

Fachmesse für mündliche Sprachkorpora

im Rahmen der Arbeitstagung zur Gesprächsforschung 2013

Institut für Deutsche Sprache, R5, 6-13, 68161 Mannheim, Vortragssaal (Erdgeschoss)

Freitag, 22. März 2013, 14.00 – 17.00 Uhr

Präsentiert werden Software, Plattformen, Datenbanken sowie Korpora von Dienstleistern und IT-Experten rund um die Erstellung und Auswertung von (mündlichen) Sprachkorpora.

Übersicht der Abstracts:

1.	audiotranskription.de – Lösungen zur digitalen Aufnahme und Transkription	1
2.	Bayerisches Archiv für Sprachsignale (BAS)	2
3.	EXMARaLDA – Transkription, Annotation und Analyse für Korpora gesprochener Sprache	2
4.	„gi – Gesprächsanalyse interaktiv“	3
5.	Korpora und Korpustechnologie im Programmbereich „Mündliche Korpora“ am IDS.....	4
	Archiv für Gesprochenes Deutsch (AGD)	4
	Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)	5
	FOLKER und OrthoNormal.....	5
	Datenbank für Gesprochenes Deutsch (DGD2).....	5
	Gesprächsanalytisches Informationssystem (GAIS)	6
6.	The Language Archive, Max-Planck-Institut für Psycholinguistik Nijmegen	6
	ELAN – Multimedia Annotator	6
	Lexus – A Web-based Lexicon Tool	7
7.	[moca2] – multimodal oral corpus administration	7
8.	trAVis – musikzentriertes Transkriptionsprogramm für audiovisuelle Medienprodukte	8



1. audiotranskription.de – Lösungen zur digitalen Aufnahme und Transkription

Thorsten Pehl

audiotranskription.de zeigt Lösungen zur digitalen Aufnahme und Transkription von Interviews, Diktaten u.v.m. Auf unserer Website bieten wir die kostenfreien Transkriptionsprogramme f4 (Windows) und f5 (Mac), das kostenfreie Praxisbuch Transkription, umfassende Test- und Vergleichsberichte zu digitalen Rekordern und diverse Tutorials für den Arbeitsalltag an.

Ansprechpartner: Thorsten Dresing, Thorsten Pehl

Web: <http://www.audiotranskription.de/>

Kontakt: info@audiotranskription.de



2. Bayerisches Archiv für Sprachsignale (BAS)

Christoph Draxler

Das BAS erstellt, pflegt und distribuiert Sprachdatenbanken und sprachverarbeitende Software, bietet Beratung und Schulung im Bereich Sprachdatenbanken, und es ist eines von derzeit neun deutschen CLARIN-Zentren. Auf der Fachmesse Gesprächsforschung präsentiert das BAS zwei aktuelle Anwendungen: WebMAUS zur automatischen Segmentierung und Etikettierung von gesprochener Sprache und die erweiterte Version von Percy für webbasierte Perzeptionsexperimente.

WebMAUS erstellt aus einer Audiodatei und der zugehörigen orthographischen Transliteration eine Segmentierung und Etikettierung auf drei Annotationsebenen: Wortsegmentierung, kanonische Aussprache sowie tatsächlich realisierte Aussprache unter Berücksichtigung von Koartikulation und phonologischen Prozessen. WebMAUS unterstützt die Sprachen Deutsch, Englisch (inkl. schottisches und australisches Englisch), Italienisch, Niederländisch, Ungarisch und Polnisch. Die öffentlich zugängliche Webseite erlaubt das Hochladen von mehreren Audio-Transliterations-Dateipaaren und liefert die Ergebnisse als TextGrid-Dateien für die Weiterverarbeitung in Praat. WebMAUS ist bereits als Funktion in andere sprachverarbeitende Software integriert, so z.B. den Annotationseditor ELAN oder den Sprachdatenbank-Viewer SpeechFinder.

Percy ist ein Softwarepaket für webbasierte akustische Perzeptionsexperimente. Um die Definition eigener Experimente zu vereinfachen ist nun ein ebenfalls webbasierter Experiment-Editor verfügbar, mit dem man Struktur und Inhalt eines Experiments über ein grafisches Interface erstellen kann. Diese Experimentdefinition wird dann auf den Percy-Server am BAS hochgeladen, geprüft und dann als Experiment öffentlich verfügbar gemacht. Der Administrator eines Experiments hat jederzeit Zugang zu seinen eigenen Daten, so dass eine kontinuierliche Kontrolle möglich ist. Das BAS bietet die Nutzung von Percy als kostenlose Dienstleistung für akademische Nutzer an.

