

Editorial

Vierter CLARIN-D- Newsletter

Draußen gehen die Eisheiligen nahtlos in die Schafskälte über, bevor dann die Novembernebel das Erscheinen des nächsten Newsletters verkünden werden ...

Eine frühes Ergebnis von CLARIN-D ist die *Federated Content Search*, eine verteilte web-basierte Textsuche in verschiedenen Ressourcen in den CLARIN-D-Zentren (weblicht.sfs.uni-tuebingen.de/Aggregator). Was liefert diese Suche nun zu den obigen Wetterbegriffen? ‚Schafskälte‘ und ‚Eisheilige‘ sind den Zentren unbekannt, aber zu ‚Novembernebel‘ haben sowohl das IDS als auch die Uni Tübingen passende Datensätze in ihren Korpora! Und dass es einmal ‚Sommerstage‘ gegeben haben muss, davon zeugen viele frühe Dokumente in den Archiven der Zentren ...

Zu diesem Newsletter: in jeder Ausgabe stellt sich ein CLARIN-D-Zentrum vor – dieses Mal ist es das Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart, das umfangreiche Korpora und Tools anbietet, den Webauftritt von CLARIN-D betreibt und betreut und das, gemeinsam mit dem Zentrum Saarbrücken CLARIN-D-Doktorandentage angeboten und erfolgreich durchgeführt hat.

Das Thema eHumanities wird in der Forschungslandschaft zunehmend wichtiger. Eine wichtige Veranstaltung zu diesem Thema war das Kickoff-Treffen der 24 Verbundprojekte in Leipzig, auf dem CLARIN-D prominent vertreten war.

Eine Besonderheit von CLARIN-D sind die Kurationsprojekte. Diese Projekte sollen in den jeweiligen wissenschaftlichen Fachgemeinden zum einen die CLARIN-D-Infrastruktur bekanntma-

chen, zum anderen aus den Fachgemeinden Bedürfnisse und Anforderungen der Nutzer an diese Infrastruktur an CLARIN-D zurückgeben. Zwei Beiträge widmen sich dem Thema Kurationsprojekte und es zeigt sich, dass diese Projekte ihren Zweck erfüllen.

Ebenfalls in Richtung Fachgemeinden geht die 2-tägige Konferenz „Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven“, die im Februar 2013 an der BBAW stattfand und auf der das gesamte Spektrum von CLARIN-D-Diensten und -Ressourcen über die Nutzung historischer Korpora bis hin zum Aufbau neuer Ressourcen abgedeckt wurde.

Die Visualisierung von Daten steht bei zwei Anwendungen, die Tom Zastrow

aus Tübingen vorstellt im Vordergrund: CiNaViz ist ein auf Google Maps basierender und frei zugänglicher Webdienst, der die geografische Verteilung von Ortsnamen und Namensprä- und -suffixen grafisch darstellt. WhoIsInTheNews ist ebenfalls ein Webdienst, und er zeigt die Häufigkeiten und geografischen Verteilungen von in den Nachrichten erwähnten Personen, Orten und Organisationen an – beides schöne Beispiele für eine gelungene Visualisierung von Daten.



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-KP-GeWiss: Das zweite Kurations- projekt der F-AG 1 „Deutsche Philologie“

Start des Kurations- projekts CLARIN- KP-GeWiss zur Bün- delung von Sprachres- sourcen

Am 01.04.2013 nahm das Kurationsprojekt CLARIN-KP-GeWiss seine Arbeit auf. Ziel ist es, die im Projekt „Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen“ (GeWiss) aufgebauten Sprachressourcen zu bündeln und in die europäische CLARIN-Infrastruktur zu integrieren. Außerdem sollen die Nutzungsmöglichkeiten der Korpusressource erweitert werden. Das Projekt ist am Herder-Institut der Universität Leipzig angesiedelt und wird in Kooperation mit den CLARIN-Zentren am IDS Mannheim, an der Universität Leipzig und an der Universität Hamburg (HZSK) durchgeführt.

Das GeWiss-Korpus

Vor dem Hintergrund zunehmender Internationalisierung und Mobilität im akademischen Bereich gewinnt die Frage nach der adäquaten Beschreibung wissenschaftssprachlicher Kompetenz in der eigenen und fremden Wissenschaftssprache und einer entsprechenden Ausbildung immer größere Bedeutung. Dies gilt auch für den Gebrauch des Deutschen als Wissenschaftssprache. Die Forschung in diesem Bereich konzentrierte sich jedoch vor allem auf die Schriftsprache, da die Erhebung und Transkription mündlicher Sprachdaten mit hohem Zeit- und Arbeitsaufwand verbunden ist und für die Untersuchung der spezifischen Konventionen des mündlichen Gebrauchs der Wissenschaftssprache bislang keine frei verfügbaren Korpusressourcen vorlagen. Um einen ersten Schritt zur Verbesserung dieser Situation zu unternehmen, wurde im Rahmen des trinationalen Forschungsprojektes GeWiss „Gesprochene Wissenschaftssprache kontrastiv:

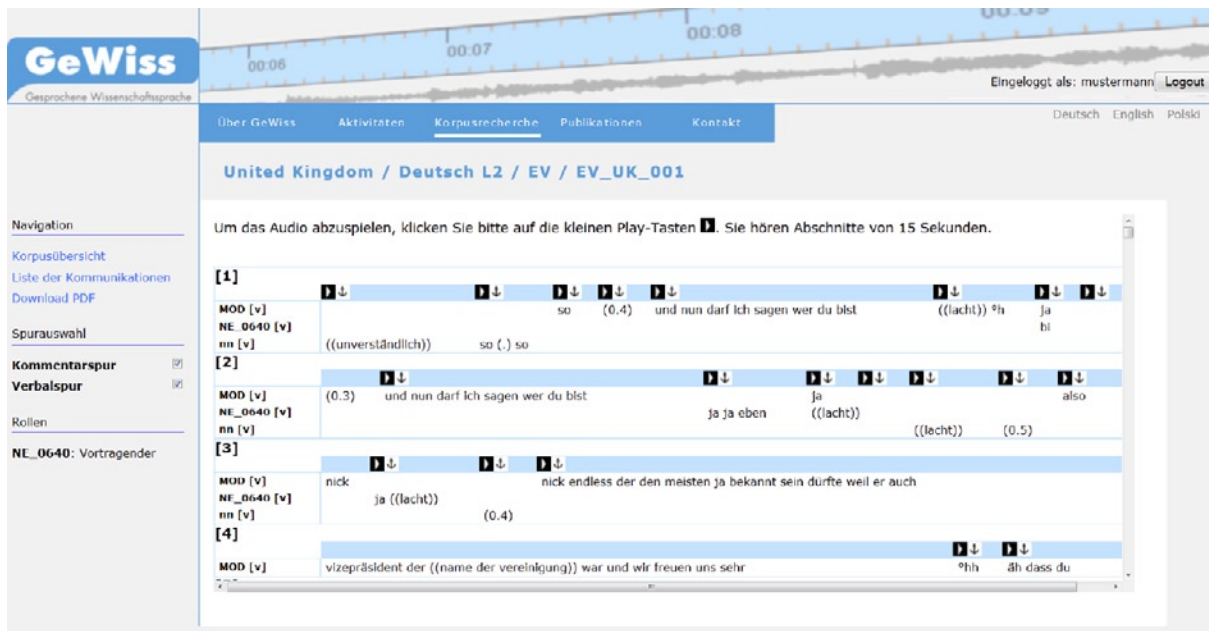


Abb. 1: Die Beta-Version des GeWiss-Korpus: Ansicht der Volltranskripte mit verknüpften Audioaufnahmen.

Deutsch im Vergleich zum Englischen und Polnischen“ von 2009 bis 2013 ein Vergleichskorpus zur gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen aufgebaut [1]. Das GeWiss-Korpus umfasst zwei zentrale Genres der mündlichen Wissenschaftskommunikation – wissenschaftliche Vorträge und Prüfungsgespräche – in den Sprachen Deutsch, Englisch und Polnisch, die in den akademischen Kontexten Deutschlands, Großbritanniens und Polens erhoben wurden. Es enthält insgesamt 126:05 h an aufgenommenen Sprachdaten bzw. 1.273.529 Token an Transkriptionen. Das Korpus umfasst 371 verschiedene Kommunikationen (58 Expertenvorträge, 89 Studentische Vorträge, 224 Prüfungsgespräche), an denen insgesamt 462 Hauptsprecher (Vortragende, Prüflinge, Prüfer und Seminarleiter) beteiligt sind.

Zum Ende der Laufzeit des Projekts konnte im März 2013 eine erste Beta-Version des GeWiss-Korpus veröffentlicht werden. Diese ermöglicht zunächst einen Zugriff auf Volltranskripte und die mit ihnen verknüpften Audioaufnahmen. Außerdem werden umfangreiche Metadaten zum Kontext der einzelnen Gesprächsereignisse und zu den Sprechern bereitgestellt. Das Korpus ist unter gewiss.uni-leipzig.de nach kostenloser Registrierung für Forschung und Lehre frei zugänglich.

Das Kurationsprojekt

Das am 1.4.2013 begonnene Kurationsprojekt wird die im Rahmen von GeWiss erarbeiteten, bereits veröffentlichten und noch unveröffentlichten Ressourcen bündeln und in CLARIN-kompatibler

[1] Das GeWiss-Projekt wurde von der VolkswagenStiftung im Rahmen der Profillinie „Deutsch plus – Wissenschaft ist mehrsprachig“ gefördert (Az.: II/83967).

Form der wissenschaftlichen Öffentlichkeit zur Verfügung stellen. CLARIN-KP-GeWiss verfolgt im Einzelnen die folgenden Ziele:

Korpusausbau

Das GeWiss-Kernkorpus bietet in zwei Dimensionen Ressourcen für die vergleichende Erforschung der gesprochenen Wissenschaftssprache Deutsch: einerseits im Hinblick auf den Gebrauch des Deutschen als fremder Wissenschaftssprache in nicht-deutschsprachigen akademischen Kontexten (exemplarisch Großbritanniens und Polens), andererseits für den Vergleich mit anderen Wissenschaftssprachen (exemplarisch Englisch und Polnisch). Im Rahmen der Kuration werden die für diese beiden Vergleichsdimensionen verfügbaren Ressourcen durch die Integration weiterer Teilkorpora verstärkt. Es soll darüber hinaus der Weg bereitet werden für das langfristige Ziel, ein Referenzkorpus der gesprochenen Wissenschaftssprache mit dem Deutschen im Zentrum zu schaffen. Hierzu werden im Projekt Dokumentationen und Workflows erarbeitet, welche für die Einbindung weiterer Ressourcen

unter gleichen Qualitätsstandards zur Verfügung stehen.

Welche zusätzlichen Teilkorpora gewinnt die Community?

