CLARIN-D
Newsletter

Nummer 3, 2012, November

PID: 11858/00-1779-0000-0009-1759-9

Editorial

Dritter CLARIN-D-Newsletter

Der dritte CLARIN-D-Newsletter steht ganz im Zeichen der Anwendung von CLARIN-D-Technologie und -Techniken in den Geisteswissenschaften. Ein herzliches Danke an die Autoren!

Eva-Maria Wunder ist Doktorandin an der Universität Augsburg. Sie erforscht Transfer-Phänomene, die beim Erlernen mehrerer Fremdsprachen auftreten. Sie hat sich dazu eine sehr spezielle Gruppe von Lernern angeschaut: Mandarin-Sprecher, die Englisch und Deutsch lernen. In ihrem Beitrag zeigt sie, welche CLARIN-Tools sie verwendet hat und wie diese sich im Einsatz bewährt haben.

Die AsiCa-Sprachdatenbank wurde am Institut für Romanistik in enger Kooperation mit der IT-Gruppe Geisteswissenschaften der LMU erstellt. Als erste große externe Anwendung der CLARIN-D-Metadatenprofile für multimodale Korpora hat das BAS diese Sprachdatenbank erfolgreich in das eigene Repository einpflegen können.

In der Rubrik „Grenzgänge“ geht es kulinarisch zu – und gleichzeitig historisch. Aus zwei alten, handgeschriebenen Kochbüchern soll eine CLARIN-D-kompatible Sprachressource werden. Der Weg dorthin erfordert viele einzelne Arbeitsschritte. In diesem Newsletter machen wir Ihnen den Mund wässrig und in loser Folge beschreiben wir dann in den nächsten Newsletter-Ausgaben die weiteren Arbeitsschritte.

Von der BBAW kommt ein Beitrag zu einem aktuellen Kurationsprojekt der Facharbeitsgruppe 1 „Deutsche Philologie“. In diesem Projekt geht es um historische Textressourcen und wie diese in die CLARIN-D-Infrastruktur eingebettet und dort langfristig verfügbar gemacht werden können.

Und natürlich gibt es auch wieder eine Zentrenbeschreibung: das Hamburger Zentrum für Sprachkorpora stellt sich mit seinen Forschungsprojekten und Sprachressourcen vor.

Weiterhin berichtet Jens Stegmann vom 7. CLARIN-D-Entwicklertreffen, das am 24. September in Stuttgart stattfand. Und Heike Zinsmeister fasst zusammen, was am gleichen Tag ebenfalls in Stuttgart bei einem Workshop zum Wortartentagging diskutiert wurde.

Zu guter Letzt kommen noch zwei Berichte vom M12-Workshop aus Leipzig. Dieser Workshop war die erste große Gelegenheit, CLARIN-D einer breiteren Fachöffentlichkeit und den internationalen Gutachtern vorzustellen und die Resonanz fiel insgesamt sehr positiv aus.

Wie üblich steht hier noch der Aufruf: **der Newsletter lebt von Ihren Beiträgen!** Berichten Sie von Ihren CLARIN-D-Aktivitäten, von Konferenzen und Workshops oder, und das ist für CLARIN-D besonders wichtig, von jungen Wissenschaftlern, die CLARIN-D-Technologien einsetzen. Zum Schluss des Newsletters gibt es, wie beim letzten Mal, die „Never-Ending-List der CLARIN-Abkürzungen“ (NELCA).

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

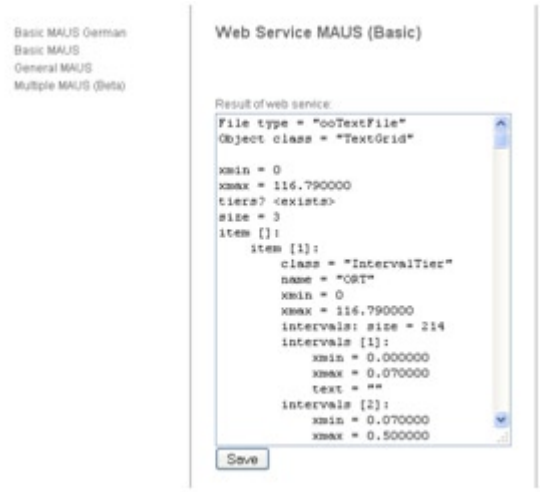
Erfahrungsbericht: CLARIN-Tools *Wiki-Speech* und *WebMAUS*

„Trifft ein Chinese auf eine Augsburgerin, die mit Münchner Software an einer Doktorarbeit in Münster arbeitet …“

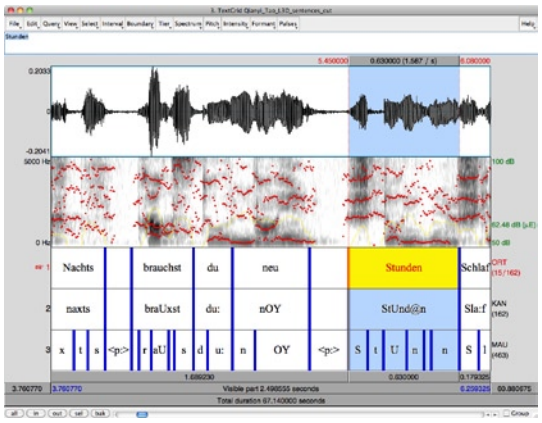
So könnte einer dieser Witze beginnen, die national-regionale Klischees bemühen. In meinem Fall ist das kein Scherz, sondern die Rahmenbedingungen meiner gerade entstehenden Dissertation. Inwiefern dies mit CLARIN zu tun hat, soll im folgenden Praxisbericht über die Arbeit mit CLARIN-Tools erläutert werden.

Für meine Dissertation mit dem Arbeitstitel „On the Hunt for Lateral Phonological Transfer in Multilinguals“ führe ich eine Längsschnittstudie durch, um besagtem lateralen Transfer (d.h. interlingualem Transfer zwischen zwei oder mehr Fremdsprachen) auf die Spur zu kommen. Wenn ein türkischer Muttersprachler beispielsweise für die Aussprache von dt. *wandern* /ˈvandɐn/ auf das Englische zurückgreift und stattdessen *[ˈwʌndəɹn] produziert, handelt es sich dabei um diese Art von Transfer. Da das Sprachenprofil meiner designierten „Versuchskaninchen“ relativ speziell ist – erwachsene Muttersprachler des Mandarin-Chinesischen (denn dieses ist maximal unähnlich im Vergleich zu den beiden Fremdsprachen zwischen denen der Transfer stattfinden soll) mit fortgeschrittener Kompetenz in der ersten Fremdsprache Englisch und Anfängerkompetenz in der Zielsprache Deutsch – antizipierte ich bereits Schwierigkeiten bei der Suche und letztendlich auch bei den Aufnahmen besagter Studienteilnehmer.

Nachdem ich bei einer *Summer School* von CLARIN erfahren hatte, wollte ich mir das Datensammeln mit Hilfe des Tools *WikiSpeech* gleich einmal erleichtern. *WikiSpeech* ermöglicht es, über die eingebaute Aufnahmesoftware Speech-Recorder sich selbst zum Beispiel vom heimischen Schreibtisch aus mit einem Headset über das Internet beim Lesen vorgegebener Texte aufzuzeichnen – an jedem Ort der Welt zu jeder beliebigen Tages- und Nachtzeit. Technisch eine wunderbare Sache für mein Projekt! Leider probierten nur wenige Kandidaten die Aufnahmen allein zu Hause aus, was vielleicht an der zurückhaltenden



WebMaus-Screenshot



Ansicht in Praat

Mentalität meiner Versuchsteilnehmer liegen könnte, an der etwas einseitigen „Kommunikationssituation" vor dem Bildschirm, oder auch etwa an Überforderung durch zu viele Stimulusmaterialien. Am Ende führte ich alle Aufnahmen persönlich mit einem transportablen Gerät durch.

