Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven

Veranstalter: Common Language Resources and Technology Infrastructure Deutschland (CLARIN-D); Deutsches Textarchiv (DTA)
Datum, Ort: 18.02.2013–19.02.2013, Berlin
Bericht von: Martina Gödel, textloop, Hamburg

Die Konferenz „Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven" fand am 18. und 19. Februar 2013, begleitet von zwei Workshops, an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) in Berlin statt. Ausgehend von einem Workshop, der in die Nutzung bereit stehender Dienste und Ressourcen in CLARIN-D einführte, widmete sich die Konferenz selbst Projekten, die wie das Deutsche Textarchiv auf den Gebieten des Auf- und Ausbaus, der Annotation und der Analyse historischer Volltextkorpora arbeiten und die dabei zum Teil bereits mit dem DTA kooperieren. Anhand zahlreicher Beispiele wurden so der Aufbau von und die Arbeit mit historischen Textkorpora anschaulich gemacht. Ein sich anschließender Workshop beleuchtete die im DTA zur Nutzung bereit stehenden Tools.[1]

Workshop I

Der der Konferenz vorangestellte Workshop des Arbeitspakets 5 („Dienste und Ressourcen") in CLARIN-D behandelte die hier zur Verfügung stehenden Dienste zur Analyse von Sprachressourcen.

Einleitend skizzierte AXEL HEROLD (Berlin) die Zielsetzungen des europäischen Projektes CLARIN[2] im Allgemeinen, sowie insbesondere der deutschen Initiative CLARIN-D[3], innerhalb derer gegenwärtig neun Institutionen mit eigens gebildeten Servicezentren kooperieren.[4] Eine service- und zentrenbasierte Infrastruktur für Sprachressourcen im weitesten Sinne solle geschaffen werden, die eine nachhaltige Speicherung von Forschungsdaten biete und es auch Nicht-Linguisten ermögliche, mit linguistischen Verfahren an den bereitgestellten Ressourcen bzw. ihrem eigenen Material arbeiten zu können. Die von verschiedenen Arbeitsgruppen angebotenen Werkzeuge sollen miteinander kombinierbar und interoperabel sein, alle Schnittstellen sollen auch externen Nutzern frei zur Verfügung stehen. Die Infrastruktur müsse aber so gestaltet werden, dass unterschiedliche rechtliche Beschränkungen beachtet werden könnten.[5]

KERSTIN ECKART (Stuttgart) wies in Ihrem Vortrag zur nachhaltigen Aufbereitung von Textressourcen auf die Notwendigkeit einer klaren Trennung von Primärdaten (d.h. der grundlegenden Version einer Ressource) und Annotationen (d.h. der Anreicherung der Ressource mit weiterführenden Informationen unterschiedlichen Inhalts und eigener Strukturierung) hin. Dementsprechend plädierte sie für eine Anreicherung der Primärdaten durch Annotationen nicht direkt in den Primärdaten, sondern im Stand-off-Verfahren. Die Verankerung der Annotationen erfolgt bei dieser Methode über Buchstaben, Wortkoordinaten oder ähnliche Referenzpunkte der Primärdaten. Für automatische Annotationen mittels linguistischer Tools sei diese Methode aufgrund von möglicherweise überlappenden Annotationen, aber auch von Korrekturen und sonstigen Veränderungen im Primärtext zu empfehlen. Annotationen sollen möglichst standardisierten Verfahren (z. B. den Richtlinien des STTS Tag Sets[6]) folgen und können dazu die von CLARIN-D bzw. vom DTA bereitgestellten Tools nutzen. Das Vorgehen müsse sich aber auch nach den Bedürfnissen des jeweiligen Projektes richten.

Im dritten Vortrag informierte AXEL HEROLD über die Lieferung von standardisier-

[1] Die Konferenzübersicht und Materialien zu einzelnen Vorträgen sind unter <http://www.deutschestextarchiv.de/doku/workshop2013> (24.04.2013) verfügbar.

[2] <http://www.clarin.eu> (24.04.2013).

[3] <http://de.clarin.eu> (24.04.2013).

[4] <http://de.clarin.eu/de/clarin-d-zentren> (24.04.2013).

[5] Siehe dazu das Benutzerhandbuch/CLARIN-D User Guide: <http://de.clarin.eu/de/sprachressourcen/benutzerhandbuch> (24.04.2013).

