CLARIN-D: Ressourcen gesprochener Sprache und Webservices des Bayerischen Archivs für Sprachsignale

Draxler, Christoph
draxler@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Schiel, Florian
schiel@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Reichel, Uwe
reichelu@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Kisler, Thomas
kisler@phonetik.uni-muenchen.de
Bayerisches Archiv für Sprachsignale, Institut für Phonetik und Sprachverarbeitung, Deutschland

Das BAS Repository

Das Repository des BAS ist in einen öffentlich zugänglichen und einen zugangsgeschützten Bereich unterteilt. Der öffentliche Bereich enthält die Metadaten der im Repository enthalten Datenbanken gesprochener Sprache, der geschützte Bereich die Annotationen und Sprachsignaldaten.

Aktuell (Stand 31.12.2015) umfasst das Repository 33 Korpora mit den Schwerpunkten Sprachtechnologie, regionale Variation, Sprechermerkmale und Grundlagenforschung. Auf der obersten Repository-Ebene werden die wichtigsten Angaben prominent platziert: Name und Eigentümer der Ressource, Audio bzw. Video, verwendete Sprache, und die Zugangsrestriktionen (s. Kap. 6).

Von den 33 Korpora enthält eines arabische Sprachaufnahmen, und enthalten zwei italienische, drei englische und 26 deutsche Sprachaufnahmen; dazu kommt ein Korpus in Deutscher Gebärdensprache. 26 Korpora enthalten nur gesprochene Sprache, 6 sowohl gesprochene Sprache als auch Video, eines nur Video. Ein Korpus ist allgemein verfügbar, 30 sind für akademische Nutzer_innen frei zugänglich, zwei nur als lizensierte Korpora.

Der Datenumfang beträgt mehr als 2,86 TB an Audio-, Video- und Sensorsignaldaten, sowie ca. 17,4 GB an Annotations- und Metadaten.

***Abb. 1**: Startseite des BAS Repository mit der Liste verfügbarer Korpora.*

Inhalt des Repository

Die überwiegende Anzahl der Ressourcen wurde vom BAS selbst oder in Kooperation mit industriellen oder akademischen Partnern erstellt. Zunehmend werden weitere, von externen Partnern erstellte Ressourcen in das Repository aufgenommen.

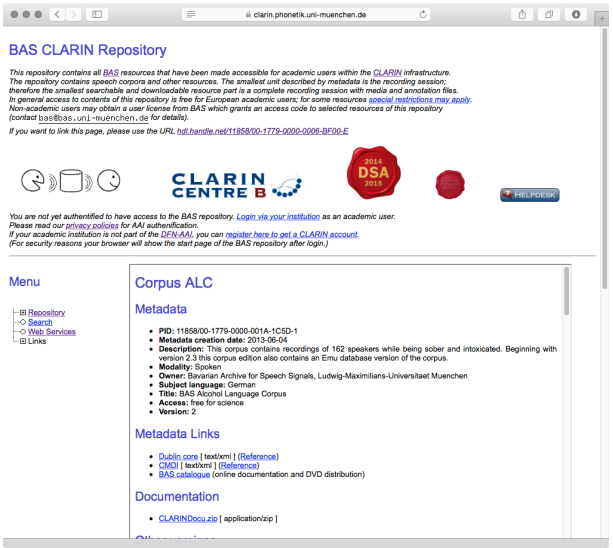

***Abb. 2**: Repositoryansicht der Sessions im Korpus ALC.*

Darüberhinaus sucht das BAS aktiv nach verwaisten Ressourcen, d. h. Sprachdatensammlungen, die im Rahmen von Projekten, Dissertationen, oder sonstigen Datenerhebungen erstellt wurden, deren Fortbestand aber aufgrund fehlender Finanzierung oder personeller Veränderungen gefährdet ist. Für diese Ressourcen bietet das BAS Unterstützung bei der Erstellung der Metadaten sowie auch die Archivierung im Repository an. Ein Leitfaden zur Aufbereitung externer Korpora ist in Vorbereitung.

Aktuell werden Metadaten für das schweizerische Jugendsprachekorpus (Tissot 2015), das WaSeP Korpus zur Analyse emotionaler Prosodie (Wendt 2007), das Sprechermerkmale-Korpus von Brüderpaaren von Hanna Feiser (Feiser 2015), das ASD-Corpus (Siebenbürger Deutsch) der Italianistik der LMU München sowie das VOYS Korpus mit Aufnahmen schottischer Jugendlicher (Dickie et al. 2009) für das Repository aufbereitet.

CMDI Metadatenformat

Grundlage des Repository sind zum einen das Metadatenformat CMDI (*Component Metadata Initiative*) (CLARIN-D AP 5 2012) , zum anderen ein in *perl* implementiertes und an die CLARIN-Anforderungen angepasstes Repositoryframework.

