Suche und Visualisierung von Annotationen historischer Korpora mit ANNIS. Kritik der korpuslinguistischen Analysemethoden in einem erweiterten Nutzungskontext

Odebrecht, Carolin
carolin.odebrecht@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Krause, Thomas
krauseto@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Guescini, Rolf
rolf.guescini@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Kühnlenz, Frank
frank.kuehnlenz@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Lüdeling, Anke
anke.luedeling@hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Dreyer, Malte
malte.dreyer@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Historische Korpora (Gippert und Gehrke 2015; Claridge 2008; Rissanen 2008) dienen in vielen geisteswissenschaftlichen Disziplinen als Analysegrundlage und können sehr unterschiedlich aufbereitet sein. Mit korpusbasierten Studien können qualitative und quantitative Analysen, die für die Überprüfung von Hypothesen über ein bestimmtes Phänomen notwendig sind, durchgeführt werden. Dem gegenüber steht methodisch die korpusgetriebene Studie, die das Korpus selbst nutzt, um Hypothesen über ein Phänomen zu generieren (vgl. McEnery und Hardie 2012; Lüdeling und Zeldes 2007). Neben diesen zwei Studientypen können mit Hilfe von Korpora auch einzelne Belege und Kontexte für die Beantwortung verschiedenster Forschungsfragen ermittelt werden.

Eine andere methodische Unterscheidung wird mit dem *close reading* und dem *distant reading* gemacht (vgl. Moretti 2016; Federico 2015; Gooding et al. 2013; Simanowski 2011). Wobei *close reading* hermeneutische, nicht zwingend digitale Methoden umfassen und eine methodische Nähe zu den korpusinformierten Belegstudien sowie zu qualitativen korpusbasierten Studien aufweisen kann. *Distant reading* ist hingegen vergleichbar mit überwiegend quantitativen, korpusgetriebenen Studien. Ein zusätzlicher Aspekt dieser Methoden ist auch die Visualisierung der Daten für *distant reading* oder des Textes für *close reading*. Verschiedene Visualisierungen der Annotationen werden für die korpusbasierten, -getriebenen und -informierten Studien eingesetzt und können so verschiedene Analysen unterstützen oder auch erst ermöglichen.

In den digitalen Geisteswissenschaften müssen daher für die jeweiligen Methoden und Forschungsdaten Analyse- und Visualisierungswerkzeuge entwickelt werden, die es den Forscherinnen und Forschern ermöglichen, für ihren jeweiligen Forschungskontext aus einem breiten methodischen Spektrum wählen zu können (vgl. für einen Überblick z.B. Kupietz und Geyken 2016). Ein solches Werkzeug ist ANNIS (Krause und Zeldes 2016), das Such- und Visualisierungstool für Annotationen, das wir in unserem Workshop den Forscherinnen und Forschern aus den Digital Humanities vorstellen möchten. ANNIS erlaubt das Durchsuchen von Korpora, die unterschiedliche Arten von Annotationen, die möglicherweise durch unterschiedliche Forschergruppen unter verschiedenen Gesichtspunkten annotiert worden, in einem Korpus vereinen. Diese Flexibilität erlaubt es, annotierte Phänomene in der Suche zu kombinieren und damit komplexere Strukturen zu finden.

Neben der Unterstützung der vielfältigen Analysemethoden ist eine weitere Herausforderung für die Analysewerkzeuge, dass historische Korpora je nach Forschungskontext und -frage unterschiedlich erstellt und aufbereitet werden (Lüdeling 2011). Dies zeigt sich unter anderen in den vielfältigen Transkriptions- und Normalisierungsverfahren (vgl. z.B. Odebrecht et al. 2016; Krasselt et al. 2015; Archer et al. 2015; Bollmann et al. 2012; Jurish 2010) und Annotationsguidelines (für z.B. Annotation von Wortarten für historisches Deutsch Coniglio et al. 2016; Dipper et al. 2013) sowie verschiedenen Formaten (z.B. Romary et al. 2015; Schmidt und Wörner 2009; Burnard und Baumann 2008; Wittenburg et al. 2006; Dipper 2005), die allein für die Erstellung von historischen Korpora eingesetzt werden.

Damit historische Korpora mit verschiedenen Methoden analysiert werden können, muss deren Wiederverwendung ermöglicht werden. Die Wiederverwendung von historischen Korpora wird durch u.a. deren freie Veröffentlichung und umfassende Dokumentation möglich (Odebrecht 2014; Borgmann 2012; Büttner et al. 2011). Weiterhin erhöht eine Wiederverwendung ihre Sichtbarkeit und stellt eine Chance zur engeren Vernetzung und Zusammenarbeit in den digitalen Geisteswissenschaften

dar. So können auch historische Korpora in unterschiedlichen Wiederverwendungsszenarien gedacht werden (vgl. Simons und Bird 2008) und als empirische Grundlage für die verschiedenen Analysemethoden dienen.

