



Nummer 5, 2013, November

PID: 11858/00-1779-0000-0022-F090-8

Editorial

Fünfter CLARIN-D-Newsletter

CLARIN-D läuft nun schon fast 2 ½ Jahre. Was haben wir erreicht? Alle Zentren haben das Data Seal of Approval bekommen. Die technische Infrastruktur ist weit gereift, alle Zentren haben Repositories eingerichtet, deren Inhalte laufend in das Virtual Language Observatory eingepflegt werden. Eine Authentifizierung über ein einziges Login erlaubt den kontrollierten Zugriff auf Ressourcen, eine erste Version der Suche parallel in allen CLARIN-D-Repositories ist implementiert, und, und, und ...

Darüber hinaus sind viele Ressourcen, namentlich Textkorpora und Sprachdatenbanken, aber auch Tools und Ergänzungen der ISOcat Definitionen, zum CLARIN-D-Bestand hinzugekommen. Eine Eigenschaft von CLARIN-D, im Marketingdeutsch ein „Alleinstellungs-

merkmal“, sind die Facharbeitsgruppen, und darin die Kurationsprojekte. Sie bieten den verschiedenen Wissenschaftsbereichen die Möglichkeit, für ihre eigenen Zwecke die CLARIN-D-Technologien und Infrastruktur zu nutzen.

In diesem Newsletter stellen gleich zwei Facharbeitsgruppen ihre Ressourcen vor: die Facharbeitsgruppe 6 „Sprache und andere Modalitäten“ hat drei multimodale Korpora aus dem Bereich der Mensch-Mensch- und Mensch-Maschine-Kommunikation erstellt und annotiert, die Facharbeitsgruppe 7 „Angewandte Sprachwissenschaft und Computerlinguistik“ hat erfolgreich das webbasierte linguistische Annotationstools WebAnno entwickelt.

Von der Digital Humanities-Fachkonferenz berichtet Erhard Hinrichs – und zieht ein sehr positives Fazit der Entwicklung in diesem Bereich. Passend dazu haben wir intern und in den Facharbeitsgruppen nach Erfolgsgeschich-

ten gefragt. Zwei davon sind in diesem Newsletter: Die Universität Graz nimmt CLARIN-D-Korpora als Vorbild und ein Startup-Unternehmen aus München verwendet WebMAUS zur Steuerung eine Spiele-Avatars.

Neu in diesem Newsletter ist die Rubrik zu den CLARIN-D-Repositories. Nachdem diese nun in allen Zentren eingerichtet wurden, ist es an der Zeit, sie etwas ausführlicher vorzustellen. Welche Ressourcen sind im jeweiligen Zentrum verfügbar, wie kommt man an die Daten heran, wie werden sie ausgeliefert?

Als fünftes CLARIN-D-Zentrum stellt sich in dieser Ausgabe das Zentrum der Eberhard Karls Universität Tübingen vor, angesiedelt am Seminar für Sprachwissenschaft. Hier ist nicht nur die Leitung von CLARIN-D beheimatet, sondern hier werden zentrale Technologiekomponenten wie WebLicht, der Aggregator für die verteilte Suche entwickelt sowie umfangreiche eigene und externe Korpora aufbereitet und verfügbar gemacht.

Und es gibt auch wieder Neuigkeiten aus der Küche: Die historischen Kochbücher sind nun eingescannt, transliteriert, teilweise schon linguistisch annotiert – und sie werden aktuell als Anwendungsbeispiel für Sprachkorpora in der Lehre verwendet.

Zu guter Letzt noch zwei Terminhinweise: am Bayerischen Archiv für Sprachsignale findet am 31.03.2014 ein eintägiger Workshop „Sprachdatenbanken – von der Aufnahme zur Publikation“ statt, in Leipzig die gemeinsam von der [European Summer School in Digital Humanities „Culture & Technology“](#) und CLARIN-D organisierte Sommerschule (21.07.-02.08.2014).



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.



CLARIN-D goes multimodal

Von multimodalen Korpora und ihren Metadaten

Im Kurationsprojekt 1 der F-AG 6 „Sprache und andere Modalitäten“ werden multimodale Ressourcen zur Integration in CLARIN-D aufbereitet.

Natürliche Kommunikation besteht nicht nur aus gesprochenen Worten, sondern auch aus deren Lautstärke, Sprechtempo oder Sprachmelodie sowie Gestik, Mimik, Blickrichtung oder Körperhaltung. Während in der lingu-

istischen Forschung korpusbasierte Methoden mittlerweile fest etabliert sind, steckt die korpusbasierte Grundlagenforschung insbesondere im Bereich multimodaler Kommunikation noch in den Anfängen. Das liegt u.a. darin begründet, dass der Aufbau entsprechender Korpora extrem zeit- und kostenaufwändig ist und bislang kaum automatische oder teil-automatische Analysemethoden zur Verfügung stehen. Verschiedene Disziplinen haben jedoch großes Interesse an multimodalen Ressourcen. So sind multimodale Daten für WissenschaftlerInnen aus Psychologie, Linguistik, Neurologie, Philologie etc. relevant, weil sie ein vollständigeres und detailliertere-



Abbl: Beispielaufnahmen aus dem SaGA Korpus, in dem von TeilnehmerInnen eine virtuelle Stadt beschrieben wird.

res Bild menschlicher Kommunikation ermöglichen. Aber auch aus technischer Sicht spielt das Thema Multimodalität eine immer größere Rolle, etwa wenn es darum geht, natürliche und intuitive Benutzerschnittstellen wie intelligente, virtuelle Agenten oder humanoide Roboter zu entwickeln.

Das Ziel des Kurationsprojekts, das im Januar 2014 abgeschlossen wird, besteht darin, den Weg für die zukünftige Bereitstellung multimodaler Daten über CLARIN-D zu ebnen. Dazu wurden drei unterschiedliche Korpora ausgewählt, an denen die Aufbereitung und Integration in die CLARIN-D-Infrastruktur exploriert wird. Bei den Korpora handelt es sich um (1) das manuell annotierte Bielefelder *Speech and Gesture Alignment Corpus* (SaGA), das aus Wegbeschrei-

bungsdialogen besteht, in denen die Kombination von natürlicher Sprache und Gestik im Mittelpunkt steht. Das an der Universität Hamburg erstellte (2) *Dicta-Sign Corpus* aus dem gleichnamigen EU-Projekt enthält sowohl Monologe als auch Dialoge in deutscher Gebärdensprache. Neben dem deutschen Teil existieren drei weitere Parallelkorpora in Britischer, Französischer und Griechischer Gebärdensprache. Das (3) *Natural Media Motion Capture*-Korpus (NM MoCap-Korpus) wurde an der RWTH Aachen aufgenommen. Es enthält monologische und dialogische Objektbeschreibungen, die mit modernsten *Motion Capturing*-Verfahren aufgenommen wurden.

Die drei Korpora wurden jeweils vor Ort, in Bielefeld, Hamburg und Aachen CLA-



Abb2: Beispielaufnahmen aus dem Dicta-Sign Korpus der deutschen Gebärdensprache.

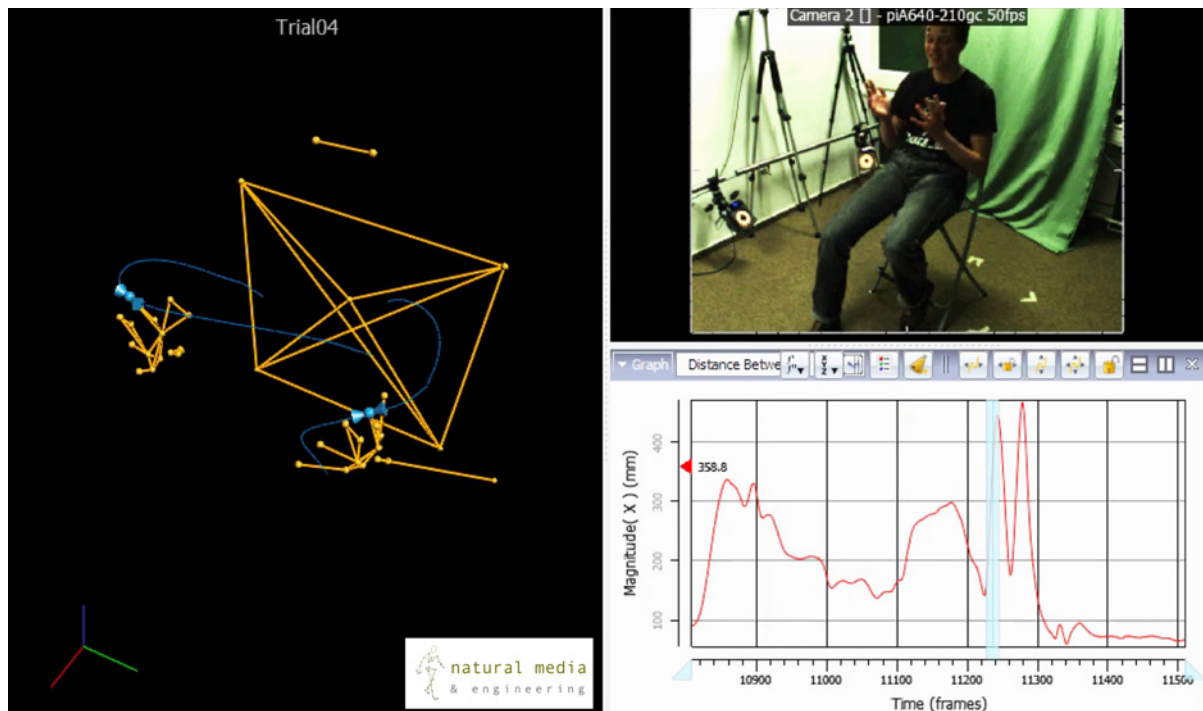


Abb3: Beispiel einer Visualisierung von *Motion Capture*-Aufnahmen einer Dombeschreibung.

RIN-D-konform aufbereitet und dokumentiert. In enger Zusammenarbeit der drei MitarbeiterInnen Farina Freigang, Rie Nishio und Matthias Priesters wurden darüber hinaus Gemeinsamkeiten und Unterschiede der drei Ressourcen herausgearbeitet, die zukünftig das Einbringen multimodaler Ressourcen in die CLARIN-D-Infrastruktur verbessern und vereinfachen sollen. Einen besonderen Schwerpunkt der gemeinsamen Arbeit bildeten Metadatenbeschreibungen. Um die Besonderheiten multimodaler Daten auch auf dieser Ebene zugäng-

lich und nutzbar zu machen, wurden ergänzend zu den bereits in CLARIN-D verwendeten CMDI-Profilen weitere Metadatenkomponenten erarbeitet, die beispielsweise Charakteristika von Zeitreihen und Videodaten beinhalten oder detaillierte Angaben über die Händigkeit der SprecherInnen. Damit gehen die neuen Metadatenprofile für multimodale Daten deutlich über bestehende Projekte hinaus und ermöglichen WissenschaftlerInnen eine gezielte Suche nach verschiedenen Aspekten natürlicher Kommunikation.



Rie Nishio
Universität Hamburg



Kirsten Bergmann
Universität Bielefeld



Farina Freigang
Universität Bielefeld



Matthias Priesters
RWTH Aachen

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

Erfassung der handgeschriebenen Rezepte

Im ersten Bericht im vorletzten Newsletter haben wir das kleine Kochbuchprojekt vorgestellt. Zur Erinnerung: Kathrin Beck hat zwei handgeschriebene Kochbücher geerbt, ein großes, repräsentatives und wenig benutztes sowie ein kleines Gebrauchskochbuch.



Abbildung 1: Buchscanner der Firma Atiz. Deutlich sichtbar die rechtwinklig montierten Glasscheiben, die die Buchseiten anpressen.

Seit dem letzten Bericht ist einiges geschehen: Das kleine Kochbuch wurde eingescannt und vollständig manuell transliteriert, das zweite ist noch in Arbeit. Einige Texte wurden bereits linguistisch annotiert, andere bilden die Basis für ein multimodales Korpus mit gesprochenen Rezepten.

