The BAS Speech Data Repository

Uwe D. Reichel, Florian Schiel, Thomas Kisler, Christoph Draxler, Nina Pörner
Research Institute for Linguistics, Hungarian Academy of Sciences,
Institute of Phonetics and Speech Processing, University of Munich
Benczúr u. 33, 1068 Budapest, Hungary, Schellingstr. 3, 80799 Munich, Germany
uwe.reichel@nytud.mta.hu, {schiel,kisler,draxler,poerner}@bas.uni-muenchen.de

Abstract

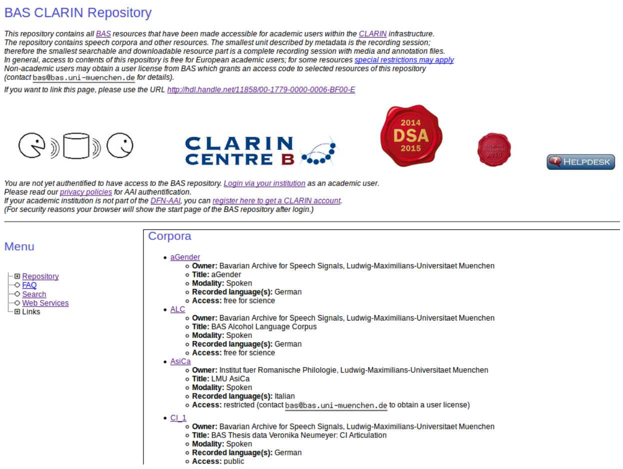
The BAS CLARIN speech data repository is introduced. At the current state it comprises 31 pre-dominantly German corpora of spoken language. It is compliant to the CLARIN-D as well as the OLAC requirements. This enables its embedding into several infrastructures. We give an overview over its structure, its implementation as well as the corpora it contains.

Keywords: Data repository, spoken language corpora, CLARIN-D

1. Introduction

The BAS CLARIN speech data repository (BAS, 2016a; Reichel, 2013) is maintained by the Bavarian Archive for Speech Signals (BAS, 2016c) in the context of the CLARIN-D project (CLARIN, 2016b). It is located at the Institute of Phonetics and Speech Processing, Ludwig-Maximilian-University in Munich, Germany. At its current state (March 2016) it comprises 31 pre-dominantly German corpora of spoken language summing up to 2.62 TB of signal and annotation files, and 16.6 GB of metadata. 29 of these corpora are freely available for academic research. The repository is compliant to the CLARIN-D as well as the OLAC requirements, so that it can be harvested by several infrastructures such as the Virtual language observatory (VLO, 2015; Goosen and Eckart, 2014) and the Open Language Archives Community (OLAC, 2011). Figures 1 and 2 show the repository's and a corpus' landing page, respectively.

This paper describes the structure of this repository, its main features, and gives an overview over the provided corpora.



BAS CLARIN Repository

This repository contains all BAS resources that have been made accessible for academic users within the CLARIN infrastructure.
The repository contains speech corpora and other resources. The smallest unit described by metadata is the recording session;
therefore the smallest searchable and downloadable resource part is a complete recording session with media and annotation files.
In general, access to contents of this repository is free for European academic users; for some resources special restrictions may apply
Non-academic users may obtain a user license from BAS which grants an access code to selected resources of this repository
(contact bas@bas.uni-muenchen.de *for details).*

If you want to link this page, please use the URL http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E

You are not yet authentified to have access to the BAS repository. Login via your institution as an academic user.
Please read our privacy policies for AAI authentification.
If your academic institution is not part of the DFN-AAI, you can register here to get a CLARIN account.
(For security reasons your browser will show the start page of the BAS repository after login.)

Menu

Corpora

- aGender
  - **Owner:** Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universitaet Muenchen
  - **Title:** aGender
  - **Modality:** Spoken
  - **Recorded language(s):** German
  - **Access:** free for science
- ALC
  - **Owner:** Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universitaet Muenchen
  - **Title:** BAS Alcohol Language Corpus
  - **Modality:** Spoken
  - **Recorded language(s):** German
  - **Access:** free for science
- AsiCa
  - **Owner:** Institut fuer Romanische Philologie, Ludwig-Maximilians-Universitaet Muenchen
  - **Title:** LMU AsiCa
  - **Modality:** Spoken
  - **Recorded language(s):** Italian
  - **Access:** restricted (contact bas@bas.uni-muenchen.de to obtain a user license)
- CI_1
  - **Owner:** Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universitaet Muenchen
  - **Title:** BAS Thesis data Veronika Neumeyer: CI Articulation
  - **Modality:** Spoken
  - **Recorded language(s):** German
  - **Access:** public

Figure 1: The repository's landing page.