Ganz anders verlief die Datenanalyse unter Zuhilfenahme des CLARIN-Tools *WebMAUS* – eine volle Erfolgsgeschichte! Mit den aufgenommenen Sprachproduktionsdaten sollten phonetische Analysen durchgeführt werden – was sich einfacher anhört, als es ist. Nicht grundlos existieren dazu nicht allzu viele empirische Arbeiten, da solche Analysen unglaublich zeitraubend sind. Deswegen griff ich auf das webbasierte, vollautomatisierte phonemische Segmentierungs- und Labelling-Tool *WebMAUS* zurück. Nachdem man eine Sprachaufnahme als wav-Datei samt orthografischer Transkription im txt-Format hochgeladen hat, gibt *WebMAUS* nach ein bisschen „Bedenkzeit" eine TextGrid-Datei (ein spezielles Format der Sprachaanalysesoftware Praat [1]) aus, die sowohl eine orthografisch transkribierte Segmentierung in Wörter beinhaltet, als auch die dazugehörige kanonische Zitierform sowie die phonemische Segmentierung in der maschinenlesbaren Aussprachenotation SAMPA. Mit der erzeugten TextGrid-Datei konnte ich sodann meine weiteren phonetischen Analysen in Praat durchführen.

SpeechRecorder

[1] *Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer [Computerprogramm], http://www.praat.org/*

Dank *Hidden Markov Modelling* segmentiert *WebMAUS* genau das, was im Sprachsignal zu hören ist – selbst wenn dies markant von der im Wörterbuch angegebenen kanonischen Aussprache abweicht. Somit kann *WebMAUS* fantastischerweise auch für Lernersprache, wie bei meinen Studienteilnehmern eingesetzt werden. Der größte Vorteil liegt jedoch zweifelsohne bei der enormen Zeitersparnis, die die Nutzung von *WebMAUS* bedeutete; statt tagelanger manueller Annotation einer Sprachaufnahme wird durch die relativ genaue automatische Segmentierung der Analysevorgang deutlich verkürzt. Insofern sind die in Münster promovierende Augsburgerin und ihre chinesischen „Versuchskaninchen" sehr zufrieden mit der Münchner **CLARIN-Software ...**

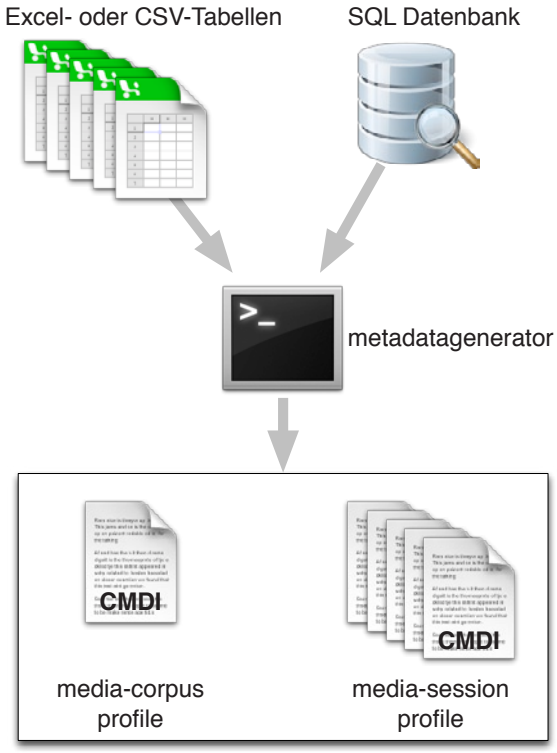
Eva-Maria Wunder, M.A.
Lehrstuhl für Englische Sprachwissenschaft, WWU Münster

AsiCa: CLARIN-D lernt Kalabrisch

Das Korpus *Atlante Sintattico della Calabria* der Münchener Italianistik wurde mit CLARIN-D-konformen Metadaten ausgestattet

Eines der großen Ziele von CLARIN ist es, die Geisteswissenschaften, sprich die Geisteswissenschaftler, davon zu überzeugen, ihre mit enormen Aufwand erstellten Sprachdaten über die CLARIN-Infrastruktur der Öffentlichkeit (und der Zeit nach dem Projektende) verfügbar zu machen. Einen ersten kleinen Erfolg in dieser Richtung konnte das CLARIN-D-Zentrum der LMU München feiern, nachdem sich die dort ansässige Italianistik (Prof. Th. Krefeld) bereit erklärt hat, ihr AsiCa-Korpus (*Atlante Sintattico della Calabria*) mit CLARIN-D-konformen Metadaten zu veröffentlichen.

AsiCa enthält Aufzeichnungen von Sprechern des süditalienischen Dialekts Kalabrisch. Schwerpunkt des Korpus sind spontan geführte Dialoge mit den Versuchsleitern, aber AsiCa enthält auch



Der Metadatengenerator

stärker kontrolliertes Material wie gelesene Sprache. Alle 60 Sprecher sind Muttersprachler des Kalabrischen, wobei genau die Hälfte der Sprecher zudem einen längeren Migrationsaufenthalt in Deutschland hatte. Die Informanten wurden kontrolliert nach Geschlecht und geographischer Herkunft rekrutiert, um eine möglichst repräsentative Stichprobe der Sprache zu erhalten. Alle Aufzeichnungen wurden orthographisch

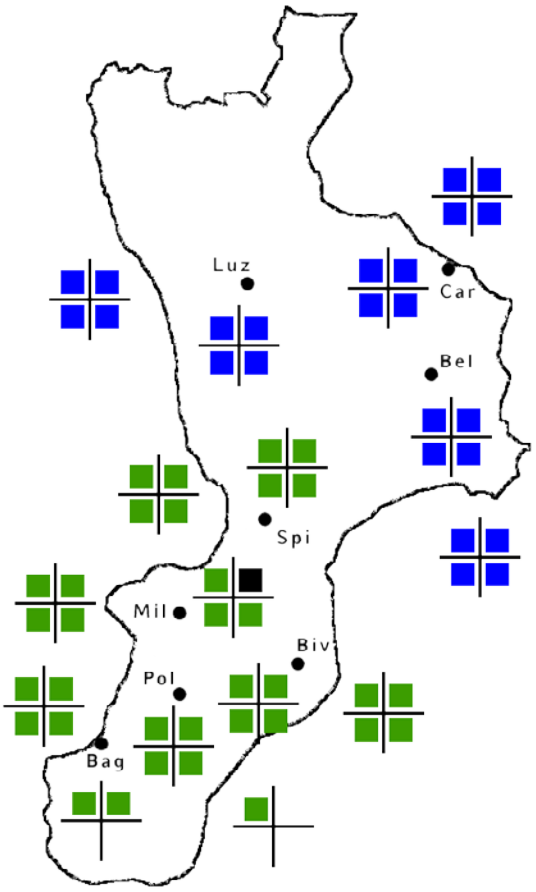
Mitmachen!

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 Beitragsgesaltung findet ihr im Wiki.

transkribiert und nach syntaktischen Ereignissen getaggt.

Die Portierung der proprietären AsiCa-Metadaten in CMDI-Profile *media-corpus-profile* und *media-session-profile* erwies sich als relativ einfach. Anhand der dokumentierten Excel-Tabellen, die CLARIN-D als Input für das am BAS entwickelten CMDI-Generatortool ‚metadatagenerator' bereitstellt, wurden vom Verantwortlichen des AsiCa-Korpus, Herrn Dr. Lücke, fünf virtuelle Tabellen in der AsicA-SQL-Metadatenbank definiert, welche genau die geforderten Metadaten der Sprachaufnahmen/Annotationen enthielten. Über das Standard-Webinterface der SQL-Datenbank wurden diese vom CLARIN-Center gelesen und sofort in CMDI-Dateien (entsprechend dem Profil *media-session-profile*) transformiert. Bis auf einige Kodierungsprobleme bei italienischen Namen lief dies vollautomatisch und ohne Probleme.