[6] <https://catalog.clarin.eu/isocat/rest/dcs/376> (24.04.2013).

ten Metadaten als eine Mindestanforderung für Projekte, die die CLARIN-D-Infrastruktur zur Speicherung und Weitergabe eigener Daten nutzen möchten. Er betonte, dass Metadaten innerhalb des CLARIN-D-Repositories nicht nur für eine bessere Sichtbarkeit der Ressourcen und facettenreichere Suchabfragen innerhalb dieser wichtig seien, sondern auch in technischer Hinsicht für die Interoperabilität von Programmen und Ressourcen (z. B. bei der Hintereinanderschaltung von Tools) unerlässlich seien. Mit der Component MetaData Infrastructure (CMDI)[7] habe der CLARIN-Verbund ein „Meta-Metadaten-Modellformat" geschaffen, das als vermittelnde Schnittstelle zwischen den in den jeweiligen Projekten eingesetzten unterschiedlichen Metadatenformaten fungieren könne. Für Projekte heißt das: Die vorhandenen Metadatenfelder müssen mit Hilfe von ISOcat[8] auf ihre semantischen Repräsentanzen[9] abgebildet werden. Im Idealfall gäbe es hier aber bereits fertige Konverter. CMDI werde dann alle Informationen aufnehmen und diese, je nach Bedarf, in einem anderen Metadatenformat wieder ausgeben können.

In der anschließenden Demonstration veranschaulichte KAI ZIMMER (Berlin) die Servicearchitektur WebLicht.[10] Nutzer können hier eigene Texte hochladen und unterschiedliche Webservices darauf anwenden bzw. in einer einfachen Weboberfläche Tools für automatisierte Annotationen in einer so genannten Tool Chain hintereinander schalten. Aus der folgenden Diskussion ergaben sich interessante Desiderata: Gewünscht wurden der Abbau von Zugangsbeschränkungen zu WebLicht, zumindest in Bezug auf die Nutzung der Open-Source-Tools, sowie die Bereitstellung bereits erprobter Anwendungsabläufe mit vorgefertigten Tool Chains für häufig ausgeführte Operationen, etwa die Identifikation von Eigennamen (NER, Named Entity Recognition). Auch an technischen Dokumentationen zu den einzelnen Tools mangele es bisher.

EDMUND POHL (Potsdam) beschrieb in der Folge einen auf der Basis der lexikalischen Datenbank dlexDB[11] in Vorbereitung befindlichen Webservice für WebLicht. Derzeit seien in WebLicht nur drei Eigenschaften von Lemmata im Text Corpus Format (TCF)[12], dem Eingangs-, Austausch- und Zielformat der einzelnen Webservices, vorbereitet. Im Sinne eines Werkstattberichts wurden mögliche Weiterentwicklungen des TCF-Formats auf der Grundlage von Anforderungen eines externen Projekts demonstriert, die auch zum Teil bereits durch CLARIN-D übernommen wurden.

Ein weiterer in Vorbereitung befindlicher Webservice wurde von THOMAS ECKART (Leipzig) skizziert: der Prototyp eines TEI-Integrators. Erleichtert werden solle die Aufnahme von Textkorpora, die den Guidelines der Text Encoding Initiative (TEI) folgend ausgezeichnet wurden. Dabei werde unter anderem die Abbildung der Metadaten auf das CMDI-Metadatenmodell ermöglicht und damit die Hürde für die Aufnahme von neuen Texten in CLARINs Virtual Language Observatory (VLO), das eine nachhaltige Speicherung und vorbereitete Präsentation bietet, verringert. Offene Probleme seien zurzeit noch, dass zwar TEI erfolgreich in TCF umgewandelt werden könne, aber anders herum noch keine Rückumwandlung aus TCF nach TEI möglich sei. Auch sei die Kodierung gegenwärtig statisch: Ändere sich also beispielsweise etwas an den TEI-Basisdateien, müssten die neuen Versionen hochgeladen und aufs Neue verarbeitet werden.

