



Nummer 2, 2012, Mai

PID: 11858/00-1779-0000-0006-AAD1-C

Editorial

Zweiter CLARIN-D-Newsletter

Die neue Ausgabe des CLARIN-D Newsletters erscheint nicht nur pünktlich, sie ist auch noch deutlich umfangreicher als die erste. Ein herzliches Danke an die Autoren!

Der Newsletter beginnt mit zwei Beiträgen zur Positionierung von CLARIN-D, sowohl in Deutschland als auch in Europa. CLARIN ist überhaupt erst „das zweite Konsortium unter allen Initiativen ... das den ERIC-Status erhalten hat“. Und wenn man jetzt weiß, dass ERIC für „European Research Infrastructure Consortiums“ steht, dann erkennt man, welche Leistung das ist und welche Anforderungen das noch an uns stellen wird.

Im Zentrenbericht stellt sich der CLARIN-D-Partner Leipzig, die Abteilung Automatische Sprachverarbeitung unter Prof. Dr. G. Heyer, vor. Diese Abteilung

wird im Juni dieses Jahres den wichtigen Meilenstein-Workshop nach dem ersten Projektjahr veranstalten. Zu diesem Workshop werden über 100 Teilnehmer aus allen CLARIN-D-Zentren erwartet, vor allem aber auch aus den Facharbeitsgruppen, die bereits jetzt, von Leipzig koordiniert, die CLARIN-D-Infrastruktur und entwickelte Dienste und Tools nutzen.

Zur Öffentlichkeitsarbeit zählen auch die Präsentation von CLARIN-D-Diensten und -Tools auf nationalen und internationalen Tagungen und Messen. So wurde CLARIN auf der „International Conference on Research Infrastructures“ in Kopenhagen vorgestellt, auf der IDS-Fachmesse zu Korpustechnologie für Sprachdatenbanken, und natürlich auf den Facharbeitsgruppen- und Arbeitspaket-Workshops in Saarbrücken im März dieses Jahres. Auch auf der kommenden „Digital Humanities“-Konferenz in Hamburg wird CLARIN-D vertreten sein.

In der neuen Rubrik „Grenzgänge“ präsentieren wir interessante und unterhal-

tende Beiträge zu linguistischen, literaturwissenschaftlichen und verwandten Themen, hier natürlich mit einem Bezug zu CLARIN-D. Es ist sicherlich schon bekannt, dass man mittels statistischer Verfahren Textsorten klassifizieren und Texte Autoren zuordnen kann – in dieser Ausgabe wendet Uwe Reichel ein solches Verfahren an, um für zwei Texte zu klären, ob sie von Shakespeare stammen. Außerdem in dieser Ausgabe enthalten sind vier Workshop-Berichte.

Wie immer steht am Schluss die Bitte, Beiträge für den Newsletter einzureichen. Wir planen für die nächste Ausgabe eine Übersicht bereits vorhandener Dienste und Tools, denn so langsam können wir auf diesem für die Außenwirkung so wichtigen Gebiet auch schöne Demonstrationen und Systeme im täglichen Einsatz zeigen! Besonders sind wir natürlich an Berichten aus der Praxis interessiert

– wo wird ein CLARIN-D Dienst bereits erfolgreich eingesetzt, wo konnte ein Tool die Arbeit wesentlich erleichtern? Da muss es doch schon etwas geben!

Und ganz neu im Newsletter: ein eigenes Abkürzungsverzeichnis! Mit etwas Ehrgeiz müssten wir es schaffen, dass diese Rubrik in den nächsten Ausgaben den Umfang des ersten Newsletters übertrifft ... 😊

Christoph Draxler & Fabian Bross



V. i. S. d. P./Impressum:

Christoph Draxler
Ludwig-Maximilians-Universität München
Institut für Phonetik und Sprachverarbeitung
Schellingstr. 3
80799 München

Telefon: +49 (0) 89 / 2180 - 2807
E-Mail: newsletter@phonetik.uni-muenchen.de

Für die Inhalte der Artikel sind die jeweiligen Autoren verantwortlich.

Alles Weitere unter:

www.clarin-d.org

Bericht über die CLARIN-ERIC-Gründungsveranstaltung in Den Haag

Liebe Kolleginnen und Kollegen von CLARIN-D,

am 29. Februar 2012 hat die Europäische Kommission CLARIN den Status eines ‚European Research Infrastructure Consortiums‘ (ERIC) verliehen. Diese neue Rechtsform wurde eigens geschaffen, um Infrastrukturkonsortien mit den gleichen Vorteilen (z.B. Steuerfreiheit und Zugang zu nationalen und europäischen Förderprogrammen) auszustatten, die auch für andere Internationale Organisationen gelten. Nach dem SHARE-ERIC ist CLARIN das zweite Konsortium unter allen Initiativen auf der Roadmap des „European Strategy Forums for Research Infrastructures“ (ESFRI), das den ERIC Status erhalten hat.

Das CLARIN-ERIC hat mit Bulgarien, der Bundesrepublik Deutschland, Dänemark, Estland, den Niederlanden, der Niederlande Taalunie (einer gemeinsam von Belgien und den Niederlanden getragenen transnationalen Organisation), Polen, Österreich und der tschechischen Republik neun Gründungsmitglieder. Interessensbekundungen weiterer europäischer Staaten (u.a. von Finnland, Frankreich, Griechenland, Italien, Kroatien, Lettland und Litauen) liegen bereits vor, so dass die Anzahl der Mitglie-

der im CLARIN-ERIC zügig ansteigen wird. Die mittelfristige Zielsetzung sieht eine Mitgliederzahl von 20 Mitgliedern innerhalb der kommenden fünf Jahre vor.

Das CLARIN-ERIC wurde am 18./19. April 2012 mit einer feierlichen Gründungsversammlung im niederländischen Ministerium für Bildung, Kultur und Forschung in Den Haag offiziell eröffnet. An dieser Gründungsversammlung haben als Vertreter der Bundesrepublik Helge Kahler (als Vertreter des BMBF) und ich (als Koordinator von CLARIN-D) teilgenommen. Die Veranstaltung bestand aus zwei Teilen: der Mitgliederversammlung am 18.4. und der konstituierenden Sitzung des ‚National Coordinators Forum‘ am 19.4.

Die Bundesrepublik Deutschland wird sich an der Leitung des CLARIN-ERIC auf verschiedenen Ebenen aktiv beteiligen. Zum Präsidenten des CLARIN-ERIC wurden auf der Mitgliederversammlung Dr. Helge Kahler (BMBF) und zum Vizepräsidenten Jacek Gierlinski (Polen) gewählt.

Ich wurde auf der Sitzung des National Coordinators Forum zum Vorsitzenden des National Coordinator Forums gewählt und damit gleichzeitig in das

Direktorium des CLARIN-ERIC aufgenommen. Meine Stellvertreterin ist Kadri Vider (Estland).

Zum Executive Director des CLARIN-ERIC wurde Steven Krauwer (Universität Utrecht) und zu seiner Stellvertreterin Bente Maegaard (Universität Kopenhagen) gewählt. Das CLARIN-ERIC Direktorium besteht z.Zt. aus Steven Krauwer, Bente Maegaard und mir. Ein weiteres Direktoriumsmitglied wird in naher Zukunft aus dem Kreis des Standing Committees der europäischen CLARIN-Zentren hinzukommen.

Aus dem Ausgang der Wahlen ist deutlich abzulesen, dass die Mitglieder des CLARIN-ERIC sich von Deutschland und somit von CLARIN-D starke Impulse für das CLARIN-ERIC erhoffen. Ich bin zuversichtlich, dass wir dieser Rolle entsprechen können.

Ich gebe abschließend einen kurzen Ausblick auf die nächsten Schritte im CLARIN-ERIC.