Die Möglichkeiten für die Untersuchung des Gebrauchs des Deutschen als fremder Wissenschaftssprache im nicht-deutschsprachigen Raum werden durch die Integration eines im bulgarischen akademischen Kontext aufgenommenen Teilkorpus deutschsprachiger studentischer Referate (GeWiss-SV-BG) erweitert. Die Daten dieser Ressource wurden in Kooperation mit Partnern der St.-Kliment-Ohridski-Universität Sofia aufgenommen. Damit wird neben den im Kernkorpus vorhandenen Daten aus dem polnischen akademischen Kontext ein zusätzliches Teilkorpus zur Verfügung gestellt, das die Germanistik in der Region Mittel-Osteuropa repräsentiert.

Die Möglichkeiten für kontrastive Untersuchungen zur gesprochenen Wissenschaftssprache werden durch die Integration eines im italienischen akademischen Kontext aufgenommenen Teilkorpus

Weitere Infos auch im Wiki:

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

italienischsprachiger Konferenzvorträge (GeWiss-EV-IT) erweitert. Die Daten dieser Ressource wurden in Kooperation mit Partnern der Universität Pisa aufgenommen. Damit werden in der GeWiss-Ressource neben germanischen und slawischen auch romanische Sprachdaten für kontrastive Untersuchungen zur Verfügung gestellt und somit Nutzungsmöglichkeiten auch für Anwender aus der romanistischen Philologie eröffnet.

Korpusaufbereitung

Neben dem Korpusausbau ist jedoch auch die Aufbereitung der Korpusdaten im Hinblick auf die Erweiterung der Abfragemöglichkeiten wichtig. Das veröffentlichte GeWiss-Kernkorpus enthält bereits Annotationen von Sprachwechselphänomenen [2]. Da für die Erforschung der gesprochenen Wissenschaftssprache insbesondere Annotationen hinsichtlich pragmatischer Aspekte interessant sind, besteht ein weiteres langfristiges Ziel des GeWiss-Projekts in der Entwicklung von entsprechenden Typologien und der Bereitstellung diesbezüglich annotierter Ressourcen. In einem ersten Schritt hierzu wurde bereits eine Typologie zu Diskurskommentierungen erarbeitet und ein Teilkorpus der deutschsprachigen Konferenzvorträge manuell danach annotiert. Bei Diskurs-

kommentierungen handelt es sich um Äußerungen, in denen ein Sprecher vom eigentlichen thematischen Inhalt abweicht und stattdessen die Gliederung seiner Präsentation, die dem Vortrag vorangehenden bzw. folgenden Diskursphasen (Vorstellung, Diskussion) oder andere Vorträge der Sektion kommentiert [3]. Im Unterschied zu den für schriftliche wissenschaftliche Texte beschriebenen Textkommentierungen [4] weisen die Diskurskommentierungen ein deutlich weiteres Spektrum auf und sind in besonderer Weise durch Merkmale der Mündlichkeit gekennzeichnet. Sie stellen daher für die Überprüfung der vor allem an der Schriftsprache entwickelten Beschreibungskategorien der Wissenschaftssprache einen besonders interessanten Phänomenbereich dar. Das nach Diskurskommentierungen annotierte Teilkorpus GeWiss-EV-DE-Meta wird im Rahmen der Kuration ebenfalls in die Ressource eingebunden und zur Verfügung gestellt.

Bessere Sichtbarkeit und erweiterte Nutzungsmöglichkeiten

Alle beschriebenen Ressourcen werden im Rahmen des Kurationsprojekts den CLARIN-Standards entsprechend auf-

[2] Siehe hierzu: Reershemius, Gertrud/Lange, Daisy (zur Veröff. angenommen): „Sprachkontakt in der mündlichen Wissenschaftskommunikation. Zur Annotation von Sprachwechsel im GeWiss-Korpus“. In: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchronverlag.

[3] Siehe hierzu: Fandrych, Christian (in Vorb.): „Metakomentierungen in Wissenschaftlichen Vorträgen“. In: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchronverlag.

[4] Siehe hierzu: Fandrych, Christian/Graefen, Gabriele (2002): *Text commenting devices in German and English academic articles*. In: *Multilingua* 21, 17-43.

bereitet. Das bedeutet insbesondere, dass die in der EXMARaLDA Korpusmanagementsoftware Coma angelegten Metadaten in das CMDI-Format (Component MetaData Infrastructure) überführt werden. Durch Verwendung eines OAI-PMH Providers (Open Archives Initiative Protocol for Metadata Harvesting) werden die Metadaten auch von außerhalb des Repositoriums zu-

gänglich und indizierbar sein, sodass die gesamte Ressource nach Abschluss des Kurationsprojekts zusätzlich über das VLO (Virtual Language Observatory) recherchierbar sein wird. Darüberhinaus erfolgt eine Registrierung von PIDs (Persistent Identifiers) für die Teilkorpora und ihre einzelnen Bestandteile, die auf diese Weise langfristig identifizierbar und zitierbar gemacht werden.

Die folgende Übersicht fasst die Ziele von CLARIN-KP-GeWiss zusammen:

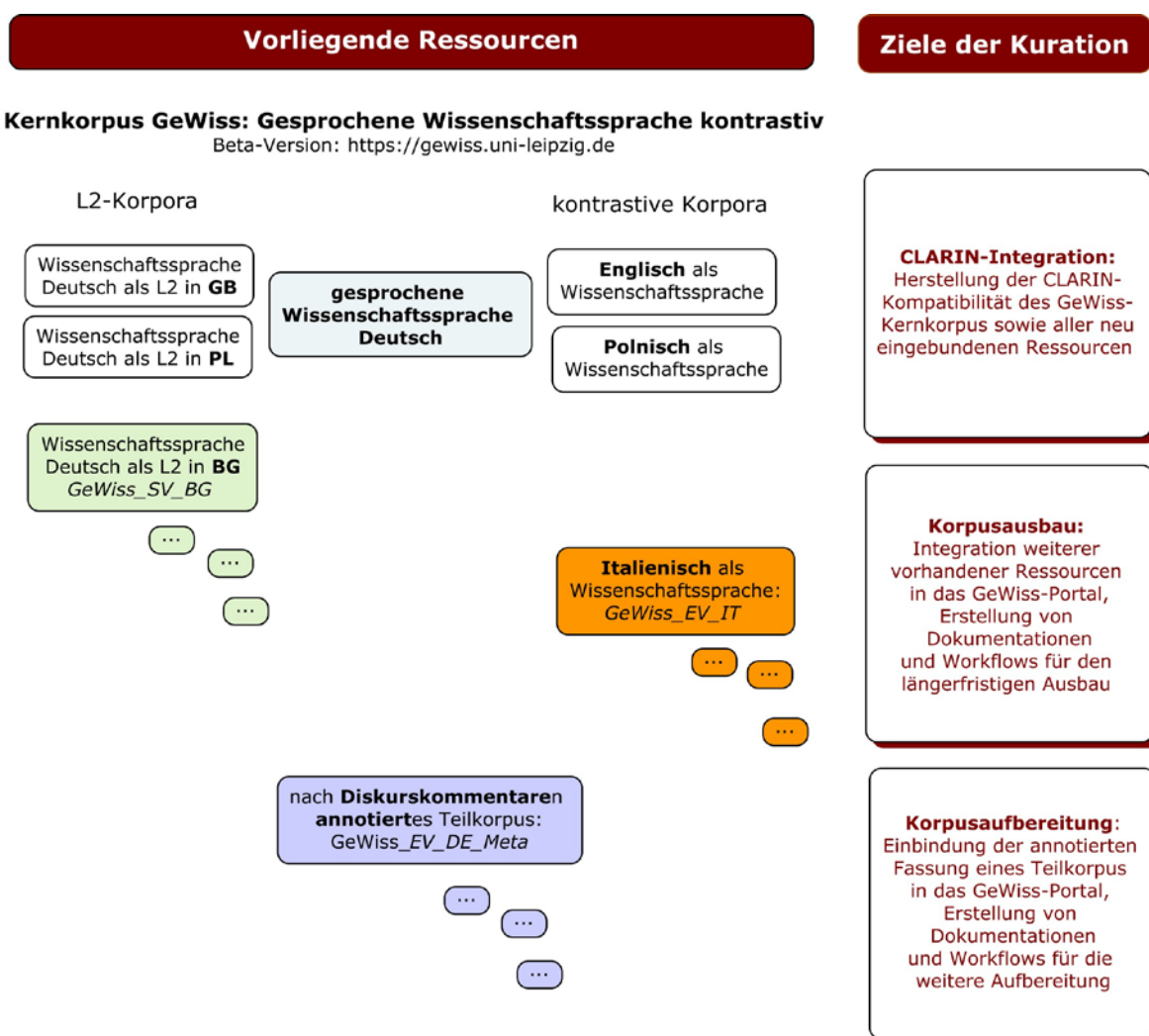


Abb. 2: Überblick über die Ziele des Kurationsprojekts CLARIN-KP-GeWiss: Die im GeWiss-Projekt aufgebauten Sprachressourcen sollen gebündelt und in die CLARIN-Infrastruktur eingebunden werden. Die Nutzungsmöglichkeiten der Korpusressource werden erweitert und es wird die Basis für den längerfristigen Ausbau und die weitere Aufbereitung gelegt.

Insgesamt möchte das Kurationsprojekt CLARIN-KP-GeWiss also zum einen eine unter dem CLARIN-Dach dauerhaft zugängliche Ressource für die vergleichende Erforschung der gesprochenen Wissenschaftssprache bereitstellen und die Nutzungsmöglichkeiten dieser Ressource weiter ausbauen und verbessern. Zum anderen will es die infra-

strukturelle Basis für den längerfristigen Ausbau und die weitere Aufbereitung der Ressourcen zu einem Referenzkorpus schaffen. Durch die bereitgestellten Dokumentationen und Workflows sollen Integrationsmöglichkeiten für weitere Datensammlungen aus diesem Forschungsbereich eröffnet werden.



Christian Fandrych
Herder-Institut,
Universität Leipzig



Daniel Jettka
HZSK,
Universität Hamburg



Cordula Meißner
Herder-Institut,
Universität Leipzig

Ein CLARIN-Zentrum stellt sich vor: das Institut für Maschinelle Sprachverarbeitung (IMS) an der Universität Stuttgart

Analysewerkzeuge aus der computerlinguistischen Forschung

Geschichte

Das IMS wurde 1987 von Christian Rohrer gegründet, der bis dahin Professor für Romanistik in Stuttgart war. Neben seinem Lehrstuhl Computerlinguistik wurde bald danach der zweite Lehrstuhl Formale Logik und Sprachphilosophie mit Hans Kamp besetzt. Heute besteht das Institut aus den Lehrstühlen Grundlagen der Computerlinguistik, Theoretische Computerlinguistik, Experimentelle Phonetik und der Juniorprofessur Computerlinguistik. Das IMS kooperierte stets eng mit dem Institut für Linguistik der Universität Stuttgart, unter anderem im Sonderforschungsbereich 340 Sprachtheoretische Grundlagen für die Computerlinguistik (1995-2000) und im laufenden Sonderforschungsbereich 732 Incremental Specification in Context (seit 2006).