Hauptkonferenz

Die eigentliche Konferenz wurde von ALEXANDER GEYKEN (Berlin) eingeleitet. Ein Ziel von CLARIN-D und des DTA sei es, unterschiedliche historische Textkorpora auf einer Plattform zusammenzuführen und über zur Verfügung stehende Tools linguistisch zu erschließen. Das DTA biete auf seiner Webseite ein stetig wachsendes historisches Textkorpus und über das Erweiterungs-Modul DTAE[13] einen bewährten Weg, eigene Texte für die Aufnahme in das Repositorium vor-

[7] <http://www.clarin.eu/node/3219> (24.04.2013).

[8] <http://www.isocat.org> (24.04.2013).

[9] <https://catalog.clarin.eu/isocat/interface/index.html> (24.04.2013).

[10] <http://de.clarin.eu/de/sprachressourcen/weblicht> (24.04.2013).

[11] <http://www.dlexdb.de> (24.04.2013).

[12] <http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format> (24.04.2013).

[13] <http://www.deutschestextarchiv.de/dtae> (24.04.2013).

zubereiten. Mit Hilfe standardisierter Metadaten, einheitlicher Textkodierung und linguistischer Erschließung der integrierten Texte gewährleiste man die nachhaltige Verfügbarkeit der Ressourcen sowie deren Interoperabilität. An einem Beispieltext aus dem Polytechnischen Journal, das jüngst als externe Ressource durch das DTA angeglichen und integriert wurde, machte Geyken anschaulich, wie über eine Normalisierung der Orthographie und die Vereinheitlichung der strukturierenden Annotationen die Ergebnisse linguistischer Abfragen erheblich verbessert werden können.

Als Auftakt des ersten Themenblocks der Konferenz zu „Aufbau und Zusammensetzung von Korpora" stellte HEIKE SAHM (Siegen) die Arbeiten des DFG-geförderten Projekts „Literaturexplosion und Intertextualität. Bedingungen und Merkmale der ‚Verschriftlichung des Lebens' in Nürnberg" vor. Dabei sprach sie über die Arbeiten zur Erschließung und Untersuchung der städtischen Literatur des 15. Jahrhunderts am Beispiel der Stadt Nürnberg, wo ab 1430 ein starker Anstieg der Literaturproduktion zu beobachten ist. Im Rahmen des Projekts wird ein Textkorpus erstellt, welches subakademische Texte (i.e. Handwerkerliteratur, geistliche Literatur sowie Sachliteratur) als Zeugnisse der städtischen Kultur enthält.

Ein weiterer, anhand thematischer Kriterien sowie aufgrund des Entstehungsortes der Textzeugen zusammengestellter Textbestand wurde von CLAUDIA RESCH (Wien) und THIERRY DECLERCK (Saarbrücken/Wien) vorgestellt. Das Projekt AbaC:us widmet sich der Bearbeitung einer Sammlung theologischer Texte des Barock zu den Themen Tod und Sterben, insbesondere von Schriften Abrahams a Santa Clara bzw. aus dessen Umfeld. Aufgrund der Erstdrucke werden die Texte mit einer möglichst hohen Fehlerfreiheit erfasst und linguistisch erschlossen. Probleme bereiten dem Projekt vor allem die Automatisierung von Tokenisierung und Lemmatisierung. Für die computergestützte manuelle Abbildung der historischen Wortformen auf das Neuhochdeutsche werden eigene Tools entwickelt, die anschließend auch im CLARIN-D Zusammenhang nachnutzbar sein sollen.

MARIA FEDERBUSCH (Berlin) stellte im Anschluss eine an der Staatsbibliothek zu Berlin durchgeführte Studie zur Beurteilung der Qualität von OCR-Ergebnissen auf der Grundlage von Drucken der frühen Neuzeit vor.[14] Dabei wurden Funeralschriften des 16. Jahrhunderts mithilfe unterschiedlicher OCR-Produkte (BIT Alpha; HK-OCR/FREngine 9) im Volltext erfasst und die Resultate verglichen. Als vorläufiges Fazit nannte sie – bei entsprechendem Training und optimaler (durchaus aufwändiger) Anpassung der Parameter der Texterkennungssoftware an die Vorlage – eine Erkennungsgenauigkeit von 97 Prozent und höher realistisch. Dies Ergebnis ließe sich durch Training der Software, den Einsatz von Listen historischer Wortformen und die Optimierung der korrekten Segmentierung von Wörtern weiter verbessern.