Grundlage der CLARIN-ERIC-Arbeiten wird ein CLARIN-ERIC-Agreement zwischen allen Mitgliedsländern sein. Über dieses Agreement hatten wir im

Vorfeld der CLARIN-ERIC-Planungen bereits im Frühjahr 2011 gesprochen und über die Inhalte Konsens erzielt. Es wird nun bis Ende Juni darum gehen, dieses Agreement zu finalisieren. Am Template für das CLARIN-ERIC-Agreement werden noch letzte kleinere Änderungen vorgenommen. Sobald es in einer von allen CLARIN-ERIC-Mitgliedern akzeptierten Form vorliegt, werde ich es an alle CLARIN-D-Zentren weiterleiten.

Ich bin sehr froh darüber, dass es mit dem CLARIN-ERIC nun auch auf europäischer Ebene wieder zügig vorangeht. Auf den Sitzungen in Den Haag herrschte eine große Aufbruchstimmung unter den CLARIN-ERIC-Mitgliedern, ein sehr angenehmes und konstruktives Arbeitsklima und Vorfreude auf die anstehenden gemeinsamen Arbeiten.



Mit besten Grüßen
Erhard Hinrichs



CLARIN – Wo stehen wir?

Infrastruktur eines europäischen Langzeitvorhabens

Nach 5 Jahren Arbeit an den CLARIN-Ideen ist es angebracht, zu hinterfragen wo wir stehen und inwieweit unsere ursprünglichen Zielsetzungen richtig waren bzw. angepasst werden müssen.

CLARIN hat seine Arbeit im Jahre 2008 aufgenommen. Die Europäische Kommission hat auf der Basis der ES-FRI-Roadmap europäische Mittel bereitgestellt und in einigen Ländern wurden nationale Mittel, darunter auch in Deutschland vom BMBF, zur Verfügung gestellt. Die entscheidenden Diskussionen zwischen einigen Fachkollegen (Linguisten und Technologen) haben allerdings schon viel früher, etwa im Jahr 2006 begonnen, und diese haben letztlich zu einem Antrag geführt.

Man kann von einem Impuls sprechen, der uns in unserer Arbeit geleitet hat – sowohl auf der europäischen als auch auf der nationalen Ebene, obwohl es bereits in der ersten Phase verschiedene Schwerpunkte in den verschiedenen nationalen CLARIN-Projekten gab. Was uns alle einigte, war der Wille, in der Vorbereitungsphase nicht nur über Standards, Vereinbarungen, Organisationsmodelle

und ähnliches zu diskutieren, sondern wir wollten konkrete Infrastruktur-Pfeiler entwerfen, implementieren und in der Praxis testen.

Impuls unserer Arbeit

Auf der Basis der allgemeinen Zielsetzungen, die Erhard Hinrichs bereits sehr gut beschrieben hat (siehe Nr. 1), wurden auf der europäischen Ebene einige Aktivitäten definiert. Dabei lag eine Auffassung des Begriffes „Infrastruktur“ zugrunde, die sich sehr schön mit einem Beispiel aus der Welt der Eisenbahnen beschreiben lässt: Die Infrastruktur, das sind das Schienennetz, die Signalanlagen und die Bahnhöfe, die es ermöglichen, neuartige, schnellere Züge mit einer höheren Dichte fahren zu lassen. Ein Reisender nutzt die Züge, um von A nach B zu kommen, und das auf möglichst komfortable und schnelle Art – welche Infrastruktur dafür notwendig ist, interessiert ihn in der Regel wenig. Die Infrastruktur bleibt unsichtbar, nichtsdestotrotz muss sie vorhanden sein und Optionen für zukünftige Entwicklung beinhalten.

Natürlich wussten wir, dass eine solche Infrastruktur nur sinnvoll aufgebaut werden kann, wenn wir auch erste Tools („Züge“) entwickeln, um dann erste Nutzer einladen zu können, die Eckpfeiler der Infrastruktur testen. Konkret haben wir an den folgenden Themen gearbeitet: (1) Definition von Anforderungen

[1] <http://www.wissenschaftsrat.de/index.php?id=345>

für Zentren, die letztlich die stabile und persistente Infrastruktur tragen müssen und die Etablierung eines offenen Archiv-Angebots. (2) Etablierung einer distribuierten Infrastruktur zur Authentifizierung und Integration aller Zentren, die es ermöglicht, alle potentiell interessierten Nutzer teilhaben zu lassen. (3) Aufbau eines Systems, das es ermöglicht, Daten-Objekte und Tools in ihren unterschiedlichen Versionen zu registrieren und somit Integrität und Authentizität über Jahre zu gewährleisten. (4) Aufbau eines Metadaten-Systems, das flexibel genug ist, die verschiedenen Daten, Kollektionen und Tools zu beschreiben und das den direkten Zugriff auf die Daten etc. erlaubt. (5) Etablierung einer Infrastruktur, um das Arbeiten mit web-basierten Tools neben dem Arbeiten mit lokalen Tools zu ermöglichen. (6) Entwicklung einer distribuierten Such-Umgebung, um Metadaten-Suchen mit Inhalts-Suchen kombinieren zu können.

Der Spagat

In all den genannten Punkten haben wir, auch durch nationale Beiträge, enorme Fortschritte gemacht. Woher wissen wir das? CLARIN ist in ständigem Austausch mit anderen Infrastruktur-Ansätzen (Natur-, Lebenswissenschaften etc.) und wir können mit Fug und Recht behaupten, dass wir in verschiedenen Punkten die Agenda bestimmen konnten und treibende Kräfte waren, um Barrieren zu überwinden (oder Gräben zu überbrücken). Was aber, wenn wir einen Linguisten mit unseren Resultaten, über die wir stolz sein mögen, konfrontieren? Eine typische Antwort könnte sein: Das,

was die da machen, hat alles nichts mit mir zu tun. Wäre sie vollkommen falsch? Nein, denn sie zeigt uns unter anderem, dass wir unsere konkreten Ziele nach den 3 Jahren Arbeit neu justieren und offensichtlich noch mehr den täglichen Bedürfnissen der Wissenschaftler näher kommen müssen.

Lasst uns wiederum das Eisenbahn-Beispiel zur Erläuterung verwenden: Wir wissen, dass kein moderner Zug ohne spezielle Brücken auskommt, aber wir können den Nutzern nicht sagen: „Schaut, hier ist die Brücke, wie gut dann doch erst die neue Eisenbahn sein muss“. Der Nutzer will dann erst wissen, ob es auch bessere und schnellere Züge gibt. Wie können wir diesen Spagat hinbekommen, dass wir einerseits die eigentlich unsichtbare und notwendigerweise langfristig angelegte Infrastruktur bauen müssen und andererseits gemessen werden an der Zufriedenheit der Nutzer über die die Infrastruktur bevölkernden Züge, die wir eigentlich nicht bauen sollten?

Justierungen

Sowohl CLARIN-D, als auch CLARIN-NL haben nunmehr eine Strategie definiert, um Nutzergruppen mittels gezielter Projekte und Diskussions-Plattformen einzubinden. Demonstratoren und Tools, die hieraus entstehen, sind unsere Probezüge auf den neuen Strecken – der Nutzen, die möglichen Resultate sind für die Adressaten und zukünftigen Anwender direkt erkennbar. Momentan intensivieren wir diese Aktivitäten insofern, als wir mehrere bereits bestehende Datenarchive und breit genutzte Tools an die Infrastruktur anpassen, um somit ihre Verfügbarkeit für die wissenschaftlichen

Nutzer zu erhöhen.

Die Kernfrage wird allerdings bleiben, wie weit eine Infrastruktur involviert sein darf in die Entwicklung neuer Tools und Daten, die direkt die wissenschaftlichen Nutzer anspricht. Dies sollte eigentlich nach wie vor innovativen Forschungsprojekten vorbehalten bleiben. Forschungsinfrastrukturen sollten sich auf ihre Kernaufgaben konzentrieren: Integration, Interoperabilität, Sichtbarkeit, Verfügbarkeit, Persistenz, etc. Nur durch eine derartige Trennung ist auch zu gewährleisten, dass die Kostenstruktur transparent bleibt und die operativen Kosten der Infrastruktur niedrig bleiben.