Abfotografieren und Scannen

Damit die Bücher transliteriert werden können, müssen sie eingescannt oder abfotografiert werden. Automatische Schrifterkennung funktioniert bei handgeschriebenen Texten, und dann noch bei historischen Schriften und hier auch noch bei fleckigem Papier und sichtbaren Schattenschriften, nicht – die Transliteration muss daher manuell erfolgen. Das Abfotografieren der Buchseiten geht mit speziellen Aufsicht-Buchscannern

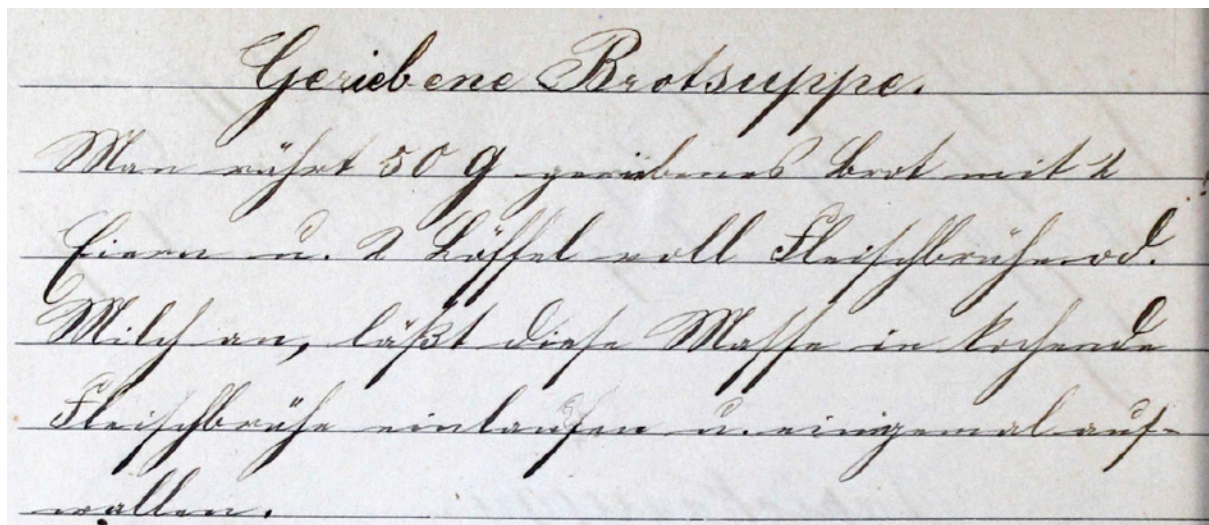


Abbildung 2: Suppenrezept aus dem großen Kochbuch. Die blasse, in die entgegengesetzte Richtung stehende Schrift ist ein Abdruck der gegenüberliegenden Kochbuchseite.

einigermaßen schnell und relativ schonend, da das Buch nicht ganz flach, sondern nur bis ca. 100 Grad aufgeklappt werden muss, wobei die Seiten mit einer Glasscheibe flachgedrückt und beide Seiten auf einmal fotografiert werden. Das IT-Zentrum der Geisteswissenschaften der LMU München hat einen solchen Buchscanner und so konnte das große Kochbuch an einem Vormittag fotografiert werden. Siehe dazu Abbildung 1.

Mit dem kleinen Kochbuch ging das nicht, denn es ist bis in den Falz hinein beschrieben – wenn die Seiten nicht komplett flach aufgeschlagen werden, kann man die Schrift ganz nah am Falz nicht lesen. Dieses Buch musste daher auf einem herkömmlichen Flachbettscanner eingescannt werden, was aufgrund der teilweise verklebten Seiten zu verrutschten Scans und Unschärfen in der Nähe des Falzes geführt hat.

Als Ergebnis stehen nun 140 jpg-Dateien der fotografierten Einzelseiten (je 2400

× 3000 px, aus technischen Gründen nur jpg-Format) des großen und 69 tiff-Dateien der gescannten Doppelseiten (je 2604 × 1947 px) des kleinen Kochbuchs zur Verfügung (Abbildung 2 und 3).

Mit einem Grafikprogramm wurden die Helligkeit reduziert und der Kontrast erhöht, damit die Schrift besser zu lesen ist.

Transliteration der Scans

Die Kochbücher sind in der alten deutschen Kurrentschrift geschrieben. Diese war bis in die Mitte des 20. Jahrhunderts in Gebrauch; sie wurde zunächst von der Sütterlinschrift und dann aufgrund des „Normalschrift-Erlasses“ 1941 von der deutschen Normalschrift ersetzt (auf Wikipedia gibt es hierzu lesenswerte Artikel).

Auf der Suche nach jemandem, der die Kurrentschrift noch lesen kann, haben wir im Internet die „Sütterlinstube Hamburg“ entdeckt. Ihre Aufgabe ist u.a. die „Unterstützung bei der Über-

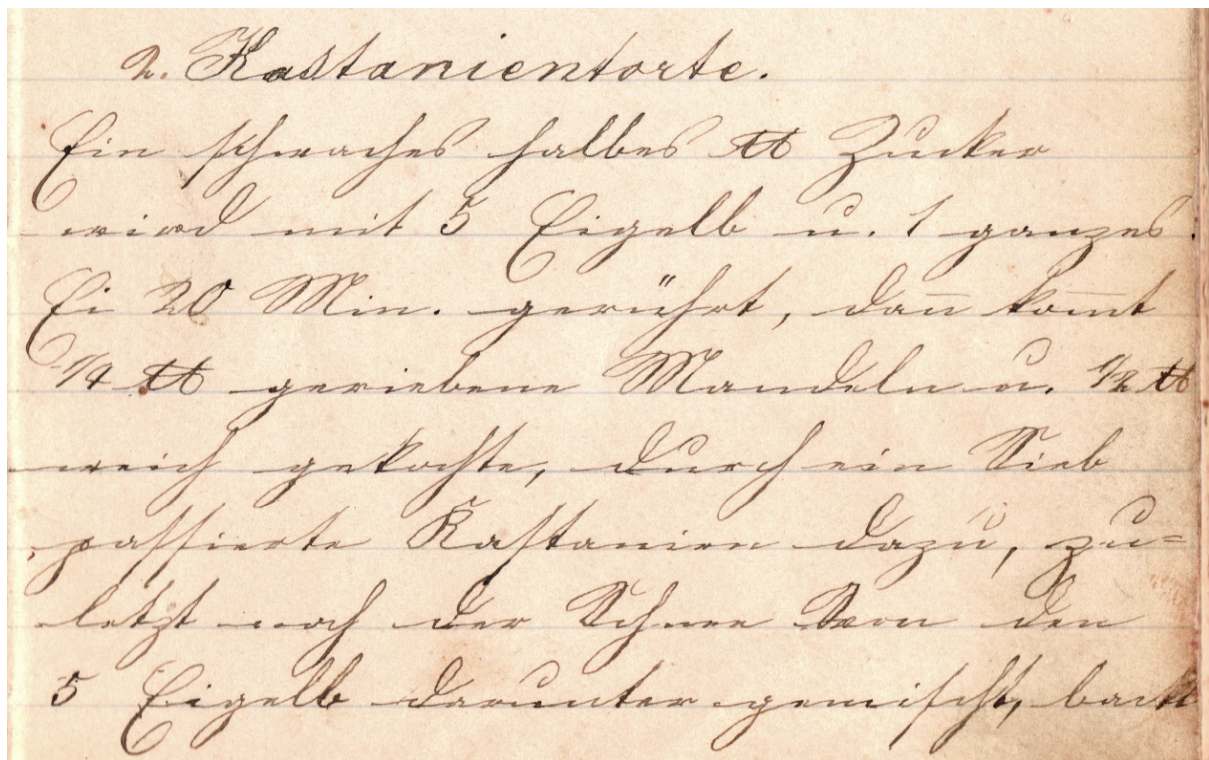


Abbildung 3: Tortenrezept aus dem kleinen Kochbuch. Die Seiten sind vom Buchfals bis zum äußeren Rand beschrieben, das Papier ist fleckig und die Ecken und Ränder sind abgegriffen.

tragung historischer Dokumente aus der deutschen in die lateinische Schrift als Beitrag für ein sinnerfülltes Leben im Ruhestand, für eine Verständigung zwischen den Generationen“ (Abb. 4).

Die Sütterlinstube Hamburg hat sich bereit erklärt, das kleine Kochbuch buchstabengetreu inklusive aller Schreib- und sonstiger Fehler zu verschriften. Dazu wurden die eingescannten Dateien auf A3 in Farbe ausgedruckt und nach Hamburg geschickt. Nach wenigen Wochen kamen die ersten Word-Dokumente sowie die Ausdrucke zurück – fantastisch! Den Mitarbeiterinnen und Mitarbeitern der Sütterlinstube dafür einen herzlichen Dank.

Die verschiedenen Word-Dokumente wurden anschließend zu einem einzigen Fließtext zusammengeführt und

als weitgehend unformatierte Textdatei im UTF-8 Format abgespeichert. Diese dient nun als Ausgangsbasis für die weitere linguistische Auszeichnung und sonstige Verarbeitung.



Abbildung 4: Webseite der Sütterlinstube Hamburg e.V.

Ausblick

Aktuell planen wir, die Rezepte von mehreren Sprecherinnen in verschiedenen Dialekten lesen zu lassen, um auch eine gesprochene Version des Kochbuchs zu bekommen.

Im Rahmen eines praktischen Korpusprojekts im Master-Seminar „Sprachdatenbanken“ am Institut für Phonetik in München werden am Beispiel des Kochbuchs alle Schritte von der Spezifikation über die Aufnahme bis zur Erstellung der Annotation und CLARIN-kompatiblen Metadaten durchlaufen.

In einem ersten Schritt wurden die einzelnen Rezepte aus dem Fließtext extrahiert und in Aufnahmeskripte für die Software SpeechRecorder übertragen. Nun wird die Aufnahmetechnik zusammengestellt und in tragbare Koffer ver-



Kathrin Beck

Seminar für Sprachwissenschaft, Universität Tübingen

packt, so dass die Sprachaufnahmen „im Feld“ bald beginnen können.

Bon appétit!



Christoph Draxler

Institut für Phonetik und Sprachverarbeitung, LMU München



Elke Teich

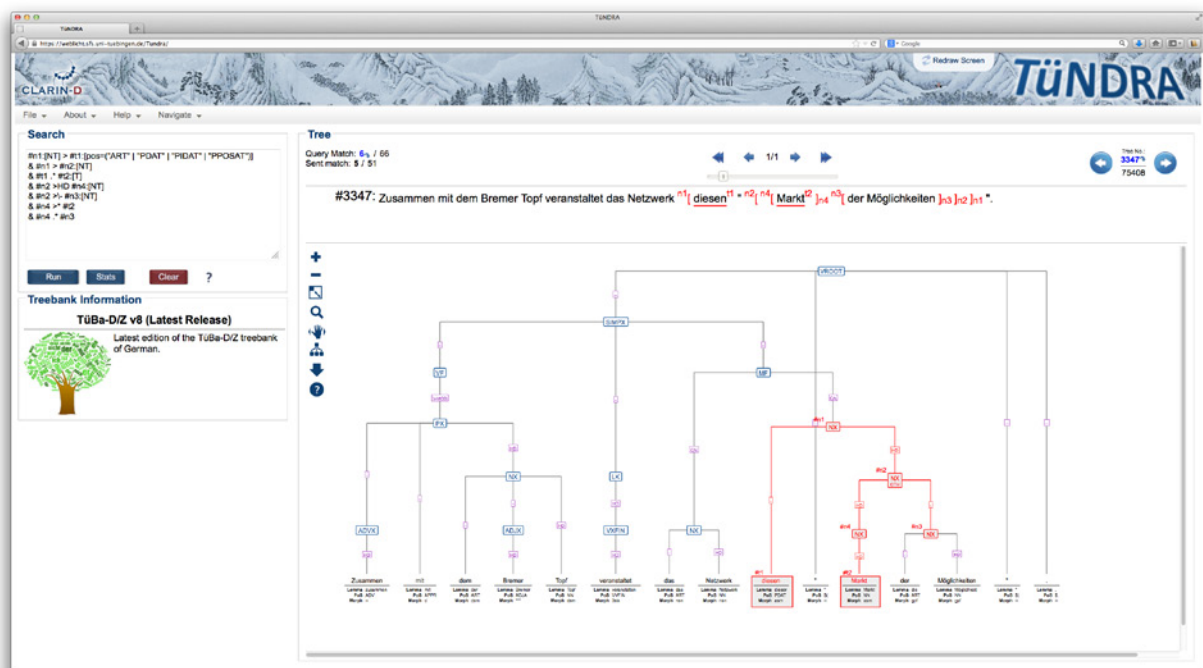
Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes

Ein CLARIN-Zentrum stellt sich vor: Das CLARIN-Zentrum an der Universität Tübingen

Von Netzen, Bäumen und Anfragen

Die Universität Tübingen hat eine lange Tradition im Bereich der *Digital Humanities*. Die in Tübingen entwickelte TUSTEP-Software [1] spielt seit den 1980er Jahre eine wichtige Rolle für die Erstellung digitaler Editionen. Seit den 1990er Jahren ist die nachhaltige Erstellung von Forschungsprimärdaten

ein fester Bestandteil im Forschungsprogramm der sprachwissenschaftlichen Sonderforschungsbereiche „Linguistische Datenstrukturen“ (SFB 441) und „Bedeutungskonstitution: Dynamik und Adaptivität sprachlicher Strukturen“ (SFB 833). Mit der Einführung von Informationsinfrastrukturprojekten in geisteswissenschaftlichen Sonderforschungsbereichen hat die Tübinger Linguistik schon früh auf die Nachnutzbarkeit von Forschungsdaten gesetzt und die dafür notwendigen Strukturen



Suche über syntaktisch annotierte Korpora mit Tundra

[1] <http://www.tustep.uni-tuebingen.de>



Gruppenfoto: das CLARIN-Team

und Einrichtungen geschaffen. Als Teil des Zukunftskonzepts der Universität Tübingen, das in der 3. Förderlinie der Exzellenzinitiative der DFG gefördert wird, spielen nachhaltige, institutionelle Strukturen für die *Digital Humanities* eine zentrale Rolle. Für den Bereich der Sprachwissenschaft ist das CLARIN-Zentrum an der Universität Tübingen fester Bestandteil dieser institutionellen Strategie. Das Tübinger CLARIN-Zentrum integriert sich so sowohl in die nationale und internationale CLARIN-Infrastruktur als auch in eine dynamische lokale Forschungsinfrastruktur. Das Tübinger CLARIN-Zentrum verfügt über besondere Expertise in den Bereichen manuell und automatisch an-

notierter Sprachkorpora, lexikalischer Ressourcen sowie Webanwendungen und Webservices für die Annotation, Visualisierung und Abfrage von Textkorpora.

Bäume und Netze: Ressourcen des CLARIN- Zentrums

Lexikalische Ressourcen in Form von Wörterbüchern und Lexika reichen bis an die Anfänger sprachwissenschaftlicher Forschung zurück. In den vergangenen Jahrzehnten wurden neue Arten von Lexika entwickelt: In Wortnetzen sind bedeutungsähnliche Wörter in sogenannten Synsets zusammengefasst.

Dabei werden diese Synsets unter dem Aspekt der zwischen ihnen bestehenden semantischen Relationen (z.B. Antonymie, Hyponymie) abgebildet. GermaNet [2] ist die Wortnetz-Ressource des Deutschen und umfasst ca. 85.000 Synsets und ca. 100.000 Relationen. GermaNet ist bereits im 8. Release verfügbar und wird von vielen Forschern im Bereich Linguistik und Sprachtechnologie verwendet. Es wird vom CLARIN-Zentrum Tübingen (weiter)entwickelt, verfügbar gemacht und archiviert .

Neben lexikalischen Ressourcen stellt das Tübinger CLARIN-Zentrum ein breites Angebot linguistisch annotierten Korpora [3] für gesprochene und geschriebene Sprache zur Verfügung. Am prominentesten ist dabei die Tübinger Baubank des Deutschen/Zeitungskorpus, ein manuell bearbeitetes syntaktisches Korpus in Form einer Baubank, die auf Zeitungstexten basiert. Weitere Baubanken für die Sprachen Deutsch, Englisch und Japanisch schließen die automatisch annotierte Texte des „Tübinger Partiiell Geparsten Korpus des Deutschen/Zeitungskorpus“ (TüPP-D/Z) sowie die Baubanken Tü-Ba-D/S, TüBa-J/S und Tü-Ba-J/S für gesprochene Sprache ein.

Das CLARIN-Zentrum in Tübingen besitzt damit wesentliche Kompetenzen für lexikalische Ressourcen und syntaktische Korpora in CLARIN. Auf diese Expertise kann auch von anderen Forschenden zurückgegriffen werden, die Ressourcen können im Rahmen entspre-

chender Lizenzen zugänglich gemacht werden.

Ohne Staub und Regale: Ein Archiv für Sprachressourcen

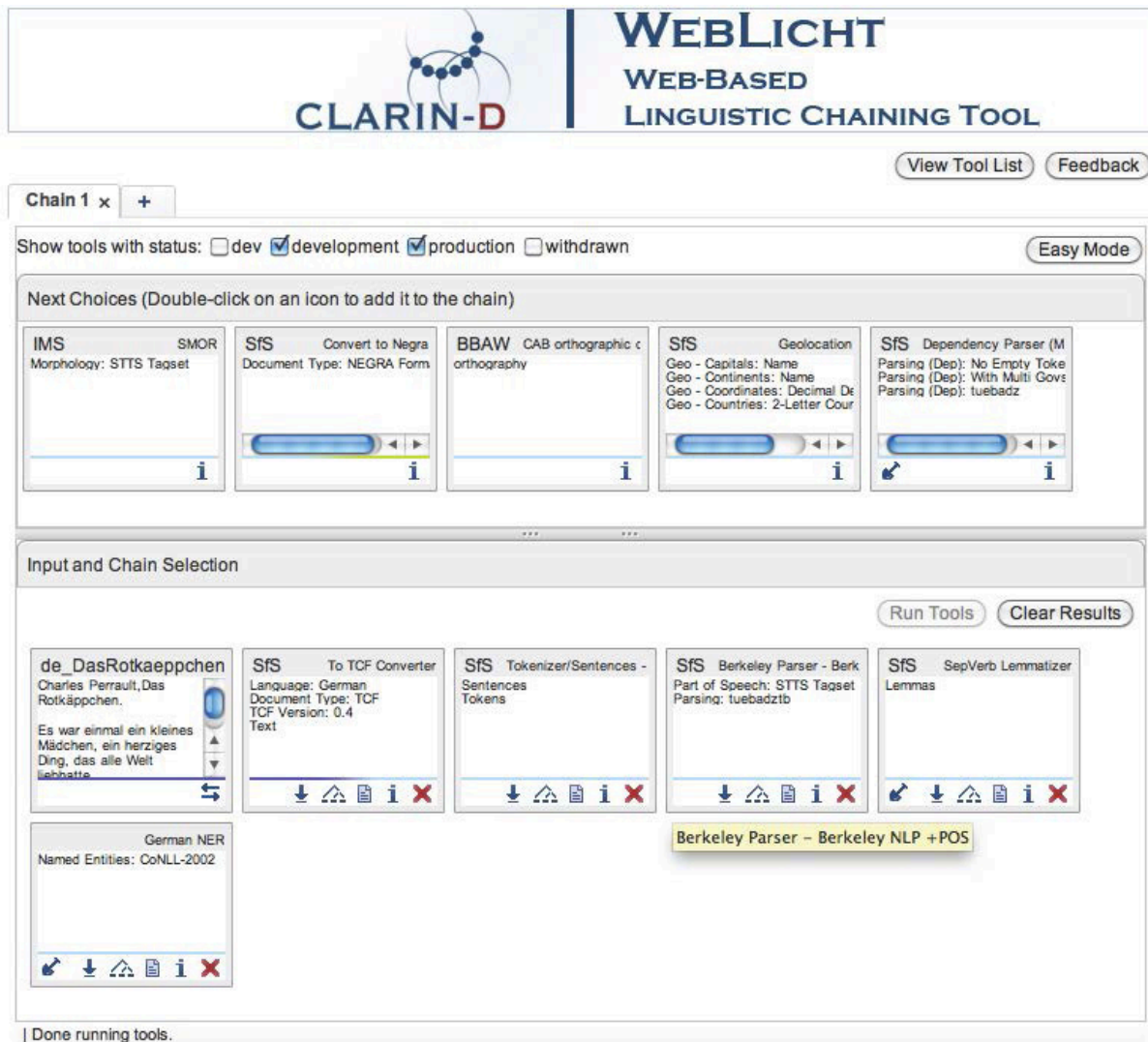
Ein wesentlicher Bestandteil der Zugänglichkeit von Sprachressourcen ist es, dass ihr Aufbewahrungsort bekannt ist und es erlaubt, die Ressourcen zu lokalisieren. Das CLARIN-Zentrum Tübingen besitzt hierfür ein Repository [4], das die strukturierte Ablage von Sprachressourcen und ihre Ergänzung um Metadaten erlaubt. Über diese Metadaten können andere Forscher die Ressource auffinden. Das CLARIN-Zentrum Tübingen setzt hierfür mit *Fedora-Commons* ein bewährtes und kontinuierlich weiterentwickeltes Open-Source-Softwaresystem ein.

Ergänzt um eine Benutzeroberfläche, können so alle Ressourcen archiviert werden, auf die sich das Tübinger Zentrum spezialisiert hat, etwa Baubanken und syntaktische Korpora. Durch die Flexibilität des Systems können aber auch andere Ressourcen aufgenommen und bereitgestellt werden. Durch die Zertifizierung als CLARIN-Zentrum vom Typ B und das *Data Seal of Approval* ist zudem gesichert, dass das CLARIN-Zentrum Tübingen höchste Qualitätsanforderungen an die Archivierung erfüllt. Durch eine Replizierung der Daten im Rechenzentrum Garching wird eine hohe Datensicherheit gewährleistet.

[2] <http://www.sfs.uni-tuebingen.de/GermaNet>

[3] <http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora.html>

[4] <http://www.sfs.uni-tuebingen.de/ascl/clarin-center/repository.html>



The screenshot shows the WEBLICHT interface with the following components:

- Header:** CLARIN-D logo and WEBLICHT WEB-BASED LINGUISTIC CHAINING TOOL.
- Navigation:** View Tool List, Feedback, Chain 1 x +, Show tools with status: dev development production withdrawn, Easy Mode.
- Next Choices:** A grid of tool cards including IMS (Morphology: STTS Tagset), SFS (Convert to Negra), BBAW (CAB orthographic), SFS (Geolocation), and SFS (Dependency Parser).
- Input and Chain Selection:** Run Tools, Clear Results, and a grid of tool cards including de_DasRotkaeppchen (text input), SFS (To TCF Converter), SFS (Tokenizer/Sentences), SFS (Berkeley Parser - Berk), SFS (SepVerb Lemmatizer), and German NER.
- Status:** Done running tools.

Kombination computerlinguistischer Analysewerkzeuge: WebLicht zur Zusammenstellung eigener Verarbeitungsketten

Das Archiv am CLARIN-Zentrum Tübingen steht dabei auch für alle andere Wissenschaftler offen, die im Rahmen von Vereinbarungen in Tübingen ihre Daten archivieren und aus diesem Archiv auch langfristig verfügbar machen möchten. Zum Beispiel wird in Tübingen bereits ein syntaktisch annotiertes Korpus der Werke von Thomas von Aquin gehostet, der *Index Thomisticus*, der an der Università Cattolica del Sacro Cuore in Mailand erstellt wird.

Stets zu Diensten: Webanwendungen und Webservices

Es gibt ein Gerücht in den Geistes- und Sozialwissenschaften: Computerlinguistische Programme sind schwer zu installieren, komplex zu bedienen und auch sonst nur von Fachleuten zu verwenden. Sie einfach kurz auszuprobieren ist darum schwierig. In CLARIN wird dieses Gerücht widerlegt.

Mit Webanwendungen, die sich auch untereinander kombinieren lassen, kann man sich so auch als Nicht-Fachmann seine benötigten computerlinguistischen Analysen individuell zusammenstellen und mit ihrer Hilfe neue Fragestellungen bearbeiten, die mit manueller Bearbeitung zu aufwendig wären. Mit Hilfe der WebLicht-Umgebung [5], die am CLARIN-Zentrum Tübingen entwickelt wird und Webservices verschiedener CLARIN-D-Partner integriert, ist es so möglich, eigene Texte zu verarbeiten, z.B. automatisch Wortarten und syntaktische Strukturen zu analysieren und diese anschließend zu visualisieren und zu durchsuchen. Speziell zur Abfrage syntaktisch annotierter Textkorpora wurde die Webplattform Tundra [6] ent-

wickelt. Für die gleichzeitige Korpusabfrage über Ressourcen verschiedener CLARIN-Partner wurde eine weitere Webanwendung entwickelt, die *Federated Content Search* (FCS) [7].