Im letzten Beitrag zu diesem Themenblock stellte MANFRED NÖLTE (Bremen) ein gemeinsam mit dem DTA geplantes Projekt der Staats- und Universitätsbibliothek Bremen vor. Grundlage bildet die Zeitung „Die Grenzboten" (Laufzeit 1841–1922), welche an der SUUB Bremen vollständig digitalisiert wurde, wobei die Bilddigitalisate mit einer Roh-OCR hinterlegt wurden.[15] Zusätzlich wurden manuell METS-Strukturdaten aller Bände erhoben, die unter anderem eine artikel- bzw. autorspezifische Navigation in den Bänden erlauben. In dem beantragten Fortsetzungsprojekt wolle man nun untersuchen, inwieweit sich die Ergebnisse der OCR automatisiert verbessern ließen. Unterschiedliche Fraktur-OCR-Software solle verglichen und mithilfe von Crowdsourcing weiter verbessert werden. Mit Unterstützung des DTA sollen in einem eigenen Arbeitspaket die Vorlagen seitenweise strukturiert (um beispielsweise Überschriften, Fußnoten und Fließtext als Textsegmente unterscheiden zu können) und die Texte linguistisch erschlossen werden. Ein Abgleich mit dem DTA-Kernkorpus als qualitativ hochwertiger Referenzbasis werde helfen, mögliche historische Schreibvarianten aus dem Grenzboten-

[14] <http://staatsbibliothek-berlin.de/die-staatsbibliothek/abteilungen/historische-drucke/aufgaben-profil/projekte/funeralschriften> (24.04.2013).

[15] <http://brema.suub.uni-bremen.de/grenzboten> (24.04.2013).

Volltext von potentiellen OCR-Fehlern zu unterscheiden.

Der zweite Themenblock, „Arbeiten mit historischen Korpora“, wurde von MARIUS HUG (Berlin) eröffnet. Bei der Arbeit mit dem heute bereits digital vorliegenden „Polytechnischen Journal“[16] ginge es nun darum, dieses über eine Vernetzung in einen größeren Kontext zu stellen. Als illustrierende Beispiele für die sich ergebenden Chancen nannte Hug die Verlinkung der Inhalte mit bereits digital vorliegenden internationalen Publikationen (neben Zeitschriften/Büchern wurde hier v.a. auf den Mehrwert einer Verlinkung mit Patentschriften hingewiesen), die Aufschlüsselung von historischen Maßeinheiten durch interne Querverweise oder externe Links zu Inhalten der Wikipedia (wobei hier perspektivisch an einen eigenen Converter für die im Journal vorkommenden Einheiten gedacht war), sowie die im Projekt angewandte Trendsuche mit einer Timemap im Ergebnis. CHRISTIAN KASSUNG betonte abschließend die Wichtigkeit der Entwicklung weiterer Schnittstellen, um das Projekt nun „ins Leben zu bringen“. So sei eine Kooperation mit dem Deutschen Technikmuseum Berlin angedacht, das gewissermaßen die Realien zu den im ‚Dingler‘ beschriebenen technischen Innovationen beherberge.

NOAH BUBENHOFER (Dresden) demonstrierte Möglichkeiten korpusstatistischer Untersuchungen von Sprachwandel anhand des homogenen, diachronen Text+Berg-Korpus.[17] Dabei handelt es sich um ein Korpus alpinistischer Literatur der Schweiz, welches die kontinuierlich erschienenen Publikationsreihen des Schweizer Alpen-Clubs „Jahrbuch des S.A.C.“ (1864-1923) und „Alpen“ (1925-heute) umfasst. Die Texte wurden mittels OCR erkannt und teils manuell, teils automatisch nachkorrigiert. Die nachfolgende linguistische Analyse umfasste POS-Tagging, Lemmatisierung und Eigennamenerkennung. Die Textstrukturierung erfolgt in einem an TEI angelehnten XML-Format. Bubenhofer stellte verschiedene Methoden der Hypothesen- oder Daten-geleiteten Auswertung dieses Textkorpus vor. Er zeigte dabei, wie sich mentalitätsgeschichtliche Umbrüche anhand alpinistischer Texte mithilfe korpusbasierter, empirischer Analysen nachvollziehen lassen.