2. Repository structure and main features

2.1. Structure

The repository is based on a file system and is hierarchically structured into *corpus, recording session* and *primary data* items. Each corpus contains one or more recording sessions, and each session comprises primary data, i.e. signal and annotation files. The repository is divided into a freely accessible and a protected part. The protected part contains the primary resources, whereas the metadata is freely accessible.

In compliance with the CLARIN-D requirements the BAS repository adheres to standardized file formats, provides metadata, supports persistent data storage and versioning, and requires user authentication for its protected part.

2.2. Metadata

Each corpus and each recording session is described by a CMDI metadata record (CLARIN, 2016d) that is dynamically rendered to a landing page for this item to be accessed by the users. Next to CMDI also Dublin Core and OLAC format (DC, 2016; OLAC, 2011) are supported. The metadata can be harvested via an OAI-PMH endpoint (BAS, 2016b).

2.3. Persistence

The BAS Repository supports a persistent storage of the contained data: each version of a repository item is permanently stored without changes. Primary data is stored together with its MD5 checksum (Rivest, 1992) so that consistency can be regularly checked.

Furthermore, each version of a corpus and of a session is assigned an ePIC Handle persistent Identifier (PID) (ePIC, 2016), by which its landing page is durably accessible via the handle system.

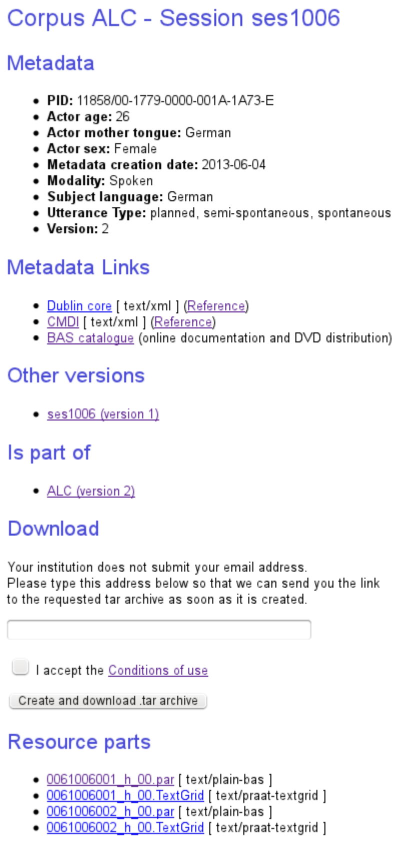
The repository is constantly backed-up by the Leibniz Rechenzentrum, Garching (LRZ, 2016) via the IBM Tivoli system.

2.4. Versioning

The insertion, editing or removal of a signal or annotation file leads to a new version of this file as well as transitively of the recording session and thus of the corpus it is part of. New versions of a session and a corpus each are assigned a new PID, by which they can be addressed uniquely. For each version of an item the versions of the super-ordinate items are stored by means of is-part-of relations. Repository internally all versions of a repository item are bundled by a unique repository ID.

2.5. User authentication

The protected primary data part of the repository can only be accessed by members of academic institutions after Shibboleth authentication (Shibboleth, 2016). The user can log in via her/his institution, provided that it is part of the CLARIN Service Provider Federation (SPF) (CLARIN, 2016c), which consists of a growing number of national Identity Federations (e.g. the DFN-AAI for Germany (DFN-AAI, 2010) or SURFconext for the Netherlands (SURFconext, 2016)). If the user's entitlement is classified as 'academic', access to the protected part of the repository is granted. If the academic institution has not joined any national federation, that is part of the SPF, the user can apply for a CLARIN-D account to get repository data access. For non-academic users access can be enabled for selected corpora after having obtained a BAS user license.