Dieser Workshop möchte ausgehend von diesen Themenkomplex mit den Teilnehmerinnen und Teilnehmern folgende Fragen diskutieren: Wie können Analysewerkzeuge den Forscherinnen und Forschern vielfältige Analysemethoden und Visualisierungsmethoden für verschiedene historische Korpora ermöglichen? Wie kann ANNIS die verschiedenen Analysemethoden bislang unterstützen? Wie kann es gelingen, auch die Vielfältigkeit der Forschungsdaten als solche zu berücksichtigen und deren Wiederverwendung zu ermöglichen? Wie können Werkzeuge spezifisch genug entwickelt werden, um genaue und für den Forschungskontext und die Forschungsdaten angepasste Analysen zu ermöglichen?

Der Workshop hat das Ziel, anhand mehrerer historischer Korpora des Deutschen das generische Such- und Visualisierungstool ANNIS (Krause und Zeldes 2016) für den Einsatz in den Digital Humanities zu diskutieren und anzuwenden, da es bislang überwiegend für korpusbasierte und korpusgetriebene Studien sowie für das Auffinden von sprachlichen Belegen eingesetzt wird.

ANNIS wird seit 2009 als ein generisches webbasiertes Such- und Visualisierungstool für verschiedene Korpustypen und Annotationskonzepte in verschiedenen Kooperationen mit der Humboldt-Universität zu Berlin und der Georgetown University und in mehreren Projekten entwickelt. Der Quellcode von ANNIS ist frei zugänglich veröffentlicht und bietet gleichzeitig eine Desktop- sowie Server-Installation. In ANNIS können Korpora mit Token-, Spannen-, Baum- und Pointingannotationen unabhängig von den einzelnen, jeweils korpusspezifischen Annotationsguidelines in ANNIS analysiert werden. ANNIS bietet weiterhin den Korpuserstellerinnen und -erstellern annotations- oder fachspezifische Visualisierungen für Korpora. Mit einer wiederum generischen und mächtigen Anfragesprache (ANNIS Query Language – AQL) können alle Korpora in ANNIS nach Annotationen und Kombinationen von Annotationen durchsucht werden. Weiterhin können die Suchergebnisse für bspw. weitere statistische Auswertungen exportiert werden. Jedes Korpus, jede Suchanfrage und jeder Beleg kann über einen permanenten Link stabil referenziert werden. Mit dem Konverterframework Pepper (Zipser und Romary 2010) werden Korpora, die in verschiedenen Formaten vorliegen können, in das ANNIS-Format überführt.

Repositorien wie das LAUDATIO-Repository (Odebrecht et al. 2015) ermöglichen einen Open Access Zugang zu verschiedensten historischen Korpora und stellen eine umfassende Korpusdokumentation (Odebrecht 2014) zur Verfügung, die eine Erschließung dieser heterogenen Datengrundlage unabhängig von den Korpuserstellerinnen und -erstellern ermöglicht. Damit wird eine Voraussetzung für die Wiederverwendung der historischen Korpora erfüllt. Für den Workshop werden aus LAUDATIO beispielhaft die Korpora „Referenzkorpus Altdeutsch“ (Donhauser 2015) und „RIDGES Herbology Korpus“ (Odebrecht et al. 2016) verwendet.

Das Referenzkorpus Altdeutsch ist ein historisches Mehrebenenkorpus der ganzen Sprachperiode des Althochdeutschen mit ca. 650.000 Wörtern (von den ersten Überlieferungen bis Mitte des 11. Jahrhunderts). Als Grundlage für die diplomatischen Transkription sind Editionen der jeweiligen Handschriften, die mit weiteren Annotationen zur Textstruktur sowie mit komplexen Wortartenannotation (Dipper et al. 2013), Annotation zu Flexionsklassen und Lemmatisierung versehen sind. Das RIDGES Korpus ist ein tief annotiertes Korpus mit Auszügen aus gedruckten Kräuterbüchern aus der Zeit zwischen 1487 und 1910, anhand derer die Entwicklung der deutschen Wissenschaftssprache auf vielen Ebenen untersucht wird. Die Drucke sind diplomatisch transkribiert (wo möglich, nach vorher digitalisierten oder durch OCR-Verfahren erstellte Vorlagen, vgl. Springmann und Lüdeling 2017). Die Daten sind mehrfach normalisiert und auf vielen Ebenen annotiert (unter anderem mit Wortart, Lemma, Informationen zu Kompositionstypen (Perlitz 2014), Dependenzsyntax, Informationen zur graphischen Struktur nach den TEI Guidelines). Dabei werden automatische und manuelle Annotationsverfahren und Prüfverfahren genutzt.