CMDI zeichnet sich dadurch aus, dass es in kontrollierter Form erweiterbar ist und selbstbeschreibend ist. Sämtliche in CMDI verwendeten Deskriptoren, Components genannt, müssen in einem öffentlich zugänglichen Register spezifiziert sein. Profile sind vordefinierte Listen von Deskriptoren für spezifische Datenbestände.

Das BAS hat zwei Profile für Ressourcen gesprochener Sprache definiert, das MediaCorpus und das MediaSession Profil. Diese Profile stellen sicher, dass das Repository

1. nur Daten enthält, die formalen Mindestanforderungen genügen,
2. von externen Diensten wie Suchmaschinen oder Informationsdiensten gelesen werden kann,
3. in alternativen Metadatenformaten ausgegeben werden kann, und
4. externe Suchanfragen nach Meta- und Inhaltsdaten unterstützt.

Um die Erstellung CLARIN-kompatibler Metadaten zu erleichtern bietet das BAS den Webservice COALA an (s. Kap. 7).

Programmierschnittstellen und Protokolle

Aktuell wird das BAS Repository regelmäßig von den Harvestern des Virtual Language Observatory (VLO) (Zinn et al. 2015) sowie vom Informationsdienst Reuters über die Standardschnittstelle OAI-PMH ausgelesen, die unterstützten Metadatenformate sind Dublin Core und OLAC, und für die Suche in den Annotationsdaten wird über die Schnittstelle SRU-CQL angeboten.

Persistent Identifiers

Das BAS Repository verwendet EPIC Handle Persistent Identifiers (PID) für die Korpus- und Session-Objekte. Zum Beispiel verweist die PID http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E auf die Webadresse des Repository.

Unterschiedliche Versionen eines Korpus bzw. einer Session erhalten jeweils eine eigene PID. Auf Signal- oder Annotationsfiles in einer Session kann mithilfe von Part-Identifiern zugegriffen werden (vorausgesetzt der User ist authentifiziert), hier z. B. ist die Part-ID m_0000000001 http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@partId=m_0000000001

Die PIDs werden von der Gesellschaft für wissenschaftliche Datenverarbeitung in Göttingen (GWDG) verwaltet.

Zugangskontrolle

CLARIN strebt eine Single Sign-On Authentifizierung an. Als Protokoll wird Shibboleth (Knight 2015) verwendet. Für den Zugriff auf geschützte Ressourcen gibt die Nutzerin ihr Login mit Passwort ein, dieses wird von der Heimatinstitution überprüft. Diese Institution teilt dem Server, der die Ressource verwaltet, mit, ob die Nutzerin bekannt ist – wenn ja, dann darf sie in CLARIN diese und alle weiteren Ressourcen nutzen, für die sie autorisiert ist.

Im Wesentlichen gibt es drei Autorisierungsstufen: ACA für akademische Nutzer, RES für beschränkten Zugriff, und PUB für den unbeschränkten öffentlichen Zugang. Im Repository sind diese Stufen gut sichtbar aufgeführt, so dass sofort klar wird, welche Ressource wie zugänglich ist.

Aktuell sind die akademischen Institutionen der 15 CLARIN Mitgliedsländer (AT, BG, CZ, DK, EE, DE, GR, LI, NL, NO, PL, PT, SL, SE, UK) sowie die Nederlandse Taalunie als länderübergreifende Einrichtung zugangsberechtigt. Für Nutzer außerhalb akademischer Einrichtungen unterhält CLARIN ein eigenes Nutzerverzeichnis, so dass ausgewählte Mitglieder ebenfalls Zugriff auf CLARIN Ressourcen haben.

BAS Webservices

Das BAS betreibt eine Reihe von phonetisch-linguistischen Webservices (Schiel 2013) sowie web-basierte Schnittstellen BAS (2011-2016) , die auf diesen Webservices basieren.