Web: WebMAUS: <http://clarin.phonetik.uni-muenchen.de/BASWebServices/>
Percy: <http://webapp.phonetik.uni-muenchen.de/WebExperiment/>

Kontakt: draxler@phonetik.uni-muenchen.de



hzsk hamburger zentrum
für sprachkorpora



3. EXMARaLDA – Transkription, Annotation und Analyse für Korpora gesprochener Sprache

Timm Lehmborg, Thomas Schmidt, Kai Wörner

EXMARaLDA ist ein System zum Erstellen, Verwalten, Auswerten und Publizieren von Korpora gesprochener Sprache. Es wurde vom Juli 2000 an im Sonderforschungsbereich 538 (Mehrsprachigkeit) der Universität Hamburg entwickelt und wird seit 2011 unter anderem im Hamburger Zentrum für Sprachkorpora (HZSK, <http://corpora.uni-hamburg.de>) weitergepflegt.

EXMARaLDA findet vornehmlich Anwendung in der Gesprächsforschung, der Spracherwerbsforschung, der Dialektologie und in der Phonetik/Phonologie. Zu den mit EXMARaLDA erstellten Korpora gehören beispiels-

weise das *Korpus der gesprochenen Sprache im Ruhrgebiet*, das *METU Corpus of Spoken Turkish* und das *Korpus Sprachvariation in Norddeutschland*.

Im Rahmen der Systemdemonstration werden die einzelnen Werkzeuge des EXMARaLDA-Systems (Der **Partitur-Editor**, das EXMARaLDA Analyse- und Konkordanzwerkzeug **EXAKT**, der Corpus-Manager **Coma** sowie weitere Komponenten) vorgestellt. Über diese Desktop-Anwendungen hinaus finden sich auf der EXMARaLDA-Homepage (<http://exmaralda.org>) noch weitere Handreichungen wie Stylesheets, Dokumentation und Tutorials sowie Veröffentlichungen rund um EXMARaLDA und verwandte Themen.

Das System bietet außerdem eine Vielzahl von Schnittstellen zu anderen Transkriptions- und Annotationssystemen und eine Reihe von Visualisierungsoptionen. Die Software ist Java-basiert und auf den verbreitetsten Desktop-Betriebssystemen (Windows, OSX, Linux) lauffähig. Programme sowie Quellcode sind frei verfügbar.

Das oben erwähnte HZSK fungiert als eines von neun Zentren im CLARIN-D-Verbund, in dem linguistische Daten, Werkzeuge und Dienste in einer integrierten, interoperablen und skalierbaren Infrastruktur für die Fachdisziplinen der Geistes- und Sozialwissenschaften bereitgestellt werden sollen.

Die im HZSK entwickelten Methoden und Werkzeuge sollen soweit wie möglich in die CLARIN-Infrastruktur überführt oder für die Benutzung mit der Infrastruktur angepasst werden. Die am HZSK vorgehaltenen Korpora sollen in ein Repository als Teil der CLARIN-D Infrastruktur überführt werden, was unter anderem die Pflege und Distribution der Daten erleichtert. Zum anderen ermöglicht es, Korpora anhand ihrer Metadaten über eine CLARIN-weite Suche auffindbar zu machen und mit den Daten die Dienste (insbesondere Webservices), der weiteren CLARIN-Zentren zu nutzen.

Web: <http://exmaralda.org>
<http://corpora.uni-hamburg.de>

Kontakt: exmaralda-support@uni-hamburg.de



4. „gi – Gesprächsanalyse interaktiv“

Katrin Lindemann, Adriano Sabini, Nicolas Wiedmer

Mit „gi - Gesprächsanalyse interaktiv“ wird das Ziel verfolgt, eine innovative Vermittlungsform zu etablieren, die passgenau auf die spezifischen Herausforderungen der Vermittlung der Gesprächsanalyse (und anderer qualitativer Forschungsrichtungen) zugeschnitten ist.

Nach dem Prinzip des kollaborativen forschenden Lernens werden den Studierenden alle wesentlichen Komponenten einer geschäftsanalytischen Untersuchung in Form angeleiteter Gruppenarbeit und internetgestützten Selbstlernens vermittelt.

In der „Vollversion“ wird „gi“ auf der Lernplattform OLAT zur Durchführung kollaborativen Forschenden Lernens in betreuten Blended-Learning-Lehrveranstaltungen eingesetzt; dies bildet den Schwerpunkt des Projekts. Zusätzlich werden einzelne Module für die Unterstützung von Präsenzveranstaltungen verwendet, und zudem wird „gi“ von Studierenden zur individuellen Prüfungsvorbereitung als Selbstlern-Tool („gi-light“) genutzt.