Für die Beschreibung des Korpus selbst (*root* CMDI, Profil *media-corpus-profile*) erzeugte ‚metadatagenerator' eine CMDI-Vorlage mit teilweise automatisch generierten Einträgen und detailliert dokumentierten Musterelementen. Diese Vorlage ermöglichte es den Korpus-Verantwortlichen ohne weitere Konsultierung der CMDI-Registry die erforderlichen Metadaten für ihre Ressource händisch einzutragen. Nach einer abschließenden



Die Region Kalabrien im Süden Italiens

formalen und inhaltlichen Validierung durch das CLARIN-D-Zentrum wurde das AsiCa-Korpus in das Repository des BAS integriert [1].

Der Verlauf dieser Korpus-Kuration zeigt, dass sich mit geeigneten einfachen Werkzeugen der Aufwand für den Erzeuger oder Inhaber einer Ressource minimieren und somit hoffentlich die Akzeptanz der CLARIN-Infrastruktur signifikant erhöhen lässt.

Florian Schiel
Institut für Phonetik und Sprachverarbeitung, LMU München

Grenzgänge

In der Rubrik Grenzgänge berichten Forscher erstaunliche, ungewöhnliche oder amüsante Ergebnisse. Dieses Mal: Kathrin Beck, Christoph Draxler und Elke Teich über **CLARIN-Tools und alte Kochbücher**

---

1. Haselnusstorte.
Man rührt 250g Zucker mit 7 Eigelb, 1 Ganzes 20 Minuten lang, 250 g geriebene Hasel- od. Baumnüsse u. die 8 Eiweiß zu Schaum geschlagen mischt man langsam bei; bäckt für ¾ Std. im nicht heißen Ofen u. überzieht sie mit Wasserglasur.

Das ist das erste Tortenrezept aus einem kleinen, handgeschriebenen Koch- und Backbuch von Anna Authaler, geboren 1885 in der Region Göppingen. Das Kochbuch scheint viel genutzt worden zu sein, denn die Seiten sind fleckig, teilweise eingerissen, und es enthält Korrekturen und Ergänzungen.

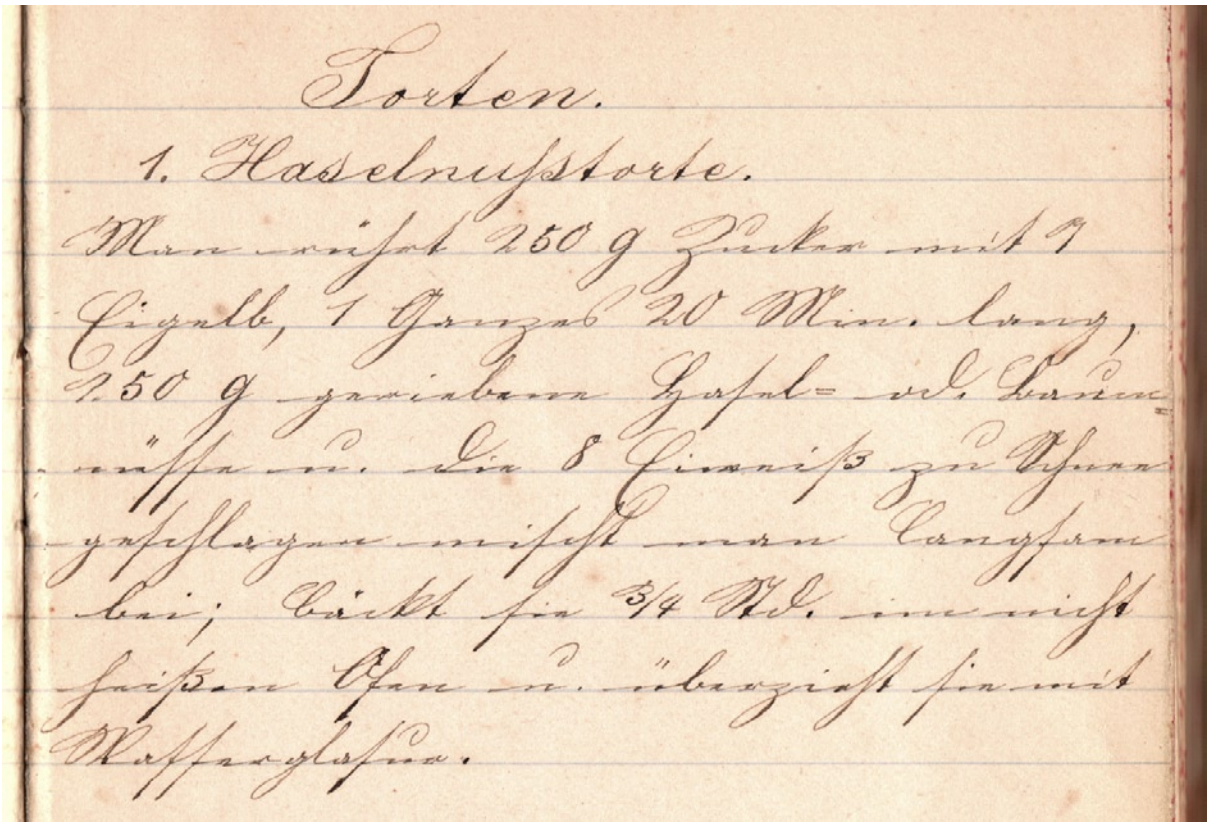
[1] *http://www.phonetik.uni-muenchen.de/BASRepository/Corpora/AsiCa/AsiCa.1.html*

Na und? Was hat das mit CLARIN zu tun?

Genau diese Fragen haben wir, Kathrin Beck, Elke Teich und ich uns auch gestellt. Die Antwort ist nicht ganz überraschend – sehr viel natürlich! Historische Dokumente wie dieses Kochbuch sind mindestens aus linguistischer, dialektologischer, soziologischer, kulturgeschichtlicher und informationsverarbeitender Perspektive interessant:

- Wie wurden Rezepte damals formuliert, können wir aus den Benennungen der Zutaten etwas über die regionale Herkunft sagen?
- Wer war die Person, die dieses Buch geschrieben und verwendet hat?
- Welche Zutaten gab es damals, welche Kochtechniken?
- In welchen sozialen Schichten wurden solche Kochbücher erstellt? Zu welchem Zweck?
- Wie können die handgeschriebenen Texte maschinenlesbar erfasst und aufbereitet werden, damit sie auch heute – z.B. für die Forschung – genutzt werden können?

In den nächsten Newsletter-Ausgaben werden wir einige Einblicke in die Aufbereitung (oder vielleicht besser: ‚Zubereitung') dieses Kochbuchs geben. Ziel ist es, entsprechend dem CLARIN-Gedanken ein kleines annotiertes Text- und Sprachkorpus zu erstellen. Das Ganze ist einerseits ein ernstzunehmendes


Torten.
1. Haselnußtorte.
Man nimmt 250 g Zucker mit 9
Eigelb, 1 Ganzes 20 Min. lang,
250 g geriebene Schok- und Zucker
nüsse u. die 8 Eiweiß zu Schnee
geschlagen mischt man Alles
bei; bäckt sie 3/4 Std. in mittel
heißen Ofen u. überzieht sie mit
Kaffeeglasur.

Das Rezept im Kochbuch

wissenschaftliches Unterfangen – und andererseits macht es einfach Spaß, die historischen Rezepte zu lesen, sich zu überlegen, ob man sie nachkocht oder ob man damit interessante Forschungsfragen stellen und beantworten kann – und über die man sich dann bei Anna Authalers Haselnusstorte vollmundig und genüsslich austauschen kann.

Katrhin Beck
Seminar für Sprachwissenschaft, Universität Tübingen

Christoph Draxler
Institut für Phonetik und Sprachverarbeitung, LMU München

Elke Teich
Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes

Hier gibt es das Rezept auch zum Anhören!

Weitere Infos auch im Wiki:

http://de.clarin.eu/mwiki

Vorstellung des Kurationsprojekts 1 der CLARIN-D-FAG 1 „Deutsche Philologie“

Ein Kurationsprojekt zur Ermittlung, Verzeichnung, Aufwertung und Integration historischer Textressourcen in einer nachhaltigen CLARIN-Infrastruktur

Das Kurationsprojekt „Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur“ [1] hat zum 1. September 2012 seine Arbeit aufgenommen. Neben der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) sind die Justus-Liebig-Universität Gießen, die Herzog August Bibliothek Wolfenbüttel (HAB) sowie das Institut für Deutsche Sprache (IDS) in Mannheim an dem Projekt beteiligt. Koordiniert wird das Kurationsprojekt am Deutschen Textarchiv (DTA) der BBAW.