Offensichtlich müssen wir nunmehr auch viel mehr Trainings- und Ausbildungs-Aktivitäten anbieten. Dabei geht es vor allem um die Vermittlung auf verschiedenen Ebenen: (1) Wir müssen alle sinnvollen Tools and Daten, die es in der Community gibt und die von den CLARIN-Zentren gepflegt werden, demonstrieren und erläutern – dies gilt für lokale Tools wie auch für web-basierte Tools. (2) Wir müssen darstellen, wie jeder einzelne Wissenschaftler seine Ressourcen und Tools in den CLARIN-Kontext einbringen kann, um sie langfristig zu sichern und sie sichtbar und wiederverwendbar zu machen. (3) Wir müssen insbesondere junge Wissenschaftler zu *hands-on*-Seminaren einladen, auf denen wir zeigen, wie zukunftsweisende Methoden angewendet werden und/oder wie neuartige Methoden hinzugefügt werden können.

Wir dürfen dabei auf keinen Fall hektisch werden, denn im Bereich der Infra-

struktur gibt es immer noch erhebliche Lücken wie z.B. bezüglich der distribuierten Authentifizierung, bei denen CLARIN von anderen abhängig ist. Bei derartigen Problemen können wir Behelfslösungen bauen, werden uns aber an die allgemeinen Trends anpassen müssen, wenn diese dann greifen.

Eine große Frage für uns alle ist die nach der europäischen Dimension. Ganz im Gegensatz zu vielen anderen Disziplinen scheinen die Linguistik bzw. die *Humanities* Bereiche zu sein, in denen viele ihr Heil in der Kleinstaaterei sehen – und dem müssen wir entgegenwirken. Seit Februar 2012 haben wir mit dem CLARIN-ERIC wiederum eine europäische Organisation. Wir werden diese engagiert und überzeugt mit Leben füllen müssen, um europäische Lösungen zu finden, d.h. auch nationale Engstirnigkeiten überwinden müssen.



Peter Wittenburg
MPI für Psycholinguistik, Nijmegen

Ein CLARIN-Center stellt sich vor: Die Abteilung Automatische Sprachverarbeitung der Universität Leipzig

Vom deutschen Wortschatz und *digital humanities*

Die Abteilung Automatische Sprachverarbeitung (ASV) [1] ist eine Abteilung des Instituts für Informatik der Fakultät für Mathematik und Informatik an der Universität Leipzig. Die Universität Leipzig wurde im Jahr 1409 gegründet und ist damit die zweitälteste Universität Deutschlands. Heute studieren über 30.000 Studenten an der Universität. Die Abteilung wurde im Jahr 1994 gegründet und steht von Beginn an unter der Leitung von Prof. Dr. Gerhard Heyer.

Die ASV versteht sich als Teil der Angewandten Informatik. Forschungsschwerpunkt ist die automatische Verarbeitung geschriebener Sprache. Hierzu zählen insbesondere Anwendungen des Textmining wie die automatische und semi-automatische Extraktion semantischer Relationen sowie deren Repräsentation und Nutzung. Hierbei werden sprachunabhängige Verfahren favorisiert. Ziel der Forschungen ist die Erstellung und Entwicklung von Daten, Verfahren und An-

wendungen, wobei das Credo „Investition in Algorithmen statt manuelle Annotation“ als grundlegendes Prinzip der Arbeiten an der ASV gelten kann.

Die ASV ist mit einem eigenen Forschungsbereich am Institut für Angewandte Informatik [2] (InfAI e.V.) vertreten. Aufgrund seiner besonderen Struktur als ein vom Senat der Universität Leipzig anerkanntes An-Institut werden dort zahlreiche industriennahe Forschungsprojekte durchgeführt.

Die ASV pflegt zudem den intensiven Kontakt zu und Austausch mit Partnern in der Industrie. Hierzu zählen unter anderem Volkswagen, SAP und DaimlerChrysler.

Projekt Deutscher Wortschatz

Eine wichtige Grundlage der Forschungen stellen die Daten des Projekts Deutscher Wortschatz [3] dar. Seit Mitte der 1990er Jahre werden im Wortschatz-Projekt die Texte großer Online-Nachrichtenportale und zahlreicher weiterer Quellen gesammelt und aufbereitet. Über das Wortschatz-Portal können auf diesen Daten basierende Informationen zu einem gegebenen Wort abgefragt werden. Hierzu zählen Frequenzen, Satz- und Nachbar-

[1] <http://asv.informatik.uni-leipzig.de/>

[2] <http://infai.org/>

[3] <http://wortschatz.uni-leipzig.de/>



Prof. Dr. Gerhard Heyer
Leiter der Abteilung Automatische Sprachverarbeitung

schaftskookkurrenzen und Beispielsätze. Das internationale Wortschatz Portal [4] erlaubt den Zugriff auf entsprechend strukturierte monolinguale Wortlisten in über 130 verschiedenen Sprachen. Über das Wortschatz-Portal [5] ist zudem der Zugang zu weiteren Diensten möglich, welche auf den Daten und Verfahren des Wortschatz-Projektes basieren oder aus diesen abgeleitet wurden. Dazu zählen die Wörter des Tages in deren Rahmen tagesaktuellen Begriffe aus ausgewählten Newsdiensten extrahiert werden. Eine weitere Anwendung stellt das Projekt FindLinks/NextLinks [6] dar. In FindLinks wird mit Hilfe eines dezentralen Webcrawlers die Struktur der Verlinkung einer großen Anzahl von Websites erfasst, aufbereitet und in NextLinks zur Erstellung einer Liste

[4] <http://corpora.informatik.uni-leipzig.de/>

[5] <http://wortschatz.uni-leipzig.de/wort-des-tages/>

[6] <http://wortschatz.uni-leipzig.de/nextlinks/>

[7] <http://wortschatz.uni-leipzig.de/Webservices/>

[8] <http://www.eaqua.net>

von Linkempfehlungen zur vom Nutzer gerade besuchten Webseite genutzt.

Ein wichtiges Anliegen der ASV ist die Bereitstellung der erstellten Daten und Verfahren zum Zweck der Forschung und Lehre. So stehen die Daten des Wortschatz-Projektes in großen Teilen für diesen Zweck frei zum Download zur Verfügung. Ein Teil der Daten ist zudem schon seit 2004 über Webservices [7] zugänglich. Belohnt wurde diese offene Politik von der Fachcommunity und aus der Wirtschaft durch die intensive Nutzung dieser Ressourcen. So wurden die Webservices in den Jahren von 2006 bis einschließlich Mitte April 2012 über 880 Millionen Mal genutzt, wobei allein 2011 über 260 Millionen Abfragen erfolgten. Für den Oktober 2012 wird erwartet, dass die 1-Milliarde-Zugriffsmarke erreicht wird.

Weitere Aktivitäten

Bereits in der Vergangenheit war die ASV in Projekten mit Bezug zu den eHumanities aktiv. Beispielhaft sei hier das Projekt eAQUA [8] (Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaften) genannt. Ziel des Projektes ist es aus antiken Quellen mittels speziell angepasster Methoden des Textminings Wissen zu extrahieren und über ein Webportal für die Forschung insbesondere in den Altertumswissenschaften in einer modernen und nachhaltigen Form zur Verfügung zu stellen. Im Zuge dieses Projektes konnte die ASV wichtige Erfahrung in der Kooperation mit den Partnern aus den Altertumswissenschaften sammeln

und neben der inhaltlichen Arbeit mit und Entwicklung von fortgeschrittenen Methoden des Textminings auch Wissen im Infrastrukturbereich aufbauen.

Momentan ist die ASV unter anderem in Form der Projekte eTRACES [9], INSEARCH [10], Visual Analytics [11] und CLARIN-D in diesem und verwandten Themengebieten aktiv. In der zweiten Hälfte des Jahres 2012 wird zudem in den Projekten ‚Post-democracy and Neoliberalism. On the usage of neoliberal argumentations in federal German politics 1949-2011‘ und ‚eXChange – Exploring Concept Change and Transfer in Antiquity‘ die Arbeit aufgenommen [12].