Mit diesen Diensten, Ressourcen und dem Repository ist das CLARIN-Zentrum Tübingen an dem nationalen und europäischen CLARIN-Projektverbund beteiligt und bringt sich so in diese Infrastruktur ein. Als Koordinatoren in CLARIN-D laufen dabei viele Handlungsstränge in Tübingen zusammen, um für die Geistes- und Sozialwissenschaften, die sich mit sprachbasierten Daten beschäftigen, neue Forschungsfragestellungen zu ermöglichen und bestehende Herausforderungen zu überwinden.



Kathrin Beck
*Seminar für
Sprachwissen-
schaft, Universi-
tät Tübingen*



Thorsten Trippel
*Seminar für
Sprachwissen-
schaft, Universi-
tät Tübingen*

[5] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

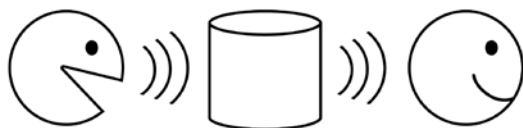
[6] <http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra>

[7] <http://weblicht.sfs.uni-tuebingen.de/Aggregator>

Ankündigung: CLARIN-D-Workshop „Sprachdatenbanken – von der Aufnahme zur Publikation“

IPS

INSTITUT FÜR PHONETIK
UND SPRACHVERARBEITUNG



Bayerisches Archiv für Sprachsignale



Der eintägige Workshop am **31.03.2014** richtet sich an Studierende und junge Wissenschaftler/innen, die Sprachaufnahmen durchführen, sie annotieren, analysieren und auswerten und die ihre Daten anschließend verfügbar machen wollen.

Im Workshop präsentieren Dozenten des Instituts für Phonetik und Sprachverarbeitung der **LMU München** die Arbeitsschritte von der Spezifikation des Korpus über die Sprachaufnahmen und die Segmentation bis hin zur Dokumentation, und sie stellen die dazu entwickelten Tools und bewährte Vorgehensweisen vor.

Bewerben Sie sich mit einem kurzen **Motivationsschreiben** für die Teilnahme am Workshop. Dieses Schreiben darf max. 150 Wörter lang sein, muss als PDF vorliegen und muss bis spätestens 16.02.2014 bei workshop@phonetik.uni-muenchen.de eingegangen sein.

Die Teilnahmegebühr für den Workshop beträgt 30 €, Studierenden kann ein Zuschuss von max. 200 € gewährt werden.

Weitere Infos auch im Wiki:

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

Erfolgreicher Abschluss der Kurationsprojekte von F-AG 7

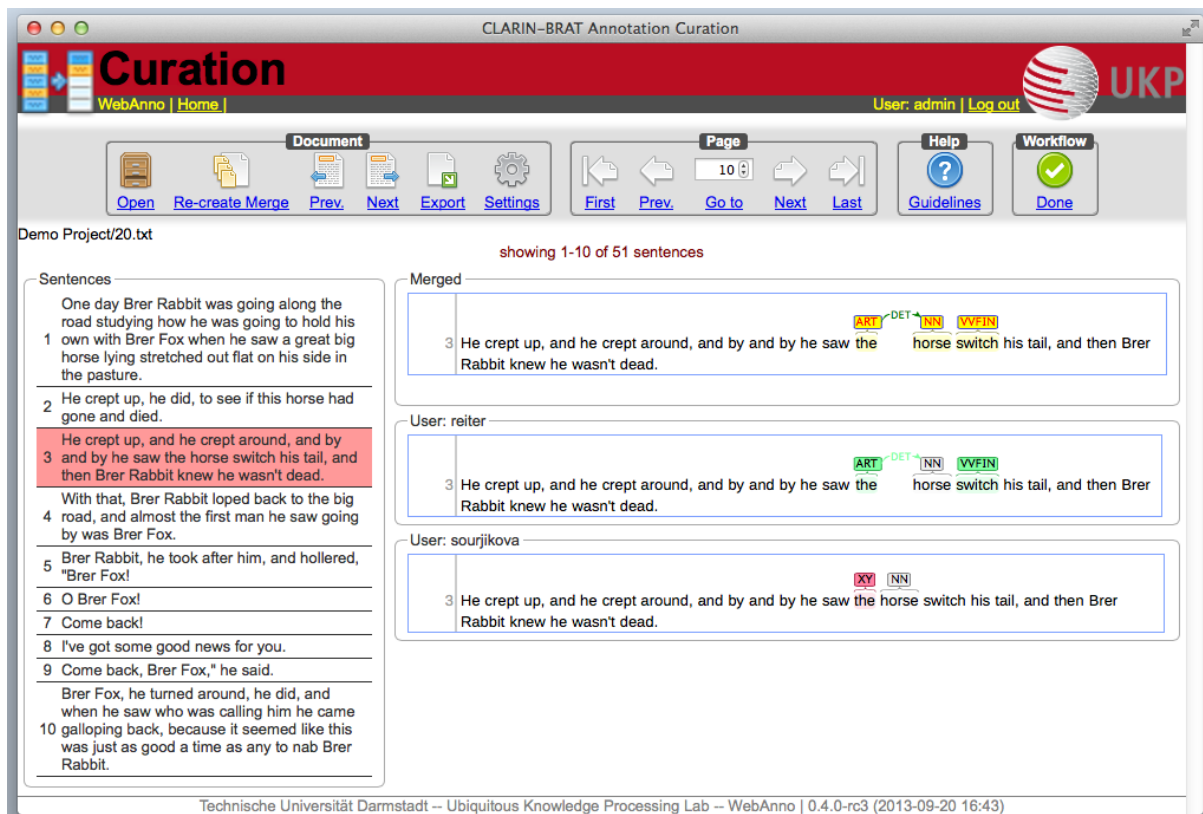
G_SCL-Arbeitskreis zu Digital Humanities gegründet, Kurationsprojekte erfolgreich abgeschlossen

Die F-AG 7 *Angewandte Sprachwissenschaft und Computerlinguistik* [1] setzt sich die Aufgabe, die Leistungsfähigkeit maschineller Sprachverarbeitung (NLP) für das Deutsche zu befördern. Besonderes Augenmerk gilt dabei der Analyse von speziellen Textsorten aus verschiedensten Domänen und Genres, die in den geisteswissenschaftlichen Fächern relevant sind. Herausforderungen bilden hierbei außergewöhnliche Texteigenschaften (z.B. Wortstellung in der Ly-

rik, umgangssprachliche Äußerungen in Online-Chats oder Äußerungen von Lernenden, historische oder fachsprachliche Termini). Die Analyse sog. „Nicht-Standard-Varietäten“ und Texte spezieller Domänen stellt für maschinelle Sprachverarbeitungssysteme noch immer eine besondere Herausforderung dar.

In der F-AG 7 haben wir uns in diesem Kontext zunächst auf den Aspekt der *linguistischen Annotation* konzentriert, da die Verfügbarkeit von linguistisch annotierten Sprachdaten einen entscheidenden Einfluss auf die Nutzbarkeit von vielen modernen (statistischen) NLP-Tools hat. Anders als im Englischen sind für das Deutsche vergleichsweise wenige und wenig variationsreiche annotierte Sprachdaten vorhanden. Es besteht daher großer Bedarf an linguistischen Annotationen auf allen sprachlichen Ebenen, z.B. Wortarten, syntaktische Strukturen oder Eigennamen. Bereits mit geringen Mengen gezielt ausgewählter Sprachdaten können existierende statistische Analysemodelle verbessert und an neue Textsorten angepasst werden. Davon profitiert gerade die linguistische

[1] F-AG 7: <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-7-computerlinguistik.html>



Die Kurationsansicht, in der parallele Annotationen verglichen und zusammengeführt werden können

Forschung zu speziellen und weitgehend unerforschten Textsorten: Durch komfortable und inkrementell lernende Analysewerkzeuge.

Diesen Überlegungen folgend wurden zwei Kurationsprojekte konzipiert, die sich auf unterschiedliche Weise mit Annotationen beschäftigen.

WebAnno: Ein web-basiertes Annotationstool

Ziel des Kurationsprojekts *WebAnno* war es, ein web-basiertes Annotationstool zu entwickeln. Die Annotation kann dabei von einem beliebigen Computer aus erfolgen. Gleichzeitig kann durch die zentrale Speicherung die Wissenschaftlerin

oder der Wissenschaftler den Fortschritt der Annotationen direkt überwachen und, z.B. wenn eine Annotatorin oder ein Annotator die Aufgabe falsch verstanden hat, direkt eingreifen. Zur Qualitätssicherung der Annotationen dient insbesondere auch eine integrierte Berechnung von Agreement-Maßen, die direkt Aufschluss darüber geben, inwieweit die Annotatorinnen und Annotatoren übereinstimmen.

Derzeit unterstützt *WebAnno* wortbasierte Annotationen (z.B. Wortarten oder -bedeutungen, Eigennamen) und relationale Annotationen (z.B. Koreferenzen, syntaktische Dependenzrelationen). Annotationskategorien können eigens für spezielle Annotationsprojekte definiert werden; in der Zukunft sollen auch frei definierbare Annotationen er-

#	Sprecher	Beitrag
25	System	Lantonie betritt den Raum
26	Lantonie	:)
27	quaki	lantonieeeeeee
28	Lantonie	Hallo. :)
29	zora	LANTOOO :)))
34	marc30	Lantöööö :o)
35	TomcatMJ	hi lanto
40	Faryen-Angle	hi Lantonie

Tabelle: Beispiel aus dem Dortmunder Chat-Korpus

möglichst werden. Durch eine Anbindung an die crowdsourcing-Plattform Crowd-Flower können auch crowdsourcing-Annotationen in WebAnno verwaltet sowie vor- und nachbereitet werden.

Das Projekt wurde an der Technischen Universität Darmstadt unter der Leitung von Prof. Dr. Chris Biemann und Prof. Dr. Iryna Gurevych mit dem Team Darina Benikova, Richard Eckart de Castilho, Benjamin Milde und Seid Muhie Yimam durchgeführt. Das Annotationstool ist bereits als *open source*-Software veröffentlicht und frei heruntergeladen werden [2]. Das 1.0-Release ist für den Monat November geplant.

NoSta-D: Ein annotiertes Korpus mit Nichtstandardvarietäten

Die Annotation von (deutschsprachigen) Nichtstandardvarietäten wurde im zweiten Kurationsprojekt angegangen. Hier war das Ziel, ein aus heterogenen Sprachvarietäten zusammengestelltes Korpus auf mehreren Sprachebenen zu

annotieren und hierbei die Anwendbarkeit existierender Annotationsrichtlinien zu erproben und vor allem Probleme ihrer Anwendbarkeit zu dokumentieren. Das Korpus (ca. 39.000 Tokens) enthält Texte aus sechs Varietäten: L2-Lerner-Aufsätze, Chat-Protokolle, historische Texte, Literarische Prosa, Dialoge zu Wegbeschreibungen und Zeitungstexte. Abgesehen von den Dialogen handelt es sich um geschriebene Sprache; die Zeitungstexte wurden zum Zwecke der Gegenüberstellung hinzugefügt, als „Standarddomäne“.

Annotiert wurde auf drei verschiedenen Ebenen: Eigennamen, syntaktische Abhängigkeiten sowie Koreferenzen, wobei existierende Annotationsrichtlinien nicht nur erprobt, sondern auch verbessert und verfeinert wurden. Insbesondere die Annotation syntaktischer Abhängigkeiten birgt eine Fülle von Herausforderungen (z.B. Selbstkorrektur in gesprochener Sprache, uneinheitliche Grammatik und Orthographie in historischen Texten), die ganz eigene Herangehensweisen für die Konzeption einer linguistischen Annotation erfordern.

[2] WebAnno: <https://code.google.com/p/webanno/>

Auch bei Eigennamen, die auf den ersten Blick unproblematisch erscheinen können, liegt in einzelnen Domänen eine hohe Variabilität vor, wie das Beispiel aus dem Dortmunder Chat-Korpus (Tabelle) zeigt.

