AP5 Fifth Year Activity Report

May 2016
Contents

1 Introduction and overview 4

2 Curation and adaptation of language resources 6
  2.1 Text corpora and archives ........................................ 6
  2.2 Speech and multimodal corpora .................................. 13
  2.3 Lexical resources ................................................ 17
  2.4 Other resources ................................................ 18

3 Integration of language tools 19
  3.1 CLARIN-D infrastructure development .............................. 19
  3.2 Tokenizers and sentence splitters .................................. 20
  3.3 Morphological analysers ............................................ 20
  3.4 Named entity recognizers .......................................... 20
  3.5 Parsers ........................................................... 21
  3.6 Spoken language tools .............................................. 21
  3.7 Other tools ..................................................... 22

Throughout this document, CLARIN-D member institutions appear abbreviated:

  BAS     Bayerisches Archiv für Sprachsignale, LMU München
  BBAW    Berlin-Brandenburgische Akademie der Wissenschaften
  HZSK    Hamburger Zentrum für Sprachkorpora, Universität Hamburg
  IDS     Institut für deutsche Sprache, Mannheim
  IMS     Institut für maschinelle Sprachverarbeitung, Universität Stuttgart
  MPI     Max Planck Institute for Psycholinguistics, Nijmegen (NL)
  UdS     Universität des Saarlandes
  ULei    Universität Leipzig
  UTüb    Universität Tübingen
1 Introduction and overview

This report documents the work and achievements of the CLARIN-D consortium with respect to the tasks defined for AP5 “Dienste und Ressourcen” (work package 5, tools and resources) during the fifth year of CLARIN-D (from June 2015 to May 2016). The work covered in this report generally builds upon the curation efforts for linguistic resources and tools (LRT) reported in the preceding four yearly activity reports R5.1–R5.4. The time period that this report covers concludes the implementation phase of CLARIN-D. Combined, the five existing reports describe all achievements of work package 5 throughout this phase.

During the fifth project year, CLARIN-D member institutions again focussed on the curation and integration of LRT that originated from different research institutions outside the consortium. However, as in previous years, all partners also continued integrating and further curating their own existing LRT. These efforts are described in more detail in sections 2 and 3.

Minor maintenance work and bug fixes of LRT already integrated during the first four years of the project are generally not mentioned explicitly. The focus of this report is decidedly on newly added LRT and significant technical or functional improvements of LRT integrated earlier.

VLO Taskforce (VLO-TF)

At the end of 2014, AP5 members formed a dedicated taskforce as a curation effort specifically targeting metadata curation with respect to its application in the Virtual Language Observatory (VLO). See report R5.4 for more detailed background information and motivation for the foundation of the VLO-TF. The VLO-TF was coordinated by BBAW and concluded its work at the end of 2015.

The work and results are documented in detail on the CLARIN community’s TRAC pages at https://trac.clarin.eu/wiki/VLO-Taskforce (login required). Many of the proposals and suggestions of the taskforce have already been implemented by the VLO development team.

The VLO-TF worked in close cooperation with related taskforces on CLARIN ERIC level, including:

**CLARIN Concept Registry Coordinators:** The CCR coordinators have been using the VLO facet definitions and the proposed ISOcat data categories for those facets as the starting point for their curation work on the CLARIN Concept Registry (CCR). Eventually, this will provide a solid semantic base for the VLO’s facet implementation.

**Metadata Curation Taskforce:** This taskforce aims at designing an evaluation scheme for metadata quality among all European CLARIN centers. The
initial use case for this taskforce are the metadata representations within the VLO based on the VLO-TF’s proposals and their implementations by the CLARIN-D centers. Some VLO-TF members have been continuing their related work in the Metadata Curation Taskforce.

CMDI-1.2 Taskforce: Many members of the VLO-TF continued developing the CMDI-1.2 specification and framework.

The VLO-TF served as a major initial force for many of the cooperating European taskforces. Its legacy will be upheld especially by the taskforce on Metadata Curation and the CCR coordinators. This effectively promotes the aims and purposes of the VLO-TF at the European level and fosters the participation of members from other CLARIN initiatives.

1.0.1 DSA General Assembly

Due to the success of the Data Seal of Approval (DSA), a General Assembly was launched. The DSA General Assembly will put the DSA on a path to becoming more community-driven and sustainable. General Assembly representatives have voting rights; they elect the DSA Board and advise the Board when needed. General Assembly representatives are also electable as DSA Board members themselves. The DSA Regulations\(^1\) provide additional details on the rights and responsibilities of the General Assembly members.

CLARIN-D is represented in the DSA General Assembly with two centers (HZSK and UdS).