CLARIN-D am IMS

Leiter des CLARIN-D-Zentrums Stuttgart ist Jonas Kuhn, der einst selbst am IMS Computerlinguistik studierte, bevor er – nach verschiedenen Stationen in den USA und Deutschland – 2010 den Lehrstuhl Grundlagen der Computerlinguistik am IMS übernahm.



Abb. 1: Prof. Dr. Jonas Kuhn, Leiter des Stuttgart CLARIN-Zentrums

	Gödel	wollte	nach	Amerika	reisen	.
reisen.1	A0		A3			

Parsing sentence required 16ms.

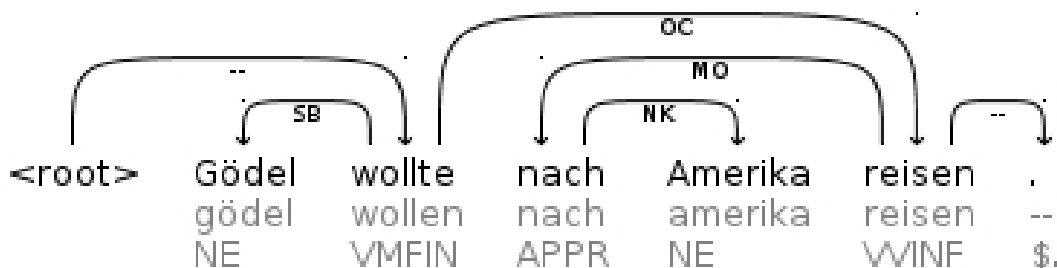


Abb. 2: Semantische und syntaktische Analyse (de.sempar.ims.uni-stuttgart.de/)

Eine wichtige Aufgabe des Stuttgarter Zentrums liegt darin, das reiche Inventar der computerlinguistischen Ressourcen des IMS der Forschungsgemeinschaft nachhaltig zur Verfügung zu stellen. Neben dem Vorhalten der Werkzeuge, Korpora und lexikalischen Daten, geschieht dies durch eine gründliche Dokumentation, Schulungen und aktiven Nutzersupport [1]. Beispiele für die Werkzeugvielfalt, die das IMS anbietet, sind der bekannte TreeTagger (Schmid 1994) und die Morphologiekomponente SMOR (Schmid, Fitschen, Heid 2004) sowie ein multilingualer syntaktischer Dependenzparser (Bohnet 2010, Bohnet und Kuhn 2012) und ein darauf aufbau-

endes Analysewerkzeug zur Kennzeichnung semantischer Rollen (Björkelund et al. 2010). Abbildung 2 zeigt wie die beiden letztgenannten Werkzeuge das Hauptverb eines Satzes und seine Argumente identifizieren sowie dem ganzen Satz eine Dependenzanalyse zuweisen. Ein etwas älteres Werkzeug ist TIGER-Search (Lezius 2002), mit dem syntaktische Konstituentenbäume systematisch durchsucht werden können. Abbildung 3 zeigt, wie Suchergebnisse visualisiert und exportiert werden können. Das Stuttgarter CLARIN-D-Team ist auch an der Schaffung neuer Werkzeuge speziell für den Bedarf der Textwissenschaften „jenseits“ der Linguistik

[1] Nutzersupport: clarin@ims.uni-stuttgart.de

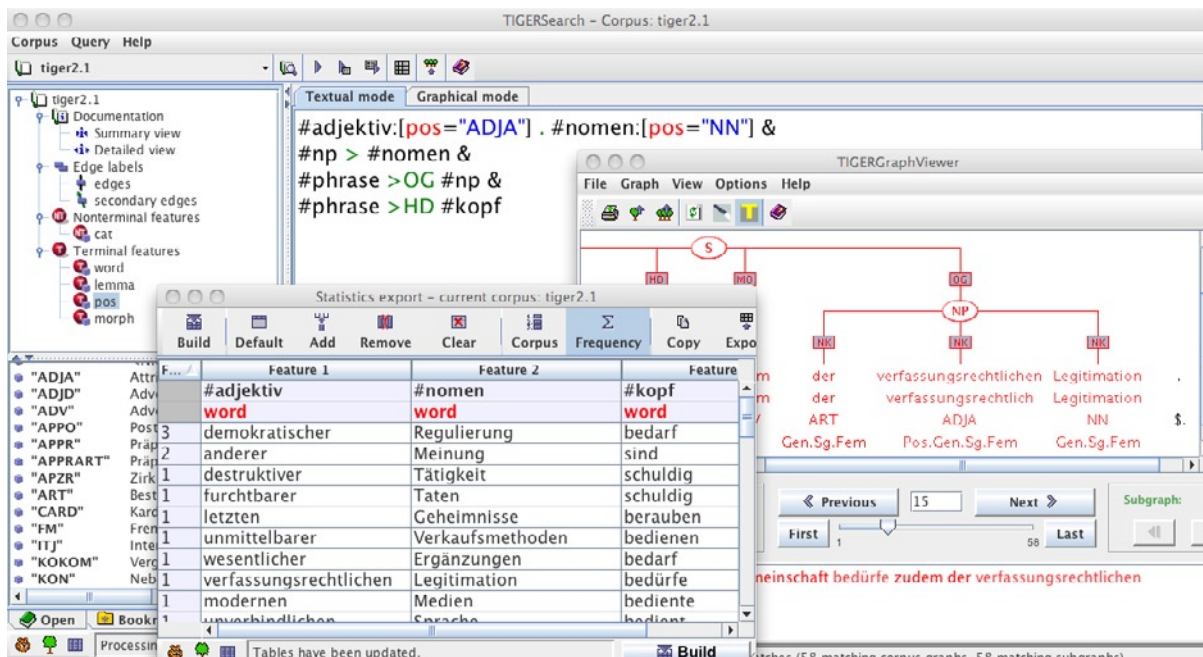


Abb. 3: Frequenzaufistung von modifizierten Genitivobjekten mit TIGERSearch

beteiligt. In Kooperation mit Stuttgarter eHumanities-Projekten koordiniert André Blessing die Entwicklung einer interaktiven Plattform zur Extraktion von Relationen aus großen Textsammlungen, die intern eine ganze Kaskade von linguistischen Analysewerkzeugen verwendet (Blessing, Stegmann und Kuhn 2012). Mit der Plattform können Fachwissenschaftler bei der Korpusanalyse die für ihre Studie relevanten inhaltlichen Kategorien und Relationen selbst definieren, interaktiv trainieren und im Studienverlauf weiter verfeinern und anpassen; Abbildung 4 illustriert ein Zwischenstadium des Trainings für die Relation ‚Schüler_von‘.

Das IMS in CLARIN-D

Das Stuttgarter Zentrum ist mit einer Reihe von Werkzeugen in der Online-

Analyseplattform WebLicht vertreten. Der oben erwähnte syntaktische Dependenzparser wurde hierfür z.B. eigens im Rechenzentrum Garching installiert; mit seinem verhältnismäßig hohen Speicherbedarf ist er ein interessanter Testfall für das Deployment von Werkzeugen auf externen Servern. Um für Nutzer von CLARIN-D auffindbar zu sein, sind die Ressourcen des IMS im Virtual Language Observatory vertreten. Grundlage für die Präsenz in den beiden Online-Plattformen sowie für die Nutzung in zukünftigen Komponenten der Infrastruktur ist das nachhaltige Vorhalten einschlägiger CMDI-Metadaten innerhalb des Stuttgarter IMS-Repository, das zur Zeit hinsichtlich eines Archivdienstes für Objektdaten ausgebaut wird. Technischer Ansprechpartner für das IMS-Repository ist Jens Stegmann, der außerdem in Kooperation mit der Tübinger CLARIN-D-Leitung und in Abstimmung mit

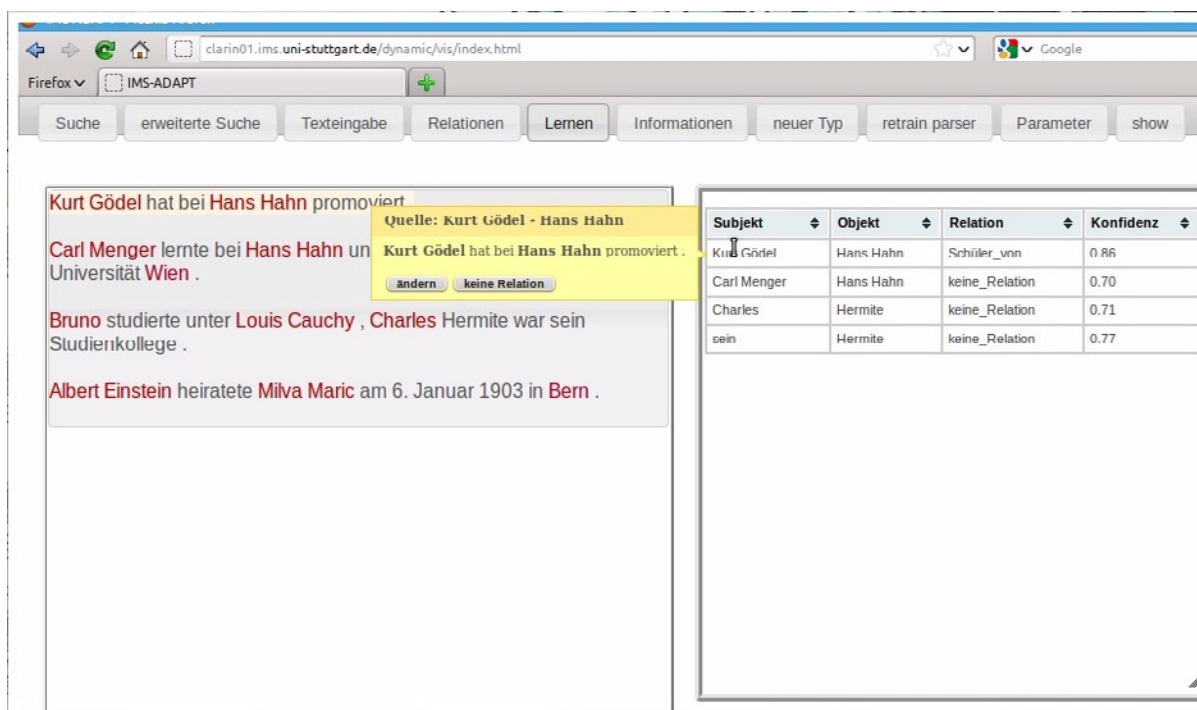


Abb. 4: Interaktives Lernen der Extraktion von „Schüler_von“.