Das eTRACES-Projekt hat das Ziel, Textwiederverwendungen (*Text Re-use*) wie bspw. Zitationsspuren, Allusionen und Schablonen in meist juristischen Texten, aber auch Phrasen und Idiome in historischen digitalen Bibliotheken zu verfolgen (engl.: *to trace*). Hierbei wird auf sozialwissenschaftlichen Texten seit Anfang des 20. Jahrhunderts (in Kollaboration mit der GESIS), deutschsprachiger Literatur des 16. bis 19. Jahrhunderts (in Zusammenarbeit mit dem Göttinger ‚Centre of Digital Humanities‘) sowie auf antiken Texten der ‚Perseus Digital Library‘ gearbeitet.

Im Rahmen des INSEARCH-Projekts soll ein Informationssystem entwickelt werden, welches kleine und mittlere Unternehmen mittels fortgeschrittener Wissensverarbeitungs-, -retrieval und -analysemethoden in Innovationsprozessen unterstützen soll. Ein besonderes Augenmerk gilt dabei dem Aufdecken von sowohl intern als auch extern

vorhandenem Wissen, welches genutzt werden kann um innovative Produkte und Prozesse zu entwickeln.

Im vom Bundesministerium für Bildung und Forschung geförderten Projekt CLARIN-D beteiligt sich die ASV am Aufbau der Infrastruktur und verfolgt das Ziel der Etablierung eines CLARIN-Ressourcenzentrums in Leipzig. Die wichtigste Aufgabe stellt jedoch die Arbeit im Arbeitspaket 4 ‚Fachspezifische Arbeitsgruppen‘ [13] des Projektes dar, mit dessen Leitung die Abteilung betraut ist. Das Leipziger CLARIN-D-Team [14] setzt sich aus vier studierten Informatikern – V. Boehlke, T. Compart, T. Eckart und I. Schuster – und einem Fachinformatiker – T. Hynek – zusammen. Die Arbeitsgruppe wird von Prof. Dr. G. Heyer geleitet und in der täglichen Arbeit von V. Boehlke koordiniert.

Die Abteilung ASV ist maßgeblich am Aufbau der eHumanities an der Universität Leipzig beteiligt, die zukünftig neben der in Besetzung befindlichen Professur für ‚Computational Humanities‘ auch eine Professur für ‚Digital Humanities‘ am Institut für Informatik vorsieht. Diese Professur für Digital Humanities soll ab 2013 mit Professor Gregory Crane als zukünftigem Humboldt-Professor an der Universität Leipzig besetzt werden.



Volker Boehlke
*Institut für
Informatik,
Universität
Leipzig*

[9] <http://asv.informatik.uni-leipzig.de/projects/25> und <http://etraces.e-humanities.net/>

[10] <http://asv.informatik.uni-leipzig.de/projects/23> und <http://www.insearch-project.eu/>

[11] <http://asv.informatik.uni-leipzig.de/projects/17>

[12] <http://www.e-humanities.net/upcoming.html>

[13] <http://de.clarin.eu/index.php/de/aktivitaeten/arbeitspakete/ap-4-fachspezifische-arbeitsgruppen>

[14] <http://asv.informatik.uni-leipzig.de/projects/24>

Die *International Conference on Research Infrastructures*

Internationale Infrastrukturen

Vom 21. bis zum 23.3. fand in Kopenhagen die große ICRI (*International Conference on Research Infrastructures*) mit ca. 600 Beteiligten aus allen Ländern der Welt statt. Die ICRI ist DIE Konferenz, auf der alle politisch und strategisch relevanten Aspekte der Forschungs-Infrastrukturen besprochen werden mit dem Ziel, klare Handlungsanweisungen vor allem für die Europäische Kommission und die europäischen Mitgliedsstaaten abzuleiten.

Mithin ist die ICRI von großer Bedeutung auch für CLARIN und das neu gegründete CLARIN-ERIC. Überhaupt kann man sagen, dass CLARIN sehr prominent vertreten war, insofern als das CLARIN-ERIC das zweite ERIC ist und mithin alle beteiligten Ministerien zu recht „stolz“ darauf sind, dass der nicht gerade leichte Entscheidungsprozess erfolgreich abgeschlossen werden konnte, wie mehrfach in den *Keynote*-Vorträgen betont wurde. Für CLARIN ist es natürlich auch toll, dass wir wiederum ausgewählt wurden, mit einem *Plenary*-Vortrag zu der Session über „Data: A Common Challenge“ beizutragen, denn damit wurde die gute technologische Arbeit in CLARIN honoriert.

Ein großer Höhepunkt des ersten Tages war der Vortrag von Hans Rosling über seine *Gapminder*-Aktivitäten, in der er im Prinzip alle offiziell verfügba-

ren Informationen zu gesellschaftlichen Entwicklungen zusammenbrachte und mittels exzellenter Visualisierungen darstellte. Man kann nur jedem Interessierten empfehlen, sich die Applikation herunterzuladen (<http://www.gapminder.org/>). Ein weiterer Höhepunkt war sicherlich der Beitrag von Frau Vierkorn-Rudolph, im BMBF verantwortlich für Forschungs-Infrastrukturen und gegenwärtig auch Chair von ESFRI. Sie gab einen Überblick, in dem u.a. auch CLARIN genannt wurde. Allerdings wurde aus dem Beitrag auch deutlich, dass ESFRI jetzt von CLARIN noch mehr erwartet (siehe unten).

Hervorragende Sprecher aus verschiedenen Ländern, u.a. der Nobelpreisträger Shechtman, führten aus, wie Forschung heute betrieben wird, welchen Stellenwert die Forschung hat und welche Rolle exzellente Infrastrukturen spielen. In der Session über Globale Forschungs-Infrastrukturen wurde ganz klar, dass internationale oder auch europäische Herangehensweisen sehr gewünscht sind. Alan Blatecky (NSF) hat z.B. einen Vorschlag für eine internationale Organisation unterbreitet, die im Bereich des Daten-Managements und Zugriffs-Harmonisierungen aktiv werden soll, der momentan intensiv diskutiert wird.

Im weiteren wurden Sessions zu Entwicklungen bezüglich der *Grand Challenges* im Bereich Gesundheit, Klima, Energie und e-Infrastrukturen (Netzwerke, HPC, Grid, Daten) durchgeführt. In der letzten Session wurde über

die große Bedeutung der Daten für die moderne Wissenschaft diskutiert. Es gipfelte in dem Ausspruch, „data become the currency of research“.

Ganz im Sinne dieser letzten ICRI-Session wurde am Tag vor der ICRI-Konferenz der sogenannte DAITF-Workshop (Data Access and Interoperability Task Force, www.daitf.org) auch mit internationaler Beteiligung durchgeführt. Hier hat sich die Keimzelle einer Gruppe von „Data Practitioners“ getroffen und über die Topics der Harmonisierung wie z.B. PIDs, AAI und *Policy-Rules* diskutiert. Die nächsten Treffen wurden bereits geplant (September 2012, Washington, Februar/März 2013, Europa) und es wird sehr bald erste Arbeitsgruppen geben. DAITF ist eine Graswurzel-getriebene Initiative ähnlich wie die IETF (Internet Expert Task Force) und alle Beteiligten waren sich einig, dass diese Gruppe sich öffnen muss, so dass sich andere „Data Practitioners“ einbringen können. Jeder Experte mit fundierten und praktisch abgesicherten Auffassungen ist letztlich zur Mitarbeit eingeladen. Die Basis von DAITF ist „robust consensus“ auf der Basis von „running code“-Beispielen. Momentan wird ebenfalls diskutiert, wie die DAITF-Idee mit dem Organisationsvorschlag von Alan Blatecky zusammengebracht werden kann. Der Hintergrund ist der, dass ein reines Bottom-up-Verfahren wie z.B. IETF komplett dadurch charakterlich auf den Kopf gestellt wurde, dass die großen Telekommunikations-Unternehmen die Arbeitsgruppen dominieren und sich daher andere frustriert abwendeten.

