Best practice: Digitale Korpora. Workshop der AG „Elektronisches Publizieren“

Veranstalter: AG „Elektronisches Publizieren“ der Akademienunion
Datum, Ort: 08.10.2013-09.10.2013, Berlin
Bericht von: Jörg Wettlaufer, Akademie der Wissenschaften zu Göttingen

Was ist ein Korpus? Wie unterscheidet sich ein Korpus von einer Sammlung? Kann und soll man den Korpusbegriff weit interpretieren? Wie können Korpora miteinander verknüpft und soweit interoperabel gestaltet werden, dass sie außerhalb der ursprünglichen Projekte ebenfalls verwendet werden können – möglicherweise unter gänzlich anderen fachwissenschaftlichen Perspektiven als ursprünglich geplant? Diese Fragen und Probleme beschäftigten die Teilnehmer des 23. Workshops der AG „Elektronisches Publizieren“ in der Akademienunion, der dieses Jahr vom 8.-9. Oktober an der Berlin Brandenburgischen Akademie der Wissenschaften (BBAW) stattfand.

Grußworte wurden von Wolf-Hagen Krauth für den Gastgeber, der BBAW, Annette Schaefgen für die Union der deutschen Akademien der Wissenschaften sowie eine Einführung in das Thema von GERHARD LAUER (Göttingen), dem Vorsitzenden der AG, gesprochen. Der Gastgeber gab in seiner Funktion als Wissenschaftsdirektor der BBAW zu bedenken, dass bislang leider 4/5 der durch die Bund-Länderkommission finanzierten Akademievorhaben keine Ausstattung zur Digitalisierung ihrer Ergebnisse zur Verfügung stünden. Dies werfe die grundsätzliche Frage auf, ob man nicht besser auf ein Neuvorhaben in Zukunft verzichten sollte, um die fehlende Mehrausstattung für die schon existierenden Akademienprojekte auf diese Weise zu finanzieren. Gerhard Lauer betonte die Notwendigkeit, in Zukunft mehr Projekte unter neuen Fragestellungen zusammenzuführen, und die bestehenden digitalen Ressourcen in neue Korpora zu gruppieren. Er vertrat eine entsprechend weite Definition des Begriffs Korpus: Tendenziell seien im digitalen Zeitalter verschiedenste Datensammlungen als Korpora ansprechbar und auch anzusprechen. Ziel sei es, die Quellen und Texte (auch multilingual und -modal) so aufzubereiten, dass sie digital zusammengeführt und somit als interoperable Korpora angesprochen und damit aus der Sicht unterschiedlicher Fachdisziplinen genutzt und analysiert werden können. Dies biete den Vorteil, dasselbe Material in völlig unterschiedlichen Fachdisziplinen mit ihren jeweils unterschiedlichen Fragestellungen zu verwenden. Zuletzt zog er den Vergleich zum Internet der Dinge und forderte dazu auf, Akademievorhaben konsequent als Teile von fächerübergreifenden Korpora zu denken und zu konzipieren.

Den ersten Vortrag zum Thema bestritt KARIN DORNHAUSER (Berlin), die zu „Korpus - Begriff und Konzept“ sprach und das DFG-geförderte Projekt „DeutschDiachronDigital“ vorstellte. Es handelt sich dabei um ein historisches Referenzkorpus der deutschen Sprache mit Schwerpunkt auf Alt-, Mittel- und Frühneuhochdeutsch. Das Projekt legt besonderen Wert auf die zeitliche und räumliche Ausgewogenheit, unter anderem durch die Einbindung niederdeutscher Texte, damit bei der Analyse keine einseitigen Verzerrungen aufgrund eines Schwerpunkts im Hochdeutschen oder einer bestimmten Mundart auftreten. Die Texte können mit dem Analysetool ANNIS computerlinguistisch ausgewertet werden. Anschließend stellte ALEXANDER GEYKEN (Berlin) das „Deutsche Text Archiv“ (DTA) als multimodal verwendbares Historisches Textkorpus für die Geistes- und Sozialwissenschaften vor. Dieses Projekt hat sich zur Aufgabe gesetzt, ein ausgewogenes historisches Korpus der deutschen Sprache für die Zeit von 1600 bis 1900 zu erstellen und für die Forschung zugänglich zu machen. Bislang konnten bereits 1.014 Werke mit über 500 Millionen Zeichen über das Portal des Projekts zur Verfügung gestellt werden. Herauszuheben ist auch die Entwicklung einer eigenen Qualitätssicherungssoftware (DTAQ), die eine kollaborative Online-Korrektur der XML Texte erlaubt. MAXIMILIAN LANZINNER (Düsseldorf) schloss die erste Sektion mit einem Bericht über die „Acta Pacis Westfalicae“, einem 1997 ins Akademienprogramm aufgenommenen Projekt, dass 2011 beendet wurde und voraussichtlich ab Anfang 2014 online zur Verfügung stehen wird. Die Umsetzung der Online-Version wird von der Bayeri-

schen Staatsbibliothek (BSB) geleistet. Im Gegensatz zu den zuerst vorgestellten Korpora handelt es sich hier um eine digitale Ressource, die sich zunächst für Historiker und nicht in erster Linie für Sprachforscher bestimmt ist.