der Münchner Leitung des Arbeitspakets 9 die Betreuung der CLARIN-D-Webseite (www.clarin-d.de) und des projekt-internen Wikis wahrnimmt. Ansprechpartnerin für die Erstellung von CMDI-Metadaten und Kuratorin der Stuttgarter Metadatensätze ist Kerstin Eckart. Sie ist außerdem Autorin der Kapitel „Concepts and data categories“ sowie „Resource annotations“ im CLARIN-D-User-Guide. Ebenfalls zum User-Guide beigetragen hat Heike Zinsmeister, die als stellvertretende Projektleiterin das Stuttgarter Zentrum bei Kooperationen mit sprachwissenschaftlichen Projekten vertritt und zusammen mit dem Saarbrücker CLARIN-D-Zentrum die ersten CLARIN-D-Doktorandentage zum Thema Corpora am 25. / 26. 03. 2013 in Stuttgart durchgeführt hat.

eHumanities-Projekte

Das gemeinsame Ziel von CLARIN-D ist die Schaffung einer technischen Forschungsinfrastruktur. Hierfür müssen Computerlinguisten und Informatiker eng mit Geistes- und Sozialwissenschaftlern zusammenarbeiten, um die Anforderungen an die Infrastruktur durch deren späteren Nutzer besser zu verstehen und gleichzeitig das Infrastrukturangebot in den Fachwissenschaften zu etablieren. Durch Jonas Kuhns Ko-Leiterfunktion in zwei BMBF-geförderten eHumanities-Projekten befindet sich das IMS in einer Vorreiterrolle für die fachübergreifende Zusammenarbeit. In e-Identity, federführend geleitet von Prof. Dr. Cathleen Kantner, untersuchen Stuttgarter Politikwissenschaftler kollektive Identi-

täten in internationalen Debatten um Krieg und Frieden. In dem von Prof. Dr. Sandra Richter federführend geleiteten Projekt ePoetics analysieren Literaturwissenschaftler Schriften zur Dichtungstheorie aus drei Jahrhunderten mittels neuer Visualisierungsmethoden. In beiden Projekten liefern computerlinguistische Methoden die Grundlagen für die fachbezogenen Analysen.

Ausblick

2012 hat die Leitung der Universität Stuttgart die eHumanities als einen von vier Themenbereichen im Kooperativen Forschungscampus (dem Zukunftsprofil der Universität Stuttgart für die kommenden Jahre) definiert. Das IMS und sein CLARIN-Zentrum freuen sich, diesen Themenbereich mitgestalten zu können. Ein methodologisches Ziel besteht darin, Textwissenschaftlern aus den verschiedenen geistes- und sozialwissenschaftlichen Disziplinen auf Dauer nicht nur den Einsatz von Analysewerkzeugen „von der Stange“ zu ermöglichen. Vielmehr sollen sie in einen kritisch reflektierten Gebrauch von komputationellen Analysemodellen mit einbezogen werden – stellen die eHumanities doch die Chance dar für eine echte Integration des vielfältigen fachwissenschaftlichen Wissens über Sprache, Text und seinen Kontext mit den technischen Möglichkeiten von automatischen Annotationsmodellen und statistischen Analyse- und Lernverfahren.

Referenzen

- Björkelund, A;** Bernd Bohnet; Love Hafdell; Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In: COLING 2010: Demonstration Volume.,
- Blessing, A.;** Jens Stegmann; Jonas Kuhn. 2012. SOA meets Relation Extraction: Less may be more in Interaction. In: Proceedings of the Digital Humanities 2012 Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts.
- Bohnet, B.** 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of COLING 2010.
- Bohnet, B.;** Jonas Kuhn. 2012. The Best of Both Worlds – A Graph-based Completion Model for Transition-based Parsers. In: Proceedings of EACL 2012.
- Lezius, W.** 2002. Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Doktorarbeit, IMS, Universität Stuttgart. Veröffentlicht als AIMS 8 (4).
- Schmid, H.;** Arne Fitschen; Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In: Proceedings of LREC 2004.
- Schmid, H.** 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. IN: Proceedings of International Conference on New Methods in Language Processing.



Dr. Heike Zinsmeister,
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

DTA-/CLARIN-D-Konferenz und -Workshops: Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven

Tagungsbericht von der DTA-/ CLARIN-D-Konferenz aus Berlin

DTA-/CLARIN-D-Konferenz und -Workshops am 18./19. Februar 2013 im Einsteinsaal der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin. Ein Tagungsbericht.

Am 18./19. Februar 2013 veranstalteten das Deutsche Textarchiv und CLARIN-D gemeinsam eine zweitägige Konferenz mit dem Titel „Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven“. Die Konferenz fand im Einsteinsaal der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin statt, der mit mehr als 90 TeilnehmerInnen sehr gut gefüllt war. Die TeilnehmerInnen und ReferentInnen kamen aus teilweise sehr verschiedenen (Teil-)Disziplinen (Informatik, Philologien und Editionswissenschaften, Korpuslinguistik, Bibliothekswissenschaften

u.a.), wodurch die angesprochenen Themen aus unterschiedlichen Perspektiven dargestellt und diskutiert wurden.

Der erste Konferenztag widmete sich nach einer Begrüßung und Einführung in das Deutsche Textarchiv von Alexander Geyken (Berlin) in einem ersten Vortragsblock Projekten, die derzeit mit dem manuellen bzw. halbautomatischen Aufbau neuer historischer Textkorpora befasst sind. Heike Sahn (Siegen) berichtete von einem Projekt zur „Erschließung Städtischer Literatur im 15. Jahrhundert: am Fallbeispiel Nürnberg“ an der Universität Siegen. Thierry Declerck (Saarbrücken/Wien) und Claudia Resch (Wien) präsentierten die Arbeiten am Austrian Baroque Corpus (ABaC:us). Aus Sicht der Bibliotheken, die zur Erfassung großer Textmengen für die Texterschließung OCR-Verfahren nutzen und optimieren, berichteten Maria Federbusch (Berlin) von einer Fallstudie zum OCR-Einsatz bei der Volltexterfassung von Quellen der Frühen Neuzeit der SBB Berlin und Manfred Nölte (Bremen) von dem Digitalisierungsprojekt „Die Grenzboten“ der SuUB Bremen. Im letzten Themenblock präsentierten Marius Hug und Christian Kassung (Berlin) mit „Dinglers Polytechnischem Jour-

nal“ sowie Noah Bubenhofer (Dresden) mit dem „Text+Berg Korpus“ Analyse-möglichkeiten von bereits digital vorliegenden Volltextkorpora. Den Abschluss des ersten Konferenztages bildete ein Beitrag von Thomas Gloning (Gießen) zur Frage, inwiefern thematische (Teil-)Korpora als Arbeitsgrundlage für die historische Lexikographie dienen könnten. Den Auftakt zum zweiten Konferenztag bildeten Beiträge von Timo Steyer (Wolfenbüttel) sowie Eva-Maria Dickhaut und Jörg Witzel (Marburg) zu Volltextsammlungen, die im Rahmen des DFG-geförderten Verbundprojekts AEDit Frühe Neuzeit entstehen. Gerhard Heyer (Leipzig) zeigte anhand der Edition der zwischen 1871 und 1933 erschienenen Leipziger Rektoratsreden Möglichkeiten des automatisierten *Information Retrieval* in einem historischen Textkorpus. Eine Infrastruktur zur linguistischen Analyse verschiedener historischer Korpora präsentierten Carolin Odebrecht, Florian Zipser (Berlin) mit dem Projekt Laudatio. Stefanie Dipper (Bochum) widmete sich in ihrem Beitrag den Möglichkeiten der orthographischen Normalisierung und des Taggings frühneuhochdeutscher Texte. Aus Sicht der Orthographiegeschichte beleuchtete abschließend Anja Voeste (Gießen) die Signifikanz graphischer Variation in historischen Texten und die daraus resultierenden Leitlinien für die Texterfassung. Die Konferenz wurde von zwei Workshops flankiert:

Workshop 1 am Vormittag des 18. Februar wurde vom Arbeitspaket 5: Dienste und Ressourcen in CLARIN-D veranstaltet und widmete sich den CLARIN-

D-Empfehlungen und -Richtlinien für die linguistische Annotation von Korpora (Kerstin Eckart, Stuttgart) und für die Erfassung von Metadaten (Axel Herold, Berlin) sowie verschiedenen CLARIN-D-Diensten für die linguistische Textanalyse mit Beiträgen von Kai Zimmer (Berlin), Edmund Pohl (Potsdam) und Thomas Eckart (Leipzig).

Workshop 2 am Nachmittag des 19. Februar wurde von Mitarbeitern des Deutschen Textarchivs veranstaltet und thematisierte Methoden des Aufbaus von Sprachressourcen am Beispiel des DTA und des Kurationsprojekts 1 der CLARIN-D F-AG 1. In Referaten und praktischen Übungen wurden die Richtlinien für den Korpusaufbau und verschiedene Services des DTA, welche die Erarbeitung von Texten anhand dieser Richtlinien unterstützen, vorgestellt und erprobt. Der Workshop wurde aufgrund der großen Nachfrage am 19. April 2013 an der BBAW in Berlin wiederholt.

Nähere Informationen zur Konferenz und den beiden Workshops, das ausführliche Programm sowie die Folien der Referenten sind auf der Konferenzwebseite www.deustextarchiv.de/doku/workshop2013/ zugänglich. Ein ausführlicher Tagungsbericht erschien auf H-Soz-u-Kult und ist einsehbar unter der Adresse hsozkult.geschichte.hu-berlin.de/tagungsberichte/id=4791.

AutorInnen: Deutsches Textarchiv (Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, Frank Wiegand)

Bericht zum Kickoff-Treffen der 24 eHumanities Verbundprojekte in Leipzig

Vorstellung der Projektverbünde

Am Montag und Dienstag den 08./09.04.2013 trafen sich über 150 Mitglieder der 24 vom Bundesministerium für Bildung und Forschung (BMBF) geförderten eHumanities Verbundprojekte, der Forschungsinfrastrukturen CLARIN-D [1] und DARIAH-DE [2] und weiterer in den eHumanities aktiver Personen am Medien-campus in Leipzig. Ziel des Treffens war die Vorstellung der Projektverbünde sowie die Präsentation und Diskussion von Themen von gemeinsamem Interesse, wobei ein besonderes Augenmerk auf die Möglichkeit zur Vernetzung zwischen den verschiedenen Projekten und Initiativen gelegt wurde.

Begrüßt wurden die Gäste durch die Rektorin der Universität Leipzig, Frau Prof. Dr. Beate Schücking, welche unter anderem auch die aktuelle Rolle und zukünftige Ambitionen der Universität Leipzig auf dem Gebiet der eHumanities in ihrer Rede hervorhob. Im Anschluss begrüßte Frau Dr. Angelika Willms-Herget (BMBF) die Teilnehmer und sprach

in ihrem Vortrag Herausforderungen und Chancen von Forschungsinfrastrukturen in den Geisteswissenschaften an. Der erste Block der Veranstaltung stand zudem im Zeichen der Begrüßung von Herrn Prof. Dr. Gregory Crane als frisch berufenem Humboldt-Professor an der Universität Leipzig. Herr Prof. Dr. Helmut Schwarz, der amtierende Präsident der Alexander von Humboldt-Stiftung referierte aus diesem Anlass zum Thema „Personenförderung, Internationalität und akademische Netzwerkbildung als Prinzipien der Humboldt-Stiftung“.