Resümee für CLARIN

Für mich ergeben sich aus der ICRI-Konferenz und auch aus dem DAITF-Workshop eine Reihe von Folgerungen:

- Die Entwicklung in Richtung Forschungs-Infrastrukturen wird sich politisch in Zukunft eher noch stärker manifestieren, d.h. CLARIN fügt sich sehr gut in das politische Klima ein.
- Die Erwartung in Richtung auf europäische und sogar internationale Kollaborationen ist sehr stark, d.h. auch CLARIN muss den nationalen Raum, in dem wir uns oftmals befinden, verlassen.
- Für die politischen Schlüssel-Experten ist wichtig, ob sich eine Forschungs-Infrastruktur auch den großen Herausforderungen dieser Zeit annimmt. Ich wurde konkret gefragt, was denn die großen Herausforderungen für uns sein könnten und ich kam auf Dinge wie z.B. „Stabilität der Gesellschaften und Köpfe“, „Evolution der Sprachen in der Vergangenheit und vor allem in Zukunft durch die weltweite Migration“, „Verbesserung der Sprachtechnologie auf der Basis eines tieferen Verständnisses der Prozesse im menschlichen Gehirn“.
- Technologisch ist CLARIN mit seinen Fragestellungen und auch Lösungsansätzen in der Spitzengruppe und dies wird auch von vielen gesehen.

Es bedarf eines großen linguistischen Entwurfes. Dabei denke ich auch an unsere Interoperabilitäts-Diskussion. Nahezu alle oft genannten FI-Initiativen arbeiten an der Interoperabilität in großem Maßstab, wissend das die Herausforderungen nicht trivial sind – aber sie packen das Problem an. In unserem Bereich sind derartige Ansätze noch viel zu halbherzig und daher nicht zukunftsweisend.

Peter Wittenburg

CLARIN-D-Panel bei der DH 2012 in Hamburg

Ein Panel zu „Large Historical Reference Corpora of German“ an der Uni Hamburg

CLARIN-D wird auf der Digital Humanities 2012 (16.–22. Juli 2012 in Hamburg [1]) mit einem Panel zum Thema „Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints“ vertreten sein [2].

Das Panel wird veranstaltet von dem Projekt ‚Deutsches Textarchiv‘ an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW), der Herzog August Bibliothek Wolfenbüttel (HAB) sowie von der *Special Interest Group Deutsche Philologie* in CLARIN-D. Es werden die Aktivitäten der Projekte DTA und AEDit (Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit) zur Zusammenführung, Homogenisierung und Analyse verstreuter Textressourcen und deren Bereitstellung im Kontext von CLARIN-D vorgestellt.

Eines der Hauptziele der *Digital Humanities* im deutschsprachigen Raum besteht darin, große Referenzkorpora für die deutsche Sprachgeschichte aufzubauen. Bereits jetzt existieren Unternehmungen, derartige Korpora aufzubauen. So erarbeitet das Projekt „Deutsches Textarchiv“ sukzessive ein Referenzkorpus des Neuhochdeutschen für den Zeitraum vom 17. bis zum 19. Jahrhundert. Die HAB Wolfenbüttel erstellt Korpora für das Frühneuhochdeutsche des 15. bis 18. Jahrhunderts. Neben diesen Unternehmungen, die normalerweise im Kontext größerer Institutionen stehen, werden vielerorts durch Einzelwissenschaftler oder kleinere Projekte weniger umfangreiche Textressourcen zu spezifischen Forschungsfragen aufgebaut. Diese finden jedoch in der Regel keinen Platz unter den öffentlich zugänglichen Recherchekorpora des Fachs.

Es ist daher notwendig, eine integrierte Korpus-Infrastruktur für große Korpora des historischen Deutschen aufzubauen. In diesem Zusammenhang stellt das Panel Bemühungen vor, zum Aufbau einer Community, zur Textakquise, zur technischen Umsetzung der Zusammenführung von Teilkorpora sowie zur diesbezüglichen Zusammenarbeit der beiden großen Infrastrukturzentren BBAW und HAB im Rahmen der Publikations-

[1] <http://www.dh2012.uni-hamburg.de/>

[2] <http://de.clarin.eu/images/veranstaltungen/dh2012panel.pdf>

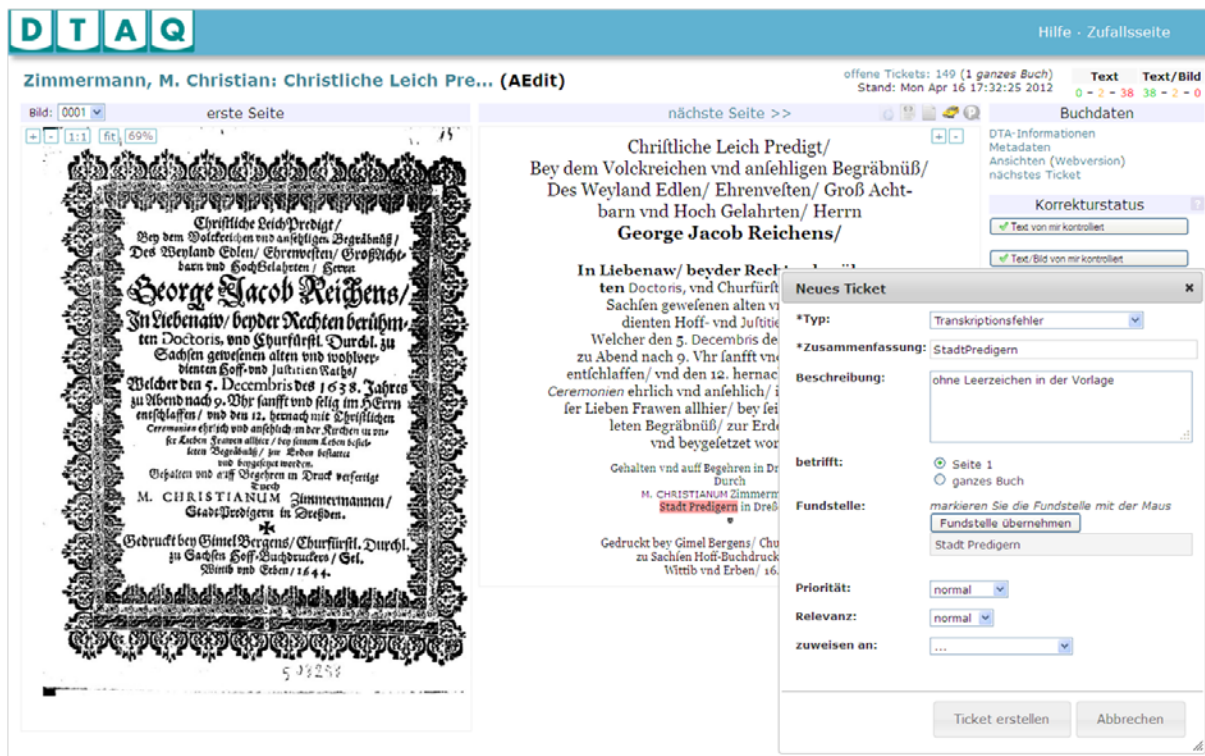
umgebungen DTA und AEDit. Darüber hinaus wird die Problematik geeigneter Standardisierungen zur Sicherstellung der Interoperabilität von Textressourcen diskutiert und werden Methoden und Tools für die Qualitätssicherung

auf der Ebene der Texttranskription und -annotation vorgestellt.

