Anbieter 3. Fachmesse zur Korpustechnologie

am 1. April 2011 von 14-17 Uhr im Vortragssaal des IDS
direkt im Anschluss an die AGF 2011

Archiv für Gesprochenes Deutsch (AGD)	Ulf-Michael Stift (Mannheim)
Datenbank Gesprochenes Deutsch (DGD 2.0)	Joachim Gasch (Mannheim) Sylvia Dickgießer (Mannheim)
Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)	Jenny Kunz (Mannheim)
Transkriptionseditor FOLKER	Wilfried Schütte (Mannheim)
Gesprächsanalytisches Informationssystem (GAIS)	Wilfried Schütte (Mannheim)
EXMARaLDA und das Hamburger Zentrum für Sprachkorpora	Thomas Schmidt (Hamburg) et al.
IDS-Programmbereich "Forschungsinfrastrukturen" (FI)	Andreas Witt (Mannheim) Oliver Schonefeld (Mannheim)
Bayerisches Archiv für Sprachsignale (BAS)	Christoph Draxler (München)
[moca2] – multimodal oral corpus administration	Daniel Alcón López (Freiburg) Oliver Ehmer (Freiburg)
ELAN – multimedia annotation tool	Han Slöetjes (Nijmegen)
ANNEX – Annotation Explorer	Lari Lampen (Nijmegen)
gi – Gesprächsanalyse interaktiv	Wolfgang Kesselheim (Zürich) Katrin Lindemann (Zürich)

Archiv für Gesprochenes Deutsch (AGD)

In der Linguistik gewinnen Sprachkorpora eine immer größere Bedeutung für die Forschung. Vor allem für Gesprächskorpora gilt jedoch, dass ihre Erstellung sehr aufwändig ist und umfassende technische und methodische Kenntnisse erfordert. Deshalb bietet das Archiv für Gesprochenes Deutsch (AGD) am IDS Mannheim vielfältige Unterstützung an, um die Erstellung und wissenschaftliche Auswertung von Gesprächskorpora in Forschung und Lehre zu fördern.

Zentrale Aufgabe des Archivs ist es, Korpora aus abgeschlossenen Forschungsprojekten zu übernehmen und für zukünftige Forschung und Lehre in der Wissenschaftsgemeinschaft zu erhalten. Auf diese Weise sind in 60 Jahren ca. 45 Korpora mit über 5000 Stunden Gesamtdauer gesammelt worden.

Um den wissenschaftlichen Bedürfnissen noch umfassender Rechnung tragen zu können, werden zur Zeit in der Abteilung "Pragmatik" zwei moderne Forschungskorpora aufgebaut. Zum einen ist dies das Korpus "Deutsch heute", das die Variation des standardnahen Deutsch, wie es heutzutage im deutschen Sprachraum gesprochen wird, auf der Basis von 170 Erhebungsorten dokumentiert. Zum anderen entsteht in den kommenden Jahren das "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK), das einen breiten Querschnitt von Transkripten, Ton- und Videoaufnahmen aus unterschiedlichsten Gesprächstypen in deutscher Sprache umfassen wird. Beide Korpora werden in den kommenden Jahren über die "Datenbank Gesprochenes Deutsch" (DGD), die zur Zeit umfassend überarbeitet und auf eine neue ORACLE-Plattform umgesetzt wird, online zur Verfügung gestellt.

Eine weitere wichtige Aufgabe des Archivs ist es, die wissenschaftliche Gemeinschaft bei der Erstellung von Gesprächskorpora zu beraten und Informationen aller Art rund um die Korpustechnologie im "Gesprächsanalytischen Informationssystem" (GAIS) anzubieten. Um Anbieter und Nutzer dieser Technologie mit einander ins Gespräch zu bringen, veranstaltet das AGD Fachmessen und Kolloquien. Darüber hinaus beteiligt es sich auch aktiv an der Entwicklungsarbeit, z.B. mit dem neuen Transkriptionseditor FOLKER, dem Metadaten-Schema für die neue DGD und der Unterstützung von GAT 2.