Wesentliche Ziele des Kurationsprojekts sind die kriteriengestützte Identifikation hochwertiger Textressourcen, die bislang über das Internet verstreut oder nur lokal in laufenden oder abgeschlossenen Forschungsprojekten und anderen Initiativen verfügbar sind und deren Integration in die CLARIN-D-Infrastruktur.
Sämtliche ermittelte Ressourcen werden in einem zentralen Verzeichnis erfasst, charakterisiert und bewertet. Die in den beteiligten Institutionen bereits bestehenden Korpora sollen durch den Beitrag des Kurationsprojekts wesentlich bereichert werden. Über die Repositorien der CLARIN-Partner sollen die kuratierten Ressourcen sukzessive auch in die CLARIN-D-Infrastruktur integriert werden. Die verteilte bzw. gespiegelte Repositorien-Struktur der Partner soll als dauerhaftes und zentrales Repositorium für historische Textquellen dienen.

Ermittlung, Aufwertung und Integration qualitativ hochwertiger Textressourcen

Anhand einer Kombination von inhaltlichen, technischen und die Qualität der Transkription sowie deren rechtliche Verfügbarkeit betreffenden Kriterien

[1] *Fortlaufend aktualisierte Informationen zum Kurationsprojekt finden Sie unter http://www.clarin-d.de/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.html [alle URLs in diesem Beitrag abgerufen am 31.10.2012].*

werden geeignete Ressourcen zur Integration in die bestehenden Korpora des DTA, der HAB und des IDS ermittelt. Die Volltexttranskriptionen werden von ihrem Ausgangsformat – sei es HTML, MS Word, XML, plain-Text o. a. – in das XML/TEI P5-basierte Basisformat des Deutschen Textarchivs (DTA) [2] konvertiert.

Die Ressourcen werden im Zuge der Konvertierung mit detaillierten Metadaten versehen, aus denen z. B. die Vorlage der Transkription, deren Urheber und ursprünglicher Entstehungskontext, die Lizenz, unter der sie bereitgestellt wird sowie die vorgenommenen Konvertierungsschritte hervorgehen. Die auf diese Weise aufgewerteten und angereicherten Volltexte werden in einem nächsten Schritt – nach Möglichkeit verbunden mit korrespondierenden Bilddigitalisaten der Vorlage – in die bestehende Korpusinfrastruktur des DTA, der HAB und des IDS integriert.
Erste Proben kuratierter Ressourcen sind bereits in DTAE, dem Erweiterungsmodul des Deutschen Textarchivs, zugänglich [3]. Die bisher verfügbaren Volltexte stammen zu einem Teil aus bestehenden Kooperationen mit einzelnen Projekten, die im Rahmen des Kurationsprojekts intensiviert und erweitert werden sollen, zum anderen aus großen Volltextsammlungen wie Wikisource oder Gutenberg.org [4].

Ein großer Teil der bislang integrierten Ressourcen stammt dabei aus der freien Quellensammlung Wikisource, wo etliche qualitativ hochwertige Volltexttranskriptionen zu finden sind. Deren Nutzbarkeit gerade für computergestützte Analysen ist allerdings durch die zum Teil unzureichende Erschließung durch Metadaten und noch mehr durch die ‚sperrige' Wiki-Syntax erschwert. Mit Stand vom 31.10.2012 wurden allein aus Wikisource bereits 67 Werke im Rahmen des Kurationsprojekts konvertiert und mit den computerlinguistischen Methoden des DTA analysiert:

Werke	**67**
Seiten	**11899**
Zeichen	**13347027**
Token	**3999952**

Statistischer Überblick zu den bisher aus Wikisource integrierten Ressourcen (Stand: 31.10.2012)

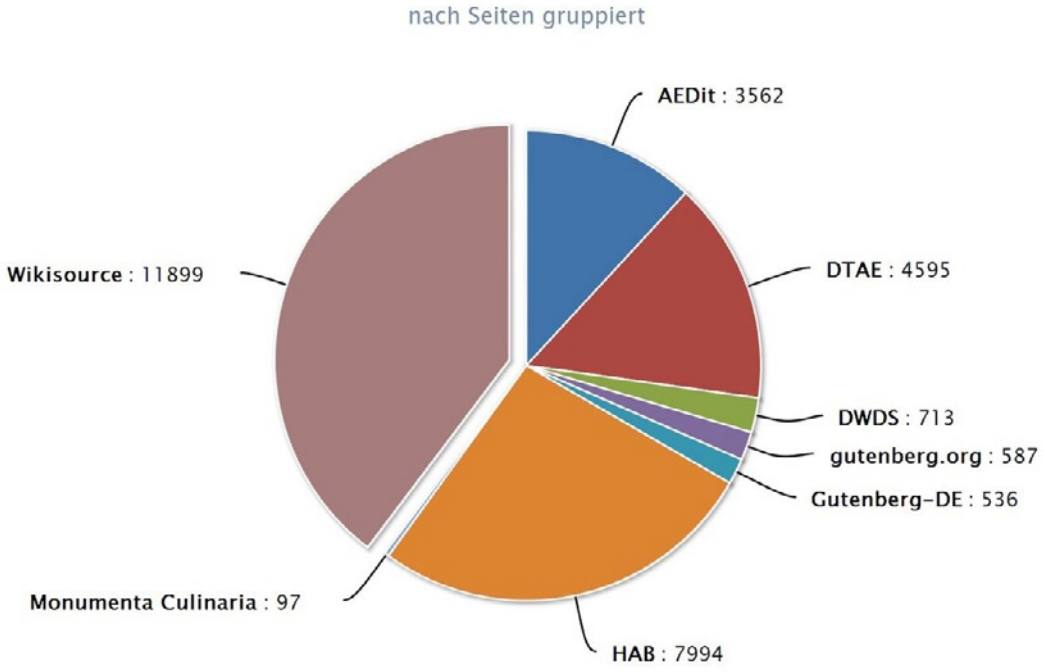
Auf dem beschriebenen Weg der Identifikation, Verzeichnung, Aufwertung und Integration werden im Kurationsprojekt die bereits vorhandenen, jedoch teilweise schwer auffindbaren und/oder in unflexiblen Formaten vorliegenden Textressourcen zu einem integrierten Textkorpus für das Frühneuhochdeutsche und das ältere Neuhochdeutsche (15.–

[2] *DTA-Basisformat, www.deutschestextarchiv.de/doku/basisformat.*
[3] *DTAE, www.deutschestextarchiv.de/dtae.*
[4] *Vgl. zu möglichen weiteren Quellen für hochwertige Textressourcen Christian Thomas, Frank Wiegand: Making great work even better: Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D. Full Paper for the International Conference „Historical Corpora 2012", 6.–9.12.2012, Goethe Universität Frankfurt, Germany, URL: http://edoc.bbaw.de/volltexte/2012/2308/, URN: urn:nbn:de:kobv:b4-opus-23081.*

19. Jh.) zusammengeführt. Die Ressourcen werden Nutzerinnen und Nutzern aller Disziplinen in einem innerhalb von CLARIN-D anerkannten, einheitlichen, ausdrucksstarken und flexiblen Format – dem Basisformat des DTA – zur vielseitigen Nutzung zur Verfügung gestellt. Dies wird die Möglichkeiten text- bzw. korpusbasierter Forschung erheblich verbessern. [5]

Mit dem Auf- bzw. Ausbau einer nachhaltigen Infrastruktur, durch die vorhandene Textressourcen in ein ständig wachsendes, zentrales und strukturiertes Korpus integriert werden können, sollen Nutzerinnen und Nutzer angeregt werden, ihre Expertise und ihre eigenen Ressourcen einzubringen und sich somit, durch die Arbeit an und mit den bereitgestellten Ressourcen, stärker als bisher als Teil einer ‚community' unter dem Dach von CLARIN-D zu verstehen.