Alexander Geyken, Thomas Gloning, Thomas Stäcker

Überblick

1. Einführung: Alexander Geyken (DTA, DWDS, CLARIN-D)
2. Das DTA ‚Basisformat‘: Susanne Haaf (DTA, CLARIN-D), Alexander Geyken (DTA, DWDS, CLARIN-D)
3. DTA Erweiterungen – DTAE: Christian Thomas (DTA)
4. DTA Qualitätssicherung – DTAQ: Frank Wiegand (DTA)
5. Das Projekt AEDit: Thomas Stäcker (HAB)
6. CLARIN-D – Historische Korpora, Kollaboration, Aufbau der Community: Thomas Gloning (Universität Gießen, CLARIN-D)



The screenshot shows the DTAQ web interface. On the left, there is a thumbnail of a historical document titled "Zimmermann, M. Christian: Christliche Leich Pre... (AEdit)". The document text is framed in a decorative border and includes the name "George Jacob Reichens" and the date "1638". On the right, the main text of the document is displayed, including the title "Christliche Leich Predigt/" and the author "George Jacob Reichens/". Below the text, there is a "Neues Ticket" form with fields for "Typ" (Transkriptionsfehler), "Zusammenfassung" (StadtPredigern), "Beschreibung" (ohne Leerzeichen in der Vorlage), "betrifft" (Seite 1), "Fundstelle" (markieren Sie die Fundstelle mit der Maus), "Priorität" (normal), and "Relevanz" (normal). The form also includes buttons for "Ticket erstellen" and "Abbrechen".


Zusammenführung historischer Korpora im Rahmen von DTAE und AEDit; Qualitätssicherung

CLARIN-D auf der Messe „Technologie für mündliche Sprachkorpora“ des IDS

Bericht von der IDS-Tagung „Das Deutsch der Migranten“ in Mannheim

Im März fand in Mannheim die Jahrestagung des Instituts für Deutsche Sprache (IDS) zu dem Thema „Das Deutsch der Migranten“ statt. Flankierend wurde die „Messe zur Technologie für mündliche Sprachkorpora“ durchgeführt. Die Teilnehmer der Jahrestagung hatten den Tag über Gelegenheit, sich an den Ständen über die verschiedenen vorgestellten Projekte zu informieren. Das CLARIN-

FORSCHUNGSINFRASTRUKTUR FÜR SPRACHRESSOURCEN IN DEN GEISTES- UND SOZIALWISSENSCHAFTEN



Das Projekt

Das Ziel des ESFRI-Projekts CLARIN-D ist der Aufbau eines mit ausgewählten Fachdisziplinen eng verbundenen Zentrenverbunds als Rückgrat einer Forschungsinfrastruktur insbesondere für Wissenschaftlerinnen in den Geistes- und Sozialwissenschaften. In der Summe decken die ausgewählten Disziplinen ein breites Spektrum der Geisteswissenschaften ab, für die Sprachressourcen eine zentrale Rolle in der Forschung spielen.

Highlights

- Metadaten für Sprachressourcen und -dienste
- Virtuelle Arbeitsumgebungen in der Cloud
- Langfristige Verfügbarkeit und Archivierung
- Kuratieren fachrelevanter Ressourcen
- Richtlinien für Urheber- und Datenschutzrecht
- Dissemination von Ressourcen und Werkzeugen
- Schulung und Ausbildung
- Help-Desk und technischer Support

Szenarien

Frage: Wie kann ich meine Ergebnisse und Rohdaten mit Kollegen teilen?

Antwort: Die sichere Authentifizierung in CLARIN-D sowie die aufeinander abgestimmten Austauschformate erlauben Kollegen den Zugriff auf freigelegte eigene Datenbestände und die Nutzung von Webdiensten.

Frage: Wie lasse ich für mich relevante Daten und Dienste?

Antwort: In CLARIN-D sind alle Daten und Dienste durch Metadaten beschrieben. Diese Metadaten können automatisch durchsucht werden. In einer virtuellen Arbeitsumgebung werden die gefundenen Daten und Dienste präsentiert und zur weiteren Auswertung angeboten.

Frage: Wie bekomme ich die Ausgabe meines Lieblingstagger in meinem Parser?

Antwort: CLARIN-D bietet eine Verknüpfung von linguistischen Tools zu einer Verarbeitungspipeline an. Damit können Nutzer auf einer Vielzahl von CLARIN-kompatiblen Tools die benötigten ausführen und sie sinnvoll verketten - die technischen Details bleiben verborgen.

eHumanities
Einsatz moderner Informationstechnologien zur Unterstützung geisteswissenschaftlicher Forschung

Daten
Textkorpora, Audio/Video

Nachhaltigkeit
Interoperabilität, Archivierung

Weblicht
<https://weblicht.stb.uni-tuebingen.de/WEBLIGHT/>

TVIEWER
<http://clarin-d.de/index.php/en/language-resources/web-light-ecv-tutorials>

Virtual Language Observatory
<http://www.clarin.eu/en/observatory.php>

Dynamic Corpus Analyzer
<http://www.fhpmw-jastrow.de/2001-00a/>

Anwendungen, die auf die CLARIN-D-Infrastruktur aufbauen

Kontakt

Projektkoordinator:
Prof. Dr. Erhard Hinrichs
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 19
72074 Tübingen
Tel.: +49 7071 29 74279
Fax: +49 7071 29 52 14
Email: erhard.hinrichs@uni-tuebingen.de

www.clarin-d.de

CLARIN-D-Zentren:

- Bayerisches Archiv für Sprachsprache, Ludwig-Maximilians-Universität München
- Berlin-Brandenburgische Akademie der Wissenschaften
- Institut für Deutsche Sprache, Mannheim
- Max-Planck-Institut für Psycholinguistik, Mönchengladbach
- Eberhard Karls Universität Tübingen, Seminar für Sprachwissenschaft
- Universität Hamburg, Zentrum für Sprachkorpora
- Universität Leipzig, Institut für Informatik
- Universität des Saarlandes, Englische Sprach- und Übersetzungswissenschaft
- Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Rechenzentren:

- Rechenzentrum Garching (RZG) der Max-Planck-Gesellschaft und des IPP
- Forschungszentrum Jülich
- Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWVDG)

Webpräsenz des europäischen Langzeitprojekts: **www.clarin.eu**

D-Projekt präsentierte sich mit einem eigenen Stand, an dem WebLicht sowie einige weitere Web-Anwendungen und -Ressourcen live und online vorgeführt wurden. Neben dem CLARIN-D-Projekt war das CLARIN-D-Logo auch an weiteren Ständen präsent – einige CLARIN-D-Partner stellten ihre individuellen Werkzeuge und Projekte auf der Messe vor: ELAN (MPI), Exmaralda (Hamburg), das TLA (BBAW und MPI), Folk & Folker (IDS) sowie LiS (Hamburg und IDS).

Die Fragen der Besucher am CLARIN-D-Stand erstreckten sich über ein weites

Spektrum. Es wurden sowohl allgemeine Fragen über das CLARIN-D-Projekt als auch konkrete Fragen zur WebLicht-Architektur gestellt. Ein wiederholt auftretendes Thema war die Verwendung von WebLicht und anderen Teilen der CLARIN-D-Infrastruktur in der akademischen Lehre. Weitere Fragen betrafen das geplante *Single-Sign-On*-Verfahren in CLARIN-D sowie die Verfügbarkeit der gezeigten Tools und Ressourcen.

Thomas Zastrow
Seminar für Sprachwissenschaft
Universität Tübingen



2. F-AG-Workshop (29.03.2012; Saarbrücken)

Aktuelle Arbeitsergebnisse der Facharbeitsgruppen

Hauptanliegen des zweiten F-AG-Workshops [1] war die Präsentation und Diskussion aktueller Arbeitsergebnisse der Facharbeitsgruppen. Ein zweites Ziel stellte die Vertiefung des Kontakts zwischen den Facharbeitsgruppen und den Mitgliedern der CLARIN-D-Ressourcenzentren dar. Die Agenda des Workshops wurde an diesen Zielen ausgerichtet und basierte auf den Rückmeldungen und Fragen, welche in den vergangenen Projektmonaten gesammelt wurden.

