard: *„Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien”"*, Journal for Language Technology and Computational Linguistics (JLCL), vol. 26, iss. 1, pp. 97-117, 2011.