Kontakt: pragmatikservice@ids-mannheim.de

Webadresse: http://agd.ids-mannheim.de

Datenbank für Gesprochenes Deutsch (DGD 2.0)

Gegenwärtig wird im Archiv für Gesprochenes Deutsch (AGD) ein neues Korpusmanagementsystem (DGD 2.0) entwickelt, von dem es zwei Instanzen geben wird. Die erste Instanz (DGD 2.0 - intern), an der wir vorrangig arbeiten, zielt auf die IDS-interne Korpusverwaltung und wird neben den Informationen, die in der DGD 1.0 bereitgehalten werden, auch Daten aus neueren IDS-Projekten umfassen. In einem zweiten Schritt wird aus der internen Instanz eine Instanz für externe Nutzer (DGD 2.0 - extern) abgeleitet, die mittelfristig die zur Zeit im Internet zugängliche DGD 1.0 ablösen soll. Für die DGD 2.0 wurde ein Metadatenstandard entwickelt, der auf einem neuen Datenmodell für die Dokumentation von Korpora der gesprochenen Sprache beruht und vier darauf aufbauende, korpusübergreifende XML-Schemata umfasst. Im Mittelpunkt der Systemarchitektur der DGD 2.0 steht eine objektrelationale XML-Datenbank (Oracle 11g). Die XML Technologie ermöglicht die schemabasierte, native Speicherung von Metadaten und Transkripten. Das Datenbanksystem unterstützt Volltextsuche und kontextsensitives Informationsretrieval sowie standardisierte, dynamisch generierte Präsentationen von XML-Dokumenten.

Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

Während inzwischen verschiedene Korpora des geschriebenen Deutsch wissenschaftsöffentlich zugänglich sind, gibt es noch keine vergleichbare Sammlung des gesprochenen Deutsch. Mit FOLK baut das IDS ein kontinuierlich wachsendes Korpus auf, welches Gespräche aus den unterschiedlichsten Bereichen des gesellschaftlichen Lebens im deutschen Sprachraum (Arbeit, Freizeit, Bildung, Medien) via Internet zugänglich macht. Für viele linguistische und gesprächsanalytische Untersuchungen wird damit die Notwendigkeit entfallen, eigene Korpora zu erstellen. Kultur- und medienwissenschaftliche Untersuchungen können durch FOLK vielfältige Einblicke in die Realität der sozialen Kommunikation im Deutschland der Gegenwart gewinnen. Der germanistischen Hochschullehre und dem Unterricht im Fach DaF bietet FOLK Anschauungsbeispiele des heutigen gesprochenen Deutsch.

Alle Aufnahmen sind nach GAT 2 transkribiert. Die Transkripte sind mit dem Ton aligniert – jede Transkriptstelle kann augenblicklich angehört werden. Jedes Gespräch verfügt über eine umfassende Metadaten-Dokumentation von Gesprächsumständen und soziodemographischen Sprecherdaten. Diese sind wie die Transkripte nach verschiedensten Parametern zu recherchieren. Suchergebnisse werden als KWIC angezeigt und können personalisiert werden. Die Zugänglichkeit der Korpusbestandteile ist nach Datenschutzerfordernissen gestuft. Ausgewählte Gespräche können vollständig heruntergeladen werden. Metadaten und Transkripte sind XML-kodiert.

FOLK wird ab 2011 über die DGD 2.0 via Internet passwortgeschützt nach personalisierter Anmeldung ausschließlich für Forschungs- und Lehrzwecke zugänglich sein.

Ansprechpartner: Jenny Kunz (kunz@pragmatik.ids-mannheim.de)

Webadresse: http://agd.ids-mannheim.de/html/folker.shtml

Transkriptionseditor FOLKER

Um Transkripte für das neue "Forschungs- und Lehrkorpus" (FOLK) schnell und komfortabel erstellen zu können und in der Datenbank DGD 2.0 optimal auswertbar und darstellbar zu machen, musste ein neuer Editor mit einem konsistenten Datenmodell und einem konsequenten XML-Format entwickelt werden. Seit 2010 steht der Editor auf der Website des AGD (agd.ids-mannheim.de) zum kostenlosen Download zur Verfügung. Die aktuelle Version mit multilingualer Benutzeroberfläche ist 1.1. Entwickelt wurde das Programm von Thomas Schmidt (Hamburg) in enger Abstimmung mit der Abteilung "Pragmatik" am IDS.