Das annotierte Korpus und die begleitende Dokumentation der erprobten oder neu konzipierten Annotationsrichtlinien wird derzeit in die CLARIN-D-Infrastruktur überführt und als CLARIN-D-Ressource verfügbar sein [3]. Das Projekt wurde durchgeführt an der Ruhr-Universität Bochum und der Humboldt-Universität Berlin von Prof. Dr. Anke Lüdeling, Prof. Dr. Stefanie Dipper und Marc Reznicek.

Neben Fachpublikationen, die in beiden Kurationsprojekten erfolgt sind, haben die Projekte ihre Ergebnisse auch sowohl beim M24-Treffen in Nijmegen als auch beim Disseminationsworkshop in Leipzig vorgestellt.

GSCL Arbeitskreis „CL für Digital Humanities“ gegründet

Um die Aktivitäten rund um Computerlinguistik in den Digital Humanities zu bündeln und in der CL-Community zu verankern, haben Prof. Dr. Anette Frank und Prof. Dr. Anke Lüdeling einen Arbeitskreis zu Digital Humanities im Rahmen der GSCL (Gesellschaft für Sprachtechnologie und Computerlin-

guistik) [4] gegründet. Der AK „CL für Digital Humanities“ wurde im Rahmen der GSCL-Konferenz im September 2013 in Darmstadt der Fachgemeinschaft vorgestellt.



Anette Frank,
Institut für Computerlinguistik, Universität Heidelberg



Nils Reiter,
Institut für Computerlinguistik, Universität Heidelberg

[3] Korpus NoSta-D: <http://hdl.handle.net/11022/0000-0000-1D83-C>

[4] GSCL-Arbeitskreis zu Digital Humanities: <http://www.gscl.org/ak-dh.html>



CLARIN-D

Erfolgsgeschichten

WebMAUS steuert Spiele-Avatar

Ein Startup-Unternehmen aus München verwendet WebMAUS zur Steuerung der Lippenbewegungen seines Spiele-Avatars. Zitat aus der E-Mail des Entwicklers: „Das Spiel ist als App für das iPad konzipiert; graphisch verwendet es einfache, kindgerechte comichafte 2D-Grafiken und simple, schrittweise Animationen. Spielerläuterungen und Hilfen für die Lernspiele werden durch animierte Figuren gegeben, die hierzu bestimmte Texte sprechen, die als einzelne Sprach-Audiodateien vorgegeben sind. Das Sprechen soll durch Animation der Münder der Figuren zusätzlich visualisiert werden. Hierbei gibt es nur zwei einfache Animationszustände: Mund geöffnet und Mund geschlossen. In einem inhaltlich verwandten Vorgängerspiel wurde dies so gelöst, dass in einem vor-

gelagerten Prozess alle Audio-Sequenzen hinsichtlich ihrer Amplituden abgetastet wurden. War die Amplitude über einem definierten Schwellwert, wurde dies als „Mund auf“ interpretiert, darunter als „Mund zu“. Die Ergebnisse dieses Präprozesses wurden in Datenstrukturen abgespeichert, die beim Abspielen der Sounds später im Spiel jeweils zusammen mit dem Audio geladen wurden, so dass die Animation synchron gesteuert wurde.

Das Ergebnis dieser Vorgehensweise ist jedoch nicht vollständig zufriedenstellend, da die Mundbewegungen leider nicht natürlich wirken. Wir haben den Eindruck, dass gerade Konsonanten in der Aussprache zwar einen Amplitudenausschlag bewirken, dass man dabei aber keinen geöffneten Mund in der Animation erwartet. „Wir haben daher beim aktuellen Spiel nach einer anderen Lösung gesucht. Als Basis liegen uns sowohl

die Sprecher-Audiodateien als auch alle entsprechenden Texte vor, nicht jedoch eine ‚Lautschrift‘-Darstellung. Unser Grundgedanke war dann, dass wir tendenziell bei allen Vokalen den Mund offen darstellen und bei allen Konsonanten geschlossen“, sagte ein Sprecher des Unternehmens.

„Bei Recherchen nach Lösungen sind wir auf MAUS gestoßen und haben mit WebMAUS für eine Beispielsequenz den Versuch unternommen, eine Audiodatei zusammen mit dem Textfile des Sprecher-Texts phonetisch segmentieren zu lassen. Mit dem Output von MAUS konnten wir mit den dort angegebenen Intervallen die Animation steuern, wobei wir den phonetischen ‚Buchstaben‘ jeweils einen der beiden Animationsszustände zugewiesen haben. Das Ergebnis war und ist verblüffend: Die Animation wirkt viel natürlicher als mit dem alten Algorithmus.“

Universität Graz nimmt deutsche CLARIN-Sprachkorpora als Vorbild und verwendet CLARIN-Tools

Prof. Martin Hagmüller der Universität Graz orientiert sich bei eigenen Aufnahmen für Österreichisches Deutsch an der Struktur des deutschen CLARIN-Sprachkorpus BAS PD1, der über das CLARIN-Repository des BAS frei verfügbar gemacht wurde. Außerdem verwendet die Arbeitsgruppe um Hagmüller das CLARIN-Aufnahme-Tool *SpeechRecorder* und erzeugt das Aussprache-Lexikon des neuen Sprachkorpus mit dem CLARIN-Tool BALLOON, das von Uwe Reichel im Rahmen des deutschen CLARIN-Projektes weiterentwickelt wurde und demnächst als frei verfügbarer Web-Service zur Verfügung stehen wird.

Alles Weitere unter:

www.clarin-d.org

Das BAS-Repository

Vorstellung des Repositoriums am Bayerischen Archiv für Sprachsignale

Aufbau

Das BAS-Repository ist über die folgende Webseite zu erreichen: <https://clarin.phonetik.uni-muenchen.de/BAS-Repository/> (alternativ über den Per-

sistent Identifier: <http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E>).

Es umfasst zum gegenwärtigen Zeitpunkt 16 multimodale Sprachkorpora mit einem Umfang von insgesamt 2.5 TByte sowie zugehörigen Metadaten im CM-DI-Format im Umfang von 13 GByte.

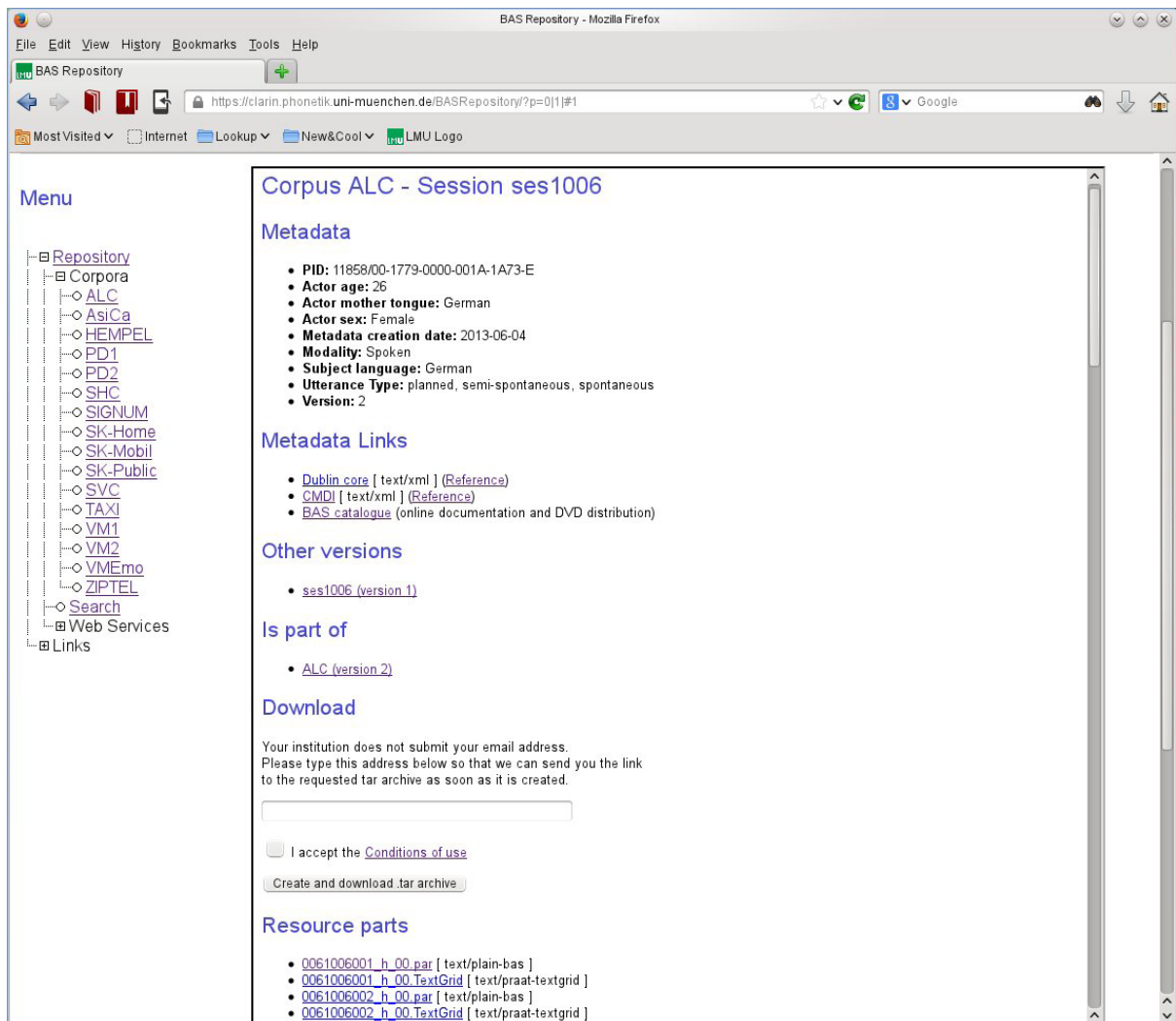
Dem Repository liegt ein Filesystem zugrunde mit einem frei zugänglichen und einem geschützten Bereich. Frei zugänglich sind die Startseiten



The screenshot shows a web browser window displaying the BAS Repository page. The URL in the address bar is <https://clarin.phonetik.uni-muenchen.de/BASRepository/?p=0#f1>. The page features the CLARIN CENTRE B logo and a navigation menu on the left. The main content area displays the following information:

- Menu:** Repository, Corpora, Search, Web Services, Links.
- PID:** 11858/00-1779-0000-0006-BF00-E
- Metadata:**
 - [CMDI \[text/xml \]](#) (Reference, Usage)
- Corpora:**
 - [ALC](#)
 - [AsiCa](#)
 - [HEMPEL](#)
 - [PD1](#)
 - [PD2](#)
 - [SHC](#)
 - [SIGNUM](#)
 - [SK-Home](#)
 - [SK-Mobil](#)
 - [SK-Public](#)
 - [SVC](#)
 - [TAXI](#)
 - [VM1](#)
 - [VM2](#)
 - [VMEmo](#)
 - [ZIPTTEL](#)
- Repository Search:**
 - [Search Form](#)
- Web Services:**
 - [WEBMAus \[text/html \]](#), [CMDI \[text/xml \]](#)

Screenshot: Repository



The screenshot shows a web browser window titled "BAS Repository - Mozilla Firefox" with the URL <https://clarin.phonetik.uni-muenchen.de/BASRepository/?p=0|1|1#1>. The page content is as follows:

Menu

- Repository
 - Corpora
 - ALC
 - AsiCa
 - HEMPEL
 - PD1
 - PD2
 - SHC
 - SIGNUM
 - SK-Home
 - SK-Mobil
 - SK-Public
 - SVC
 - TAXI
 - VM1
 - VM2
 - VMEmo
 - ZIPTTEL
 - Search
 - Web Services
 - Links

Corpus ALC - Session ses1006

Metadata

- **PID:** 11858/00-1779-0000-001A-1A73-E
- **Actor age:** 26
- **Actor mother tongue:** German
- **Actor sex:** Female
- **Metadata creation date:** 2013-06-04
- **Modality:** Spoken
- **Subject language:** German
- **Utterance Type:** planned, semi-spontaneous, spontaneous
- **Version:** 2

Metadata Links

- [Dublin core](#) [text/xml] (Reference)
- [CMDI](#) [text/xml] (Reference)
- [BAS catalogue](#) (online documentation and DVD distribution)

Other versions

- [ses1006 \(version 1\)](#)

Is part of

- [ALC \(version 2\)](#)

Download

Your institution does not submit your email address.
Please type this address below so that we can send you the link to the requested tar archive as soon as it is created.