The General Assembly met for the first time on September 15th, 2015. A Webinar for new DSA reviewers was held on January 26th, 2016.

2 Curation and adaptation of language resources

In this section we summarise all LRT that were made available to CLARIN during the fifth year of the project. We also included LRT that were already mentioned in previous reports if significant additional adaptations or improvements were made.

2.1 Text corpora and archives

2.1.1 CoInCo Corpus (IMS)

The lexical substitution corpus CoInCo (Concepts in Context) was developed at the Institute for Computational Linguistics, University of Heidelberg, and is based on contiguous texts provided in MASC (the manually annotated subcorpus of the Open American National Corpus, OANC). It contains substitute words for every content word in selected (complete) text files. These substitute words were collected via crowdsourcing (using Amazon Mechanical Turk) by six different annotators per token. Annotators saw a target sentence and the preceding and following sentence in the original text file as context. The dataset comprises more than 150,000 responses for around 15,000 targets in about 2,500 sentences containing approximately 35,000 words. Targets are roughly balanced across the genres “news” and “fiction”. For more information see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/coinco.html. A CMDI metadata description for the corpus was created and the respective PIDs were registered.

2.1.2 Corpus of Comparisons in Product Reviews (IMS)

The corpus contains annotations for 1,707 sentences from camera reviews annotated with comparisons. The sentences were taken from the epinions.com dataset by Branavan et al. (2009). For more information see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reviewcomparisons/index.html. A CMDI metadata description for the corpus was created and the respective PIDs were registered.

2.1.3 DDR-Presseportal (BBAW)

In cooperation with the Staatsbibliothek zu Berlin and the Zentrum für Zeithistorische Forschung (ZZF, Potsdam), the data of the DDR-Presseportal was curated and made available to CLARIN through the DWDS website hosted at the BBAW. This archive comprises complete corpora of official press publications from the German Democratic Republic:

- Neues Deutschland (ND, 1946–1990, approximately 445 million tokens)
• Berliner Zeitung (BZ, 1945–1993, approximately 500 million tokens)
• Neue Zeit (NZ, 1945–1994, approximately 400 million tokens)

The corpora are lemmatized and part-of-speech-annotated. For more information on the DDR-Presseportal see http://sbb.berlin/ddrpresse.

2.1.4 DeReKo (IDS)

The German Reference Corpus (DeReKo) is the largest collection of electronic corpora of contemporary written German. As of March 2016, it comprises approximately 28.5 billion words. New newspaper sources have been added in 2015 including Süddeutsche Zeitung (from 1992), Die ZEIT (from 1953), NZZ, FOCUS, Falter, Profil, Weltwoche, and Luxemburger Tag (each from 2000). A new compilation of Wikipedia corpora of 9 languages (German, English, French, Hungarian, Norwegian, Spanish, Croatian, Italian, Polish) from 2015 dumps, and a Usenet newsgroup corpus with 374 German speaking newsgroups have also been added.

The Corpus Search, Management, and Analysis System (Cosmas II) at the IDS cannot handle an archive of this size with reasonable indexing and query response times. Therefore, only subsets of DeReKo are included in Cosmas II. Because of this limitation, IDS has been developing a new corpus analysis platform KorAP since 2011 (see section 3.7.6).

2.1.5 Deutscher Wortschatz (ULei)

For many years, the project Deutscher Wortschatz has been offering large corpora and statistical data on its corpus portal. Currently, these corpora are available in more than 200 different languages. The data are accessible via a web portal (http://corpora.informatik.uni-leipzig.de/), SOAP web services (http://wortschatz.uni-leipzig.de/Webservices/), and some of the corpora are available for direct download (http://corpora.informatik.uni-leipzig.de/download.html). A formal description of the data and a specification of the Wortschatz corpus format is also available. For further information on the Deutscher Wortschatz project see also report R5.2.

The software tool that allows the automatic deployment of Wortschatz corpora (including supporting web services and metadata) into the CLARIN infrastructure was further stabilized and developed to accommodate changes in the Wortschatz project’s architecture.

By the end of the fifth year, the number of available datasets in the CLARIN infrastructure was further increased, e.g., for metadata search in the VLO, querying via the Federated Content Search (FCS) and for processing with WebLicht. Corpora for a few missing European languages, newer corpora for highly demanded languages
(like English and German), and corpora for other important languages of the world (according to the number of speakers) were checked for quality and introduced into the CLARIN ecosystem.