THOMAS GLONING (Gießen) illustrierte am Beispiel der Frühgeschichte des Fußballspiels die Notwenigkeit einer Neuorientierung der historischen Lexikographie. Da jedes Thema, über das gesprochen werde, wichtige Komponenten für die Organisation des Wortschatzes einer Sprache enthalte, müsse zukünftige Wortforschung mehr als bisher darauf bezogen werden. Solche thematischen Prägungen ließen sich etwa durch Ontologien oder themenbezogen ermittelte Texttypen dokumentieren. Wichtig für die Wortforschung sei es, Schlüsseltexte zu einzelnen Themen zu ermitteln und diese verschiedentlich zu erschließen (durch Volltexterfassung, lexikologische Untersuchungen und die lexikographische Aufarbeitung des Wortschatzes). Gloning veranschaulichte diese Schritte am Beispiel früher Diskurse zum Fußballspiel.

Der zweite Konferenztag wurde mit der Präsentation des Verbundprojektes AEDit[18], das eine „Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit“ schaffen will, eröffnet. Der Beitrag TIMO STEYERs (Wolfenbüttel), behandelte das Vorgehen: Die Partner stellen unterschiedliche Textkorpora, wie etwa handschriftliche Briefe, frühe Drucke und Leichenpredigten zur Verfügung und erhalten dabei technische Unterstützung. Es wurde festgehalten, dass das in der HAB Wolfenbüttel genutzte TEI-XML-Format bereits sehr nah an dem DTA-Basisformat sei und bei den vom DTA kommenden Leichenpredigten bereits alle Konvertierungsschwierigkeiten erfolgreich behoben werden konnten.

Nähere Informationen zur interessanten Quellengruppe der Leichenpredigten wurden im Anschluss von EVA-MARIA DICKHAUT und JÖRG WITZEL (Marburg) geliefert. Schon auf der Basis von 75 der mehr als 200 im Projekt vorgesehenen Leichenpredigten aus der ehemaligen Stadtbibliothek Breslau, die bislang im DTA als elektronische Volltexte vorliegen, ließen sich interessante biographische, sozial- und wirtschaftsgeschicht-

[16] <http://www.polytechnischesjournal.de/> (24.04.2013).

[17] <http://www.textberg.ch> (24.04.2013).

[18] <http://www.hab.de/de/home/wissenschaft/projekte/aedit-fruehe-neuzeit-archiv–editions–und-distributionsplattform-fuer-werke-der-fruehen-neuzeit.html> (24.04.2013).

liche Schlüsse ziehen und perspektivisch verallgemeinern. Der Bestand der Leichenpredigten im DTA wird im Rahmen des DFG-geförderten AEDit-Projekts in den nächsten Monaten auf mehr als 300 Volltexte erweitert werden.

Mit „LAUDATIO“ wurde von CAROLIN ODEBRECHT und FLORIAN ZIPSER (Berlin) eine Infrastruktur zur linguistischen Analyse historischer Korpora vorgestellt.[19] Mit SaltNPepper werde hier an einem Pipelinekonverter gebaut, der von einem Anntotationsformat in das andere übersetzen könne. Der Einsatz verlange von den beteiligten Projekten aber ein hohes Maß an Methodenbewusstsein hinsichtlich des Formats ihrer Metadaten.

GERHARD HEYER präsentierte Arbeiten am Korpus der „Leipziger Rektoratsreden“, die zwischen 1871 und 1933 jährlich anlässlich des Wechsels des Rektors der Universität gehalten wurden und jeweils von einem Jahresbericht begleitet waren. Das Korpus, das 123 Texte enthält, erschien als Ergebnis eines entsprechenden Editionsprojektes in zwei Bänden im Druck. Es wurde nun aus den PDF-Dateien extrahiert und mittels automatischer Methoden analysiert. Die Analyse umfasste etwa die Eigennamenerkennung, die Aufschluss über in diesem Kontext relevante Personen und deren Beziehungen zueinander erlaubt.

STEPHANIE DIPPER (Bochum) präsentierte Verfahren zur orthographischen Normalisierung sowie zum Tagging historischer Texte des Frühneuhochdeutschen. Zugrunde gelegt wurde das sogenannte Anselm-Korpus, ein Textkorpus aus 43 Handschriften und 7 Drucken des 14.-16. Jahrhunderts, welches Überlieferungen der „Fragen an Maria“ Anselms von Canterbury enthält. Dipper stellte ein mehrstufiges Verfahren vor, in welchem die vielfältigen historischen Schreibweisen in diesen Texten automatisch normalisiert werden. Die so modernisierten Formen werden im Anschluss durch automatisches POS-Tagging mit Wortarteninformationen angereichert. Während das Tagging bereits gute Ergebnisse liefert, erwies sich die Schreibweisen-Normalisierung aufgrund der stark variablen Vorlagen noch als relativ fehleranfällig. Zur Optimierung müssten nun landschafts- und zeittypische Phänomene stärker miteinbezogen werden.