I accept the [Conditions of use](#)

Resource parts

- [0061006001_h_00_par](#) [text/plain-bas]
- [0061006001_h_00_TextGrid](#) [text/praat-textgrid]
- [0061006002_h_00_par](#) [text/plain-bas]
- [0061006002_h_00_TextGrid](#) [text/praat-textgrid]

Screenshot: ALC

sowie die Metadaten der Repository-Objekte. Im geschützten Bereich befinden sich die Ressourcen, also die Signaldateien und Annotationen.

Die Objekte des BAS-Repositories sind Versionen von Korpora und Aufnahmesessions, wobei ein Korpus aus einer oder mehreren Sessions besteht. Jedes dieser Objekte wird durch ein eigenes CMDI-File beschrieben und ist über eine eigene Startseite zugänglich, die dynamisch aus dem CMDI-File erzeugt wird. Die Metadaten können über eine OAI-

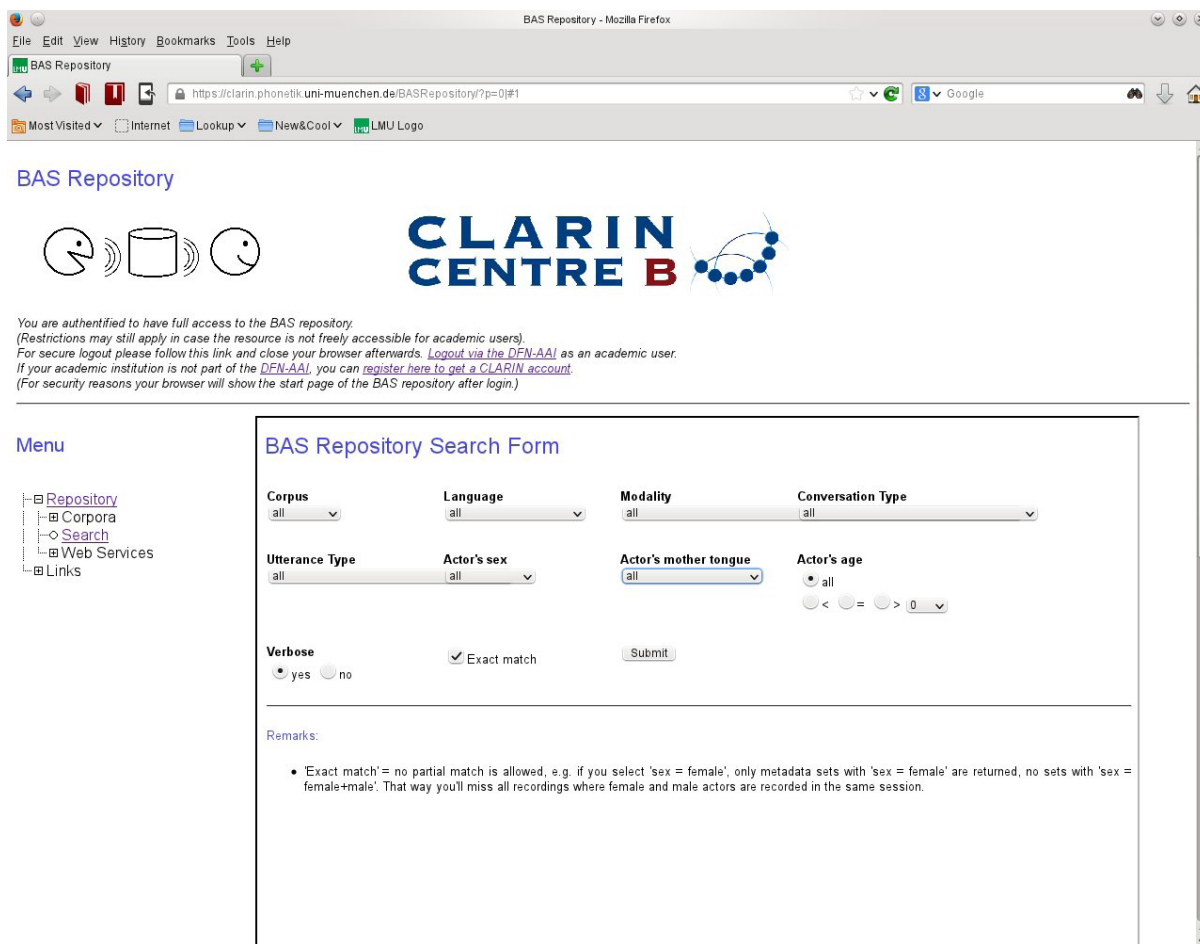
PMH-Schnittstelle geharvestet werden: <http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl?verb=Identify>

Zugang

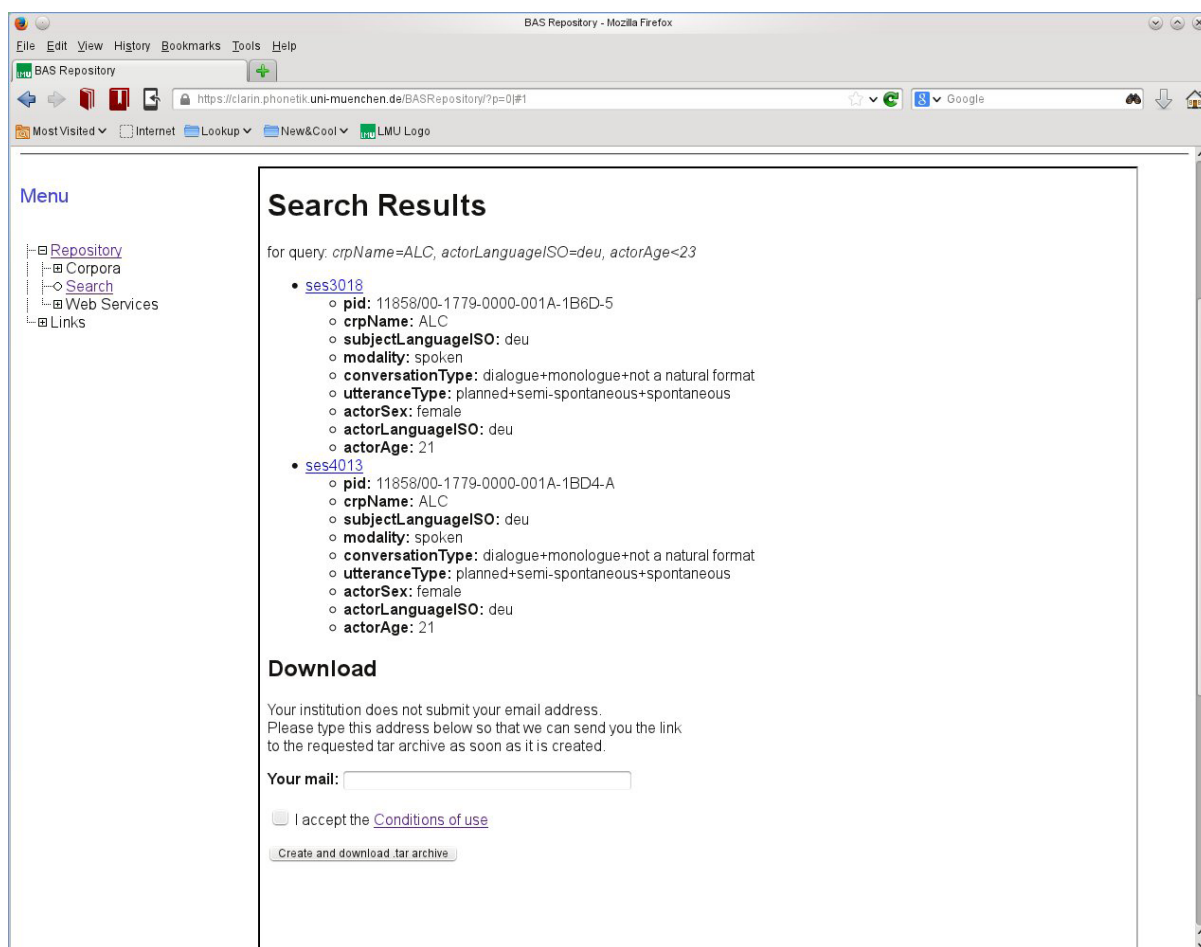
Jedem Repository-Objekt ist ein EPIC Handle Persistent Identifier (PID) zugeordnet, über den es dauerhaft zugänglich ist, wie beispielsweise die Startseite der Session 1006 des Korpus ALC: <http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3>

Neben einer Kurzbeschreibung findet sich hier auch ein Link zu den kompletten Metadaten. Diese lassen sich für eine automatisierte Verarbeitung auch über zwei direkte Wege abrufen: Zum einen ist die Verwendung eines Part-Identifiers @format=cmdi möglich: <http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@format=cmdi> Über diesen Weg können die Metadaten auch im kompakteren Dublin-Core-Format mittels @format=dc angezeigt werden. Ein weiterer direkter Zugriff auf Metadaten wird mittels *Content negotiation* ermöglicht, indem client-seitig der Accept Header auf application/x-cmdi+xml gesetzt wird.

Im Gegensatz zu den frei zugänglichen Startseiten und Metadaten liegen die primären Ressourcen in einem Shibboleth-geschützten Bereich, der erst nach entsprechender Autorisierung zugänglich ist. Den Nutzern werden drei Möglichkeiten geboten, sich zu authentifizieren. Alle akademischen Nutzer, deren Institution Teil des DFN-AAI-Netzwerks ist, können sich über die DFN-AAI authentifizieren. Akademische Nutzer anderer Institute beantragen eine Clarin-Kennung und authentifizieren sich über diese. Nicht-akademische Kunden können sich nach Zulegung einer Clarin-Kennung selektiv Rechte für Einzelkorpora erwerben.



Screenshot: Suchmaske



Screenshot: Suchergebnisse

Nach erfolgter Autorisierung werden die Links zu den Ressourcen auf der Startseite angezeigt sowie ein direkter Zugriff über den Part-Identifizierer `@partId` ermöglicht, dessen Werte den Resource-Proxy-Ids in den CMDI-Files entsprechen. Beispiel:

http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@partId=m_0000000001

Zudem erhält der autorisierte Nutzer die Möglichkeit, das entsprechende Re-

pository-Objekt als komprimiertes tar-Archiv herunterzuladen.

Suchmaske

Über eine Suchmaske kann der Nutzer korpusübergreifend Aufnahmesessions für spezielle Forschungsfragen zusammenstellen, wie beispielsweise im Hinblick auf Geschlecht oder Muttersprache der Sprecher.

Nach erfolgter Autorisierung lässt sich die gewünschte Auswahl als komprimiertes tar-Archiv herunterzuladen.

Aufnahme neuer Daten

Die Aufnahme eines neuen Korpus in das BAS-Repository verläuft vollautomatisch in folgenden Schritten:

1. Die CMDI-Files werden validiert, eingelesen und mit dem Inhalt einer Repository-Content-Tabelle abgeglichen, um festzustellen, ob es sich um neue Daten oder ein Update bereits gespeicherter Daten handelt.
2. Für alle neuen, beziehungsweise aktualisierten Sessions sowie das Corpus werden Persistent Identifier beantragt. Jede Version eines Corpus und einer Session erhält somit einen eigenen Identifier.
3. CMDI-Files werden in den frei zugänglichen Bereich kopiert und angepasst. Die Ressourcen werden in den geschützten Bereich kopiert. Für regelmäßige Konsistenzprüfungen und für die Versionierung werden Checksums ermittelt.
4. Abschließend erfolgt eine Aktualisierung der Suchdatenbank sowie der an der OAI-PMH-Schnittstelle gespeicherten Daten.

Auch BAS-externe Korpora können auf diese Weise gehostet werden. Alle Daten werden regelmäßig durch das Leibniz-Rechenzentrum in Garching durch Backups gesichert.

Software

Für das BAS-Repository wurde eine proprietäre Softwarelösung in Perl und PHP entwickelt. Voraussetzungen sind ein CGI- und PHP-fähiger Server, SQLite als Backend der Suchmaske sowie frei erhältliche Tools zur XML-Validierung, Metadaten-Transformation und Checksum-Ermittlung. Für die OAI-PMH-Schnittstelle wurde der frei erhältliche OAI-PMH2 XMLFile Datenprovider angepasst.