Um die eingangs formulierten Fragen adressieren zu können, wird der Workshop zwei Schwerpunkte enthalten. Der erste Schwerpunkt wird die Einführung in die Funktionen und Suchanfragesprache von ANNIS sowie die damit verbundene Vorstellung der zwei historischen Beispielkorpora umfassen. Wir wollen den Teilnehmerinnen und Teilnehmern die verschiedenen Analyse- und Visualisierungsmöglichkeiten online und hands-on vorstellen. Über die Vorstellung zweier historischer Korpora mit dem generischen ANNIS können bereits die Herausforderungen der heterogenen Datengrundlage in den digitalen Geisteswissenschaften für Analysetools herausgearbeitet werden.

Der zweite Schwerpunkt soll Raum für eine Diskussion mit den Teilnehmerinnen und Teilnehmern sowie auch die Möglichkeit geben, weitere Korpora in ANNIS – geleitet von den Forschungsinteressen der Teilnehmerinnen und Teilnehmern – zu durchsuchen. Mit diesem Workshop wollen wir uns gemeinsam mit den Teilnehmerinnen und Teilnehmern kritisch mit den Anforderungen an ein Analysetool für verschiedene Methoden zur Analyse und Visualisierung von historischen Korpora auseinandersetzen und prüfen, in wie weit ANNIS bereits einige dieser Anforderungen erfüllen kann. So wollen wir ANNIS in einem neuen Forschungskontext der Digital Humanities diskutieren und dabei neue Nutzerszenarien für die weitere Entwicklung erarbeiten.

Zeitplan:

- 1,5 Stunden Online Hands-on-Einführung in das Such- und Visualisierungstool, der Anfragesprache mit dem Referenzkorpus Altdeutsch und RIDGES Korpus

- 1,5 Stunden Teilnehmergeleitete Anfragen, weitere Korpora und Diskussion der Anforderungen

Technische Anforderungen:

- Für den Workshop werden ein Raum mit Beamer und Zugang zu eduroam, ggf. einzelne WLAN-Zugänge für die Teilnehmerinnen und Teilnehmer benötigt.
- Alle Teilnehmerinnen und Teilnehmer benötigen eigenes Notebook.

Teilnehmeranzahl:

- max. 30

Bibliographie

Moretti, Franco (2016): *Distant Reading*. Konstanz: Konstanz University Press.

Romary, Laurent / Zeldes, Amir / Zipser, Florian (2015): "<tiger2/>. Serialising the ISO SynAF syntactic object model", in: *Language Resources and Evaluation* 49 (1), S. 1–18. 10.1007/s10579-014-9288-x.

Archer, Dawn / Kytö, Merja / Baron, Alistair / Rayson, Paul (2015): "Guidelines for normalising Early Modern English corpora. Decisions and justifications", in: *ICAME Journal* (39), 5–24.

Bollmann, Marcel / Dipper, Stefanie / Krasselt, Julia / Petran, Florian (2012): "Manual and semi-automatic normalization of historical spelling. Case studies from Early New High German", in: *Proceedings of KONVENS 2012* 342–350. http://www.oegai.at/konvens2012/proceedings/51_bollmann12w/ [letzter Zugriff am 24.08.2016].

Borgmann, Christine L. (2012): "The conundrum of sharing research data", in: *Journal of the American Society for Information Science and Technology* 63 (6), 1059–1087. 10.2139/ssrn.1869155.

Burnard, Lou / Baumann, Sid (eds.) (2008): *TEI P5. Guidelines for Electronic Text Encoding and Interchange*. Oxford. http://www.tei-c.org/Guidelines/P5/ [zuletzt geprüft am 11.11.2015].

Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (2011): "Research Data Management", in: Büttner, Stephan / Hobohm, Hans-Christoph / Müller, Lars (eds.): *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen, 13–23.

Claridge, Claudia (2008): "Historical Corpora", in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook*, Bd. 1. 2. Berlin: De Gruyter (1), 242–259.

Coniglio, Marco / Donhauser, Karin / Schlachter, Eva / Rasskazova, Oxana / Odebrecht, Carolin / Wirth, Matthias / Miltenberger, Anke (2016): *Historisches Predigtenkorpus zum Nachfeld (HIPKON Version 1.0).* Technischer Bericht, Humboldt-Universität zu Berlin. 10.18452/13681.

Dipper, Stefanie (2005): "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation", in: *Proceedings of Berliner XML Tage 2005 (BXML 2005). Berlin*, 39–50.

Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): "HiTS. Ein Tagset für historische Sprachstufen des Deutschen“, in: Zinsmeister, Heike / Heid, Ulrich / Beck, Kathrin (eds.): Das Stuttgart-Tübingen Wortarten-Tagset. Stand und Perspektiven", in: *Journal for Language Technology and Computational Linguistics, 28(1)*, 85–137.