Corpus ALC - Session ses1006

Metadata

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

Metadata Links

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

Other versions

- ses1006 (version 1)

Is part of

- ALC (version 2)

Download

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

□ I accept the Conditions of use

Create and download .tar archive

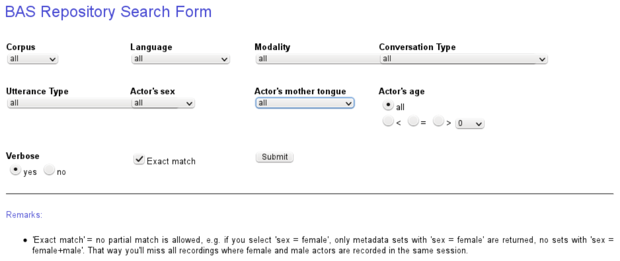
Resource parts

- 0061006001_h_00.par [text/plain-bas]
- 0061006001_h_00.TextGrid [text/praat-textgrid]
- 0061006002_h_00.par [text/plain-bas]
- 0061006002_h_00.TextGrid [text/praat-textgrid]

Figure 2: A corpus landing page.

2.6. Cross-corpora metadata and federated content search

The user can collect recording sessions across corpora depending on the respective research question related to metadata such as modality, speaker sex or mother tongue. After successful authentication the collection can be downloaded as a zipped archive. The back-end of this search engine is implemented as an SQLite database which is also used by the SRU endpoint (BAS, 2015) for federated content search (CLARIN, 2016a). The latter allows users to search CLARIN corpora world-wide across different centers. The search engine's front-end and an example search result are shown in Figures 3 and 4, respectively.



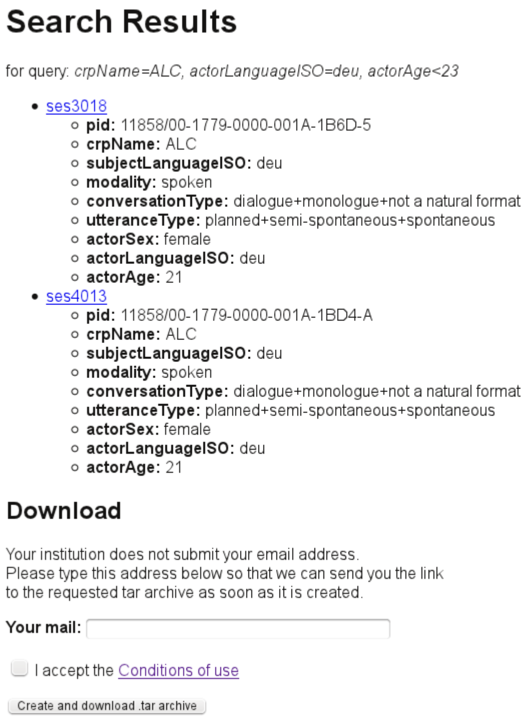
BAS Repository Search Form

Corpus: all
Language: all
Modality: all
Conversation Type: all
Utterance Type: all
Actor's sex: all
Actor's mother tongue: all
Actor's age: ● all ○ < ○ = ○ > 0
Verbose: ● yes ○ no
☑ Exact match

Remarks:

- 'Exact match' = no partial match is allowed, e.g. if you select 'sex = female', only metadata sets with 'sex = female' are returned, no sets with 'sex = female+male'. That way you'll miss all recordings where female and male actors are recorded in the same session.

Figure 3: The repository's search form.



Search Results

for query: *crpName=ALC, actorLanguageISO=deu, actorAge<23*

- ses3018
  - **pid:** 11858/00-1779-0000-001A-1B6D-5
  - **crpName:** ALC
  - **subjectLanguageISO:** deu
  - **modality:** spoken
  - **conversationType:** dialogue+monologue+not a natural format
  - **utteranceType:** planned+semi-spontaneous+spontaneous
  - **actorSex:** female
  - **actorLanguageISO:** deu
  - **actorAge:** 21
- ses4013
  - **pid:** 11858/00-1779-0000-001A-1BD4-A
  - **crpName:** ALC
  - **subjectLanguageISO:** deu
  - **modality:** spoken
  - **conversationType:** dialogue+monologue+not a natural format
  - **utteranceType:** planned+semi-spontaneous+spontaneous
  - **actorSex:** female
  - **actorLanguageISO:** deu
  - **actorAge:** 21

Download

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

Your mail:

□ I accept the Conditions of use

Create and download .tar archive

Figure 4: Result of a cross-corpora search.