Im letzten Vortrag der Konferenz hielt ANJA VOESTE (Gießen) ein anschauliches Plädoyer für den Erhalt und die konsequente Einbeziehung des Bildmaterials der Vorlage, ergänzend zum digitalisierten Volltext. Wichtige, ggf. später auszuwertende Informationen zu Schreibern, Werkstätten aber auch im Graphischen zu suchende Beweggründe für Wortgrenzen, gingen andernfalls als Quelle verloren.

Workshop II

Im Anschluss an die Konferenz veranstalteten Mitarbeiter des DTA einen Workshop, in welchem Tools und Arbeitsmaterialien des DTA den Nutzern vorgestellt und deren Anwendung eingeübt wurde. Externen Nutzern stünden über das DTAE-Modul zahlreiche Hilfsmittel zur Transkription, Annotation und linguistischen Auswertung ihrer Texte zur Verfügung. MATTHIAS SCHULZ, SUSANNE HAAF und FRANK WIEGAND stellten die DTA-Richtlinien zur Transkription und das DTA-Basisformat zur Textannotation und Metadatenerfassung vor. Zur Unterstützung der Texterstellung entsprechend den DTA-Richlinien stellt das DTA ein Zoning-Tool für die Vorstrukturierung anhand der Bilddateien, ein oXygen-Framework für die Textannotation sowie ein Webformular zur Metadatenerfassung zur Verfügung. Besonders hingewiesen wurde auf die Wichtigkeit der Erfassung von DTA-/CLARIN-D-kompatiblen Metadaten und des korrekten Einsatzes von Unicode zur Kodierung von Sonderzeichen. Für die (retrospektive) Qualitätssicherung steht die Qualitätssicherungsplattform DTAQ bereit. CHRISTIAN THOMAS demonstrierte abschließend die linguistische Suchmaschine DDC, die komplexe Suchanfragen über die DTA-Korpora ermöglicht.[20]

Resümee

[19] <http://www.laudatio-repository.org/> (24.04.2013).

[20] Aufgrund der großen Nachfrage bietet das DTA am 19.4.2013 einen zweiten Workshop zum Thema „Aufbau von Sprachressourcen am Beispiel des Deutschen Textarchivs“. Siehe auch die Ankündigung auf H-Soz-u-Kult <http://hsozkult.geschichte.hu-berlin.de/termine/id=21419> .

Workshops und Konferenz ermöglichten spannende Einblicke in unterschiedlichste Projekte und ihre historischen Textkorpora. Anschaulich wurde gezeigt, wie bereichernd die Arbeit mit kleineren oder bisher weitgehend unbearbeiteten Quellen für Verbundprojekte sein kann. Darüber hinaus wurde deutlich, dass die Kooperation mit dem DTA und CLARIN-D zahlreiche Hilfen für die Arbeit an dem eigenen Material und neue Perspektiven bieten kann, sowohl in Bezug auf eigene Texte bzw. Korpora, als auch durch deren Analyse in einem größeren Zusammenhang. DTA und CLARIN-D bieten Standardtools zur linguistischen Analyse und Schnittstellenformate als Brücke zwischen unterschiedlichen Projekten und den von ihnen eingesetzten Formaten für Meta- und Objektdaten. Die kooperative, an Standards und ‚Best Practices' orientierte, Erstellung und Bearbeitung von Korpusdaten und deren Verbreitung in verteilten Umgebungen wie CLARIN-D, ermöglicht die Nachnutzung durch andere Forscher, so dass sich die Projekte gegenseitig bereichern können.

Auf übergeordneter Ebene wurde deutlich, dass für die überaus wünschenswerte Zusammenführung unterschiedlicher Textkorpora ein hohes Methodenbewusstsein in den einzelnen Projekten hinsichtlich der Objekt- und Metadaten erforderlich ist. In technischer Hinsicht zeigte sich, dass für linguistische Analysen das XML-Format der TEI nicht allein ausreichend ist und linguistische Ergebnisse am geeignetsten über Stand-off Annotationen hinzugefügt werden können.