Uwe Reichel
Institut für
Phonetik und
Sprachverarbeitung,
LMU
München

Bericht zum zweiten Disseminationsworkshop der CLARIN-D F-AGs

Am Montag, dem 30.09.2013 fand in Leipzig der zweite Disseminationsworkshop der Facharbeitsgruppen statt. Ziel dieser jährlich stattfindenden Workshops ist es, über Fortschritte und Aktivitäten in CLARIN-D auf allen für die Fachcommunities relevanten Gebieten zu berichten. Auf diese Weise soll mit den Anwendern in Kontakt getreten sowie den Fachcommunities die Möglichkeit für wichtige Rückmeldungen an CLARIN-D gegeben werden.

Im Rahmen der ersten Präsentationen wurden zunächst einige der CLARIN-D-Showcases vorgestellt und zum aktuellen Stand der Arbeiten an der CLARIN-D-Infrastruktur berichtet. Dieter van Uytvanck gab zunächst einen allgemeinen Überblick zu Fortschritten bezüglich der Spezifikation, Implementierung und Weiterentwicklung wichtiger Eckpfeiler der Infrastruktur. Darauf folgte die Präsentation des Showcase „*Globale Elemente der CLARIN-D Infrastruktur*“, welcher die gezielte Suche und Auswahl von Ressourcen anhand ihrer Metadaten

über das Virtual Language Observatory [1], die inhaltliche Suche über die CLARIN-D-*Federated Content Search* [2] und die Weiterverarbeitung mit WebLicht [3] zum Gegenstand hat. Abgerundet wurde dieser Themenkreis durch die Showcases „*WebMaus – automatisches Segmentieren und Etikettieren über das Web*“ und „*TEI-Integrator*“ (Integration von TEI-Kollektionen in CLARIN).

In der folgenden Präsentation der F-AG 2 [4] und 4 [5] (Prof. Mair, Dr. Rücker) mit dem Titel „*Verstetigung – Oder wie verankern wir die Infrastrukturproblematik nachhaltig in den Fachcommunities?*“ wurde das für CLARIN-D wichtige Thema der Nachhaltigkeit besprochen. Insbesondere im Rahmen der abschließenden, engagiert geführten Diskussion dieses Themas wurde mehrfach deutlich, welchen hohen Stellenwert die Bemühungen um eine nachhaltige Verankerung des Infrastrukturgedankens im Allgemeinen und der Angebote von CLARIN-D im Besonderen in der Zukunft spielen müssen.

[1] <http://www.clarin.eu/vlo>

[2] <http://www.clarin.eu/node/3449> und <http://weblicht.sfs.uni-tuebingen.de/>

[3] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

[4] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-2-andere-philologien.html>

[5] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-4-altertumswissenschaften.html>

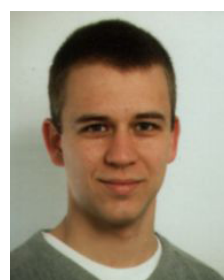
Anschließend fand eine Unterteilung des Workshops in zwei verschiedene thematische Schwerpunkte statt: „Anwendungsszenarien“ und „Dokumentation, Metadaten und Community Best Practices“. Im Bereich der Anwendungsszenarien wurden bereits in CLARIN-D integrierte Ressourcen vorgestellt. Hierzu zählen beispielsweise die Webannotationsplattform *WebAnno* [6], das *OpenScience-Portal* [7] oder die Ressourcen des *GeWiss-Projektes* [8]. Auch das Kurationsprojekt *„Erschließung digitaler Textarchive über Metadaten und Lemmata“* [9] wurde in diesem Rahmen vorgestellt.

Der zweite Themenschwerpunkt, *„Dokumentation, Metadaten und Community Best Practices“*, sollte dazu dienen, den aktuellen Stand und die Notwendigkeit weiterer Arbeiten in CLARIN-D auf diesem Gebiet zu diskutieren. Hierzu zählten Präsentationen zu generellen Themen wie die *„Dokumentation von Ressourcen und ihrer Anwendungsmöglichkeiten“* (Prof. Gloning, F-AG 1 [10]) oder *„Standards zur inhaltlichen Bewertung & nicht-technische Metadaten“* (Axel Herold, Arbeitspaket 5 [11]), aber auch spezifische Fragestellungen wie *„Guide-*

lines und Best Practices zur linguistischen Annotation von Nichtstandardvarietäten“ (Prof. Lüdeling, Marc Reznicek) [12] oder die adäquate *„Beschreibung von multimodalen Korpora“* (Farina Freigang, F-AG 6 [13]).

Der Workshop wurde durch eine Poster- und Demosession abgeschlossen. Hier konnten die in den Vorträgen besprochenen Themen vertieft, weitere wichtige Aspekte diskutiert und persönliche Kontakte geknüpft werden.

Für die engagierte Beteiligung, die gehaltenen Vorträge und die aufschlussreichen Anregungen und Diskussionen möchten wir uns an dieser Stelle noch einmal recht herzlich bei allen Teilnehmern bedanken. Die Folien aller Präsentationen stehen inzwischen auf der Webseite [14] zum Workshop zur Verfügung.



Volker Boehlke
Institut für Informatik,
Universität Leipzig

[6] <http://code.google.com/p/webanno/>

[7] <https://openscience.uni-leipzig.de/>

[8] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-2.html>

[9] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-2-andere-philologien/kurationsprojekt-2.html>

[10] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie.html>

[11] <http://de.clarin.eu/de/home/arbeitspakete/ap-5-dienste-und-ressourcen.html>

[12] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-7-computerlinguistik/kurationsprojekt-2.html>

[13] <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-6-sprache-und-andere-modalitaeten.html>

[14] <http://clarin2013.informatik.uni-leipzig.de/>

Kurz vor dem Ziel: Das Kurationsprojekt 1 der F-AG1 „Deutsche Philologie“

Das Kurationsprojekt „Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur“ geht in die letzte Phase.

Seit September 2012 laufen die Arbeiten an dem Kurationsprojekt, das im Deutschen Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) koordiniert wird. Neben der BBAW sind die Herzog Au-

gust Bibliothek Wolfenbüttel (HAB), die Justus-Liebig-Universität Gießen (JLU) und das Institut für Deutsche Sprache (IDS) noch bis Ende Februar 2014 gemeinsam an dem Projekt beteiligt.

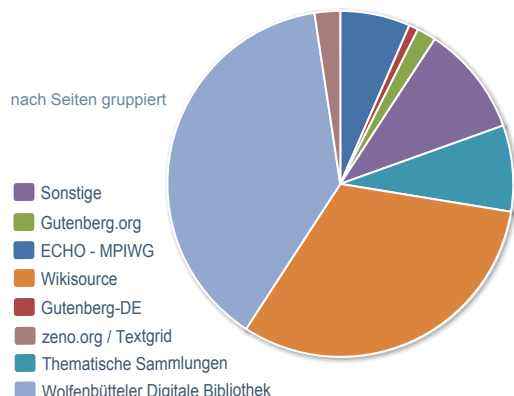
Ziel des Kurationsprojekts ist die Integration von Textressourcen des Frühneuhochdeutschen und des späteren Neuhochdeutschen (1400–1900) in die Korpora des DTA, der HAB bzw. des IDS und sukzessive in die CLARIN-D-Infrastruktur. Zu diesem Zwecke wer-



Mitarbeiter des Kurationsprojekts beim M24-Workshop im Juni 2013 in Nimwegen. Von links: Thomas Gloning (Leiter der F-AG1), Jurgita Barauskaite, Stefanie Seim, Frederike Neuber, Frank Wiegand.

Stand der Integration

Werke	433	Tokens	15 882 180
Seiten	61 214	Zeichen	102 027 903



Verteilung der Quellen im Kurationsprojekt nach Seiten gruppiert (Stand: 30.10.2013, Quelle: http://www.deutschestextarchiv.de/doku/clarin_kupro_texte).

den im Internet vorhandene oder lokal gespeicherte digitale Volltexte identifiziert und kriteriengestützt ausgewählt. Anschließend werden die Texte schrittweise aufbereitet bzw. aufgewertet, was ihre Konvertierung in das XML/TEI P5-konforme Basisformat des Deutschen Textarchivs (DTABf) [1] – das CLARIN Best Practice Format zur Auszeichnung historischer Drucke – sowie die Zuordnung der entsprechenden Imagedigitalisate beinhaltet. In einem weiteren Schritt werden die Metadaten der Ressourcen aufgenommen. Neben bibliografischen Metadaten, werden Angaben zu Erfassungsprinzipien und Nutzungsbedingungen sowie forschungsorientier-

te Metadaten verzeichnet. Um dabei Einheitlichkeit und Vollständigkeit zu gewährleisten, wurde im Laufe des Kurationsprojekts das CLARIN-D-Metadatenformular [2] entwickelt, womit aus einem benutzerfreundlichen Webformular ein DTABf-kompatibler TEI-Header generiert werden kann.

Bei anschließender Überführung in die Qualitätssicherungsumgebung des Deutschen Textarchivs DTAQ [3] werden die Texte durch CAB (Cascaded Analysis Broker) [4] linguistisch analysiert. Des Weiteren stehen die Textressourcen in DTAQ, neben anderen Formaten (XML/TEI, HTML, Reintext) in den CLARIN-Formaten TCF und CMDI, zum Download zur Verfügung. Eine OAI-PMH-Schnittstelle sichert die Indexierung der Daten für die Federated-Content-Search sowie ihre Aufnahme in das Virtual Language Observatory.

Kollaborative Kuration

Sobald sich die Texte in der Qualitätssicherungsumgebung DTAQ befinden, beginnt die eigentliche „Kuration“: Nacharbeit und Pflege an den Ressourcen. Nach dem Crowdsourcing-Prinzip werden die Texte von der Community der Plattform Korrektur gelesen (auf Transkriptions- und Auszeichnungsebene) und ggf. überarbeitet. Die Idee des kollaborativen Arbeitens hat sich im

[1] DTA-Basisformat: www.deutschestextarchiv.de/doku/basisformat [alle URLs in diesem Beitrag abgerufen am 30.10.2013].

[2] CLARIN-D Metadatenformular: www.deutschestextarchiv.de/dtae/submit/clarin.

[3] DTAQ: www.deutschestextarchiv.de/dtaq.

[4] CAB: www.deutschestextarchiv.de/doku/software#cab.

Projektzeitraum bewährt und zur Maximierung der Ergebnisse beigetragen. Zum gegenwärtigen Zeitpunkt wurden bereits mehr als 60.000 Seiten aus 433 Texten aufbereitet und in die CLARIN-D-Infrastruktur integriert, die über die projekteigene und kürzlich freigeschaltete Website [5] zugänglich sind.

Quellen: Sammlungen, Editionen, Einzelfunde

Im Wesentlichen wurden Textressourcen aus drei Typen von Quellen integriert: Große Textsammlungen, digitale Editions- und Forschungsprojekte sowie Einzelfunde aus Initiativen von Privatpersonen.

Im ersten Projektabschnitt lag der Fokus der Integration vor allem auf großen Volltextsammlungen wie der freien Quellensammlung Wikisource, aus der 148 Texte [6] übernommen wurden. Momentan noch größtenteils über die Qualitätssicherungsumgebung DTAQ verfügbar, werden die Wikisource-Texte demnächst auch über die Webseite des Deutschen Textarchivs [7] erreichbar sein. Mit der Wikisource-Community konnte der Kontakt intensiviert werden, als das

Kurationsprojekt im Oktober in Leipzig auf dem 1. Wikisource-Treffen vertreten war und seine Arbeit vorstellte. Ein weiterer Austausch sowohl von Textressourcen als auch von wissenschaftlicher und technischer Expertise ist bereits in Planung. Eine erfolgreiche Kooperation mit einem kleineren und themenspezifischerem Projekt kam mit den Initiatoren des „Gutzkow Editionsprojekts“ [8], der kommentierten digitalen Gesamtausgabe der Werke und Briefe Karl Gutzkows, zustande. Bislang konnten vier Bände der ursprünglich in HTML ausgezeichneten Gesamtausgabe durch das Kurationsprojekt eine Überführung nach XML (DTABf) erfahren und werden zukünftig so auch in das Portal rücküberführt werden. Die Übernahme weiterer Bände ist bereits in Planung. Ein Beispiel für die gelungene Integration eines einzelnen Textes aus einer privaten Initiative ist die Übernahme der drei Bände des Vergleichenden Handbuchs der Symbolik der Freimaurerei [9] von Joseph Schauberg aus dem Portal „Internetloge.de“.