Eine Präsentation zum ersten offiziell genehmigten Kurationsprojekt eröffnete den Workshop. Prof. H. Baayen und I. Schuster (F-AG 5) stellten das Projekt „Open Science“ vor. Im weiteren Verlauf des Workshops folgten Beiträge der Facharbeitsgruppen zu den Themen „Anregungen zur Kuration von NLP Werkzeugen in CLARIN-D“ (Prof. Frank; F-AG 7), „Lizenzierte Ressourcen – Umgang mit rechtlichen Fragen“ (Prof. Schubert F-AG 4) und „Dokumentation digitaler

Ressourcen im Clarin-Kontext“ (Prof. Gloning; F-AG 1). Die Zentren informierten zu den Themen „CLARIN-D Infrastruktur“ (D. v. Uytvanck; Technical Management / AP3), „Rechtliche und ethische Fragen“ (Erik Ketzan; AP6) und „Erstellung eines Evaluationshandbuchs“ (A. Herold; AP5).

Zwischen diesen Präsentationen fand eine Demosession [2] statt. Im Rahmen dieser Session wurde ein Teil der momentan an den Ressourcenzentren für CLARIN-D kuratierten bzw. in Beziehung zu CLARIN-D stehenden Tools und Ressourcen vorgestellt. Die Workshopteilnehmer hatten die Gelegenheit die Ressourcen und Tools in Aktion zu erleben und mit den jeweiligen Entwicklern ins Gespräch zu kommen. Wir streben an, diese Art der Interaktion in einer ähnlichen gestalteten Session im Rahmen des M12-Workshops in Leipzig [3] erneut zu ermöglichen, um die zukünftigen Anwender der CLARIN-D-Infrastruktur auch über die jeweiligen inhaltliche Themen der verschiedenen Fachwissenschaften für das Projekt zu interessieren.

Aus unserer Sicht stellte der Workshop einen Erfolg dar. Es war klar erkennbar, dass die Facharbeitsgruppen in-

[1] <http://fr46.uni-saarland.de/index.php?id=f-ag-workshop>

[2] http://fr46.uni-saarland.de/fileadmin/user_upload/lehrstuehle/Teich/Kolloquium/CLARIN-D_Tage_in_SB_2012/Demo_All.pdf

[3] <http://clarin.informatik.uni-leipzig.de/>

Mitmachen!

Liebe Leser des CLARIN-D-Newsletters, wenn ihr Ideen für einen kurzen Beitrag zu diesem Newsletter habt oder dringend einen Gedanken loswerden wollt, schickt euren kurzen Artikel samt Bild an newsletter@phonetik.uni-muenchen.de. Hinweise zur Beitragsgestaltung findet ihr im Wiki.

zwischen ein tiefes Verständnis für die in CLARIN-D adressierten technischen und rechtlichen Probleme und den jeweils angestrebten Lösungswegen gewonnen haben. Eine Entwicklung, die sich ebenfalls klar in den Anträgen zu CLARIN-D-Kurationsprojekten niederschlägt. Die zunehmende Verflechtung der Facharbeitsgruppen und der Ressourcenzentren wurde ebenso deutlich. Bereits im Vorfeld des Workshops haben die Facharbeitsgruppen und Zentren in verschiedenen Fragen erfolgreich kooperiert. Die gegenseitige Information über diese und weitere Tätigkeiten im Rahmen von Veranstaltungen wie dem zweiten F-AG-Workshop soll den Mitgliedern der Zentren und Facharbeitsgruppen ermöglichen, sich an diesen Aktivitäten zu

beteiligen oder neue Kooperationen einzugehen. Aufgrund der zahlreichen Anregungen und Diskussionen im Rahmen des Workshops gehen wir davon aus, dass analog zum ersten F-AG-Workshop auch dieses Mal neue Initiativen zu aktuellen Fragestellungen geformt werden können, welche das CLARIN-D-Projekt entscheidend voranbringen werden.

Wir möchten allen Vortragenden und allen Teilnehmern für die interessanten Beiträge und Diskussionen danken. Besonderer dank gilt zudem unseren Gastgebern aus Saarbrücken für die hervorragende Organisation und die gute Zusammenarbeit im Vorfeld der Veranstaltung.

Das Leipziger CLARIN-D-Team

Weitere Infos auch im Wiki:

<http://de.clarin.eu/mwiki>

Grenzgänge

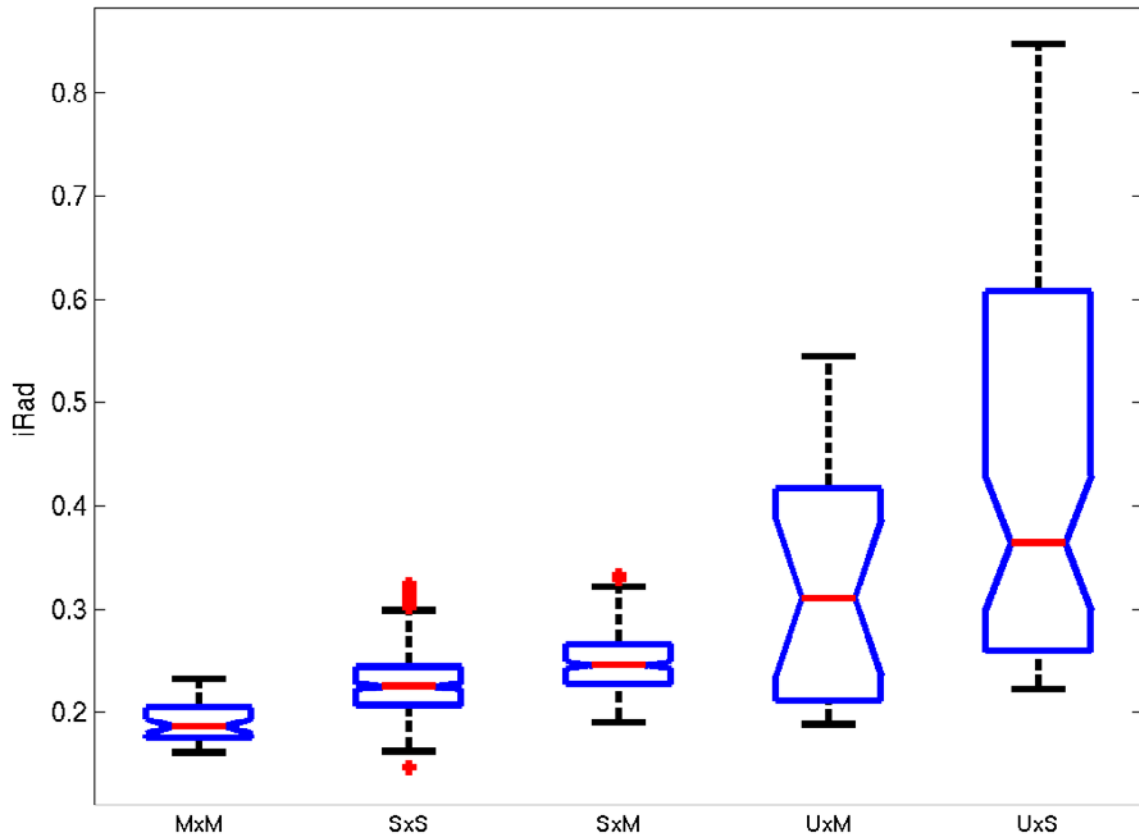
In der Rubrik Grenzgänge berichten Forscher erstaunliche, ungewöhnliche oder amüstante Ergebnisse. Dieses mal: Uwe Reichel über **Shakespeare und Wahrscheinlichkeit**

Eines der Ziele von CLARIN-D ist es, Daten und Werkzeuge zu deren maschineller Analyse für die geisteswissenschaftliche Forschung zusammenzubringen. Das Erkenntnispotential, das sich aus dieser automatisierten Analyse großer Datenmengen ergibt, soll hier an einem Beispiel aus der Shakespeare-Forschung dargestellt werden.