Dem einleitenden Block folgte eine Keynote zum Thema „Rich and linked media in the eHumanities“ von Herrn Prof. Dr. Stefan Wrobel (Institutsleiter des Fraunhofer IAIS) in deren Rahmen unter anderem über die Arbeit mit und Nutzung von audiovisuellen Medien in den eHumanities berichtet wurde. Die Keynote wurde von einem thematischen Block mit dem Titel „Erfahrungsberichte & Diskussionen zum Thema Kooperationen von Geisteswissenschaftlern und Informatikern“ gefolgt. Herr Prof. Dr. Gregory Crane und Herr Prof. Dr. Kurt Gärtner (Akademie der Wissenschaften

[1] de.clarin.eu
[2] de.dariah.eu

und der Literatur Mainz) gaben einen Einblick in ihre bisherigen Projekte und der in diesem Rahmen gesammelten Erfahrungen sowie einen Ausblick auf zukünftige Vorhaben.

Eines der Kernelemente des Kickoff-Treffens war die Präsentation der 24 Projektverbände im Rahmen einer mehrstündigen Poster-/Demosession am Nachmittag des ersten Veranstaltungstages. Mit Hilfe der vorab erstellten und an die Teilnehmer verteilten Exposé wurde eine gezielte Vorbereitung auf diese Session ermöglicht. Neben Angaben zur disziplinären Verortung, dem wissenschaftlichen *use case* und der am Verbund beteiligten Projektpartner wurden auch die im Rahmen der Projekte genutzten Ressourcen und Methoden benannt sowie die entstehenden Daten und Verfahren beschrieben. Diese durch die Projektverbände erstellten Kurzbeschreibungen stehen auf der Webseite [3] des Kickoff-Treffens zum Download zur Verfügung. Auf dieser Grundlage war es im Rahmen persönlicher Gespräche möglich, das Potential für nutzbare Synergien, den nötigen Erfahrungsaustausch oder auch für sinnvolle Kooperationen auszuloten. Insbesondere war es möglich die erstaunliche Bandbreite der in den eHumanities in Deutschland vorhandenen Erfahrungen und aktuell be-

arbeiteten Themen hautnah zu erfahren. Die Projekte umfassten unter anderem:

die Entwicklung von Methoden und Verfahren „zur Analyse von Bedeutungsverschiebungen und Diskursstrukturen in großen Textmengen für politikwissenschaftliche Fragestellungen“ [4]

den Aufbau eines Recherchesystems welches „Bedeutungsverschiebungen semantischer Räume nach Ort und Zeit visualisiert“ [5]

die bibliografische Erfassung von „frühe[n] Texte[n] der deutsch- bzw. polnischsprachigen Holocaust- und Lagerliteratur von 1933 bis 1949 [...] in einer Online-Datenbank“ [6]

die Umsetzung „einer genuin digitalen Musikedition“ am „Beispiel von Carl Maria von Webers Oper Der Freischütz“ [7]

Den Aufbau eines „neuen Forschungswerkzeugs für die Archäologie [...] welches es erlaubt, 3D-Modelle und Funktionen von Geographischen Informationssystemen (GIS) für die Dokumentation und Analyse archäologischer Fundstätten auf einer Internet-Plattform zusammenzuführen“ [8]

Die Entwicklung eines „automatisierten textanalytischen Verfahrens“ welches

[3] ehumworkshop.informatik.uni-leipzig.de/

[4] Kurzbeschr. „Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der bundesdeutschen Politik 1949-2011“

[5] Kurzbeschr. „eXChange – Exploring Concept Change and Transfer in Antiquity“

[6] Kurzbeschr. „GeoBib – Georeferenzierte Online-Bibliografie früher Holocaust und Lagerliteratur“

[7] Kurzbeschr. „Freischütz Digital – Paradigmatische Umsetzung eines genuin digitalen Editionskonzepts“

[8] Kurzbeschr. „MayaArch3D – Ein webbasiertes 3D-GIS zur Analyse der Archäologie von Copan, Honduras“

„mit Hilfe von Methoden aus der Politikwissenschaft, Linguistik und Informatik [...] neue Einsichten in die Funktionsweise deliberativer politischer Kommunikation liefern soll.“ [9]

Der erste Tag des Treffens schloss mit Vorträgen zum Thema Visualisierung: Herr Prof. Dr. Daniel Keim, Sprecher des Schwerpunktprogramms *Scalable Visual Analytics*, berichtete zum „State of the Art in der Visualisierung“ und Herr Prof. Dr. Christopher Culy (Universität Tübingen) hielt einen Vortrag mit dem Titel „Letters to the visualization field“.

Der zweite Tag begann mit dem Themenschwerpunkt „Rechtliche Probleme, Fragen und Perspektiven“ in dessen Rahmen auf Probleme bei der Nutzung existierender und der (Online-)Bereitstellung neuer Ressourcen hingewiesen werden sollte. Herr Prof. Dr. Gerhard Heyer (Universität Leipzig) gab einen Erfahrungsbericht im „Umgang mit Urheber- und Persönlichkeitsrecht im Projekt Deutscher Wortschatz“, welcher auch weniger bekannte und vorhersehbare Themen wie die Problematik von Gemeinschaftsmarken und Gattungsbezeichnungen beinhaltete. Die Schilderung konkreter rechtlicher Probleme wurde mit kurzen Hinweisen zu technischen und organisatorischen Maßnahmen abgerundet, welche sich im Wortschatz-Projekt in der Praxis als nützlich zur Problemlösung bzw. -vermeidung erwiesen haben. Im Anschluss berichte-

te Erik Ketzan im Vortrag „Legal Issues for Language Resources: Best Practices and Developments in EU Law“ über die rechtliche Situation bei der Nutzung von Sprachressourcen und aktuelle Entwicklungen auf EU-Ebene.

Ein weiteres Schwerpunktthema des zweiten Tages waren die Forschungsinfrastrukturen CLARIN-D und DARIAH-DE. Frau Dr. Heike Neuroth (wissenschaftliche Koordinatorin DARIAH-DE) und Herr Prof. Dr. Erhard Hinrichs (wissenschaftlicher Koordinator CLARIN-D) stellten dem Publikum die Ziele und bereits heute vorhandenen Angebote beider Infrastrukturen vor und luden die Teilnehmer dazu ein, mit Vertretern beider Projekte im Rahmen einer anschließenden Demosession in direkten Kontakt zu treten. Die große Anzahl der anwesenden Teilnehmer und die Vielzahl und Bandbreite der vertretenen Projekte machte deutlich, welches Nutzerpotential allein in Deutschland bereits vorhanden und durch die Forschungsinfrastrukturen adressiert und bedient werden kann.

Der letzte Block der Veranstaltung beinhaltete zwei Vorträge zum Thema „Geo- und GIS-Referenzierung“. Herr Prof. Dr. Andreas Henrich (Universität Bamberg) stellte in seinem Vortrag „Möglichkeiten der Nutzung von Geo/GIS Daten in den eHumanities“ vor und gab zudem einen Einblick in die in diesem Umfeld gesammelten Praxiserfahrungen. Im

[9] Kurzbeschr. „*VisArgue - Wie und wann überzeugen Argumente? Analyse und Visualisierung von politischen Verhandlungen*“

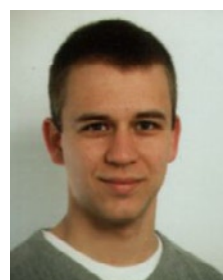
Mitmachen!

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.

Anschluss referierte Herr Prof. Reinhard Förtsch (DAI Berlin) zum Thema „Kontextualisierung von Daten/Objekten mit Geo/GIS-Daten als infrastrukturelles Problem“. Die Veranstaltung schloss mit einer Podiumsdiskussion in deren Rahmen einige Themen des Workshops erneut aufgegriffen wurden und unter anderem die Frage diskutiert wurde, in welche Richtung sich die eHumanities in Deutschland in den nächsten Jahren entwickeln, welche Probleme thematisiert und welche Lehren aus bisherigen Erfahrungen und den im Rahmen des Treffens besprochenen Themen gezogen werden sollten.

Wir danken allen Projektverbänden, Vortragenden, und Teilnehmern für ihre Mühe bei der Vorbereitung und

Durchführung der Poster-/Demosession bzw. der Vorträge und hoffen, dass wir mit dieser Veranstaltung einen Beitrag zur wichtigen Vernetzung – sowohl in der Breite als auch in der Tiefe – in den eHumanities leisten konnten. Wir wünschen allen 24 Projekten viel Erfolg bei der Bearbeitung Ihrer Forschungsfragen und würden uns freuen auch in Zukunft wieder zahlreiche Gäste im Rahmen von Veranstaltungen in den eHumanities in Leipzig begrüßen zu dürfen.



Volker Boehlke
*Institut für Informatik,
Universität Leipzig*

Die Kurationsprojekte der CLARIN-D F-AG 2 „An- dere Philologien“

Das erste Kurationsprojekt der F-AG 2 baute auf bereits weit gediehenen Vorarbeiten auf und konnte deshalb nach acht Monaten Laufzeit bereits im März 2013 abgeschlossen werden. Im März bzw. April 2013 haben zwei weitere Kurationsprojekte die Arbeit aufgenommen.

Verantwortlich für die Realisierung des ersten Kurationsprojekts „Erstellung einer webbasierten Plattform für die strukturierte Dokumentation von Sprachen im mobilen Zeitalter“ waren Prof. Dr. Jürgen Handke und Dr. Peter Franke von der Universität Marburg. Ein wesentliches Ziel des Projekts war die Erstellung einer webbasierten Plattform, des sogenannten *Language Index* [1], auf der gesprochene Daten und Informatio-

nen zu Sprachen und Varietäten gespeichert und über eine interaktive und multimediale Schnittstelle den Benutzern zugänglich gemacht werden. Außerdem war die Integration der Plattform in die CLARIN-D-Infrastruktur von großer Bedeutung.