Frederike Neuber, SHK BBAW
Christian Thomas, BBAW, Berlin



Verteilung der Texte in DTAE, nach Seiten gruppiert. Das Segment der im Rahmen des Kurationsprojekts aus Wikisource integrierten Texte ist hervorgehoben. (Stand: 31.10.2012, Quelle: www.deutschestextarchiv.de/dtae)

[5] *Dieser Ansatz wurde im Rahmen des gemeinsamen CLARIN-D-Panels der HAB, der Universität Gießen und des DTA „Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of Publication of Digitized Historical Prints" auf der DH2012 im Hamburg vorgestellt. Siehe auch CLARIN-Newsletter No. 2, http://www.clarin-d.de/images/newsletter/clarin-d-newsletter_2.pdf, S. 13f. bzw. die Aufzeichnung des Panels unter http://lecture2go.uni-hamburg.de/konferenzen/-/k/13952.*

Ein CLARIN-Center stellt sich vor: das Hamburger Zentrum für Sprachkorpora

Von gesprochener Sprache und Mehrsprachigkeit

Das Hamburger Zentrum für Sprachkorpora (HZSK) entstand 2011 aus einem Teilprojekt des Sonderforschungsbereich 538 „Mehrsprachigkeit" der Universität Hamburg. Zu der ursprünglichen Aufgabe, die Daten und Werkzeuge des Sonderforschungsbereiches nachhaltig verfügbar zu machen, sind inzwischen einige weitere hinzugekommen.

Kontinuität

Nach zwölf Jahren beendete der Sonderforschungsbereich 538 „Mehrsprachigkeit" [1] an der Universität Hamburg 2011 planmäßig seine Arbeit. Das Forschungsinteresse lag insbesondere in der Erfassung, Dokumentation, Beschreibung und Analyse der sprachlichen Formen des multilingualen Sprachgebrauchs. Alle Projekte des Sonderforschungsbereiches arbeiteten empirisch. Einige mit schriftlichen, die meisten jedoch mit mündlichen Korpora, die innerhalb der Projekte erhoben und ausgewertet wurden – allesamt wertvolle Ressourcen, die auch für andere Fragestellungen interessantes Ausgangsmaterial bieten [2].

Diesen Datenschatz nicht sich selbst zu überlassen war zudem eine wesentliche Forderung der Deutschen Forschungsgemeinschaft. Auch die Forscherinnen und Forscher des Sonderforschungsbereiches hatten ein starkes Interesse, „ihre Daten" für die Zukunft in sicheren Händen zu sehen.

So wurde mit Mitgliedern des SFB-Teilprojektes „Erfassung und Analysemethoden von multilingualen Daten" das „Hamburger Zentrum für Sprachkorpora" gegründet. Dieses hat sich neben der Pflege der SFB-Daten und -Werkzeuge auch auf die Fahnen geschrieben, als Anlaufstelle für alle Angehörigen der Universität Hamburg zu dienen, die im weitesten Sinne mit Sprachkorpora arbeiten [3].

[1] *http://www.uni-hamburg.de/sfb538/*
[2] *http://www.corpora.uni-hamburg.de/sfb538/de_overview.html*
[3] *http://www.corpora.uni-hamburg.de/*

Zu diesem Zweck finden regelmäßig Mitgliederversammlungen und Workshops statt, um das vorhandene Know-How zwischen den einzelnen Projekten und Wissenschaftlern zu vermitteln und Synergien zu entwickeln. Die Mitglieder des HZSK kommen sowohl aus dem technischen als auch aus dem geisteswissenschaftlichen Bereich, vor allem Sprachwissenschaftler aus der Germanistik, Dialektologie, Skandinavistik oder Finno-Ugristik sind unter ihnen vertreten.

Vernetzung

Zeitgleich mit seiner Etablierung starteten im HZSK zwei drittmittelgeförderte Projekte: „Etablierung eines Schwerpunktes ‚Mehrsprachigkeit'", für zwei Jahre von der DFG gefördert sowie das CLARIN-D-Projekt, in dem das HZSK eines der neun Zentren in Deutschland stellt.

Im CLARIN-D-Verbund spezialisiert sich das HZSK auf gesprochensprachliche, mehrsprachige Ressourcen und Werkzeuge zu deren Bearbeitung und Analyse.

Beide Projekte arbeiten gemeinsam daran, ein Repositorium für die Daten des Zentrums zu erstellen. Da die meisten dieser Daten – vor allem in Bezug auf die Rechte der aufgenommenen Personen – besonders schutzbedürftig sind, können sie nur auf Anfrage und nach Unterzeichnung von Nutzungsvereinbarungen herausgegeben werden. Das Repositorium soll helfen, diesen Freigabeprozess zu erleichtern. Es soll aber auch die Integration in die CLARIN-D-Infrastruktur sicherstellen, indem es die in CLARIN festgelegten technischen Standards umsetzt, also z.B. PIDs für Ressourcen und Werkzeuge festlegt, eine einheitliche Metadatenschnittstelle anbietet und ein einfaches Login-Verfahren für Nutzer der Infrastruktur vorhält.

So können Forscher, die dem HZSK Daten zur Aufbewahrung anvertrauen, diese auf einfache Weise auch in die CLARIN- Infrastruktur integrieren – so wie es beispielsweise zur Zeit mit den Daten des Verbundprojektes „Sprachvariation in Norddeutschland" [4] geschieht.

Eigene Projekte

Zu seinen Aufgaben zählt das HZSK zudem die Erstellung eigener Referenz- und Beispielkorpora wie zuletzt des „Hamburg Map Task Korpus" und aktuell des „Hamburg Modern Times Korpus." Auch die Pflege und Anpassung der Werkzeug- und Methodensammlung EXMARaLDA [5], die am Sonderforschungsbereich von Thomas Schmidt und Kai Wörner entwickelt wurde, wird am HZSK in Kooperation mit dem Institut für Deutsche Sprache [6] in Mannheim geleistet. Dies umfasst natürlich auch die Anpassung dieser Werkzeuge an die Anforderungen von CLARIN.

EXMARaLDA-Werkzeuge unterstützen Forscher bei der computergestützten

[4] *http://sin.sign-lang.uni-hamburg.de/drupal/startseite.html*
[5] *http://exmaralda.org/*
[6] *http://agd.ids-mannheim.de/index.shtml*

Transkription und Annotation gesprochener Sprache, sowie beim Erstellen und Auswerten von Korpora gesprochener Sprache. Die meisten Korpora des Zentrums wurden mit EXMARaLDA erstellt und liegen im zugrundeliegenden Dateiformat vor.

Support

EXMARaLDA hat auch außerhalb des Sonderforschungsbereiches und der Universität Hamburg weite Verbreitung und Akzeptanz gefunden – und damit auch zu einem hohen Bedarf an (technischem sowie methodischem) Support und Anwenderschulungen geführt. Die Erfahrungen, die dort gesammelt wurden, fließen nun in den Aufbau eines Support- und Helpdesk-Angebotes in CLARIN-D ein, das vom Hamburger Zentrum für Spachkorpora aus koordiniert wird.

Mitarbeiter

Das Hamburger Zentrum für Spachkorpora wird von Kai Wörner geleitet, Prof. Dr. Kristin Bührig ist Direktorin und Leiterin des CLARIN-D-Projektes. Die Mitarbeiter, die sich für das Foto versammelt haben, haben allesamt einen linguistischen Hintergrund (außer Carl).

Die Mitarbeiter des HZSK (v.l.n.r): Timm Lehmberg, Yael Dilger, Kai Wörner, Fideniz Ercan, Daniel Stein, Daniel Jettka, Hanna Hedeland (mit Carl).

A report on the CLARIN-D M12 Workshop in Leipzig

***CLARINification**, or "How to become a member of the CLARIN-family", was a recurring theme of the CLARIN-Germany (CLARIN-D) workshop on the 27th and 28th of June.*

Roughly a hundred people came to Leipzig for the CLARIN-D first year wrap-up meeting. The Leipzig Mediencampus hosted for them a varied program of demos and presentations of current and planned curation projects. It was a great combination of a solid agenda, outstanding organization, good catering and summery weather.