ULRIKE HENNY (Köln) stellte nach einer Pause den Aufbau eines digitalen Textzeugenarchivs zum Altägyptischen Totenbuch vor. Die Projektdaten wurden für die Online-Bereitstellung umfassend in XML konvertiert und werden nun unter konsequentem Einsatz von X-Technologien auf dem Portal präsentiert. Anschließend präsentierte GERFRIED MÜLLER (Würzburg / Mainz) das Hethitologie-Portal. Angesichts der Spezifik der beiden vorgestellten Korpora erhebt sich hier besonders die Frage nach den praktischen Möglichkeiten der von Gerhard Lauer geforderten Verknüpfung und Interoperabilität, die nur auf der Grundlage einer gemeinsamen Sprache und entsprechender Standards realisierbar erscheint.

Der Abendvortrag der Tagung wurde von GREGORY RALPH CRANE (Boston / Leipzig), dem Begründer der Perseus Digital Library <http://www.perseus.tufts.edu/>, zum Thema „Building a New Philology for Germany in the Twenty-First Century“ gehalten. Er berichtete über das Open Philology Project, dass er als Humboldt Professor in Leipzig in den kommenden fünf Jahren durchführen möchte. Die Philologie sei seit Jahrhunderten ausschließlich mit Hilfe arbeitsintensiver, manueller und hoch komplexe Expertenanalysen betrieben worden. Solche traditionellen Analysen blieben auch weiterhin von entscheidender Bedeutung, aber müssten nun durch andere, offene Formen der Arbeit ergänzt werden, um den neuen Herausforderungen gerecht zu werden: (1) die Verwaltung von Korpora der historischen Sprachen mit Milliarden statt Millionen von Wörtern, (2) Bereitstellung eines direkten Zugangs zu Primärquellen – auch ohne Kenntnisse der Originalsprache, und (3) Integration von Quellen aus vielen verschiedenen Korpora und Sammlungen.

Der zweite Tag begann mit einem Vortrag von JOST GIPPERT (Frankfurt am Main) zu dem diachronen Korpus „Thesaurus Indogermanischer Text- und Sprachmaterialien“ (TITUS). Er führte die Teilnehmer in die grundlegende Unterscheidung von synchronen, historischen und diachronen Korpora ein und erläuterte diese dann an Beispielen zu Sanskrit, Persisch und Georgisch. Für dieses Projekt wurde keine Textauszeichnung in TEI gewählt, da die Texte hierfür zu heterogen seien. Diese Bandbreite an unterschiedlichen Sprachen führte auch zur Verwendung von ISO 639-6 für die eindeutige Kennzeichnung von Sprachen im Projekt. Anschließend stellten VOLKER HARM und INGO KOTTSIEPER (beide Göttingen) zwei Wörterbücher vor, die von der Nutzung unterschiedlicher digitaler Workflows geprägt sind und gut den in den Grußworten angesprochenen Unterschied zwischen Alt- und Neuprojekten hinsichtlich der Berücksichtigung von Digitalisierungsstrategien verdeutlichen. Während das Grimmsche Wörterbuch (Harm) auf eine lange Tradition des analogen Arbeitens zurückblickt (Neubearbeitung seit 1960) ist das Qumran-Lexikon (Kottsieper) seit Projektbeginn 2007 digital angelegt und nutzt für den Aufbau des Lexikons intensiv eine Belegdatenbank. Obwohl es bislang noch nicht gelungen ist, die Neubearbeitung des Grimmschen Wörterbuchs selber online zur Verfügung zu stellen, so hat die Arbeitsgruppe doch schon Arbeitsmittel (z.B. eine Bibliographie) bereitstellen können (http://gso.gbv.de/DB=1.71/). Beide Referenten äußerten sich auch grundsätzlich zum Verhältnis von Korpus bzw. Belegarchiv und Wörterbuch, wobei sie Unterschiede und Gemeinsamkeiten herausarbeiteten. Nur die semantische Aufbereitung der Lemmata im Wörterbuch erlaube die Identifizierung von Phraseologismen, Erst- und Letztbezeugungen sowie eine Beschreibung der Textsortenverteilung. Für Nicht-Fachleute ist die Nutzung von Wörterbüchern zudem wesentlich einfacher als die direkte Arbeit mit den Korpora. Dabei können aber auch Korpus und Wörterbuch durchaus miteinander digital verschmelzen, wie das Beispiel des Qumran-Lexikons zeige.