3. Accessing primary and metadata

All corpora and sessions can be accessed via their landing pages and as mentioned in section 2.3. by the PID of the respective repository item. To give an example, the landing page of session 1006 of the corpus ALC is accessible by its PID as follows: *http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3*

Next to a short description this landing page provides the link to the item's metadata. For an automatized processing the metadata can be accessed directly by two methods:

- by using a part identifier *@format=cmdi*, as in *http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@format=cmdi*
- by means of content negotiation. In this case on the client's side the Accept Header has to be set to 'application/x-cmdi+xml'.

After successful authentication the links to the primary data items are shown on the landing page as well, and a direct access via the part identifier @partId is enabled. The values of this identifier are given by the Resource Proxy IDs in the CMDI metadata. To give an example: *http://hdl.handle.net/11858/00-1779-0000-0006-BDA2-3@partId=m_0000000001* directs the user to an annotation file referenced by the ID *m_0000000001* in the CMDI metadata file.
Furthermore, the authorized user is enabled to download the corpora or single sessions as zipped archives, as is shown in Figure 2.

4. Corpus Ingest

The fully automatized ingest of a new corpus into the BAS repository consists of the following steps:

1. CMDI files are validated and compared with a repository content table in order to find out whether new data is ingested or already stored data is updated.
2. For all new or updated sessions and transitively for the corpus PIDs are retrieved from the GWDG PID Handle Service (GWDG, 2009). Each version of a corpus and a session thus receives its own unique identifier.
3. CMDI files are copied to the public space and adjusted. Resources are copied to the protected area. For regular consistency checks and for versioning checksums are calculated.
4. The search database and the OAI-PMH interface are updated.

5. Software

A proprietary repository software was developed in Perl and PHP. The requirements are: a server supporting CGI and PHP, SQLite as the search engine backend, as well as freely available tools for XML validation and transformation, and for checksum calculation. We use *xmllint, xsltproc*, and *md5sum*, respectively. For the OAI-PMH interface we adapted the freely available file based OAI-PMH2 XMLFile data provider version 2.1 (Suleman, 2002).

6. Corpora

6.1. Overview

At its current state the BAS repository provides 31 corpora which are introduced in table 1 and shortly described in section 6.2.. 21 of these corpora have been produced by the Bavarian Archive for Speech Signals; the third party corpora are the Natural Media Motion-Capture Corpus by the RWTH Aachen University, the Bielefeld Speech and Gesture Alignment Corpus by the University of Bielefeld, the German sign language corpus SIGNUM by the University of Aachen, the corpus of spoken Calabrese of the University of Munich, the Cochlear Implant Speech Corpus CI Articulation, the Siemens Hearing Aid corpus HOESI, the Italian CLIPS corpus, the L2 German learners corpus SC10 and the corpora aGender and VERIF1DE provided by the Deutsche Telekom Laboratories.

6.2. Corpus descriptions

This section gives short descriptions about the corpora stored in the repository. For more detailed information please see the descriptions and the documentation zip archives on the repository corpus landing pages (BAS, 2016a). All but one corpus (AsiCa) are owned by the Bavarian Archive for Speech Signals. All corpora contain signal and annotation files.

aGender This corpus contains recordings of 945 native German speakers over public telephone lines with read and semi-spontaneous speech. The recordings were carried out by the German Telekom Labs for the purpose of gender and age classification.

ALC – Alcohol Language Corpus This corpus contains recordings of 162 speakers while being sober and intoxicated. Beginning with version 2.3 this corpus edition also contains an Emu database version (Winkelmann, 2015).

AsiCa This corpus is a documentation of the South Italian dialect Calabrese (AsiCa, 2006). It contains recordings of 60 speakers with read and spontaneous speech. A part of the speakers has migration experience in Germany. Owner is the Institute of Romance philology, University of Munich.