One noteworthy point was the highly visible major role given to researchers, something anchored in the structure of CLARIN-D, which contains seven workgroups formed around specific academic disciplines like German philology, archaeology or natural language processing. Last year, each group selected a lead researcher who then assembled a team that selected resources and tools from the research community for curation and inclusion in CLARIN. Volker Boehlke from Leipzig University, who oversees these workgroups, and Erhard Hinrichs of Tübingen, national coordinator and German counterpart of Dutch coordinator Steven Krauwer, are very excited about the enthusiasm and initial output of these workgroups. Various speakers from the digital humanities expressed their aspirations for CLARIN: easy-to-use but adaptable tools for researchers that are, ideally, suited for education.And if the tool is itself the product of user research - like a project in techniques for the effective visualization of argument structures in texts - then its users should, for example, be able to share their experiences in an online forum.

The cross-discipline Legal Help Desk was officially opened during the conference. Housed at the *Institut für Deutsche Sprache*, this service will inform researchers among other things about so-called "implied licenses": what should you do with information downloaded from the Internet, or conversely, what can others do with what you leave online? Also of general interest is a handbook of good practices and tools for estimating what a particular curation project will cost. The Berlin-Brandenburg Academy of Sciences is coordinating the first version for delivery by the end of this year. The workshop participants immediately commented that this should not be restricted to CLARIN-D.

There was at any rate a good deal of attention paid to "the extended CLARIN family." This was visible in the number of participants from countries that take part or want to take part in CLARIN-ERIC, like Denmark, the Netherlands,

Lithuania and Norway. It was especially the message of Rüdiger Klein, speaking on behalf of ALLEA, the federation of European academies of science, at the end of the first day. He made a powerful and critical - perhaps a bit too critical - call for cooperation:

- with all the humanities and social sciences, not just "the usual suspects";
- with other sciences - they use texts and language too;
- with other research infrastructure consortia: The EU expects ERICs to find common ground and develop a vision that goes beyond the duration of that ERIC. See, for example, the report of Peter Wittenburg (Max Planck Institute, Nijmegen) on the international conference of research infrastructure consortia in the previous CLARIN-D newsletter.

And finally, "Show the general public what digital humanities is!"

Marjan Grootveld

Bericht zum CLARIN-D-M12-Workshop in Leipzig

Vor etwas mehr als einem Jahr startete das CLARIN-D-Projekt [1]. Zu diesem Anlass fand am 27. und 28. Juni 2012 der CLARIN-D-M12-Workshop [2] in Leipzig statt. Ziel war es CLARIN-D einem breit gefassten Publikum aus Geistes- und Sozialwissenschaftlern, in den eHumanities engagierten Forschern und (Promotions-)Studenten sowie Informatikern mit Interesse an Aufbau und/oder Nutzung von Forschungsinfrastrukturen vorzustellen. Deutlich über 100 nationale und internationale Gäste aus zahlreichen europäischen Staaten folgten dem Ruf nach Leipzig, um an diesem ersten Disseminationsworkshop des CLARIN-D-Projekts teilzunehmen.

Eröffnet wurde der Workshop durch die Rektorin der Universität Leipzig, Prof. Dr. Schücking. In ihrer Rede unterstrich Sie die Bedeutung der Verständigung zwischen der Informatik und den Geisteswissenschaften. Sie schloss mit dem Motto der Universität Leipzig „Aus Tradition Grenzen überschreiten", welches auch die Vision hinter dem CLARIN-D-Projekt und insbesondere dem Ziel der Verständigung zwischen bisher zumeist isoliert arbeitenden Wissenschaftszweigen charakterisieren könnte. Im Anschluss folgte eine Präsentation durch Prof. Crane (*Perseus Digital Library Project* [3]) in welcher unter anderem die Bedeutung der Nachwuchsförderung, auch in schwach besetzten bzw. nachgefragten „Nischen" und des spielerischen Heranführens an aktuelle Forschungsfragen angesprochen wurde.

In diesem Zusammenhang besonders hervorzuheben ist die hohe Beteiligung von Nachwuchsforschern am Workshop. Erklärtes Ziel der Veranstaltung war es, neben etablierten Größen und wichtigen Multiplikatoren in den Fachcommunitys auch diejenigen Forscher anzusprechen, welche in Zukunft in den Geistes- und Sozialwissenschaften und den *eHumanities* tätig sein und deren Zukunft mitgestalten werden. Dies wurde auch in den Vorträgen des ersten Blocks deutlich: Hier ging es darum, die Anforderungen und Erwartungen von Forschern aus den *eHumanities* an Infrastrukturprojekte wie CLARIN-D noch einmal in Erinnerung zu rufen. Neben Frau Prof. Storrer (Universität Dortmund), welche in dieser Hinsicht als etablierte Größe gelten kann, stellten mit Annette Hautli (Universität Konstanz) und Eva Maria Wunder (Universität Augsburg) auch zwei

[1] *http://de.clarin.eu*
[2] *http://clarin.informatik.uni-leipzig.de*
[3] *http://www.perseus.tufts.edu*

junge Wissenschaftlerinnen ihre aktuellen Projekte und Forschungstätigkeiten vor und formulierten eine persönliche „Wunschliste“ an CLARIN-D.

Mit diesen Erwartungen im Hinterkopf wurde nun im weiteren Verlauf des Workshops versucht, diese und weitere Fragen gezielt zu adressieren und dabei auch die Ziele und den aktuellen Stand des CLARIN-D-Projektes zu vermitteln. Begonnen wurde mit einem Block, in dem das CLARIN-D-Projekt und die im Aufbau befindliche Infrastruktur vorgestellt wurde. Hier wurde unter anderem die Frage beantwortet, welche Infrastrukturkomponenten und Services schon heute verwendbar bzw. in naher Zukunft verfügbar sein werden.

In einem weiteren Block stellte sich ein Großteil der CLARIN-D-Facharbeitsgruppen vor und berichtete über die nun in der Durchführung befindlichen Kurationsprojekte. Hier werden zumeist etablierte, für die einzelnen Communitys wichtige Ressourcen aufbereitet, teilweise auf neuartige Weise verknüpft und in die CLARIN-D-Infrastruktur integriert. Anschließend folgte ein weiterer Block, in dessen Rahmen unter anderem das CLARIN-D-Evaluationshandbuch kurz vorgestellt und um Mitarbeit an selbigem gebeten wurde. Der Tag wurde durch einen Vortrag von Dr. Rüdiger Klein (ALLEA; ALL European Academies [4]) abgerundet, in dessen Rahmen Herr Klein seine persönlichen Eindrücke vom ersten Workshoptag zusammenfasste, CLARIN-D in den europäischen Kontext einordnete, wichtige Perspektiven verdeutlichte (etwa die weitere Zusammenarbeit mit den geisteswissenschaftlichen Fachgesellschaften im Hinblick auf neue Forschungsfragen) und Denkanstöße für zukünftige Aktivitäten (etwa im Bereich *Training* für die *Digital Humanities*) gab.

Der zweite Veranstaltungstag begann mit einem Ausflug zu einem inzwischen abgeschlossenen Projekt: eAQUA [5]. Das Projekt wurde durch Frau Prof. Schubert (Universität Leipzig) kurz vorgestellt und im Anschluss von Thomas Eckart (Universität Leipzig; CLARIN-D Leipzig und ehemals eAQUA) ein möglicher Fahrplan für eine Integration in CLARIN-D aufgestellt. Es wurde insbesondere verdeutlicht, welche Bedeutung bestimmte Architekturentscheidungen für die Flexibilität des technischen *Backends* in einem *eHumanities*-Projekt haben können und welche Aspekte unter anderem für eine Integration in die CLARIN-Infrastruktur ausschlaggebend sind. Abgeschlossen wurde dieser Block durch eine Präsentation von Marco Büchler (Universität Leipzig, eTraces [6] und ehemals eAQUA), welche neben technischen Aspekten auch Fragen der Arbeitsorganisation und Kommunikation im Projekt adressierte.