Zwei landessprachlich geprägte Korpora bzw. Wörterbücher stellten KLAUS PUSCH (Freiburg) und MARTIN GRAF (Zürich) vor. Das „CIEL-F: Corpus international écologique de langue française“ sowie das „Schweizerische Idiotikon“ haben jeweils ein genau um-

rissenes Aufgabenfeld, wobei das Schweizer Projekt auch historisch angelegt ist und als Nachweis der Schweizerdeutschen Sprache vom Spätmittelalter bis heute fungiert, während CIEL-F ein synchrones Korpus des heute in allen Teilen der Welt gesprochenen Französisch ist. Zum Schluss der Tagung berichteten MICHAEL MARX (Berlin) über das „Corpus coranicum“ und HEIKO WEBER (Göttingen) über „Digitale Textkorpora und Analysewerkzeuge im Projekt „Johann Friedrich Blumenbach-online“. Die Möglichkeit der interdisziplinären Nutzung hochwertig aufbereiteter Texte und Daten wird gerade bei dem letzten Projekt deutlich, dass sich als Fachportal vor allem an Wissenschaftshistoriker wenden wird, dessen in TEI-P5-kodierte Texte aber schon jetzt aufgrund einer Kooperationsvereinbarung zwischen der Akademie der Wissenschaften zu Göttingen und der BBAW teilweise (in Hinsicht auf die deutschsprachigen Texte von Blumenbach) auch im Deutschen Text Archiv der BBAW mit sprachwissenschaftlichem Fokus nachgenutzt werden.

Am Ende der Tagung blieben viele Fragen der praktischen Umsetzung und der Probleme, die eine interdisziplinäre Verknüpfung und Nutzung von Korpora aufwirft, noch ohne eindeutige Antwort. Die durchaus repräsentative Vorstellung von Akademie- und Universitätsprojekten demonstrierte eindrücklich die Bandbreite der möglichen Disziplinen und Fragestellungen. Im Fachbereich der Philologie und der Linguistik sind es sicher gemeinsame Fragestellungen, die zur Einbeziehung unterschiedlicher Korpora in den Forschungsprozess führen können. Disziplinübergreifend wird eine solche Nachnutzung existierender digitaler Korpora mit den bislang vorhanden Zugangsmöglichkeiten wohl kaum gelingen – zu unterschiedlich sind die Fragen, zu speziell sind die Benutzerschnittstellen mit Blick auf die Bedürfnisse der Fachdisziplinen zugeschnitten. Dies trifft aber nicht unbedingt auf die Rohdaten selber zu, wie die Kooperation zwischen Blumenbach-online und DTA nahe legt. Auch in anderer Hinsicht stellen diese Projekte schon jetzt eine gewisse Ausnahme dar, da durch die Bereitstellung hervorragender Retrievalmöglichkeiten durch das DTA und in Zukunft auch durch das Portal von Blumenbach-online deren Korpora schon jetzt auch durchaus interdisziplinär eine interessante Ressource sind.

Bei allen „exotischen“ oder ausgestorbenen Sprachen scheint es dem gegenüber unerlässlich, Übersetzungen in modernen Sprachen anzubieten, die mit einfachen Mitteln und ohne umfangreiches Hintergrundwissen verwendet werden können, damit überhaupt eine fachübergreifende Benutzung möglich wird. Trotz der an sich noch unzureichenden multimodalen Verwendung von Korpora wurde durch die Berliner Tagung der AG „Elektronisches Publizieren“ ein wichtiger Schritt in die richtige Richtung getan, dem nun weitere Taten – sprich Kooperationen – zwischen Projekten unterschiedlicher Disziplinen folgen sollten.

Konferenzübersicht:

I. Sektion

Karin Donhauser (HU Berlin) Korpus – Begriff und Konzept
http://www.deutschdiachrondigital.de/home/

Alexander Geyken (Berlin): Historische Textkorpora für die Geists- und Sozialwissenschaften. Das Beispiel „Deutsches Textarchiv“
http://www.deutschestextarchiv.de

Maximilian Lanzinner (Düsseldorf): Acta pacis Westphalicae
http://www.pax-westphalica.de

II. Sektion

Ulrike Henny (Köln): Aufbau eines digitalen Textzeugenarchivs zum Altägyptischen Totenbuch
http://totenbuch.awk.nrw.de

Gerfrid Müller (Würzburg / Mainz): Das Hethitologie-Portal
http://www.hethport.uni-wuerzburg.de/HPM/hethportlinks.html

Abendvortrag

Prof. Dr. Gregor Crane (Tufts / Leipzig): Building a New Philology for Germany in the Twenty-First Century
http://www.dh.uni-leipzig.de/wo/

III. Sektion

Jost Gippert (Frankfurt am Main): Korpus Titus
http://titus.uni-frankfurt.de/indexd.htm

Volker Harm / Ingo Kottsieper (Göttingen): Das Grimmsche Wörterbuch / Das Qumran-Wörterbuch
http://www.uni-goettingen.de/de/118878.html

IV. Sektion

Claus Pusch (Freiburg): CIEL-F: Corpus international écologique de langue française
www.ciel-f.org

Martin Graf (Zürich): Schweizerisches Idiotikon
http://www.idiotikon.ch

V. Sektion

Michael Marx (Berlin): Corpus coranicum
http://koran.bbaw.de

Heiko Weber (Göttingen): Digitale Textkorpora und Analysewerkzeuge im Projekt „Johann Friedrich Blumenbach – online“
http://www.blumenbach-online.de

Tagungsbericht *Best practice: Digitale Korpora. Workshop der AG „Elektronisches Publizieren“*. 08.10.2013-09.10.2013, Berlin, in: H-Soz-u-Kult .