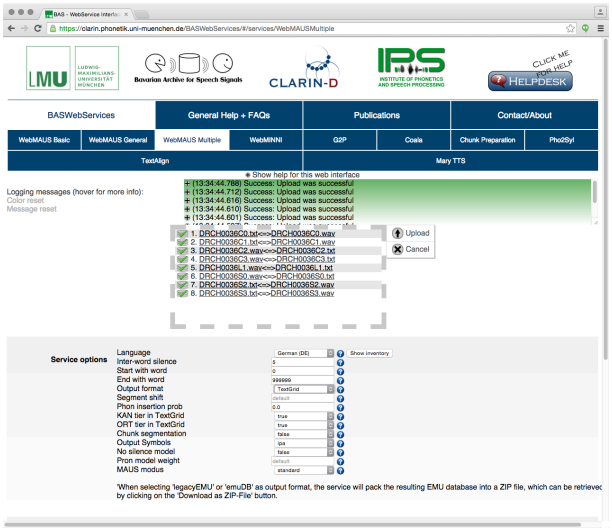

***Abb. 3**: WebMAUS mit 4 hochgeladenen Dateipaaren für die automatische Segmentation.*

Ein Webservice erlaubt einem User oder einer Applikation mittels des REST-Protokolls automatische Verarbeitungen von Daten auf einem dedizierten Server durchzuführen. Zum Beispiel benutzt das Annotations-Tool ELAN (vgl. TLA: Elan) einen Webservices des BAS um während der Bearbeitung einer Sprachdatei eine vollautomatische Segmentierung anzustoßen (Kisler et al. 2012). Web-basierte Schnittstellen dagegen erlauben die benutzerfreundliche interaktive Verwendung solcher Webservices über einen Standard Web-Browser.

Das BAS hat in den vergangenen zwei Jahren seine vorhandenen Webservices erweitert, verbessert und neue Funktionalitäten bereitgestellt. Fast alle interaktiven Services erlauben jetzt auch das Batch-Processing von großen Datensammlungen.

Das bekannte System MAUS (vgl. Schiel 1999) zur vollautomatischen Segmentierung von Sprache wurde auf derzeit 17 Sprachvarianten erweitert. MAUS liest ein Sprachsignal-File und die dazugehörige Transkripton und berechnet daraus die Phonem/Wort-Segmentierung in Form einer hierarchischen Annotation.

Hervorzuheben sind dabei

1. die neue Verarbeitung von Schweizer Deutsch ausgehend von der sog. Dieth-Kodierung,
2. die vier Sprachvarianten amerikanisches, australisches, neuseeländisches und britisches Englisch,
3. sowie als neueste Sprachen Russisch und Französisch.

WebMAUS unterstützt seit kurzem auch das neue EMU Datenbank-System-Format und die sofortige Visualisierung der Ergebnisse mit Hilfe des Javascript-basierten Labeller-Tools EMU-webApp (Winkelmann et al. 2016) .

Ein neu entwickelter Webservice WebMINNI erlaubt die phonetische Segmentierung und Transkription von Medienfiles ohne vorhandene Verschriftung in sieben Sprachvarianten. Die stochastisch-phonologische Komponente des MAUS Systems wurde dabei ersetzt durch ein phontaktisches Bigram-Modell der jeweiligen Sprache.

Das BALLOON-Werkzeug G2P zur automatischen Generierung von Aussprache-Kodierung auf der Basis von orthographischem Text-Input wurde auf nunmehr 18 Sprachen erweitert (Reichel 2012).

Ein neu entwickeltes Tool Pho2Syl erlaubt die automatische Syllabifizierung von Transkriptionen. Syllabifizierungen werden in vielen Disziplinen der empirischen Linguistik benötigt. Pho2syl ist das erste verfügbare Werkzeug, das sowohl phonologische als auch phonetische Transkripte für 18 Sprachvarianten erlaubt.

Ein weitere Neuentwicklung ist TextAlign, ein allgemein einsetzbares Werkzeug zur optimalen symbolischen Alignierung (Reichel 2012).

Eine typische Anwendung ist die Alinierung von orthographischen zu phonologischen Symbol-Ketten. TextAlign bietet sowohl eine Vielzahl von vortrainierten Kostenfunktionen für verschiedene Sprachvarianten als auch Mechanismen zur Abschätzung der optimalen Kostenfunktion aus den Input-Daten.

Für alle web-basierte Schnittstellen wurde die Benutzerfreundlichkeit kontinuierlich verbessert. Die BAS Webservices bieten dem Benutzer jetzt sowohl einen online Help Desk über die CLARIN Infrastruktur, eine FAQ-Sammlung als auch Tooltips und Usecase Beschreibungen direkt auf der Web-Seite.

Fazit und Ausblick

Das CLARIN Repository des BAS ist seit Ende 2012 in Betrieb und wird laufend erweitert. Der mit Abstand meistgenutzte Webservice ist WebMAUS - hier hat die Diskussion mit Anwendern auch dazu geführt, dass einzelne Bestandteile von MAUS mittlerweile als eigene Webservices verfügbar sind, z. B. G2P oder WebMINNI. Die zunehmende Erfahrung im Umgang mit Webservices hat dazu geführt, dass auch komplexe Arbeitsabläufe wie die Erstellung von Metadaten, oder früher wenig genutzte Dienste wie die Sprachsynthese, nun als Webservice zugänglich und damit wesentlich einfacher zu nutzen sind.

