Entwicklung und Nutzung interdisziplinärer Repositorien für historische textbasierte Korpora

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

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

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

Zielke, Dennis
dennis.zielke@cms.hu-berlin.de
Humboldt-Universität zu Berlin, Deutschland

Ziele des Workshops

Der Workshop setzt sich das Ziel, die Möglichkeiten, Aufgaben und Herausforderungen bei der Wiederverwendung von historischen Korpora zu identifizieren und zu diskutieren. Insbesondere sollen dabei deren Architektur, Dokumentation, Veröffentlichung und Speicherung betrachtet werden. So wollen wir versuchen, Methoden und Strategien für das interdisziplinäre Forschungsparadigma der Digital Humanities zu entwickeln und diese in den Fragestellungen der Konferenz der DHd 2016 zu verorten. Der Fokus wird auf die spezifischen und fächerübergreifenden Anforderungen historischer Texte in Bezug auf deren Aufbereitung und Speicherung in Repositorien zum Zweck der Wiederverwendung gelegt. Damit richtet sich der Workshop gleichermaßen an Korpuserstellende, Entwickelnde und an Betreiber_innen von Repositorien und deren Nutzer_innen.

Forschungsfragen und Kontextualisierung des Workshops

Historische Texte bilden den Forschungsgegenstand verschiedener geisteswissenschaftlicher Fächer wie der Linguistik, der Geschichtswissenschaft, der Literaturwissenschaft und vieler weiterer. Jede Disziplin hat dabei ihre eigenen Forschungsfragen und Arbeitsweisen, die sich in beispielsweise den genutzten Formaten und Annotationsweisen zeigen, wie beispielsweise die TEI Guidelines und deren TEI-XML-Format (TEI Consortium 2015) für digitale Editionen oder das Stand-Off-Format PAULA (Dipper 2005) für linguistische Korpora. Dennoch gibt es Ähnlichkeiten bei der Textauswahl und -aufbereitung, die eine gemeinsame Nutzung der vorhandenen Daten sinnvoll erscheinen lassen. In vielen Fällen wird zwischen den digitalisierten historischen Texten – den Primärdaten – und den hinzugefügten Interpretationen in Form von Metadaten und Annotationen unterschieden. Diese Begriffe – Primärdatum, Annotation, Metadatum – werden sowohl fachübergreifend aber auch innerhalb einer Disziplin oft sehr unterschiedlich genutzt. Dadurch entsteht der Eindruck, dass die Daten der Disziplinen grundverschieden sind. Da sich die oft unterschiedlichen Korpusarchitekturen jedoch auf den gleichen Forschungsgegenstand, nämlich den gleichen Text, beziehen können, gibt es durchaus große Schnittmengen zwischen den verschiedenen Disziplinen.

Ein Beispiel für eine ähnliche Textauswahl sind historische Zeitungen, auf deren Grundlage ganz unterschiedliche Fragestellungen adressiert werden (für einen kleinen Überblick siehe bspw. Burr et al. 2015). Die korpusbasierten Aufbereitungsarten reichen bei diesem Register beispielsweise von digitalen Editionen (z. B. „ Korpus romanischer Zeitungssprachen “, Burr 1994-2007), über registerspezifische Referenzkorpora (z. B. „ German Manchester Corpus “, Bennett et al. 2008) bis hin zu syntaktisch tief annotierten Korpora (z. B. Mercurius Baumbank , Demske 2003-2005). Ein weiteres Beispiel sind historische Briefe, die als digitale Edition literaturwissenschaftlich untersucht werden (z. B. „Briefe und Texte aus dem intellektuellen Berlin um 1800“, Baillot / Seiffert 2013b), oder die als linguistisch annotierte Korpora zur Untersuchung von historischen Sprachständen dienen (z. B. „ Kasseler Junktionskorpus “, Ágel / Hennig 2007-2009).