Donhauser, Karin (2015): "Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten", in: Gippert, Jost / Gehrke, Ralf (eds.): *Historical Corpora.* Tübingen: Narr (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 5), 35–49.

Federico, Annette (2015): *Engagements with Literature. Engagements with Close Reading.* Florence: Routledge.

Gippert, Jost / Gehrke, Ralf (eds.) (2015): *Historical Corpora*. Tübingen: Narr (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, 5).

Gooding, Paul / Terras, Melissa / Warwick, Claire (2013): “The myth of the new. Mass digitization, distant reading, and the future of the book”, in: *Literary and Linguistic Computing* 28 (4), 629–639. 10.1093/llc/fqt051.

Jurish, Bryan (2010): “More than Words: Using Token Context to Improve Canonicalization of Historical German”, in: *Journal for Language Technology and Computational Linguistics* 25 (1), 23–40.

Krasselt, Julia / Bollmann, Marcel / Dipper, Stefanie / Petran, Florian (2015): *Guidelines für die Normalisierung historischer deutscher Texte*. Bochumer Linguistische Arbeitsberichte, 15. urn:nbn:de:hebis:30:3-419680.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3. “A new architecture for generic corpus query and visualization”, in: *Digital Scholarship in the Humanities* 31 (1), 118–139. 10.1093/llc/fqu057.

Kupietz, Marc / Geyken, Alexander (eds.) (2016): “Corpus Linguistic Software Tools”, in: *Journal for Language Technology and Computational Linguistics* 31(1).

Lüdeling, Anke (2011): “Corpora in Linguistics. Sampling and Annotation”, in: Grandin, Karl (ed.): *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. New York: Science History Publications (Nobel Symposium, 147), 220–243.

Lüdeling, Anke / Zeldes, Amir (2007): “Three Views on Corpora. Corpus Linguistics, Literary Computing, and Computational Linguistics”, in: *Jahrbuch für Computerphilologie* (9), 149–178.

McEnery, Tony / Hardie, Andrew (2012): *Corpus Linguistics. Method, Theory and Practice*. Cambride [u.a.]: Cambridge University Press (Cambridge Textbooks in Linguistics).

Odebrecht, Carolin (2014): "Modeling Linguistic Research Data for a Repository for Historical Corpora", in: *Digital Humanities Conference Abstracts*. Lausanne 284–285.

Odebrecht, Carolin / Belz, Malte / Zeldes, Amir / Lüdeling, Anke / Krause, Thomas (2017): „RIDGES Herbology. Designing a Diachronic Multi-Layer Corpus", in: *Language Resources and Evaluation 51 (2)* First Online 2016, 695-725 . 10.1007/s10579-016-9374-3.

Odebrecht, Carolin / Krause, Thomas / Lüdeling, Anke (2015): "Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository", *37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*, DGfS-CL Poster Session, Leipzig. http://conference.uni-leipzig.de/dgfs2015/fileadmin/zusatzdokumente/dgfs-tagung-2015-final.pdf [letzter Zugriff am 17.08.2017].

Perlitz, Laura (2014): *Konkurrenz zwischen Wortbildung und Syntax. Historische Entwicklung von Benennung.* Bachelorarbeit. Humboldt-Universität zu Berlin, Berlin. 10.18452/14232

Rissanen, Matti (2008): "Corpus Linguistics and Historical Linguistics", in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook.* 2 Bände. Berlin: De Gruyter (1), 53–68.

Schmidt, Thomas / Wörner, Kai (2009): "EXMARaLDA. Creating, analysing and sharing spoken language corpora for pragmatic research", in: *Pragmatics* 19 (4), 565–582.

Simanowski, Roberto (2011): *Digital Art and Meaning. Reading Kinetic Poetry, Text Machines, Mapping Art, and Interactive Installations* (Electronic Mediations). Minnesota: University of Minnesota Press.

Simons, Gary / Bird, Steven (2008): "Toward a global infrastructure for the sustainability of language resources", in: *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*. Cebu City, 87–100.

Springmann, Uwe / Lüdeling, Anke (2017): "OCR of historical printings with an application to building diachronic corpora. A case study using the RIDGES herbal corpus", in: *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html [letzter Zugriff am 12.09.2017].

Wittenburg, Peter / Brugmann, Hennie / Russel, Albert / Klassmann, Alex / Sloetjes, Han (2006): "ELAN. A Professional Framework for Multimodality Research", in: *Proceeedings of LREC. Language Resources and Evaluation Conference*. Genoa 1556–1559. http://www.lrec-conf.org/proceedings/lrec2006/ [letzter Zugriff am 23.12.2016].

Zipser, Florian / Romary, Laurent (2010): "A model oriented approach to the mapping of annotation formats using standards", in: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. http://hal.archives-ouvertes.fr/inria-00527799/en/ [letzter Zugriff am 12.11.2014].