Die Idee des *Language Index* ist vergleichbar mit Googles *Endangered Languages Project* (ELP) [2]: Sprecher auf der ganzen Welt laden selbständig ihre Sprachdaten hoch und machen sie damit für andere Nutzer verfügbar. Grundlage für das Hochladen der Daten in den *Language Index* sind Datenblätter, die bereits für 120 verschiedene natürliche Sprachen und ihre Varietäten zur Verfügung stehen und von der Website heruntergeladen werden können. Diese strukturierte Dokumentation der Da-



The Language Index

The Place for Learning about the Languages of the World



[Explore](#) [Contribute](#) [Learn](#) [About](#) [Contact](#)

[1] www.languageindex.org

[2] www.endangeredlanguages.com

mationen zu den dort verorteten Sprechern und Sprachen, kann Videomaterial anschauen oder Audiodaten anhören. Mit ihrer Verknüpfung gehen *Language Index* und CLARIN-D eine aussichtsreiche Beziehung ein, von der beide Seiten profitieren werden. Der *Language Index* kann auf eine 15-jährige Geschichte als Teil des *Virtual Linguistics Campus* (VLC) [3] (eine e-Learning Plattform für Linguisten, aufgebaut von Jürgen Handke und seinem Team an der Universität Marburg) zurückblicken, bevor er nun als eigenständige Plattform aus dem VLC herausgelöst wurde. Von dieser erfolgreichen, jahrelangen Etablierung kann CLARIN-D profitieren, auf das an prominenter Stelle auf der Website des *Language Index* hingewiesen wird. Außerdem soll der *Language Index* in CLARINs *Virtual Language Observatory* (VLO) und dem vom Saarbrücker Zentrum im Aufbau befindlichen Teaching Hub *TeLeMaCo* verlinkt werden. Um die Integration in die Netzwerke zu ermöglichen, wurden alle Metadaten im CLARIN-D-konformen CMDI-Format erstellt.

Das zweite Kurationsprojekt der F-AG „Erschließung digitaler Textarchive über Metadaten und Lemmata“ wird von Prof. Dr. Roland Meyer von der Humboldt-Universität zu Berlin koordiniert. Ziel des Projekts, das im März 2013 seine Arbeit aufgenommen hat, ist die Entwicklung eines Werkzeugs, das eine bessere Erschließung ausgewählter historischer Archive ermöglichen und den Weg zur Erschließung weiterer Ar-

chive ebenen soll. Das Suchwerkzeug soll eine Abfrage nach Lemmata bzw. (modernen) Grundformen möglich machen und auf der Grundlage von Metadaten ein geeignetes Korpus zusammenstellen. Die dazu benötigten Sekundärdaten und Werkzeuge (Datenbanken, Lexika, morphologische Analysierer) sollen über Webservices bereitgestellt werden.

Das Projekt wird am Beispiel des Polnischen ausgeführt, ein mit seinen zahlreichen Flexionen und der starken orthographischen Variation eher schwieriges Fallbeispiel. Damit sollte aber die Übertragbarkeit auf andere Sprachen garantiert sein, was das Werkzeug auch für Wissenschaftler anderer Disziplinen interessant macht.

Ein wichtiger Aspekt für den Erfolg des Projekts ist die Verknüpfung von sprach- und literaturwissenschaftlichem Wissen mit der computer- und korpuslinguistischen Kompetenz des CLARIN-D-Netzwerkes. Der Zusammenarbeit mit verschiedenen CLARIN-D-Zentren wird deshalb eine zentrale Bedeutung zukommen. So werden das Saarbrücker Zentrum und die Tübinger Werkzeugsammlung *WebLicht* für die Anwendung und Weiterentwicklung typisch computerlinguistischer Werkzeuge (morphologische Analysierer, Lemmatisierer, digitale Lexika) eine wichtige Rolle spielen; das MPI in Nijmegen wird ein wertvoller Ansprechpartner im Bereich des zeitgemäßen Umgangs mit Metadaten sein; und das Leipziger Zentrum wird aufgrund seiner Erfahrungen und Kompetenzen aus dem Projekt *Wortschatz* ein bedeutender Kooperationspartner sein.

[3] linguistics.online.uni-marburg.de

Die Leitung des dritten Kurationsprojekts „Parallelx – Web-basierter Workflow zur Erstellung und Nutzung von Paralleltexten“ liegt bei Dr. Christof Schöch von der Universität Würzburg. Ein wesentliches Ziel des Projekts, das seit April 2013 läuft, ist die Entwicklung eines web-basierten Workflows, mit dem Fachwissenschaftler aus den sie interessierenden Texten mehrsprachige Paralleltexte erstellen und analysieren können. Besonderen Wert wird auf die Transparenz des Verfahrens, die einfache Nutzbarkeit und die ausführliche Dokumentation gelegt, was den Work-



Christian Mair
Englisches Seminar
Albert-Ludwigs-Universität Freiburg

flow auch für technisch weniger versierte Fachwissenschaftler attraktiv macht. Für den Webservice werden vorhandene Services und Tools in WebLicht zusammengeführt, soweit notwendig erweitert und durch weitere Services ergänzt. Außerdem wird der Workflow an zwei bis drei modellhaften Beispielen umgesetzt, um die Abläufe bei der Erstellung und Abfrage der Paralleltexte zu illustrieren.



Claudia Winkle
Englisches Seminar
Albert-Ludwigs-Universität Freiburg

Von Bächen und Beckern

Wie man Städtenamen sichtbar macht

Lambach, Neudeutenbach, Mittelweißbach – betrachtet man eine Deutschlandkarte, so finden sich viele mehr oder weniger große Städte und Dörfer deren Namen auf „-bach“ enden. Allerdings scheint es so, dass im südlichen Bereich eine Häufung entsprechender Ortsnamen auftritt. Diese und andere Fragen können mit „CiNaViz“ (kurz für „City Name Visualization“), einer Webapplikation zur Visualisierung europäischer Ortsnamen beantwortet werden. Im Zugriff sind über 1.2 Millionen geographische Namen, deren Verteilung über Europa anhand regulärer Ausdrücke visualisiert werden kann. Abbildung 1 zeigt, dass Orte die auf „-bach“ enden, fast nur südlich der sogenannten Benrath-Linie auftreten. Nördlich davon treten Orte mit „-beck“ an deren Stelle und im nördlichsten Bereich, nach Dänemark hin, verschwindet das „c“ („Lübek“).

Anhand der mittels CiNaViz erstellten Karten lassen sich Wanderungsbewe-

gungen, kulturelle sowie topographische Gegebenheiten anhand der Distribution von Ortsnamen aufzeigen. So finden sich Orte, deren Name mit „Sankt“ beginnt hauptsächlich in Österreich und den katholischen Gebieten Deutschlands (Abbildung 2). Die Endung „-ham“ ist sowohl in England als auch im Südosten Bayerns gebräuchlich (Abbildung 3).

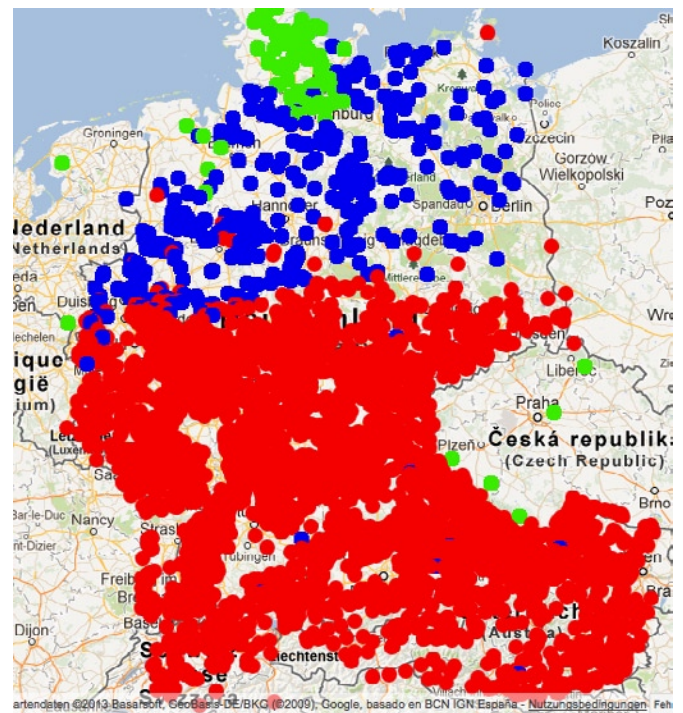


Abb. 1: Die Verteilung von Ortsnamen, die auf „-bach“ (rot), „-beck“ (blau) und „-bek“ (grün) enden

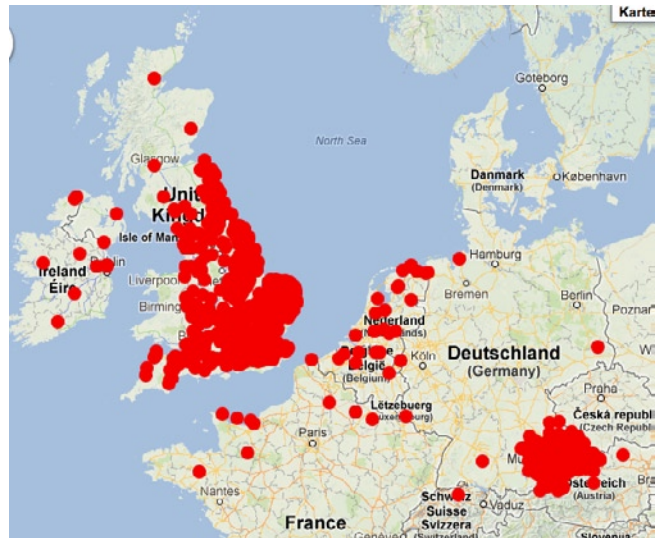
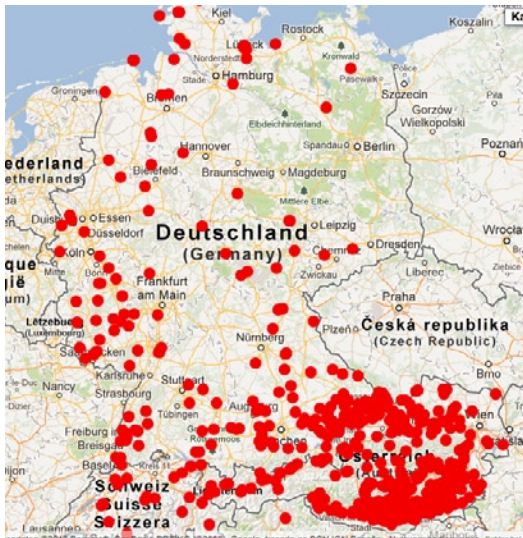


Abb. 2 (links): Ortsnamen die mit „Sankt“ beginnen

Abb. 3 (rechts): Ortsnamen die mit „-ham“ enden finden sich in England, den Niederlanden sowie Belgien und dem Südosten Bayerns

CiNaViz ist Teil der CLARIN-D-Infrastruktur und frei zugänglich unter:

weblicht.sfs.uni-tuebingen.de/CityViz/

Thomas Zastrow
Seminar für Sprachwissenschaft
Universität Tübingen



Webpräsenz des europäischen Langzeitprojekts:

www.clarin.eu

Von Menschen, Orten und Organisationen

Wie man Menschen, Orte und Organisationen sichtbar macht

WhoIsInTheNews ist eine Webanwendung, die aus zwei Teilen besteht. Der erste Teil lädt seit November 2011 jeden Tag um 15 Uhr die Newsticker überre-

gionaler deutscher Zeitungen herunter. Anschließend werden mit Hilfe von WebLicht-Webservices die enthaltenen Eigennamen extrahiert. Diese werden in einer Datenbank gespeichert und können mit dem zweiten Teil der Software, einem graphischen Benutzerinterface, analysiert und visualisiert werden. Insgesamt wurden so bis jetzt mehr als 300.000 Vorkommen von Eigennamen,