CI Articulation This corpus contains German speech recordings of 48 cochlear implant users and 48 speakers without hearing impairment. It consists of five subcorpora with focus on vowel, consonant, and VOT production, each comparing the utterances of the hearing impaired and the control group. The database is distributed in emuDB format (EMU, 2010; Winkelmann, 2015).

CLIPS MT MANUAL This corpus is part of the Italian CLIPS corpus (CLIPS, 2004) that covers 15 maptask dialogs recorded in different locations in Italy in 2000-2004.

FORMTASK FORMTASK is a German telephone speech database of prompted descriptions of typical forms found in everyday life (e.g. public transport tickets, money transfer form).

HEMPEL HEMPEL is a collection of more than 3900 spontaneous speech items recorded as extra material during the German SpeechDat-II project. Speakers were asked to report what they had been doing during the last hour. A more detailed description can be found in Draxler and Schiel (2002).

HOESI – Siemens Hoergeraete Corpus This corpus contains spontaneous speech dialogs in German. Each pair of dialog partners is recorded conversing under real-noise conditions (in a noisy cafeteria and in a car going at different velocities), as well as in a studio at various levels of Lombard noise played directly into the subjects' ears.

Title	Modality	Language	Access
aGender	spoken	German	free for science
Alcohol Language Corpus	spoken	German	free for science
AsiCa	spoken	Italian	restricted
CI Articulation	spoken	German	public
CLIPS_MT_MANUAL	spoken	Italian	free for science
FORMTASK	spoken	German	free for science
HEMPEL	spoken	German	free for science
Siemens Hoergeraete Corpus	spoken	German	free for science
Natural Media Motion-Capture Corpus	spoken, gestures	German	free for science
PhonDat 1	spoken	German	free for science
PhonDat 2	spoken	German	free for science
Ph@ttsessionz Adolescents Speech Corpus	spoken	German	free for science
Regional Variants of German 1	spoken	German	free for science
Regional Variants of German J – Juveniles	spoken	German	free for science
Bielefeld Speech and Gesture Alignment Corpus	spoken, gestures	German	free for science
SC1	spoken	L2 German	free for science
SC10	spoken	L2 German	free for science
Strange Corpus 2 Noises	spoken	German	free for science
SmartWeb Handheld	spoken	German	free for science
SIGNUM	signed	German sign language	free for science
SmartWeb Motorbike Corpus	spoken	German	free for science
SmartKom Home	spoken, gestures, facial expression	German	free for science
SmartKom Mobil	spoken, gestures, facial expression	German	free for science
SmartKom Public	spoken, gestures, facial expression	German	free for science
SmartWeb Video	spoken, eye-gaze	German	free for science
TAXI	spoken	German, English	free for science
VERIF1DE	spoken	German	restricted
Verbmobil 1	spoken	German, English, Japanese	free for science
Verbmobil 2	spoken	German, English, Japanese	free for science
Verbmobil Emotion	spoken	German	free for science
ZipTel	spoken	German	free for science

Table 1: Overview over the 31 corpora currently provided by the BAS repository.

NM-MoCap – Natural Media Motion-Capture Corpus This corpus comprises audio, video and motion capture recordings of spontaneous speech and gestures for 18 subjects. It was curated for CLARIN as part of Curation Project 1 "Editing and Integration of multimodal resources in CLARIN-D" by the CLARIN-D Working Group 6 "Speech and Other Modalities".

PD1 – PhonDat 1 The corpus contains read German speech of 201 different speakers who were recorded at four different sites in Germany (Kiel, Bonn, Bochum, and Munich).

PD2 – PhonDat 2 The corpus contains German read speech recordings of 16 speakers in from a train query task. They were recorded at three different sites in Germany (Kiel, Bonn, and Munich).

PHATTSESSIONZ – Ph@ttsessionz Adolescents Speech Corpus This speech database contains recordings of 1019 adolescent speakers of German (age range 12-20). The recordings were performed via the WWW in public secondary schools (*Gymnasium*) in 45 locations in Germany.

RVG-1_CLARIN – Regional Variants of German 1 The corpus is a collection of more than 500 speakers of different dialect regions of Germany. It contains read and spontaneous speech recorded by four different low- and high-quality microphones in normal office environments.

RVG-J – Regional Variants of German J – Juveniles The corpus contains read and non-scripted German utterances by adolescent speakers between 13 and 20 years of age recruited in public schools in or near Munich.