Ein Großteil des zweiten Tages war den Präsentationen und einer ausgedehnten Demosession zu CLARIN-D-Infrastrukturkomponenten und Ressourcen aus dem CLARIN-D-Umfeld gewidmet. Hier hatten die Workshopteilnehmer die

[4] *http://www.allea.org*
[5] *http://www.eaqua.net*
[6] *http://etraces.e-humanities.net*

Chance mit den Entwicklern in direkten Kontakt zu treten und in individuellen Gesprächen gezielt Fragen zu klären und Nutzungsperspektiven auszuloten. Die letzten Vorträge des zweiten Tages beschäftigten sich mit Kooperationen, welche CLARIN-D im nationalen und internationalen Umfeld eingeht und adressierte rechtliche Aspekte. Abgeschlossen wurde der Workshop durch eine Podiumsdiskussion in der erneut einige der immer wiederkehrenden Themen des Workshops offen diskutiert wurden.

Wir möchten uns bei allen Vortragenden für die durchweg engagierten Beiträge und bei allen Teilnehmern für die wichtigen Anregungen, kritischen Fragen und produktiven Diskussionen im Rahmen des Workshops bedanken. Es wurde insbesondere deutlich, dass Infrastrukturprojekten wie CLARIN-D neben einer Katalysatorfunktion durch die Schaffung von technischen Lösung für zahlreiche wiederkehrende Probleme auch eine starke integrative Funktion zukommt. Konferenzen, auf denen junge Nachwuchswissenschaftler aus dem Bereich der Informatik mit Interesse am Aufbau von Infrastruktur, der Philologie bzw. den Angewandten Sprachwissenschaften, Computerlinguisten mit Schwerpunkt im Bereich der Visualisierung und Politik/-Sozialwissenschaftler (um nur einige wenige Beispiele zu nennen) gleichermaßen vertreten sind und sich aktiv und mit Begeisterung austauschen, dürften noch immer äußerst rar gesät sein.

Es war sehr inspirierend den verschiedenen Gesprächen auch am Rande der Konferenz zu lauschen und zu beobachten, wie neue Kontakte geknüpft sowie Erfahrungen und Meinungen zwischen diesen verschiedenen Disziplinen ausgetauscht wurden. Es muss, um erneut das Motto der Universität Leipzig zu zitieren, zu unserer Aufgabe werden „Aus Tradition (diese) Grenzen (zu) überschreiten“. Teilweise ist dies im Rahmen des Workshops gelungen. Nun liegt es an uns daraus eine Tradition werden zu lassen.

Volker Boehlke
Institut für Informatik, Universität Leipzig

Gruppenphoto vom 7. CLARIN-D-Entwicklertreffen (siehe den Bericht auf der nächsten Seite)

Bericht vom 7. CLARIN-D-Entwicklertreffen

Internationale Infrastrukturen

Am 24. September 2012 fand am Institut für Maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart – parallel zu einem in Kooperation mit dem IWiST Hildesheim gemeinsam von den Tübinger und Stuttgarter Zentren organisierten STTS-Workshop sowie einen Tag vor dem wie immer zeitnah am gleichen Ort veranstalteten Konsortiumstreffen – das 7. Entwicklertreffen des CLARIN-D-Projekts statt.

Die Veranstaltung fand in den im Frühjahr 2012 neu bezogenen Räumlichkeiten des IMS im Forschungszentrum Informatik auf dem Vaihinger Campus statt. Bei dieser Auflage, der etwa alle 3 Monate stattfindenden Zusammenkunft, nahmen 20 technische Experten aller 9 deutschen Zentren-Standorte die Gelegenheit wahr, sich zu einschlägigen Fragestellungen sowie den anstehenden Arbeitsschritten auszutauschen und diese zu diskutieren. Inhaltlich wurden die folgenden Punkte behandelt: Berichte aus den Zentren, *Wrap-Up* zur Webseite, Stand des technischen Arbeitspakets AP3, *Federated Content Search* (GUI, *Libraries*, Formate, *Endpoints*), M24-Demonstrator, TCF-Lexikonformat, Vorstellung der TüNDRA-Webanwendung, *Corporate Design* für CLARIN-D, aktuelle AAI-Sicherheitsprobleme sowie die obligatorische Abschlussdiskussion. Mittels zahlreicher Präsentationen, Beiträge und Diskussionen wurde somit die technische Planung für die nächsten Monate vorangebracht. Neben konkreten Entscheidungen zu technischen Details sind u. a. neue Arbeitsgruppen, z. B. zum Lexikonformat und *Best Practices* für Repositorien, sowie Planungen für spezifische Workshops, z. B. zur Zertifizierung der CLARIN-D-Zentren zu nennen. Das nächste Treffen der Entwicklergruppe soll am 28. November 2012 stattfinden und wird vom HZSK der Universität Hamburg ausgerichtet werden.

Jens Stegmann
Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart

Bericht zum CLARIN-D-Workshop „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven"

Die Annotation mit Wortarten gehört zur grundlegenden Aufbereitung von Texten für die weitere (computer-)linguistische Auswertung. Auch Analysen, die keine unmittelbar linguistische Fragestellung verfolgen, wie zum Beispiel Relationsextraktion („Wer hat im 18 Jahrhundert welches Patent veröffentlicht?"), profitieren von der Anreicherung mit Wortartenanalysen, da diese eine erste Disambiguierung erlauben (z.B. meinen *als Possessivpronomen oder Verb,* Fischer *als Eigenname oder normales Nomen) und als eine erste grammatische Abstraktion für die weitere Analyse dienen.*

Das Wortartentagging scheint eines der gelösten Probleme in der Computerlinguistik zu sein. Dass in der Praxis aber schon in Bezug auf das zugrundeliegende Tagset noch viele Fragen und Bedürfnisse offen sind, hat der eintägige Workshop „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven" gezeigt, der am 24. September 2012 in Stuttgart stattfand. Der Workshop wurde von Kathrin Beck und Heike Zinsmeister von den CLARIN-D-Zentren Tübingen und Stuttgart in Kooperation mit Ulrich Heid von der Universität Hildesheim (Mitglied der CLARIN-D-F-AG 7) organisiert. 26 Teilnehmer von zehn verschiedenen Einrichtungen beteiligten sich an der Diskussion.

Die Motivation, den Workshop durchzuführen, war eine Aufarbeitung der Nutzung des Stuttgart-Tübingen Tagsets (STTS), das in den 1990er Jahren von den Universitäten Stuttgart und Tübingen gemeinsam entwickelt wurde (vgl. Schiller et al. 1999). Es hat sich seither mehr oder weniger zu einer *de facto*-Norm für die morpho-syntaktische Annotation deutscher Texte entwickelt. Gleichzeitig existieren aber verschiedene, zum Teil nicht vollständig (öffentlich) dokumentierte Varianten („Dialekte") von STTS, und es gibt viele Vorschläge zur Verbesserung und Ergänzung des Tagsets [1]. Ziel des Workshops war es, eine Bestandsaufnahme zu erstellen und Änderungsvorschläge zu sammeln und zu diskutieren, mit einem besonderen

[1] *Eine Aufstellung verschiedener Tagset-Varianten finden Sie auf www.ims.uni-stuttgart.de/projekte/corplex/german-tagsets.shtml.*

Schwerpunkt auf der Anwendung des STTS auf Nicht-Standard-Domänen wie internet-basierte Kommunikation, historische Sprachstufen oder Lernersprache.