Die Beantwortung der jeweiligen Forschungsfrage stützt sich dann häufig auf Interpretationen in Form von Annotationen in einem Korpus, deren Formen sich disziplinübergreifend ähneln können. Dennoch existieren vielfältige manuell zu erstellende oder automatisch generierbare Annotationsarten wie zum Beispiel Named-Entity-Recognition, Referenzierung auf Personendatenbanken (wie z. B. die Gemeinsame Normdatei), GEO Tagging (vgl. z. B. Elliot / Gillies 2009), Lemmatisierung (z. B. Schmid 1994) oder syntaktisches Parsing (z. B. Malt Parser), die die interpretatorischen Analysegrundlagen stellen. Da die Digitalisierung der historischen Texte (auf Grundlage von Handschriften, Drucken oder Editionen) und deren Annotation aufwändig und vielschichtig sind (vgl. u. a. Rissanen 2008, Kytö / Pahta 2012) kann die Wiederverwendung von historischen Korpora von Vorteil sein. Die Vorstellung der verschieden disziplinspezifischen Sicht- und Zugriffweisen auf solche textbasierten Daten (vgl. Pitti 2004) und deren

Wiederverwendung ist ein zentraler Themenbereich des Workshops.

Damit diese heterogenen historischen Korpora von unterschiedlichen Disziplinen genutzt und wiederverwendet werden können, müssen sie über eine gemeinsame Plattform zugreifbar sein. Diese Plattform muss das Durchsuchen der Daten sowie der Metadaten, ggf. das Evaluieren sowie das Anreichern mit weiteren Annotationen und erneute Hochladen der Daten ermöglichen. Idealerweise können Repositorien diese Funktion übernehmen. Sie funktionieren dann wie eine Art Marktplatz, auf dem historische Korpora fachübergreifend ausgetauscht und mit Informationen angereichert werden können.

Der Workshop nimmt diesen Startpunkt, dessen Voraussetzungen und Konsequenzen zum Thema und begreift ihn als einen Teilbeitrag zu einem Fragenkomplex der DHd-Konferenz:

> „Was sind die Daten der Geisteswissenschaften? Wie müssen die Daten der Geisteswissenschaften (digitalisierte bzw. digitale Texte, Bilder, Musik, Audio, Filme / Videos etc.) aufgearbeitet und vorgehalten werden, um sie über die Fächer hinweg nicht nur für unterschiedliche, sondern auch derzeit noch unbekannte Fragestellungen nutzen zu können?“

Der Workshop versucht für historische Korpora zu ergründen, wie und welche Wiederverwendungsszenarien unter welchen Voraussetzungen möglich sind und was der aktuelle Stand der Forschung ist. Dabei ist es enorm wichtig, dieses Thema vielschichtig und aus mehreren Perspektiven zu beleuchten. Fallstudien für die Wiederverwendung historischer Daten können exemplarisch Erfahrungen, Herausforderungen und Aufgaben thematisieren. Anhand von Korpusarchitekturen, die die Wiederverwendung unterstützen, können wichtige Konzepte und Modelle diskutiert und verglichen werden. Die Beschreibung von konkreten Technologien für die Umsetzung eines Repositoriums erlaubt es, die theoretischen Datenmodelle auf ihre Praxistauglichkeit zu untersuchen. Die Nutzer dieser Technologien tragen durch ihre Erfahrungen über die potentiellen Vorteile der Wiedernutzung und die Bereiche, in denen sie Sinn machen, maßgeblich zur Diskussion bei.

Damit stellen sich folgende Fragen in Bezug auf die historischen Korpora, deren Aufbereitung, die Repositorien bzw. Technologien und deren Nutzung:

- Können dieselben Primärdaten unter verschiedenen Forschungsfragen unterschiedlich genutzt werden?
- Welche Gemeinsamkeiten, welche Unterschiede weisen die Korpora hinsichtlich ihrer umfangreichen Aufbereitung historischer Texte auf.
- In wie weit fördern / erschweren die Annotationen als theoretische Konzepte und Interpretationen eine Wiederverwendung?
- Welche Arten von Annotationen und Analysen können wie wiederwendet werden?
- Welche Arten der Wiederverwendung können sich ergeben?
- Wie unterschiedlich bewerten Disziplinen die Qualität eines Korpus?
- Welche interdisziplinären Nutzer- und Nutzungsszenarien ergeben sich?
- Welche Anforderungen ergeben sich hinsichtlich der Korpusarchitektur inklusive Annotationsarten und Format?
- Welche Speicherformate eignen sich für die Wiederverwendung von Forschungsdaten?
- Wie können Lizenzen den Austausch und die Wiederverwendung fördern?
- Was sind die relevanten Metadaten über ein Korpus?
- Welche Art von Zugriff auf die Korpora ist notwendig, um eine Wiederverwendung zu erleichtern? Wie müssen Repositorien beschaffen sein?
- Welche Vor- und Nachteile besitzen disziplinspezifische / interdisziplinäre oder / und formatabhängige oder -unabhängige Repositorien?
- Eine Diskussion und mögliche Beantwortung dieser Fragen wollen wir durch einen fächerübergreifenden Austausch von Entwicklern, Korpuserstellern und Nutzern im Rahmen des Workshops ermöglichen.