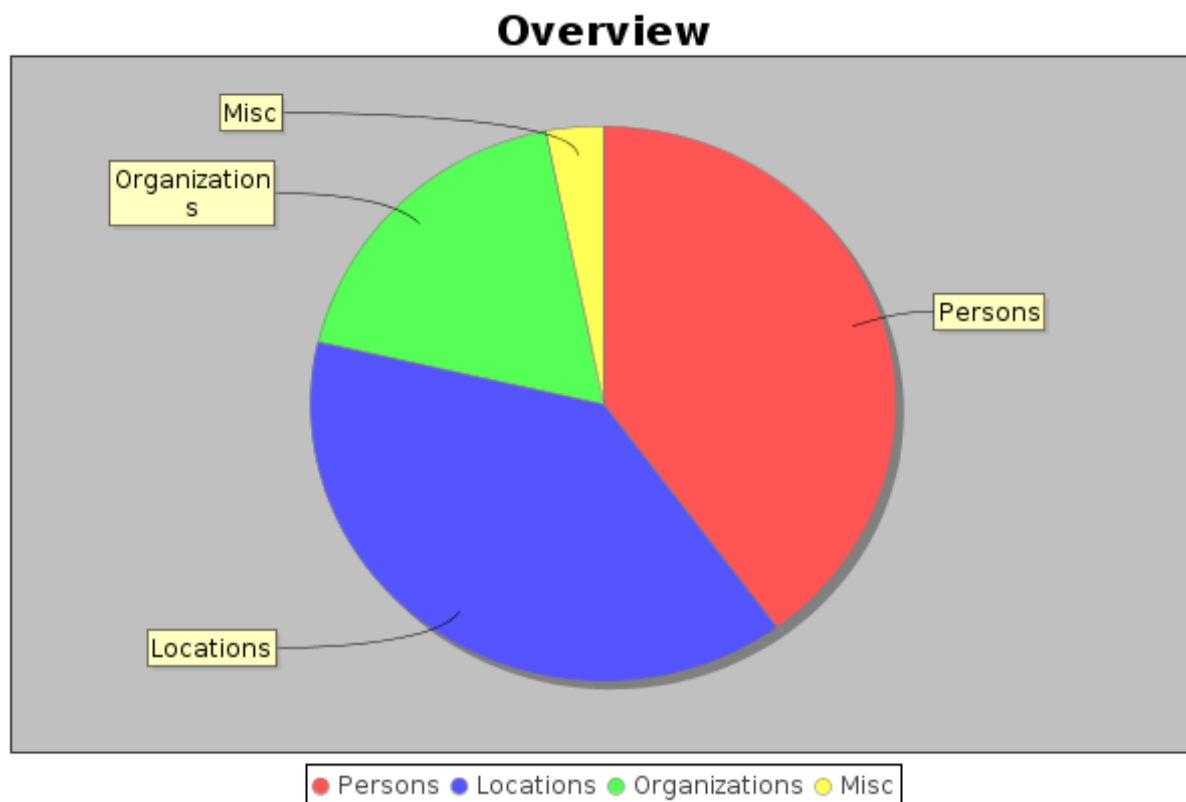


Abb. 1: Aufteilung der gesammelten Eigennamen in die Rubriken „Person“, „Ort“, „Organisation“ und „Sonstiges“



Abb. 2: Nennung von „Benedikt“ und „Franziskus“ in deutschen Nachrichtenmagazinen, Anfang 2013

unterteilt in die Kategorien „Person“, „Ort“, „Organisation“ und „Sonstiges“ gesammelt (Abbildung 1).

Abbildung 2 zeigt eine Analyse der Vorkommen der Eigennamen „Benedikt“ (blau) und „Franziskus“ (rot), entsprechend der Namensgebung des aktuellen Papstes (Franziskus) und seines Vorgängers (Benedikt). Der blaue Höhepunkt im Februar markiert den Tag der Ankündigung des Rücktritts Benedikt, wohingegen der Name seines Nachfolgers („Franziskus“) erst mit dessen Wahl einige Wochen später in den Nachrichten erscheint.

Nicht nur Personennamen, auch die Nennung von Orten lässt sich in WhoIsInTheNews analysieren und visualisieren. So sind die Kriegshandlungen in der syrischen Stadt Aleppo in Juli 2012 bzw. die Zerstörung historischer Gebäude

Ende September, Anfang Oktober 2012 erkennbar als Ausschläge nach oben in Abbildung 3 [1]. In Abbildung 4 sind die 1.000 häufigsten Orte in der Datenbank visualisiert.

WhoIsInTheNews ist Teil der CLARIN-D Infrastruktur und nach Anmeldung zugänglich unter weblicht.sfs.uni-tuebingen.de/ne/.



Thomas Zastrow
Seminar für Sprachwissenschaft
Universität Tübingen

[1] Zitat Wikipedia, Eintrag Aleppo: „Im Zuge des Bürgerkrieges in Syrien kam es im Juli 2012 in Aleppo zu heftigen Kämpfen, bei denen Raketenwerfer, Hubschrauber und Kampfflugzeuge eingesetzt wurden. In der Nacht vom 28. auf den 29. September 2012 wurde der historische Basar, weltgrößtes überdachtes altes Marktviertel und Teil des UNESCO-Welterbes, durch ein Großfeuer weitgehend zerstört, das offenbar auf Kampfhandlungen beruhte.“

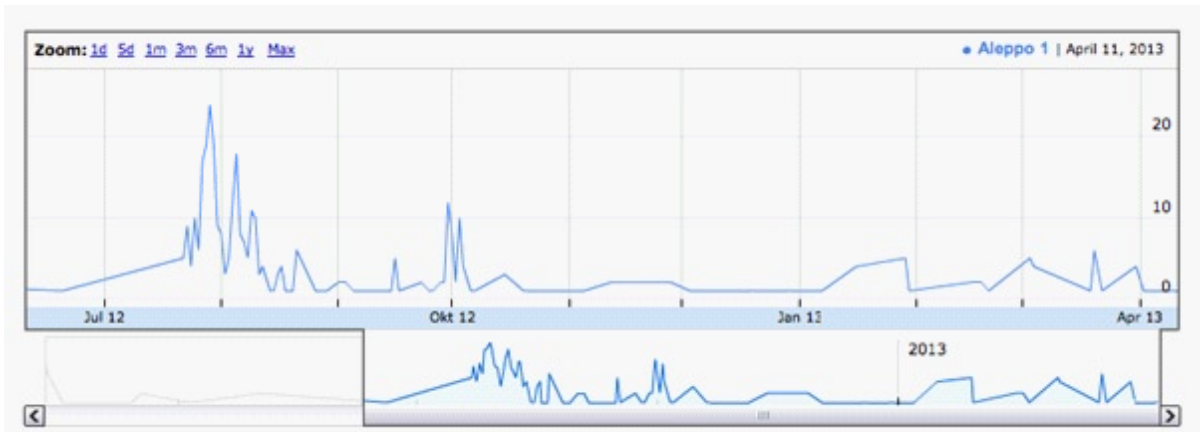


Abb. 3: Der Stadtname „Aleppo“ in den deutschen Nachrichten



Abb. 4: Die 1.000 häufigsten geographischen Orte in der WhoIsInTheNews Datenbank

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) in der DGD2

„Auf *wegen* folgt der Genitiv“, „In Nebensätzen mit *weil* steht das flektierte Verb an letzter Stelle“ – das sind Regeln, die man in jeder Schulgrammatik finden kann und die sich durch einen Blick in ein schriftsprachliches Korpus (wie DeReKo oder das DWDS-Kernkorpus) auch weitestgehend bestätigen lassen. Wie das folgende Beispiel zeigt, sind es aber auch Regeln, an die sich ein Sprecher des Deutschen nicht unbedingt halten muss:

nur Aufschluss darüber geben, inwieweit schriftsprachliche Normen in der gesprochenen Sprache eingehalten werden. Es kann beispielsweise auch verwendet werden, um Mechanismen des mündlichen Formulierens (man betrachte z.B. den Einschub in Zeile 06 und das darauf folgende Wiederansetzen des *dass*-Satzes) oder der Gesprächsorganisation (man betrachte z.B. wie die Sprecher SZ und BS sich die Arbeit teilen, die Frage nach dem Judotraining zu stellen und zu mo-

01	SZ	Aber wie isch_en des
02		ähm hascht du nochmal mit der mudder gesprochn wegen dem judo jetzt
03		äh zum beispiel (.) soll des weiter stattfinden
04	HM	des judo soll weiter stattfinden
05	BS	weil gestern hat er zu mir im bus gesagt
06		dass er also so hat_s rausgeklungen °h
07		dass er halt äh total (.) donnerstags total überfordert is

Ausschnitt aus FOLK_E_00026 (Meeting in einer sozialen Einrichtung), leicht vereinfacht.

Das Beispiel stammt aus dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), das seit 2008 in der Abteilung Pragmatik des Instituts für Deutsche Sprache aufgebaut wird. FOLK ist ein Gesprächskorpus, d.h. eine Sammlung von Aufnahmen authentischer Gespräche, die für eine sprachwissenschaftliche Analyse aufbereitet werden. Ein solches Korpus kann nicht

tivieren) zu untersuchen – beides sind typische Fragen der Gesprächsforschung. Verglichen mit schriftsprachlichen Ressourcen, bei denen analysierbares Material mittlerweile weitestgehend automatisiert akquiriert und für die Nutzung aufbereitet werden kann, ist der Aufbau eines solchen Gesprächskorpus ist mit sehr hohem Aufwand verbunden. Um überhaupt Aufnahmen authentischer