Konferenzübersicht

Workshop 1: CLARIN-D, Arbeitspaket 5: Dienste und Ressourcen

Kerstin Eckart (Stuttgart): Korpusannotation: Vom nachhaltigen Aufbereiten einer Ressource

Axel Herold (Berlin): Nutzen und Nutzung von Metadaten in CLARIN-D

Kai Zimmer (Berlin): Demonstration der WebLicht-Services

Edmund Pohl (Potsdam): Integration einer lexikalischen Datenbank in WebLicht am Beispiel von dlexDB

Thomas Eckart (Leipzig): Integrationsunterstützung für TEI-kodierte Textdokumente in CLARIN-D

DTA-/CLARIN-D-Konferenz

Alexander Geyken (Berlin): Begrüßung und Einführung

Themenblock 1: Aufbau und Zusammensetzung von Korpora

Heike Sahm (Siegen): Städtische Literatur im 15. Jahrhundert: Erschließung und Bewertung am Fallbeispiel Nürnberg

Thierry Declerck (Saarbrücken/Wien), Claudia Resch (Wien): ABaC:us – Austrian Baroque Corpus: Aufbau, Annotationen und Anwendungen

Maria Federbusch (Berlin): OCR-Einsatz bei der Volltexterfassung von Quellen der Frühen Neuzeit. Eine Fallstudie anhand von Funeralschriften aus dem Bestand der Staatsbibliothek zu Berlin

Manfred Nölte (Bremen): Das Digitalisierungsprojekt „Die Grenzboten": Methoden der Nachbesserung und Nachstrukturierung von OCR-Volltext in der bibliothekarischen Praxis und im Kontext von CLARIN-D

Themenblock 2: Arbeiten mit historischen Korpora

Marius Hug / Christian Kassung (Berlin): „... daß eine wohlthätige Wechselwirkung zu immer segenreichern Resultaten zu leiten vermag" – Arbeiten an und mit dem digitalisierten „Polytechnischen Journal"

Noah Bubenhofer (Dresden): Das diachrone Text+Berg-Korpus alpinistischer Texte: Aufbau und Analysemöglichkeiten

Thomas Gloning (Gießen): Thematische Teilcorpora und historische Lexikographie: Am Beispiel der Frühgeschichte des Fußballspiels

Timo Steyer (Wolfenbüttel): Von Briefen, Predigten und Traktaten: Integration und Modellierung frühneuzeitlicher Texte im AEDit-Projekt

Eva-Maria Dickhaut / Jörg Witzel (Marburg): Von der Katalogisierung zum Volltext: Leichenpredigten aus der ehemaligen Stadtbi-

bliothek Breslau im Projekt AEDit

Carolin Odebrecht / Florian Zipser (Berlin): LAUDATIO – Eine Infrastruktur zur linguistischen Analyse historischer Korpora

Gerhard Heyer (Leipzig): Leipziger Rektoratsreden 1871–1933. Einblicke in sechs Jahrzehnte wissenschaftlicher Praxis

Stephanie Dipper (Bochum): Von Sankt Anselm zu Ente Apfelmus: Normalisierung und Tagging frühneuhochdeutscher Texte

Anja Voeste (Gießen): Von Schlaufen, Typen und von Schreibarbeit. Graphische Fragen und historische Korpora

Workshop 2: Aufbau von Sprachressourcen am Beispiel des DTA

Matthias Schulz / Christian Thomas (Berlin): Einführung zu den DTA-Korpora und -Tools

Matthias Schulz (Berlin): Erarbeiten einer verlässlichen Transkription

Susanne Haaf (Berlin): TEI-Textstrukturierung mit dem DTA-Basisformat

Susanne Haaf / Frank Wiegand (Berlin): Erfassung von DTA-/CLARIN-D-kompatiblen Metadaten

Frank Wiegand (Berlin): Arbeiten mit der Qualitätssicherungsumgebung DTAQ

Christian Thomas (Berlin): Arbeiten mit den DTA-Korpora

Tagungsbericht *Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven*. 18.02.2013–19.02.2013, Berlin, in: H-Soz-Kult 07.05.2013.