Form des Workshops

Der Workshop soll bestehend aus zwei impulsgebenden Keynotes und sechs Vorträgen an einem Tag vor der DHd-Konferenz stattfinden. Eine Keynote wird das Thema des interdisziplinären Zugangs und der Wiederverwendung zu historischen Daten allgemein thematisieren und problematisieren (Lüdeling und Dreyer, Projektleiter des LAUDATIO -Repositoriums für historische Texte.

Eine zweite Keynote wird die Frage nach einem Qualitätsmanagement im Rahmen von Wiederverwendungsszenarien, dessen Umfang und Zweck aufwerfen und diskutieren (Laurent Romary, DARIAH Director .

Die Vorträge, die aus dem offenen Call des Workshops hervorgehen, sollen Korpuserstellende, Repositorienentwickler_innen und -nutzer_innen aus ganz verschiedenen Fachbereichen die Gelegenheit geben, die oben aufgeworfenen Fragen aufzunehmen und aus einer notwendigerweise interdisziplinären Sicht die Möglichkeiten, Herausforderungen und Lösungen für die Wiederverwendung von historischen Korpora zu diskutieren. Die Keynotes erhalten je 30 Minuten Redezeit sowie 15 Minuten Diskussion und die Vorträge je 20 Minuten und 10 Minuten Diskussion. Eine Diskussion wird den Workshop abschließen. Die primäre Sprache des Workshops ist Deutsch.

Fußnoten

1. Zur Diskussion über Primärdatum, Transkriptionen, Normalisierungen siehe bspw. Claridge 2008, Himmelmann 2012, Kramer 2014; über Metadaten siehe bspw. Odebrecht 2015, Haynes 2004; über Annotationen siehe bspw. Lüdeling 2011, Kübler / Zinsmeister 2015).
2. „Die Gemeinsame Normdatei (GND) ist eine Normdatei für Personen, Körperschaften, Konferenzen, Geografika, Sachschlagwörter und Werktitel, die vor allem zur Katalogisierung von Literatur in Bibliotheken dient, zunehmend aber auch von Archiven, Museen, Projekten und in Webanwendungen genutzt wird.“ (DNB 2015).
3. Tool zum automatischen Annotieren von syntaktischen Dependenzen (Hall et al. 2014).
4. Siehe den Call for Papers DHd 2016 http://www.dhd2016.de/node/9 [letzter Zugriff: 12.September 2015].

Bibliographie

Ágel, Vilmos / Hennig, Mathilde (2007-2009): *KAJUK* (Version 1.1). Justus-Liebig-Universität Gießen http://www.uni-giessen.de/kajuk/index.htm , http://hdl.handle.net/11022/0000-0000-2102-8 [letzter Zugriff 10. September 2015].

Baillot, Anne / Seifert, Sabine (2013a): „The Project "Berlin Intellectuals 1800–1830" between Research and Teaching“ in: *Journal of the Text Encoding Initiative* 4. DOI: 10.4000/jtei.707.

Baillot, Anne / Seifert, Sabine (2013b): *Briefe und Texte aus dem intellektuellen Berlin um 1800* http://tei.ibi.hu-berlin.de/berliner-intellektuelle/ [letzter Zugriff: 22.Dezember 2015].

Bennett, Paul / Durrell, Martin / Ensslin, Astrid / Scheible, Silke / Whitt, Richard (2008): *GerManC Project* (Version 1.0), University of Manchester http://www.llc.manchester.ac.uk/research/projects/germanc/ , http://hdl.handle.net/11022/0000-0000-2D1B-1 [letzter Zugriff 10.September 2015].