Ausserdem bieten wir „gi - light“ für externe Interessierte zum Selbststudium an. Zielgruppe sind hierbei sowohl Studierende aller Fachrichtungen als auch Doktorierende oder Dozierende aus dem In- und Ausland. Eine weitere Möglichkeit für externe Gäste ist die Teilnahme an einem Blended-Learning-Kurs der „gi“-Vollversion, mit Teilnehmern von verschiedenen Standorten und Anleitung durch einen Dozenten (für die Nutzung von „gi“ durch Externe fällt eine geringe Gebühr an).

„gi“ wird über die Lernplattform OLAT realisiert. Die multimedialen Lerneinheiten sind als CP-Bausteine implementiert und können auch in andere Lernplattformen implementiert werden. Beim Erstellen der Lerneinheiten kam die Auszeichnungssprache eLML (eLesson Markup Language) zum Einsatz.

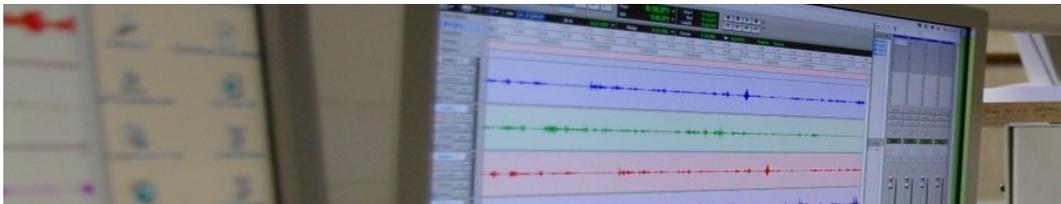
Web: <http://www.ds.uzh.ch/gi>

Kontakt: gi@ds.uzh.ch

5. Korpora und Korpustechnologie im Programmbereich „Mündliche Korpora“ am IDS

Joachim Gasch, Thomas Schmidt, Wilfried Schütte, Jenny Winterscheid

Im Programmbereich „Mündliche Korpora“ der Abteilung Pragmatik des Instituts für Deutsche Sprache Mannheim (IDS) werden Korpora gesprochener Sprache erstellt und archiviert (FOLK und AGD) sowie Werkzeuge zur Erstellung und Bearbeitung von Korpora entwickelt (FOLKER und OrthoNormal). Der Publikation von Korpora und der Recherche darin dient die Datenbank für Gesprochenes Deutsch (DGD2). Für die Dokumentation von „Best Practice“-Verfahren zur Korpuserstellung und für ein Handbuch zur Arbeit mit mündlichen Sprachkorpora steht das Gesprächsanalytische Informationssystem (GAIS) zur Verfügung.



Archiv für Gesprochenes Deutsch (AGD)

Das Archiv für Gesprochenes Deutsch (AGD) ist die zentrale Sammelstelle für Korpora des gesprochenen Deutsch. Die Korpora werden im AGD aufbereitet und für Forschung und Lehre bereitgestellt. Im AGD werden Tonaufnahmen digitalisiert und maskiert, Dokumentationsdaten nach einem einheitlichen Schema erfasst und Transkripte in nachhaltig nutzbare Formate konvertiert. Aus dem Gesamtbestand von über 40 Korpora sind derzeit 17 Bestandskorpora (mit ca. 9000 dokumentierten Ereignissen) für die DGD2 fertiggestellt. Aktuell werden im AGD z.B. Korpora zu „Jugendsprache“ und „Deutsch in Australien“ bearbeitet. Unter der Adresse agd@ids-mannheim.de beantwortet das AGD Serviceanfragen zu seinen Beständen und zur Korpustechnologie.

Web: <http://agd.ids-mannheim.de>

Kontakt: agd@ids-mannheim.de



Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

Mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) baut das IDS ein kontinuierlich wachsendes Korpus auf, um Gesprächsdaten aus unterschiedlichen Bereichen des gesellschaftlichen Lebens (Arbeit, Freizeit, Bildung, Medien) im deutschen Sprachraum via Internet zugänglich zu machen. Die Datenakquise umfasst eigene Aufnahmen, eine Übernahme aus anderen IDS-Projekten (z.B. „MapTask“-Aufnahmen und Interviews aus „Deutsch heute“), Datenspenden aus externen Projekten (u.a. von „Gesprochene Wissenschaftssprache kontrastiv“ (GeWiss) und „Sprachvariation in Norddeutschland“ (SiN)), sowie Rundfunk-Mitschnitte. Bei der Datenaufbereitung werden Aufnahmen und Transkripte maskiert, die Metadaten nach DGD-Schema eingegeben, Transkriptionen nach cGAT-Konventionen für das Minimaltranskript mit dem Editor FOLKER erstellt und die Transkripte mit OrthoNormal orthographisch normalisiert sowie mit dem TreeTagger lemmatisiert. 99 Ereignisse (ca. 70 Aufnahmestunden) sind für die DGD2 aufbereitet worden; weitere 110 Aufnahmestunden liegen vor und werden bearbeitet.