Neben Expertenurteilen können auch mathematische Ansätze zur Klärung umstrittener literarischer Urheberchaften beitragen, beispielsweise durch Heranziehen des Informationsradius (IRad). Der IRad ist ein quantitatives Maß, mit dem sich die Unähnlichkeit zweier Wahrscheinlichkeitsverteilungen ermitteln lässt. Berechnet man beispielsweise für zwei Theaterstücke jeweils eine Wahrscheinlichkeitsverteilung über die vorkommenden Wörter, drückt der IRad die Verschiedenheit der Stücke im Hinblick auf ihr Vokabular aus. Diese Eigenschaft haben wir zur Untersuchung

zweier Theaterstücken genutzt, die ursprünglich Shakespeare, mittlerweile aber eher dessen Zeitgenossen Thomas Middleton zugeordnet werden: *A Yorkshire Tragedy* und *The Puritaine Widdow*. Hierfür wurden für diese Stücke maschinell IRad-Werte ermittelt in Paarungen mit Shakespeare-Stücken ($U \times S$) und mit Middleton-Stücken ($U \times M$). Zusätzlich wurden die Shakespeare-Stücke untereinander ($S \times S$), die Middleton-Stücke untereinander ($M \times M$), sowie alle Shakespeare- und Middleton-Stücke kreuzweise ($S \times M$) verglichen. Die resultierenden IRad-Werte wiesen folgende allesamt statistisch signifikante Unterschiede auf: $M \times M < S \times S < S \times M < U \times M < U \times S$, woraus sich mehrere Schlussfolgerungen ziehen lassen:

1. $M \times M, S \times S < S \times M$: Die beiden Autoren sind anhand ihres Vokabulars gut auseinanderzuhalten.
2. $M \times M < S \times S$: Shakespeares Vokabular ist variationsreicher als Middletons, oder



Die resultierenden iRad-Werte aus dem Vergleich der verschiedenen Werke

es stammen nicht alle betrachteten Werke von ihm.

3. $S \times S < U \times S$: Die Zweifel an der Urheberschaft Shakespeares sind für die besagten Stücke berechtigt.

4. $U \times M < U \times S$ ist als Indiz für die Autorenschaft Middletons zu werten, aber

5. $M \times M < U \times M$: mit gewisser Vorsicht auf Grund nicht allzu stark ausgeprägter Ähnlichkeit.

Das maschinell ermittelbare iRad-Maß liefert damit Ergebnisse, die im Einklang mit der gängigen Expertenmeinung ste-

hen, und bietet darüberhinaus flexible Einsatzmöglichkeiten in der geisteswissenschaftlichen Forschung, beispielsweise zum Vergleich der Verwendung bestimmter Stilfiguren.



Uwe Reichel
*Institut für Phonetik und Sprachverarbeitung,
 LMU München*

Zweiter AP 8-Workshop am 30.3.2012 in Saarbrücken

Die Integration von Sprachressourcen und Sprachwerkzeugen

Im Anschluss an den F-AG-Workshop fand am Freitag den 30. März 2012 in Saarbrücken der zweite Workshop des Arbeitspaketes 8 ‚Schulung und Ausbildung‘ statt. Thema war zum einen die Kooperation mit DARIAH sowie mit AP7 ‚Support und Helpdesk‘ und zum anderen die Integration von Sprachressourcen und Sprachwerkzeugen in die universitäre Lehre.

Teilnehmer waren Vertreter von DARIAH aus Darmstadt, Mainz und Würzburg sowie des AP7s aus Hamburg, F-AG-Mitglieder, -Mitarbeiter und -Dozenten der Universität des Saarlandes sowie zwei eingeladene Sprecherinnen: Caroline Sporleder (Universität des Saarlandes) und Sabine Bartsch (Universität Darmstadt).

Es gab kurze Vorträge zu den Arbeitspaketen im Bereich Schulung und Ausbildung von CLARIN-D- und DARIAH-Vertretern. Es wurde festgestellt, dass Ideen und Probleme oft ähnlich sind, dass die unterschiedlichen Zielgruppen im Detail jedoch unterschiedliche Vorgehensweisen bedingen. Möglichkeiten zur Vernetzung und Kooperation wurden herausgearbeitet, z.B. die Koordi-

nation von Schulungsmaßnahmen bei wichtigen Veranstaltungen (Tagungen, Sommerschulen) oder gegenseitige Besuche von Arbeitstreffen.

In Bezug auf die Integration von Sprachtechnologie und Sprachressourcen in die Lehre wurden verschiedene Möglichkeiten diskutiert. So berichtete Caroline Sporleder (Universität des Saarlandes) über ihr Projektseminar ‚Text Mining in Historical Documents‘, das sie in Kooperation mit dem Institut für Geschichte durchführt und das sowohl Studierende der Geschichte, der Computerlinguistik als auch der Informatik besuchen. Sabine Bartsch (TU Darmstadt) stellte das Linguisticsweb Portal [1] vor, auf dem für Studenten der Linguistik sowie Wissenschaftler Informationsmaterial zu Methoden und Techniken in der Sprachverarbeitung bereitgestellt wird und das in enger Zusammenarbeit mit Studenten für Studenten aufgebaut wurde und weiter gepflegt wird.

Die sehr konstruktiven Diskussionen ergaben unter anderem die Notwendigkeit gezielt Dozenten anzusprechen, ihnen die Möglichkeiten zur Integration von Sprachtechnologie in die universitäre Lehre aufzuzeigen und Hilfestellung bei der Umsetzung zu bieten (Schulungen, Lehrmaterialien, aber auch Materialien zum Selbststudium). Im Rahmen von CLARIN-D wird eine Sammlung von

[1] <http://www.linguisticsweb.org/>

Materialien für Lehre und Selbststudium aufgebaut, die von den verschiedenen Communities mit getragen werden muss und eng mit der Infrastruktur für Support und Helpdesk verknüpft sein soll. Es wurde angemerkt, dass die Fachverbände bei der Bewertung und Einordnung aber auch bei der Erstellung und Verbesserung des angebotenen Ma-

terials eine Rolle spielen sollten.

Die erhaltenen positiven Rückmeldungen und unsere eigenen Eindrücke zeigen, dass der Workshop für alle Beteiligten neue Impulse gegeben hat. Wir danken allen Vortragenden und Teilnehmern und freuen uns auf die weitere Zusammenarbeit.

Das Saarbrücker CLARIN-D-Team

Abkürzungsverzeichnis

AAI	Authentication and Authorization Infrastructure
AEDit	Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit
AP	Arbeitspaket
BAS	Bayerisches Archiv für Sprachsignale (München)
BBAW	Berlin-Brandenburgische Akademie der Wissenschaften
BMBF	Bundesministerium für Bildung und Forschung
CLARIN	Common Language Resources and Technology Infrastructure
DAITF	Data Access and Interoperability Task Force
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DH	Digital Humanities
DTA	Deutsches Text Archiv
DTAQ	DTA-Qualitätssicherung
ELAN	EUDICO Linguistic Annotator
eAQUA	Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum for Research Infrastructure
EUDICO	European Distributed Corpora Project
F-AG	Fachspezifische Arbeitsgruppen
FI-Initiative	Forschungs-Infrastruktur-Initiative
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
HAB	Herzog August Bibliothek Wolfenbüttel
HPC	High Performance Cluster
ICRI	International Conference on Research Infrastructures
IDS	Institut für Deutsche Sprache (Mannheim)
IETF	Internet Expert Task Force
IMS	Institut für Maschinelle Sprachverarbeitung (Stuttgart)
InfAI e.V.	Gemeinnütziger Verein des Instituts für Angewandte Informatik in Leipzig
LiS	Literatur- und Informationsversorgungssysteme
MPI	Max-Planck-Institut
NSF	National Science Foundation (USA)
PID	Persistent Identifier
SHARE	Survey of Health, Ageing and Retirement in Europe
TLA	The Language Archive