Der Wegfall der Notwendigkeit von Softwareinstallationen sowie die konsequent auf Nutzerfreundlichkeit ausgerichtete Gestaltung von Webservices hat dazu geführt, dass ganz neue Nutzerkreise erschlossen wurden: Ethnolog_innen lassen mit WebMAUS erste Rohsegmentationen bedrohter Sprachen erstellen, Toolentwickler_innen binden die Werbservices in ihre Tools ein, usw. Neben dem erwähnten ELAN nutzen in Studentenprojekten entwickelte innovative Prototypen von Sprachlerntools die MAUS-Segmentation für eine grafisch ansprechende Gegenüberstellung von Wort-, Silben- und Lautdauern von Muttersprachler- und Lerneräußerungen.

Zum Schluss ein Aufruf: das BAS sucht weiterhin Ressourcen gesprochener Sprache für das Repository.

Insbesondere von Interesse sind Ressourcen gesprochener Sprache, deren Fortbestand gefährdet ist, oder die aus neuen, bislang unbekannten Forschungs- und Anwendungsbereichen stammen.

Bibliographie

BAS: Bavarian Archive for Speech Signals(2011-2016): *BAS WebService 2.06*. Institute of Phonetics and Speech Processing at the Ludwig-Maximilians-Universität München https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services [letzter Zugriff 08. Januar 2016].

CLARIN-D AP 5 (2012): *CLARIN-D User Guide: The Component Metadata Initiative (CMDI)*Clarin-D http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml [letzter Zugriff 08. Januar 2016].

Dickie, Catherine / Schaeffler, Felix / Draxler, Christoph (2009): "Speech Recordings via the Internet: An Overview of theVOYS project in Scotland", in: *Proc. Interspeech 2009* 1807-1810.

TLA:Elan (o. J.): *TLA (The Language Archive) Tools: Elan*. Nijmegen, Netherlands: Max Planck Institute for Psycholinguistics https://tla.mpi.nl/tools/tla-tools/elan/ [letzter Zugriff 08. Januar 2016].

Feiser, Hanna (2015): *Untersuchung auditiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Brüderpaaren unter forensischen Aspekten*. Frankfurt am Main: Verlag Polizeiwissenschaft

GWDG: *Gesellschaft für wissenschaftliche Datenverarbeitung mbH*. Göttingen: Georg-August-Universität Göttingen - Stiftung Öffentlichen Rechts, Max-Planck-Gesellschaft https://www.gwdg.de/ [letzter Zugriff 08. Januar 2016].

Kisler, Thomas / Schiel, Florian / Sloetjes, Han (2012): "Signal processing via web services: the use case WebMAUS", in: *Proceedings Digital Humanities 2012, Hamburg, Germany*: 30-34.

Knight, Justin (ed.) (2015): *Shibboleth* http://shibboleth.net/ [letzter Zugriff 08. Januar 2016].

Krefeld, Thomas / Stephan Lücke, Stephan / Mages, Emma(2009-2013): *Audioatlas Siebenbürgisch-Sächsischer Dialekte (ASD)*. Ludwig-Maximilians-Universität München [letzter Zugriff 08. Januar 2016].

Reichel, Uwe Dieter (2012): "PermA and Balloon: Tools for string alignment and text processing", in: *Proc. Interspeech. Portland, Oregon*: paper no. 346.

Schiel Florian (1999): "Automatic Phonetic Transcription of Non-Prompted Speech", in: *Proc. of the ICPhS 1999. San Francisco*: 607-610.

Schiel, Florian (2013): *BAS: Bavarian Archive for Speech Signals Webservices*. Universität Müchen http://www.phonetik.uni-muenchen.de/forschung/Bas/BasWebserviceseng.html [letzter Zugriff 08. Januar 2016].

Tissot, Fabienne (2015): *Gemeinsamkeit schaffen in der Interaktion* Diskursmarker und Lautelemente in zürichdeutschen Erzählsequenzen (= Sprache in Kommunikation und Medien 9). Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Wendt, Beate (2007): *Analysen emotionaler Prosodie*Hallesche Schriften zur Sprechwissenschaft und Phonetik. Bern / Berlin / Frankfurt am Main / New York / Paris / Wien: Peter Lang.

Winkelmann, Raphael / Raess, Georg / Jochim, Markus (2016):*EMU-webApp* Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München https://github.com/IPS-LMU/EMU-webApp [letzter Zugriff 08. Januar 2016].

Zinn, Claus / Duin, Patrick / Stehouwer, Herman / Eckart, Thomas / Looij, Kees Jan van de / Goosen, Twan / Uytvan, Dieter van (2015): *Virtual Language Observatory 3.3.2* https://vlo.clarin.eu [letzter Zugriff 08. Januar 2016].