Web: <http://agd.ids-mannheim.de/folk.shtml>



FOLKER und OrthoNormal

FOLKER, Akronym für FOLK-EditoR, ist als Transkriptionseditor optimiert für die Arbeit in FOLK, mit GAT und mit FOLK-Daten. FOLKER basiert auf EXMARaLDA und ist hiermit sowie mit Praat, ELAN etc. interoperabel. FOLKER steht zum kostenlosen Download bereit; seit März 2009 gibt es über 2500 Registrierungen. FOLKER wird in der Version 1.2 eine Syntaxkontrolle für GAT-Minimal- und Basistranskripte anbieten; dafür wird derzeit schon ein Preview angeboten. OrthoNormal ist ein Tool zur orthographischen Normalisierung in FOLK, allgemeiner: zum manuellen Annotieren auf Tokenebene.

Web: <http://agd.ids-mannheim.de/folker.shtml>



Datenbank für Gesprochenes Deutsch (DGD2)

Die Datenbank für Gesprochenes Deutsch (DGD2) präsentiert eine Auswahl aus den AGD-Korpora und den gesamten FOLK-Bestand, ermöglicht ein Browsing in Metadaten, Aufnahmen, Transkripten, Zusatzmaterial sowie mehrere Recherchemöglichkeiten: eine Volltextrecherche in Metadaten und Transkripten sowie eine tokenbasierte Recherche in Transkripten, jeweils mit Rückgriff auf Transkripte und Aufnahmen. Die DGD bietet zudem einen Download ausgewählter kompletter Datensätze.

Web: <http://dgd.ids-mannheim.de>

Gesprächsanalytisches Informationssystem (GAIS)

GAIS bietet Informationen für die wissenschaftliche Arbeit in der Gesprächsforschung und über die wissenschaftliche Gemeinschaft. Derzeit wird GAIS ausgebaut zu einem Online-Handbuch zu den Themen gesprächsanalytische Feldforschung, Datenerhebung, Transkription, Korpuserstellung, computergestützte Korpusverwaltung und computergestützte Analyseverfahren. Dabei ergeben sich Synergieeffekte: GAIS präsentiert (generalisierte) Antworten auf Serviceanfragen an das AGD, arbeitet zusammen mit Seminaren zum Thema „Feldforschung und Transkription“ im Mannheimer Masterstudiengang „Sprache und Kommunikation“ und dient als Publikationsbasis für das Projekt FOLK.

Web: <http://gais.ids-mannheim.de>



6. The Language Archive, Max-Planck-Institut für Psycholinguistik Nijmegen

Shakila Shayan, Han Sloetjes



ELAN – Multimedia Annotator

ELAN ist ein Multimedia-Annotations-Tool, das an der Abteilung „The Language Archive“ (Spracharchiv) des Max-Planck-Instituts für Psycholinguistik entwickelt und unterhalten wird. Es ist erhältlich für Windows, Mac OS X und Linux, kann frei heruntergeladen werden und ist quelloffen (Open Source).

ELAN ist ein Desktop-Werkzeug für das manuelle Annotieren von Audio- und/oder Video-Aufnahmen. Es unterstützt bis zu 4 Videos pro Annotations-Dokument. Es ist ein vielschichtiges Annotations-System: Annotationen sind in sogenannten „Tiers“ (Zeilen) enthalten, und diese Zeilen können Teil einer größeren Zeilenhierarchie sein (d.h. Zeilen können abhängige untergeordnete Zeilen haben).

Die Annotationen enthalten Unicode-Text und die Annotations-Dokumente sind in XML-Dateien (EAF) gespeichert. Das Programm erlaubt das Suchen in einer einzelnen Datei oder in einem benutzerdefinierten Satz aus Dateien (Lokaler Korpus). Andere multiple Datei-Operationen sind möglich, unter anderem das Erstellen, Konvertieren, Editieren und Exportieren von Dateien en masse. Gerade wurden verschiedene Arbeitsmodi hinzugefügt, jedem Modus fällt dabei eine spezifische Aufgabe im Transkriptionsprozess zu. Neu sind der Segmentie-

rungsmodus, für das schnelle Erstellen von Segmenten, und der Transkriptionsmodus, für das schnelle und bequeme Einfügen von Text in existierende Annotationen.