Das Programm umfasste zehn Kurzvorträge und eine lange Diskussions- und Planungseinheit. Ulrich Heid und Heike Zinsmeister präsentierten eine Zusammenfassung eines früheren Workshops von 2004 und des aktuellen Diskussionsstands in der Literatur. Kathrin Beck motivierte die Einbindung der STTS-Kategorien in das *Data Category Registry* ISOCat – eine Referenz-Implementierung des ISO-Standards 12620:2009 von CLARIN. Heike Telljohann (Tübingen) berichtete vom Wortartentagging Tübinger Korpora, denen verschiedene Textgenres zugrunde liegen. Neben modernen Zeitungstexten beinhalten sie auch gesprochene Quellen (basierend auf Interviews aus dem Verbmobil-Projekt), Web-News und historische Texte. Trotz mancher Abweichungen hat man in Tübingen sehr gute Erfahrungen mit dem STTS-Tagset gemacht. Das Tagging extremerer Formen der Mündlichkeit wurde in den Beiträgen von Swantje Westpfahl und Thomas Schmid (IDS Mannheim) sowie von Ines Rehbein (Potsdam) diskutiert. Das am IDS entstehende „Forschungs- und Lehrkorpus Gesprochenes Deutsch“ (FOLK) beinhaltet Mitschnitte von spontanen Gesprächen. Die Auswertung von Taggingexperimenten zeigte, dass insbesondere bei Partikeln sehr viele Fehler gemacht werden, gefolgt von Verben und Pronomina. Charakteristika gesprochener Sprache wie Wortabbrüche oder Buchstabiertes werden nicht

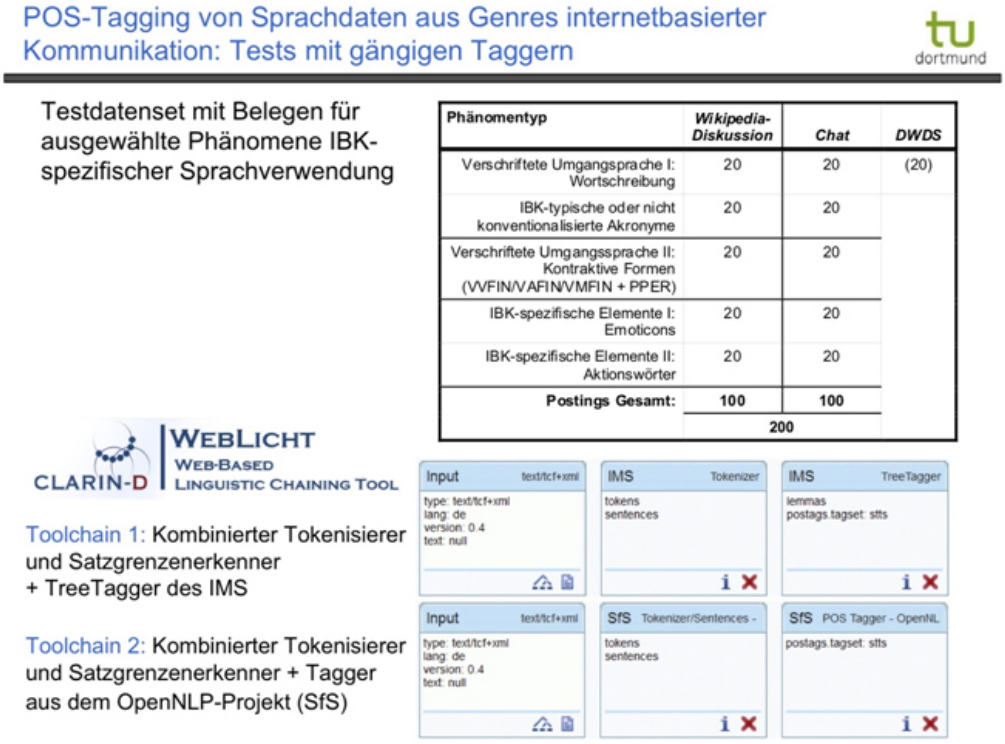


Abb. 1: Tagging des Dortmunder Chat-Korpus mit WebLicht-Webservices (Beißwenger et al. 2012).

erkannt, sondern als ‚Nichtwort' (nach STTS z.B. 3:7) klassifiziert. Das STTS bietet hierfür keine passenden Kategorien an. Experimente mit dem Potsdamer „KiezDeutsch-Korpus" KiDko deuteten auf ähnliche Probleme hin. Das Projekt schlägt daher eine Erweiterung des Tagsets für Partikeln, Pausen und ähnliche Elemente vor, die charakteristisch für gesprochene Spontansprache sind. Zwei Beiträge befassten sich mit internetbasierter Kommunikation. Josef Ruppenhofer (Hildesheim) stellte stellvertretend für Gertrud Faaß eine Untersuchung zum Tagging von Modeblogs vor. Michael Beißwenger, Angelika Storrer und Thomas Bartz (Dortmund) motivierten anhand von Untersuchungen des Dortmunder Chat-Korpus und von Wikipedia-Diskussionen mit WebLicht eine Erweiterung des Tagsets um Tags für Emoticons, Aktionswörter und Adressierungsausdrücke, die als typische Stilelemente in internetbasierter Kommunikation gelten (siehe auch Abb. 1). Mark Reznicek (Berlin) diskutierte die Verwendung von STTS im Lernerkorpus Falko und wie man Fehler von Lernenden, die oftmals zu widersprüchlichen Analysen auf den Ebenen der Morphologie, Distribution und Lexik führen, am besten abbilden kann. Im letzten Vortrag stellte Stefanie Dipper (Bochum) das „DeutschDiachronDigital-Tagset" (DDDTS) vor, das sowohl den Anforderungen der philologischen Analyse als auch der Anbindbarkeit an das STTS und computerlinguistische Taggingtools genüge leistet.

In der Diskussion zur Weiterentwicklung der Dokumentation von STTS und seiner Varianten wurden mehrere Arbeitsgruppen zu den offenen Diskussionspunkten eingerichtet. Die Teilnehmenden möchten zunächst virtuell über ein gemeinsames Wiki zusammenarbeiten; im Frühjahr 2013 ist ein Anschlussworkshop geplant. Die Ergebnisse der beiden Workshops und der gemeinsamen Arbeitsgruppen sollen längerfristig in einer gemeinsamen Publikation Niederschlag finden.

Die Einbindung des Workshops in CLARIN-D bietet eine gemeinsame Struktur im Hintergrund an, die dieses Vorhaben ermöglicht, das nur durch die Zusammenarbeit der auf verschiedene Institutionen verteilten Kompetenzen gelingen kann und das den Nachhaltigkeitszielen von CLARIN-D entspricht.

Heike Zinsmeister

Abkürzungsverzeichnis (NELCA)

AAI	Authentification and Authorization Infrastructure
AEDit	Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit
ALLEA	ALL European Academies
AP	Arbeitspaket
AsiCa	Atlante Sintattico della Calabria
BAS	Bayerisches Archiv für Sprachsignale (München)
BBAW	Berlin-Brandenburgische Akademie der Wissenschaften
BMBF	Bundesministerium für Bildung und Forschung
CMDI	Component MetaData Infrastructure
CLARIN	Common Language Resources and Technology Infrastructure
DAITF	Data Access and Interoperability Task Force
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DDDTS	DeutschDiachronDigital-Tagset
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
EXMARaLDA	Extensible Markup Language for Discourse Annotation
F-AG	Fachspezifische Arbeitsgruppen
FI-Initiative	Forschungs-Infrastruktur-Initiative
FOLK	Forschungs- und Lehrkorpus gesprochenes Deutsch
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
HAB	Herzog August Bibliothek Wolfenbüttel
HPC	High Performance Cluster
HZSK	Hamburger Zentrum für Sprachkorpora
ICRI	International Conference on Research Infrastructures
IDS	Institut für Deutsche Sprache (Mannheim)
IETF	Internet Expert Task Force
IMDI	ISLE Meta Data Initiative
IMS	Institut für Maschinelle Sprachverarbeitung (Stuttgart)
InfAI e.V.	Gemeinnütziger Verein des Instituts für Angewandte Informatik in Leipzig
IWiST	Informationswissenschaft und Sprachtechnologie (Hildesheim)
KiDko	KiezDeutsch-Korpus
LiS	Literatur- und Informationsversorgungssysteme
MPI	Max-Planck-Institut

NELCA	*Never-Ending-List* der CLARIN-Abkürzungen
NSF	National Science Foundation (USA)
PID	Persistent Identifier
SHARE	Survey of Health, Ageing and Retirement in Europe
STTS	Stuttgart-Tübingen Tagsets
TCF	Text Corpus Format
TLA	The Language Archive
TüNDRA	Tübingen aNnotated Data Retrieval Application
WADL	Web Application Description Language