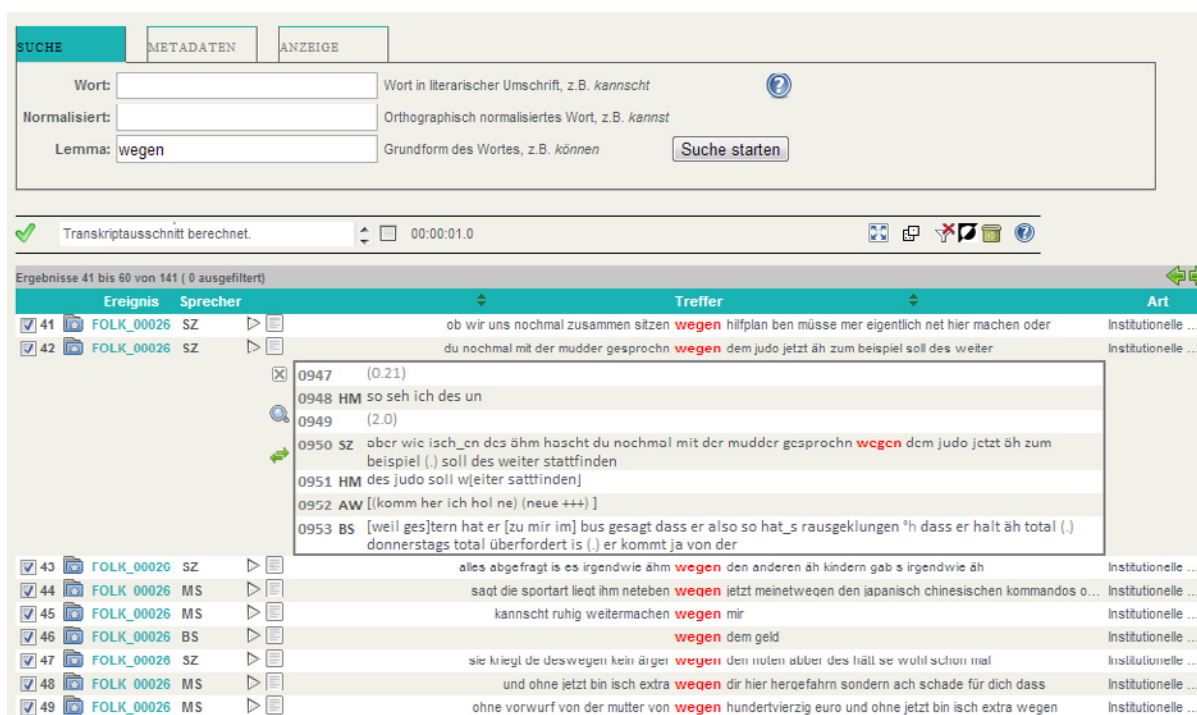
The screenshot shows the DGD website interface. At the top, there are navigation tabs: ÜBER DIE DGD, KORPORA, RECHERCHE, DOWNLOAD, HILFE, FAO, and ABMELDEN. The main header area contains the DGD logo and the text 'DATENBANK FÜR GESPROCHENES DEUTSCH'. Below this, there is a sidebar with a list of corpora (BB, DS, EK, FOLK, FR, HI, IS, ISW, ISZ, KII, MV, OS, PF, SA, SR, SV, SW, ZW) and their descriptions. The main content area displays the metadata for 'Ereignis FOLK_E_00026'. It includes a table with the following data:

Basisdaten	
Beschreibung	Meeting in einer sozialen Einrichtung
Sonstige Bezeichnungen	FOLK_MEET_01_A02
Datum	2009-03-05
Ort	Land: Deutschland Region: Hessische Sprachregion
Institution / Räumlichkeiten	Soziale Einrichtung / Büro
Aufnahmebedingungen	Nicht dokumentiert
Sprechereignisse und Sprecher	
1 Sprechereignis	FOLK_E_00026_SE_01 (Institutionelle Kommunikation: Meeting in einer sozialen Einrichtung)
Themen	vgl. FOLK_E_00026_SE_01_Z_01 ▶
5 dokumentierte Sprecher	FOLK_S_00046 ▶ (Mitarbeiterin; Arbeitskollegin in FOLK_E_00026_SE_01) FOLK_S_00047 ▶ (Gruppenleiter in FOLK_E_00026_SE_01) FOLK_S_00048 ▶ (Mitarbeiterin; Arbeitskollegin in FOLK_E_00026_SE_01) FOLK_S_00049 ▶ (Mitarbeiterin; Arbeitskollegin in FOLK_E_00026_SE_01) FOLK_S_00060 ▶ (Freiwillige in FOLK_E_00026_SE_01) FOLK_S_00061 ▶ (Praktikantin in FOLK_E_00026_SE_01)
Korpusbestandteile	

Abbildung 1: Anzeige von Metadaten zum einem Gesprächsereignis aus FOLK in der DGD2

Gespräche machen zu können, muss zunächst für einen geeigneten „Feldzugang“ gesorgt werden, d.h. es müssen Personen gefunden werden, die bereit sind, sich bei interessanten Gesprächsanlässen aufnehmen zu lassen und diese Aufnahmen der Wissenschaft zur Verfügung zu stellen. Ist eine solche Aufnahme dann gemacht, muss sie im Projekt „von Hand“ transkribiert und (aus Gründen des Datenschutzes) so maskiert werden, dass kein direkter Rückschluss auf die Identität der aufgenommenen Personen mehr möglich ist. Insgesamt stecken somit in einer Stunde Gesprächsaufnahme zwischen 50 und 100 Stunden manueller Aufbereitungsarbeit. FOLK ist daher mit seinen 100 h Aufnahmen, die Ende

2013 erreicht werden sollen, bereits eines der größten Korpora seiner Art, auch wenn es sich mit ca. 1 Millionen transkribierter Wörter im Vergleich zu einem schriftsprachlichen Korpus (das DWDS-Kernkorpus umfasst z.B. 100 Millionen Wörter) eher klein ausnehmen mag. Beim Aufbau von FOLK wird das Ziel einer „breiten Stratifizierung“ in Bezug auf Gesprächstypen verfolgt, d.h. das Hauptaugenmerk liegt darauf, möglichst viele unterschiedliche Gesprächstypen (statt z.B. möglichst viele Instanzen eines bestimmten Typs) abzudecken. Derzeit enthält FOLK vornehmlich Gespräche aus der Alltagskommunikation (z.B. Paargespräche, Tischgespräche, Eltern-Kind-Interaktionen, Gespräche bei All-



The screenshot shows the DGD2 search interface. At the top, there are tabs for 'SUCHE', 'METADATEN', and 'ANZEIGE'. The search form contains fields for 'Wort:' (filled with 'wegen'), 'Normalisiert:', and 'Lemma:'. A 'Suche starten' button is visible. Below the search form, there is a status bar indicating 'Transkriptausschnitt berechnet.' and a timer '00:00:01.0'. The main area displays search results for 'wegen' in the FOLK corpus. The results are organized into columns: 'Ergebnis', 'Sprecher', 'Treffer', and 'Art'. A detailed view of a transcript segment is shown, listing time points (e.g., 0947, 0948, 0949) and speaker initials (e.g., HM, SZ, AW, BS) along with their corresponding speech segments. The word 'wegen' is highlighted in red in the transcript snippets.

Abbildung 2: Resultat einer DGD2-Suche nach „wegen“ in FOLK

tagsaktivitäten) und aus institutioneller Kommunikation (z.B. Teambesprechungen bei der Arbeit, Schulunterricht, Prüfungsgespräche an der Hochschule). Dazu kommen demnächst auch Aufnahmen aus der öffentlichen Kommunikation, z.B. Mitschnitte der Schlichtungsverhandlungen zu Stuttgart 21. Da das Ziel einer breiten Stratifizierung kaum einzuhalten ist, wenn alle Aufnahmen im Projekt selbst gemacht werden, lebt FOLK auch von Datenspenden aus anderen Forschungsprojekten. So wurden zum Beispiel die FOLK-Daten zu den Prüfungsgesprächen an der Hochschule vom Leipziger GeWiss-Projekt (jetzt auch ein CLARIN-Kurationsprojekt) zur Verfügung gestellt, und das Projekt „Sprachvariation in Norddeutschland“ (jetzt betreut vom CLARIN-Zentrum Hamburg) hat Daten von Tischgesprä-

chen aus dem norddeutschen Raum gespendet, die derzeit aufbereitet werden. FOLK wird interessierten Forschenden und Studierenden über die Datenbank für Gesprochenes Deutsch (DGD2, dgd.ids-mannheim.de) zur Verfügung gestellt. Die DGD2 ermöglicht registrierten Benutzern ein Browsing in allen zum Korpus gehörigen Datentypen – also Aufnahmen, Transkripten, Dokumentationen von Sprechern und Gesprächsereignissen etc. – sowie die gezielte Recherche in Metadaten und Transkripten. Neben FOLK enthält die DGD2 weitere 17 Korpora des gesprochenen Deutsch, darunter mit dem „Freiburger Korpus“ und dem Korpus „Dialogstrukturen“ aus den 1970er Jahren auch die beiden ersten Gesprächskorpora des Deutschen überhaupt, sowie zahlreiche „Klassiker“ der Dialekt- und Variationsforschung

wie das Korpus „Deutsche Mundarten“ (auch bekannt als Zwirner-Korpus) und das Korpus „Deutsche Umgangssprachen“ (auch bekannt als Pfeffer-Korpus).

Als Zentrum im CLARIN-Verbund wird das IDS auch FOLK und die DGD2 schrittweise in die CLARIN-Infrastruktur integrieren. Metadaten zu Korpora und Gesprächen werden dann beispielsweise als CMDI-Daten in CLARIN-Katalogen wie dem VLO zur Verfügung stehen, und die DGD2 wird Anfragen aus einer Federated Content Search verarbeiten können. FOLK und die DGD2 sind allerdings auch schon vor dieser direkten Anbindung an CLARIN ein Beispiel dafür, wie eine digitale Ressource in den Geisteswissenschaften davon profitiert, dass andere Standorte ihre digitalen Tools und Ressourcen der Community zur Verfügung stellen. So basiert der Editor FOLKER, der im FOLK-Projekt für die Transkription verwendet wird, auf dem vom CLARIN-Zentrum HZSK (Hamburg) bereitgestellten EXMARaLDA-System und nutzt zum Abspielen von Audios einen vom CLARIN-Zentrum BAS München entwickelten Player. Die Lemmatisierung und das POS-Tagging von FOLK werden nach dem Stuttgart-Tübingen-Tagset (STTS) mit Hilfe des am CLARIN-Zentrum IMS Stuttgart entwickelten TreeTaggers vorgenommen. Und nicht zuletzt kommen beim orthographischen Normalisieren der Daten die DeReWo-Wortlisten zum Einsatz, die das CLARIN-Zentrum IDS Mannheim auf seinen Webseiten anbietet. Selbstverständlich wird daher auch

das FOLK-Projekt nicht nur das Korpus selbst, sondern auch die im Projekt entwickelten Tools (neben FOLKER derzeit auch das Tool OrthoNormal zum orthographischen Normalisieren) in der CLARIN-Infrastruktur anbieten. Das Projekt beteiligt sich darüber hinaus zusammen mit anderen Partnern aus dem CLARIN-Verbund auch an Standardisierungsinitiativen im Bereich mündlicher Korpora.

Um sich für die Nutzung von FOLK in der DGD2 zu registrieren, besuchen Sie bitte die DGD2-Website unter dgd.ids-mannheim.de. Die im FOLK-Projekt entwickelten Tools sind über agd.ids-mannheim.de/folker.shtml erhältlich. Weitere Informationen zum FOLK-Korpus erhalten Sie auch unter agd.ids-mannheim.de/folk.shtml oder über die E-Mail-Adresse folk@ids-mannheim.de.

Thomas Schmidt,
Institut für Deutsche Sprache



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
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
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
NSF	National Science Foundation (USA)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
PID	Persistent Identifier
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
VLC	Virtual Linguistic Campus
VLO	Virtual Language Observatory
WADL	Web Application Description Language