In den Editor sind ein Audioplayer und eine Darstellung des Sprachsignals als Oszillogramm eingebaut, um so effizient Segmente zum Transkribieren auswählen und beim Transkribieren automatisch Zeitmarken für diese Segmente setzen zu können. FOLKER überprüft segmentweise die Texteingabe auf Einhaltung der GAT-Konventionen und Integrität der zeitlichen Strukturen (keine Überlappungen bei Segmenten eines Sprechers). Im Editor kann das Transkript wahlweise als Folge von Segmenten, als Partitur (in Anlehnung an die Visualisierung im EXMARaLDA-Editor) oder als Folge von Sprecherbeiträgen dargestellt werden, bei denen einem Sprecher zugeordnete und aufeinander folgende Segmente zusammengefasst werden. Diese Darstellungsmöglichkeiten passen zu einer sinnvollen Abfolge von Arbeitsschritten beim Transkribieren: Die Ersteingabe eines Rohtranskripts ohne Berücksichtigung der genauen Extension von Überlappungen findet zweckmäßiger im Segment-View statt, im Partitur-View können dann in einer Korrektur-Durchsicht die Feinheiten von Überlappungen korrigiert werden; der Beitrags-View dient dem abschließenden Korrekturhören.

Ansprechpartner: Dr. Wilfried Schütte (schuette@ids-mannheim.de)

Webadresse: http://agd.ids-mannheim.de/html/folker.shtml

Gesprächsanalytisches Informationssystem (GAIS)

Das Gesprächsanalytische Informationssystem GAIS ist ein wissenschaftliches Fachinformationssystem für die Arbeit in der Gesprächsforschung und für die wissenschaftliche Gemeinschaft. GAIS wurde in der Abteilung Pragmatik des Instituts für Deutsche Sprache (IDS) aufgebaut und wird dort inzwischen vom Archiv für Gesprochenes Deutsch betreut. Adressat von GAIS ist die wissenschaftliche Gemeinschaft von fortgeschrittenen Studierenden bis Lehrstuhlinhabern und über Fachgrenzen hinweg (Linguisten, Soziologen, Sprechwissenschaftler, Psychologen, Pädagogen). Nutzerbefragungen und Erfahrungen mit Service- und Beratungsanfragen an das AGD und an die Mailliste „Gesprächsforschung" haben ergeben, welche Informationen in diesem Kreis immer wieder nachgefragt werden. Dazu gehören Informationen über die Gemeinschaft selbst (Neuigkeiten, Personen, Veranstaltungen, Stellenmarkt, Mailliste, Projekte, Korpora) als auch über ihre Arbeitsweise (Aufnahme- und Korpustechnologie). GAIS bietet weiterführende Informationen in Form einer Bibliographie und einer Linksammlung an.

GAIS umfasst u.a. drei Informationsbereiche:

- „Community" mit Informationen aus der wissenschaftlichen Gemeinschaft (Veranstaltungen, aktuelle Meldungen, Sammlung von persönlichen Webseiten, Stellenmarkt und Mailliste zur Gesprächsforschung mit inzwischen über 1200 Teilnehmern).
- Bibliographie zur Gesprächsforschung (BGF) mit inzwischen über 20.000 Einträgen und einer Sammlung mit relevanten Links.
- Technik mit Informationen zur Aufnahmetechnik (Audio, Video, Mikrofone, Kabel und Stecker), zur Korpusbearbeitung (Hard- und Software mit einem Tutorium zu Praat, Digitalisierung), Transkription (Konventionen und Editoren) und Korpusverwaltung (Datenbanken und Konvertierprogramme). Häufige Fragen werden unter FAQ beantwortet und technische Fachbegriffe unter „Fachbegriffe" erklärt.