SaGA – Bielefeld Speech and Gesture Alignment Corpus The corpus is made up of 25 dialogs of 50 interlocutors, who engage in a spatial communication task combining direction-giving and sight description. It contains annotated audio and video recordings. This Corpus was curated for CLARIN as part of the Curation Project "Editing and Integration of Multimodal Resources in CLARIN-D" by the CLARIN-D Working Group 6 "Speech and Other Modalities".

SC1 The corpus contains German read speech of 88 speakers, 16 native German L1 speakers and 72 L2 speakers born and educated in other countries. All speakers were reading Aesop's fable "Der Nordwind und die Sonne" ("The north wind and the sun", Wikipedia (2016)).

SC10 This corpus contains read and non-prompted German and mother tongue speech of 70 different speakers from 17 mother tongues in a variety of speaking styles e.g. reading, retelling, free talk etc. Recorded languages are: Arabic, Dutch, English, Finnish, French, German, Hungarian, Italian, Japanese, Modern Greek, Polish, Portuguese, Russian, Spanish, Swedish, and Turkish.

SC2 – Strange Corpus 2 Noises The corpus contains German read speech of 10 different car experts with screen prompted automobile diagnosis phrases recorded under real conditions in two different car maintenance halls.

SmartWeb Handheld (SHC), Motorbike (SMC), and Video (SVC) Corpus The SmartWeb UMTS data collection consists of three corpora *SHC*, *SMC*, and *SVC*, and comprises a collection of German user queries to a naturally spoken Web interface with the main focus on the soccer world series in 2006. The recordings include field recordings using a hand-held UMTS device (SmartWeb Handheld Corpus SHC), field recordings with video capture of the primary speaker and a secondary speaker (SmartWeb Video Corpus SVC) as well as mobile recordings performed on a motorbike (SmartWeb Motorbike Corpus SMC).

SIGNUM – BAS Database for Signer-Independent Continuous Sign Language Recognition The SIGNUM Database contains video recordings of both isolated and continuous utterances of 25 native signers. For quick access to individual frames, each video clip is stored as a sequence of images. The vocabulary comprises 450 basic signs in German Sign Language (DGS) representing different word types. The SIGNUM Database was created within the framework of a research project at the Institute of Machine Interaction, located at the RWTH Aachen University in Germany.

SmartKom SK Home, SK Mobile, SK Public The SmartKom (SK) data collection consists of three corpora *Home*, *Mobile*, and *Public*. Naive users were asked to test a prototype of an intelligent communication device for a market study not knowing that the system was in fact controlled by two human operators in a Wizard of Oz setting. Recorded and annotated modalities are emotional-state, facial expressions, gestures, and speech. SK Home and SK Mobil contain multi modal recordings of 65 and 73 subjects, respectively. Experiments were not performed in the field but rather in a studio-like environment. SK Public contains multi modal recordings of 86 subjects who use the SmartKom system.

TAXI The TAXI dialog was created in collaboration with the DFKI, Saarbruecken. It contains 86 recorded dialogs between a cab dispatcher and a client recorded over public phone lines (network and GSM). The dispatcher always spoke German, while the clients always spoke English.

VERIF1DE The German VERIF1DE speaker verification database consists of 150x20 phone calls and is a subset of the VERIDAT speaker verification database collected by T-Nova.

VM1 – Verbmobil 1 The Verbmobil 1 dialog database is a collection of German, American, and Japanese dialog recordings in the appointment scheduling task. 885 speakers participated in 1422 recordings.

VM2 – Verbmobil 2 The Verbmobil 2 dialog database is a collection of German, American, Japanese, and mixed language dialog recordings. 401 speakers participated in 810 recordings. The domain is appointment scheduling, travel planing, leisure time planing.

VMEmo – Verbmobil Emotion This database contains speech signals of dialogs in which a subject was recorded during a conversation via a spontaneous speech translation system. The response of the system was designed to invoke emotions in the subjects. VMEmo is part of the larger Verbmobil 2 speech data collection.

ZipTel The ZipTel telephone speech database contains recordings of people applying for a SpeechDat prompt sheet via telephone. The calls were recorded by an automatic telephone server. The database consists of 1957 recording sessions.