ELAN wird in einer Vielzahl von Forschungsgebieten innerhalb der Linguistik (und darüber hinaus) angewandt, wie z.B. Feld-Linguistik, Konversations-Analyse, Gestik-Studien und Untersuchung von multimedialen Interaktionen.

Während der Fachmesse wird ELAN als Demo präsentiert.

Web: <http://tla.mpi.nl/tools/tla-tools/elan/>

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



Lexus – A Web-based Lexicon Tool

LEXUS ist ein flexibles, web-basiertes Lexikon-Tool, das von der Abteilung „The Language Archive“ (Spracharchiv) des Max-Planck-Instituts für Psycholinguistik in Nijmegen entwickelt und gepflegt wird. Es ist plattformunabhängig und man kann von überall auf der Welt durch einen normalen Web-Browser darauf zugreifen.

LEXUS wurde ursprünglich im Rahmen der Dokumentation bedrohter Sprachen entwickelt, aber es kann problemlos auch in anderen linguistischen Kontexten verwendet werden. Eines der Alleinstellungsmerkmale von LEXUS ist, dass es dem Nutzer keine feste Lexikonstruktur vorschreibt, gleichzeitig wird aber das Erstellen einer ISO-LMF-konformen Struktur unterstützt. Zudem ist die Verwendung von standardisierten ISO-Datenkategorien vorgesehen.

LEXUS ermöglicht eine Vielfalt an Möglichkeiten zur nutzerdefinierten Visualisierung der Inhalte. Die eingebaute Multimedia-Unterstützung erlaubt das Einbetten von lokalen oder archivierten Film- oder Audio-Clips oder von Bildern in beliebige Elemente eines Lexikoneintrags. Des Weiteren erlaubt die Unterstützung für flexibel anpassbare Views eine Vielzahl verschiedener Darstellungsvarianten desselben Lexikons.

Außerdem ist es in LEXUS ebenfalls möglich anderen Nutzern Zugriff auf ein bestimmtes Lexikon zu geben, was die Software ideal für kollaboratives Arbeiten einer Gruppe von Forschern macht, indem sie alle von unterschiedlichen Orten über das Internet auf dasselbe Lexikon zugreifen und es gemeinsam bearbeiten können.

All die genannten Funktionen machen aus LEXUS eine vielversprechende Lexikon-Software, die mit ihrem Fokus auf kollaborativem Arbeiten und Multimedia-Unterstützung ideal für Sprachdokumentationsprojekte ist.

Web: <http://tla.mpi.nl/tools/tla-tools/lexus/>

Kontakt: Shakila.Shayan@mpi.nl



7. [moca2] – multimodal oral corpus administration

Daniel Alcón López, Stefan Pfänder

[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 personalisierten Zugang zu den Korpora zu gewährleisten. Dabei unterstützt das System eine unbegrenzte Anzahl von Nutzern, denen individuell der Zugriff auf bestimmte 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
Stefan Pfänder stefan.pfaender@romanistik.uni-freiburg.de

trAVis musikzentriertes transkriptionsprogramm
für audiovisuelle medienprodukte

8. trAVis – musikzentriertes Transkriptionsprogramm für audiovisuelle Medienprodukte

Daniel Klug, Axel Schmidt

Die Web-Applikation trAVis ist ein musikzentriertes Transkriptionsprogramm für audiovisuelle Medienprodukte, das bildbezogene und texthermeneutische Ansätze mit musikwissenschaftlichen Zugängen verbindet. Mit trAVis kann das komplexe Zusammenspiel von Bild, Text und Ton/Musik interdisziplinär transkribiert, analysiert und interpretiert werden.

Im Einzelnen umfasst dies:

- das Erstellen eines Transkripts in einem frei editierbaren Pattern- und Zeilen-System;
- das Visualisieren von audiovisuellen Korrespondenzen mit Hilfe von Zeichenfunktionen;
- das Erstellen einer individuellen Arbeitsfläche durch verschiedene Darstellungs- und Skalierungsoptionen;
- das Exportieren des Transkripts als Text und/oder Grafik;
- das Transkribieren von Musik in westlicher Notenschrift;
- die synchronisierte Darstellung der Musiknotation im Verhältnis zu Zeit- und Taktleiste;
- das individuelle Generieren von Standbildern;
- das Transkribieren textlicher Ereignisse im Metrum der Musik, u.v.m.

Web: <http://www.travis-analysis.org/>

Kontakt: info@travis-analysis.org