Kontakt: Dr. Wilfried Schütte (schuette@ids-mannheim.de)

Webadresse: http://gais.ids-mannheim.de

EXMARaLDA und das Hamburger Zentrum für Sprachkorpora

EXMARaLDA ist ein System zu Erstellen, Verwalten, Auswerten und Publizieren von Korpora gesprochener Sprache. Es wird seit Juli 2000 im Sonderforschungsbereich 538 Mehrsprachigkeit der Universität Hamburg entwickelt. EXMARaLDA findet vornehmlich Anwendung in der Gesprächsforschung, in der Spracherwerbsforschung, in der Dialektologie und in der Phonetik/Phonologie. Zu den mit EXMARaLDA erstellten Korpora gehören beispielsweise das Korpus der gesprochenen Sprache im Ruhrgebiet, das METU Corpus of Spoken Turkish und das Korpus „Sprachvariation in Norddeutschland".

Auch am SFB selbst wird EXMARaLDA zu Erstellen und Aufbereiten mehrsprachiger Korpora verwendet. Zum Abschluss des SFB im Juni 2011 werden dessen Datenbestände insgesamt 17 Korpora gesprochener Sprache umfassen. Um diese wertvollen Ressourcen weiterhin verfügbar zu halten, wurde an der Universität Hamburg das Hamburger Zentrum für Sprachkorpora gegründet, das sich mittelfristig als Zentrum in der CLARIN-Infrastruktur etablieren soll.

Das Angebot auf der Fachmesse umfasst zum einen Demonstrationen der Werkzeuge des EXMARaLDA Systems (Partitur-Editor, Corpus Manager, Analysewerkzeug EXAKT). Zum anderen werden Konzept und Datenbestände des Hamburger Zentrums für Sprachkorpora vorgestellt.

Forschungsinfrastrukturen (FI)

Der vorgestellte Programmbereich befasst sich mit der Optimierung bei der Erstellung von und dem Umgang mit digitalen Sprachdaten sowie der Förderung der Nachhaltigkeit in diesem Bereich. Dabei setzt er sich für eine intensivierte Zusammenarbeit mit vergleichbaren Kompetenzzentren ein und repräsentiert das IDS als Kooperationspartner in verschiedenen Verbundinitiativen. Zu diesen zählen mehrere BMBF-geförderte Projekte wie z.B.

* D-SPIN, die Deutsche Sprachressourcen-Infrastruktur im CLARIN-Verbund, die sich mit der Entwicklung von Grundlagen für eine stabile und nachhaltige Infrastruktur von Sprachressourcen und Sprachtechnologien durch Errichtung von dedizierten Zentren, deren Versorgung mit Meta- und/oder Primärdaten sowie der Pflege dieses Struktur-Netzwerks beschäftigt,

* TextGrid, eine vernetzte Forschungsumgebung in den eHumanities und Teil der Deutschen Grid-Initiative D-Grid gGmbH, deren erklärtes Ziel es ist, eine Internet-Plattform aufzubauen, die Wissenschaftlern Werkzeuge und Dienste für die Auswertung von textbasierten Daten in unterschiedlichen digitalen Archiven unabhängig von Datenform, Standort oder Softwareausstattung bietet,

* WissGrid, eine Initiative, die sich dafür einsetzt, eine nachhaltige Etablierung von organisatorischen und technischen Strukturen für den akademischen Bereich durchzusetzen, heterogene Anforderungen aus verschiedenen wissenschaftlichen Disziplinen zu bündeln, die Entwicklung konzeptioneller Grundlagen für die nachhaltige Nutzung der Grid-Infrastruktur sowie IT-Lösungen voranzutreiben und somit eine Dachfunktion innerhalb der Deutschen Grid-Initiative D-Grid gGmbH besetzt, oder

* Nestor, das deutsche Kompetenznetzwerk zur digitalen Langzeitarchivierung, ein Kooperationsverbund, in dem Bibliotheken, Archive, Museen sowie führende Experten gemeinsam zum Thema Langzeitarchivierung und Langzeitverfügbarkeit digitaler Quellen zusammenarbeiten.