7. Acknowledgments

We thank the European CLARIN ERIC and the German CLARIN-D consortium funded by the German Ministry of Science and Education for establishing the infrastructure. The current work of the first author is funded by the Alexander von Humboldt society.

8. Bibliographical References

AsiCa. (2006). Last update 7 Mar 2006.

BAS. (2015). BAS federated content search endpoint. https://clarin.phonetik.uni-muenchen.de/BASSRU. Last update 18 Sep 2015.

BAS. (2016a). BAS clarin Repository. https://clarin.phonetik.uni-muenchen.de/BASRepository. Last update 3 Feb 2016.

BAS. (2016b). BAS OAI-PMH endpoint. http://www.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl?verb=Identify. Last update 18 Jan 2016.

BAS. (2016c). Bavarian archive for speech signals. http://www.bas.uni-muenchen.de/Bas/BasHomeeng.html. Last update 18 Feb 2016.

CLARIN. (2016a). CLARIN-D – Federated Content Search (CLARIN-FCS). https://www.clarin.eu/content/federated-content-search-clarin-fcs. Last update 10 Mar 2016.

CLARIN. (2016b). CLARIN-D web page. http://eu.clarin-d.de/index.php/en/. Last update 10 Mar 2016.

CLARIN. (2016c). CLARIN service provider federation. http://www.clarin.eu/content/service-provider-federation. Last update 10 Mar 2016.

CLARIN. (2016d). CMDI – Component metadata. http://www.clarin.eu/cmdi. Last update 10 Mar 2016.

CLIPS. (2004). CLIPS corpus. Last visit 10 Mar 2016.

DC. (2016). Dublin Core Metadata Initiative. http://dublincore.org/. Last update 1 Mar 2016.

DFN-AAI. (2010). DFN-AAI – authentication and authorization infrastructure. https://www.aai.dfn.de/en/. Last update 11 Sep 2010.

Draxler, C. and Schiel, F. (2002). Three New Corpora at the Bavarian Archive for Speech Signals - and a First Step Towards Distributed Web-Based Recording. In *Proc. LREC*, pages 21–24, Las Palmas, Gran Canaria, Spain.

EMU. (2010). The EMU Speech Database System. http://emu.sourceforge.net/. Last update 25 May 2010.

ePIC. (2016). epic – persistent Identifiers for eResearch. http://www.pidconsortium.eu/. Last update 10 Mar 2016.

Goosen, T. and Eckart, T. (2014). Virtual Language Observatory 3.0: What's New? In *CLARIN Annual Conference*, page 4 pages, Soesterberg, Netherlands.

GWDG. (2009). PID Handle Service. http://handle.gwdg.de:8080/pidservice/. Last update 9 Nov 2009.

LRZ. (2016). Leibnitz Rechenzentrum. https://www.lrz.de/. Last update 8 Mar 2016.

OLAC. (2011). OLAC – Open Language Archives Community. http://www.language-archives.org. Last update 24 Feb 2011.

Reichel, U. (2013). Das BAS-Repository. CLARIN-D Newsletter 5, pp. 22–26.

Rivest, R. (1992). The MD5 Message Digest Algorithm. Internet RFC 1321; http://people.csail.mit.edu/rivest/Rivest-MD5.txt.

Shibboleth. (2016). Shibboleth. https://shibboleth.net/. Last update 25 Feb 2016.

Suleman, H. (2002). OAI-PMH2 XMLFile File-based Data Provider. http://www.dlib.vt.edu/projects/OAI/software/xmlfile/xmlfile.html. Last update 12 Dec 2002.

SURFconext. (2016). Surfconext. https://www.surf.nl/diensten-en-producten/surfconext/index.html. Last update 25 Jan 2016.

VLO. (2015). Virtual Language Observatory. https://vlo.clarin.eu. version 3.3.2, last update 3 Nov 2015.

Wikipedia. (2016). The north wind and the sun. https://en.wikipedia.org/wiki/The_North_Wind_and_the_Sun. Last update 4 Mar 2016.

Winkelmann, R. (2015). Managing speech databases with emuR and the EMU-webApp. In *Proc. Interspeech*, Dresden, Germany.