Das Kurationsprojekt ist dem Anspruch von CLARIN-D, außerhalb der Fach-Community praktische Hilfestellung im Umgang mit linguistischen Ressourcen zu leisten, gerecht geworden.

[5] Fortlaufend aktualisierte Informationen zum Kurationsprojekt sowie die integrierten Textressourcen finden Sie unter www.deutschestextarchiv.de/doku/clarin_kupro_index.

[6] Wikisource-Texte im Kurationsprojekt:

http://www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=wikisource

[7] DTA: www.deutschestextarchiv.de.

[8] „Gutzkow Editionsprojekt“: <http://projects.exeter.ac.uk/gutzkow/Gutzneu/edition/index.htm>

[9] Schauberg, Joseph: Vergleichendes Handbuch der Symbolik der Freimaurerei, Bd. 1. Schaffhausen, 1861.

Im DTA: http://www.deutschestextarchiv.de/book/show/schauberg_freimaurerei01_1861 ;Vorlage der Internetloge: <http://www.internetloge.de/symhandb/symb.htm>.

Durch zahlreiche Kooperationen konnten qualitativ hochwertige Volltexte aus Projekten, von Wissenschaftlern und auch von Privatpersonen ohne wissenschaftliche oder technische Expertise in ein standardisiertes Format (XML/TEI-DTABf) überführt und ggf. auch in diesem aufgewerteten Format wieder an die Urheber zurückgegeben werden. Des Weiteren kann durch die Überfüh-

rung in die CLARIN-D-Infrastruktur die langfristige Archivierung und Verfügbarkeit der Daten garantiert werden; eine Anforderung, die kleinere Projekte und Privatpersonen oft nicht alleine bewältigen können.

Frederike Neuber, Christian Thomas
Berlin-Brandenburgische Akademie der Wissenschaften

Alles Weitere unter:

www.clarin-d.org

Some things I learnt at Digital Humanities 2013

This past week, I attended the [Digital Humanities 2013](#). [Digital Humanities \(DH\)](#) is an annual conference organized by the [Alliance of Digital Humanities Organizations](#). DH 2013 was held on July 16-19 at the University of Nebraska in Lincoln. Local organization was provided by the internationally renowned [Center for Digital Research in the Humanities](#).

Following the highly successful [DH 2012 conference](#) held at the University of Hamburg, DH 2013 once again attracted a high number of participants from around the world. Without a doubt, [Digital Humanities](#) has become the premier international conference for reporting on cutting-edge Digital [Humanities](#) research and for providing a comprehensive overview of the field. This year's conference program featured on-going trends in Digital [Humanities](#) research, including the increasing use of mobile devices as tools for data collection and for providing easy access to DH data and tools; the importance of data visualization for Digital [Humanities](#) research; the challenges by Big Data; the opportunities and limitations of crowdsourcing techniques.

History, Linguistics, and Literary Studies were the three humanities disciplines

that stood out in terms of the number of papers on the DH 2013 program. In order to be able to accommodate all long and short papers accepted for presentation at DH 2013, the conference program was organized in multiple, parallel sessions. I can therefore only comment on those papers that I was able to attend personally.

Among the paper presentations that I attended in the field of history, I found the following particularly interesting: The paper on *ChartEx: a project to extract information from the content of medieval charters and create a virtual workbench for historians to work with this information* discussed how tools for language technology can play an important role in the [workflow](#) of historians. The paper *Opening Aladdin's cave or Pandora's box? The challenges of crowdsourcing the Medici Archives* provided a very useful and convincing case study of the potential and the limitations of crowdsourcing methods in the construction of a high-quality archive of historical manuscripts. The paper on *Slave Biographies: Atlantic Database Network* showed the added value of powerful data visualization techniques for the analysis of complex and geographically distributed historical data. *Visualizing Centuries: Data Visualization and the Comédie-Française*

Registers Project was another paper that highlighted the importance of data visualization – in this case for historical data in the area of theatre studies. Finally, the panel session *Center for Historical Information and Analysis: Big Data in History* provided a highly informative overview of an ambitious, NSF-funded infrastructure project in the field of history.

CLARIN researchers contributed an impressive number of papers to the main program of the DH 2013 conference. The member countries of CLARIN [ERIC](#) were largely responsible for making language resources and tools for the annotation and visualization of language data a highly visible topic at DH 2013. In their paper entitled *The German Language of the Year 1933. Building a Diachronic Text Corpus for Historical German Language Studies*, Hanno Biber and Evelyn Breiteneder, colleagues from CLARIN-AUT, reported on their on-going efforts to extend their National Academy Corpus. Douwe Zeldenrust and Marc van Oostendorp, researchers at the CLARIN-NL Center at the Meertens Institute in Amsterdam, described the integration of the resource *Speaking Map of the Netherlands* into the CLARIN infrastructure in their paper *Combining tailor made research solutions with big infrastructures*. Researchers from the CLARIN Center at the Max-Planck Institute for Psycholinguistics in Nijmegen presented on-going work on their tools LEXUS and ELAN, which are much in demand world-wide. Shakila Shayan, Andre Moreira, Menzo Windhouer, Alexander König, and Sebastian Drude reported on

LEXUS 3 - a collaborative environment for multimedia lexica. Han Sloetjes, Aarthi Somasundaram, Sebastian Drude, Herman Stehouwer, and Kees Jan van de Looij presented a co-authored paper entitled *Expanding and connecting the annotation tool ELAN*. A second paper about ELAN on the topic of *Automatic annotation of linguistic 2D and Kinect recordings with the Media Query Language for Elan* was contributed by Anna Lenkiewicz, and Sebastian Drude. The CLARIN-D Centers at the Universities of Leipzig and Tübingen reported on their research on combining language technology with data visualizations. Marc Küster (University of Leipzig) presented *Agents for Actors: A Digital [Humanities](#) framework for distributed microservices for text linking and visualization*. I presented joint research at the CLARIN-D Center Tübingen with Thomas Zastrow, Marie Hinrichs, and Kathrin Beck in our paper *Scientific Visualization for the Digital [Humanities](#) as CLARIN-D Web Applications*.

[CLARIN member](#) institutions also participated in the DH 2013 poster session. Bastian Entrup, Maja Bärenfänger, Frank Binder, and Henning Lobin from the University of Giessen, another [CLARIN member](#) institution from Germany, presented a poster on *Introducing GeoBib: An Annotated and Geo-referenced Online Bibliography of Early German and Polish Holocaust and Camp Literature (1933–1949)*. Dana Dannélls, Lars Borin, and Leif-Jöran Olssen from the University of Gothenburg had a poster on *MapServer for Swedish Language Technology*.

Sometimes you have to travel to far places to be able to fully appreciate the richness of the humanities research that CLARIN members are involved in. A case in point is the collaborative research project of the poets Katharine Coles and Julie Lein at the University of Utah with a team of computer scientists and corpus linguists at Oxford University, which includes Martin Wynne, co-director of CLARIN [ERIC](#). Their highly innovative approach to visualizing poetry was presented in their co-authored paper *Freedom and Flow: A New Approach to Visualizing Poetry*. Katharine and Julie also spoke very eloquently about the challenges and fruits of their interdisciplinary interactions in their presentation *Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research*. This very captivating presentation made me aware of the wonderful poetry of Louise Bogan, an American poet of the 20th century.

My apologies to any CLARIN colleagues who also presented their work at DH 2013 and whom I may have inadvertently left out in this piece. For more information about the above papers, but also about all papers, posters, panel presentations, and workshops offered at DH 2013 please consult: <http://dh2013.unl.edu/abstracts/>.

Digital [Humanities](#) research on language data was also very much present in other ways at DH 2013. The current issue (Vol. 28.1; April 2013) of LLC, the Journal of the Alliance of Digital [Humanities](#) Organizations, which was on display at the conference, is a special issue, guest-edited by John Nerbonne and William

A. Kretzschmar Jr, on the topic of ‘Dialectometry ++’. It contains a collection of papers on the application of computational techniques to the study of language variation. William Kretzschmar, who was present at the conference, presented as co-author a paper entitled *Simulation of the Complex System of Cultural Interaction* that nicely demonstrated the predictive value of such dialectometric methods.

Three keynote addresses brought together all conference participants in plenary sessions and provided special highlights throughout the conference. David S. Ferriero, Archivist of the United States, opened the conference with his inspiring keynote address *Harnessing the Wisdom of the Crowd: The Citizen Archivist Program at the National Archives*. Isabel Galina, Researcher at the Instituto de Investigaciones Bibliográficas at the National University of Mexico (UNAM), closed the scientific program of the conference with her thoughtful and forward-looking address *Is there anybody out there? Building a global DH community*.

The main conference program was preceded by two days of tutorials and workshops. I greatly enjoyed attending Mia Ridge’s tutorial on crowdsourcing. Thanks to this very informative and nicely presented tutorial, I am now aware of digital humanities projects such as *Digital Harlem* (<http://acl.arts.usyd.edu.au/harlem/>) and *Old Weather* (oldweather.org/), where crowdsourcing methods are an integral part of the project.

The Digital [Humanities](#) conferences also provide a forum to honor outstand-

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.

ing scholars for their achievements in the field. The Roberto Busa Award is named in honor of Father Busa and given in recognition of outstanding lifetime achievements in the application of information and communication technologies to humanistic research. At DH 2013 the award was presented to Willard McCarthy, Professor of [Humanities Computing](#) in the Department of Digital [Humanities](#) at King's College London and Fellow of the Royal Anthropological Institute (London). In his acceptance address *Getting there from here: Remembering the future of digital humanities*, Willard McCarthy reflected on the origins and the early days of what is now referred to as Digital [Humanities](#).

Finally, it is time to look toward the future. The plans for the next Digital [Humanities](#) conferences have already been made. [DH 2014](#) (<http://dh2014.org/>) will be hosted by the University of Lausanne on July 6-12, 2014. And for 2015, the conference will move to the University

of Sydney, Australia. I can only recommend attending these conferences and urge you to submit your Digital [Humanities](#) research to them. The *Digital Humanities* conference has a special tradition of entertaining participants with an extensive social program, including a welcome reception, a Busa Award reception, a special dinner for first-time participants, the DH fun run, and a banquet.



Erhard Hinrichs
Seminar für Sprachwissenschaft
Universität Tübingen

Abkürzungsverzeichnis (NELCA)

AAI	Authentication and Authorization Infrastructure
ABaC:us	Austrian Baroque Corpus
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
CiNaViz	City Name Visualization
CMDI	Component MetaData Infrastructure
CLARIN	Common Language Resources and Technology Infrastructure
DAI	Deutsches Archäologisches Institut
DAITF	Data Access and Interoperability Task Force
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DDDS	DeutschDiachronDigital-Tagset
DH	Digital Humanities
DSA	Data Seal of Approval
DTA	Deutsches Textarchiv
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
FCS	Federated Content Search
FI-Initiative	Forschungs-Infrastruktur-Initiative
FOLK	Forschungs- und Lehrkorpus gesprochenes Deutsch
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
GeWiss	Gesprochene Wissenschaftssprache kontrastiv
GIS	Geographisches Informationssystem
HAB	Herzog August Bibliothek Wolfenbüttel
HPC	High Performance Cluster
HZSK	Hamburger Zentrum für Sprachkorpora

IAIS	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme
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	<i>Never-Ending-List</i> der CLARIN-Abkürzungen
NMMoCap-Korpus	Natural Media Motion Capture-Korpus
NSF	National Science Foundation (USA)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PID	Persistent Identifier
SaGA	Speech and Gesture Alignment Corpus
SHARE	Survey of Health, Ageing and Retirement in Europe
SuUB Bremen	Staats- und Universitätsbibliothek Bremen
STTS	Stuttgart-Tübingen Tagsets
TCF	Text Corpus Format
TeLeMaCo	Teaching and Learning Materials Collection
TLA	The Language Archive
TüNDRA	Tübingen aNnotated Data Retrieval Application
TüPP-D/Z	Tübinger Partiiell Geparstes Korpus des Deutschen/Zeitungskorpus
TUSTEP	Tübinger System von Textverarbeitungs-Programmen
VLC	Virtual Linguistic Campus
VLO	Virtual Language Observatory
WADL	Web Application Description Language