Das Bayerische Archiv für Sprachsignale (BAS)

Das Bayerische Archiv für Sprachsignale (BAS) stellt Sprachdatenbanken und Tools zur Verarbeitung gesprochener Sprache zur Verfügung und bietet sprachbezogene Dienstleistungen an. Auf der diesjährigen Fachmesse "Korpustechnologie" stellt das BAS aktuelle Sprachdatenbanken sowie das Softwarepaket WebExperiment vor.

Mit WebExperiment können auf einfache Weise online-Perzeptionsexperimente mit Audio (und Video) durchgeführt werden. WebExperiment zeichnet sich dadurch aus, dass es ausschließlich auf offen spezifizierten Standards (HTML5) basiert und in jedem modernen Browser ohne zusätzliche Erweiterungen läuft. Damit ist es möglich, das Experiment auch auf neuartigen Eingabegeräten wie Smartphones, Tablet-Rechnern oder Internet-TV Geräten durchzuführen und somit neue Teilnehmerkreise zu erschließen.

Auf der Fachmesse stellen wir das online Experiment zur regionalen Zuordnung von Lautmerkmalen in gesprochenen Einzelziffern vor. Probanden hören sich ca. 40 Ziffern an und entscheiden, welche Lauteigenschaft sie gehört haben, z.B. "ß" wie in "reißen" oder "s" wie in "reisen". Ergebnis des Experiments ist eine Karte, in der die Eingaben farblich kodiert sind – in der Regel ergeben sich klar abgegrenzte Regionen.

Die Software ist frei verfügbar. Altenativ bietet das BAS an, online Experimente für akademische und industrielle Forschungsprojekte durchzuführen.

[moca2] – multimodal oral corpus administration

[moca2] ist ein Online-System zur Verwaltung mündlicher Sprachkorpora. In [moca2] werden Audio- und/oder Videoaufnahmen sowie zugehörige Transkripte gespeichert. Die Transkripte liegen in alignierter Form vor, was bedeutet, dass mit dem Text der Sprechbeiträge auch die Sprecher- und Zeitinformation erfasst wird. Hierdurch ist es möglich, in einem Internetbrowser direkt die entsprechende Aufnahme zu einer Transkriptstelle als Mediastream abzuspielen. Neben den Transkripten können auch soziolinguistische Metainformationen zur Aufnahmesituation und den beteiligten Sprechern strukturiert verwaltet werden. Über die Vergabe sogenannter Labels für Äußerungen (manuelles Tagging) können umfangreiche Kollektionen eines linguistischen Phänomens erstellt und ausgewertet werden.

Detaillierte Suchmöglichkeiten erlauben es, bestimmte Aufnahmen, Sprecher, Transkriptausschnitte und Labels zu finden. Beispielsweise ist es möglich, aus den vorhandenen Daten Aufnahmen aus einer bestimmten Region auszuwählen, um Analysen darauf zu beschränken, oder nach Sprechern zu suchen, die einer bestimmten Altersgruppe angehören. Darüber hinaus ist es möglich, in Transkripten nach Intonationsphrasen zu suchen, die bestimmte (Kombinationen oder Teile von) Wortformen enthalten.

Ziel von [moca2] ist dabei, einen intuitiven, sicheren und personifizierten Zugang zu den Korpora zu gewährleisten. Dabei unterstützt das System eine unbegrenzte Anzahl von Nutzern, denen individuell der Zugriff auf bestimme Daten gestattet oder verweigert werden kann. [moca2] kann von praktisch jedem internetfähigen Computer genutzt werden, ohne dass besondere technische Anforderungen oder Kenntnisse erforderlich sind.

Kontakt:

Daniel Alcón López daniel.alcon@romanistik.uni-freiburg.de
Oliver Ehmer oliver.ehmer@romanistik.uni-freiburg.de

URL: http://moca.phil2.uni-freiburg.de/moca_test

ELAN

ELAN is a multimedia annotation tool that it is being developed and maintained by the “The Language Archive” department of the Max Planck Institute for Psycholinguistics in Nijmegen. It is available for Windows, Mac OS X and Linux, can be downloaded free of charge and is open source.