Burr, Elisabeth (1994-2007): *Korpus Romanischer Zeitungssprachen*. Duisburg - Bremen - Leipzig http://www.uni-leipzig.de/~burr/CorpusLing/ [letzter Zugriff 10. Sepzember 2015].

Burr, Elisabeth / Burkhardt, Julia / Potapenko, Elena / Sierig, Rebecca / Concepción Durán, Arámis (2015): „Das Duisburg-Leipzig Korpus romanischer Zeitungssprachen und sein Textmodell“, in: *Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum*, DHd 2015, Graz. https://www.conftool.pro/dhd2015/index.php/Burr-Das_Duisburg-Leipzig_Korpus_romanischer_Zeitungssprachen-1771016.pdf?page=downloadPaper&filename=Burr-Das_Duisburg-Leipzig_Korpus_romanischer_Zeitungssprachen-1771016.pdf&form_id=177 [letzter Zugriff: 11.September 2015].

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

Demske, Ulrike (2003-2005): *Mercurius* (Version 1.1). Universität Potsdam http://www.uni-potsdam.de/guvdds/projekte/abgproj.html , http://hdl.handle.net/11022/0000-0000-467D-6 [letzter Zugriff: 10.September 2015].

Dipper, Stefanie (2005): “XML-Based Stand-Off Representation and Exploitation of Multi-Level Linguistic Annotation”, in: *Proceedings of Berliner XML Tage Berlin* 39-50.

DNB = Deutsche Nationalbiliothek (2015): *Gemeinsame Normdatei (GND)* http://www.dnb.de/gnd [letzter Zugriff 10.September 2015].

Elliott, Tom / Gillies, Sean (2009): “Digital Geography and Classics. Changing the Center of Gravity”, in: *Digital Humanities Quarterly* 3, 1 http://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html [letzter Zugriff 10. September 2015].

Hall, Johan / Nilsson, Jens / Nivre, Joakim (2014): *MaltParser* http://www.maltparser.org/ [letzter Zugriff 10. September 2015].

Haynes, David (2004): *Metadata for information management and retrieval*. London: Facet publishing.

Himmelmann, Nikolaus. P. (2012): “Linguistic Data Types and the Interface between Language Documentation and Description”, in: *Language Documentation and Conservation* 6: 187-207.

Kramer, Michael J. (2014): “Defining Data for Humanists: Text, Artifact, Information or Evidence?”, in: *Journal for Digital Humanities* 3, 2 http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/ [letzter Zugriff: 10. September 2015].

Kübler, Sandra / Zinsmeister, Heike (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. London / New York: Bloomsbury.

Kytö, Merja / Pahta, Päivi (2012): “Evidence from historical corpora up to the twentieth century” in: Nevalainen, Terttu / Traugott, Elizabeth C. (eds.): *The Oxford Handbook of the History of English.* Oxford o.a.: Oxford University Press 123-133.

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

Odebrecht, Carolin (2015): „Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung“, in: *Von Daten zu Erkenntnissen. 2. Jahrestagung des Verbandes der Digital Humanities im deutschsprachigen Raum* DHd 2015, Graz. https://www.conftool.pro/dhd2015/index.php/

Odebrecht-Interdisziplin
%C3%A4re_Nutzung_von_Forschungsdaten
_mithilfe_einer_technisch-
abstrakten_Modellierung-63110.pdf?
page=downloadPaper&filename=Odebrecht-Interdisziplin
%C3%A4re_Nutzung_von_Forschungsdaten
_mithilfe_einer_technisch-
abstrakten_Modellierung-63110.pdf &form_id=63 [letzter Zugriff 12.September 2015].

Pitti, Daniel V. (2004): "Designing Sustainable Projects and Publications" in: Schreibman, Susan / Siemens, Ray / Unsworth, John (eds.): *A Companion to Digital Humanities.* Oxford: Blackwell 471–487.

Rissanen, Matti (2008): "Corpus Linguistics and Historical Linguistics" in: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook.* Volume 1. Berlin: Mouton de Gruyter 53-68.

Schmid, Helmut (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees", in: *Proceedings of International Conference on New Methods in Language Processing*, 1994. Manchester.

TEI Consortium (eds.) (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 2.8.0. 2015-04-06.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/ [letzter Zugriff 11. August 2015].