ELAN is a desktop tool for manually annotating audio and/or video recordings. It supports up to 4 videos per annotation document. It is a multi-layered annotation system: annotations are contained in tiers and the tiers can be part of a tier hierarchy (i.e. tiers can have depending tiers). The annotations contain Unicode text and the annotation documents are stored in XML files (EAF). It provides facilities for searching in a single file or in a user definable set of files (local corpus). Some other multiple file operations are available as well and new ones are in preparation. Recently a preliminary level of interaction with the MPI lexicon tool Lexus has been implemented. ELAN is complemented by Annex, an online tool for web-based archive exploration.

ELAN is applied in a variety of research areas in linguistics (and beyond) such as sign language research, field linguistics, conversation analysis, gesture studies and multimodal interaction research. At the Fachmesse ELAN is presented as a demo.

Web: http://www.lat-mpi.eu/tools/elan/

Contact: Han Sloetjes (han.sloetjes@mpi.nl)

ANNEX

Annex (Annotation Explorer) is an online tool for web-based archive exploration. It is part of the web-based tool framework of the Max Planck Institute for Psychonliguistics, along with tools for lexicon exploration and creation (Lexus), ingestion of files in the archive (Lamus) and browsing the hierarchical archive of metadata records (IMDI Browser).

Annex provides different views of the annotations, analogous to the stand-alone tool Elan (also featured at the Fachmesse), and can stream media inside a browser window using the Adobe Flash plugin. It allows sections of annotations to be selected and played. Being web-based, it provides a quick and easy way to view annotations in the archive without having to download software or files from the archive.

For more information, see the web site at: http://www.lat-mpi.eu/tools/annex/

gi – Gesprächsanalyse interaktiv

„gi – Gesprächsanalyse interaktiv“ ist ein E-Learning-Angebot, das das Prinzip des kollaborativen forschenden Lernens einsetzt, um Studierende in die Gesprächsanalyse einzuführen. „gi“ wird seit September 2008 am Deutschen Seminar der Universität Zürich, Lehrstuhl Hausendorf, entwickelt und dort seit 2010 in der Lehre eingesetzt. Mit „gi“ wird eine Vermittlungsform etabliert, die den besonderen Anforderungen gerecht wird, die sich aus den methodologischen Grundannahmen der Gesprächsanalyse ergeben.

In „gi“ lernen die Studierenden die Bestandteile des gesprächsanalytischen Forschungsprozesses kennen, indem sie in begleiteten Arbeitsgruppen ein eigenes kleines Forschungsprojekt in all seinen Teilschritten durchführen. Mit seiner modularen Struktur bildet der Kurs den Forschungsprozess ab und setzt auf eigenes Forschen anstelle des reinen Nachvollziehens von Forschungsergebnissen. Ein wichtiger Mehrwert liegt in der größeren Nachhaltigkeit des Gelernten sowie einer in der Präsenzlehre kaum zu erreichenden Übersicht über den Forschungsprozess.

In „gi“ werden den Studierenden multimediale Lehrtexte mit authentischen Audio-/Videoaufnahmen zur Verfügung gestellt, mit denen sie theoretisches Wissen und praktische Fertigkeiten im Selbststudium erwerben können. Zentral ist jedoch etwas anderes: „gi“ nutzt die Möglichkeiten des Internets, um *kollaboratives* Forschen zu ermöglichen, und legt den Fokus auf interaktives Lernen. Für jede Forschungsphase stellt „gi“ daher Werkzeuge zur Verfügung, die das gemeinsame Arbeiten strukturieren (z.B. Wiki, Forum, Chat).

„gi“ verwendet die *open source* Lernplattform OLAT und kann daher auch außerhalb der Universität Zürich – sowohl im universitären als auch im außeruniversitären Bereich – eingesetzt werden. Die multimedialen Lerninhalte können auch in andere Lernplattformen integriert werden.

Kontakt:

katrin.lindemann@ds.uz.ch

wolfgang.kesselheim@ds.uzh.ch

„gi“ finden Sie unter www.ds.uzh.ch/gi