2.1.6 Dortmunder Chatkorpus 2.0 (IDS, BBAW)

The Dortmunder Chatkorpus was built at TU Dortmund from 2002 to 2008. It contains 478 chat recordings from leisure time, teaching and consulting chats comprising one million tokens in total. IDS and BBAW actively contributed to a CLARIN-D curation project aiming to integrate the corpus in the CLARIN infrastructure. The integration includes:

- modeling a scheme for the resource representation in accordance with the guidelines of the Text Encoding Initiative (TEI),
- conversion of the resource into the TEI target format,
- basic linguistic annotation (tokenization, POS-Tagging, lemmatization) using the annotation pipeline for computer-mediated communication texts built by the University of Saarland in the course of the BMBF project “Schreibgebrauch” and building on the extended STTS Tagset 2.0-BETA,
- manual adjudication of tagger outputs with the STTS Tagset 2.0-ALPHA,
- clarification of the legal terms of the provision,
- anonymization of participants’ names, and
- the actual integration of the resources in the BBAW and IDS repositories and corpus query engines.

For the adjudication task, the annotation environment OrthoNormal (see section 3.7.9) was adapted for use on the curation project’s target format.

The target TEI scheme for representing computer-mediated communication was developed in close cooperation with the TEI special interest group on computer-mediated discourse (http://www.tei-c.org/Activities/SIG/CMC/), in which representatives from both BBAW and IDS are active members.

2.1.7 DTA (BBAW)

The Deutsches Textarchiv (DTA, German Text Archive) provides a broad selection of more than 1300 significant German works of various disciplines. Their time of origin ranges from approximately 1600 to 1900. Due to the DTA’s primary focus on
the history of the German language, the earliest editions accessible are digitized, and the full text transcriptions meticulously document the original historic texts.

All DTA texts can be downloaded or harvested via OAI/PMH in different formats (including TEI/XML and TCF) and may be used under a Creative Commons license. For further background information on the DTA see the previous reports (R5.3, R5.4) and http://www.deutschestextarchiv.de/.

As of April 2016, the DTA core corpus comprises digitized and structurally as well as linguistically annotated full texts of 1357 volumes with 504,538 pages (approximately 122 million tokens).

During the report period, the DTA corpus was again further extended on the basis of cooperations with partner institutions outside of CLARIN-D:

Funeralschriften: In cooperation with the Staatsbibliothek zu Berlin, approximately 20,000 funeral eulogies from the 16th to the 19th century were curated into the DTA base format. The quality of the initial OCR results was considerably increased. For more information see http://www.deutschestextarchiv.de/doku/textquellen#sbb_funeralschriften.

Alexander von Humboldts unselbstständige Schriften: In cooperation with the University of Potsdam, 164 essays and scientific papers of Alexander von Humboldt were digitized (double-keying) and annotated according to the DTA base format. For more information see http://www.deutschestextarchiv.de/doku/textquellen#avh.

GEI digital: In cooperation with the Georg-Eckert-Institut for International Textbook Research and in coordination with a CLARIN-D curation project (“Realkundliches- und naturwissenschaftliches Wissen für Dilettanten und Experten zwischen Aufklärung und Moderne”), a selection of historical textbooks were curated and integrated. The quality of the initial OCR results was considerably increased and the texts were annotated according to the DTA base format. For more information see http://www.deutschestextarchiv.de/doku/textquellen#gei.

Magazin zur Erfahrungsseelenkunde: In cooperation with the project “Kritische Karl Philipp Moritz-Gesamtausgabe” hosted at the BBAW and the library at the University of Bielefeld, the data of the earlier digital edition of this publication (1783–1793) were transformed into the DTA base format and aligned with existing scans. This corpus comprises 30 volumes. For more information see http://www.deutschestextarchiv.de/doku/textquellen#mze-moritz.

In total, DTA and its extension corpora currently (April 2016) comprise approximately 200 million tokens.
All data and metadata are publically available at the DTA website at http://www.deutschestextarchiv.de/dtaq/ as well as from the CLARIN repository at the BBAW.

2.1.8 GRUG (UdS)

GRUG (Georgian Russian Ukrainian German parallel treebank) consists of four monolingual treebanks for Georgian, Russian, Ukrainian, and German and of four parallel Treebanks (German–Georgian, German–Russian, German–Ukrainian, Georgian–Ukrainian). The parallel texts comprise German sentences and their translations into Georgian and Russian compiled for the GREG NLP lexicon project. The languages (except German) involved in the project are considered “lesser-resourced” languages from the computational viewpoint.

GRUG was already curated and archived in the repository at UdS in 2012. The second release is an extension of the monolingual German and Georgian treebanks together with their corresponding parallel German-Georgian treebank. 124 new sentences have been added to the German-Georgian component meaning an increase of almost 300% in size.

2.1.9 Index Thomisticus (UTüb)

A new release of the Index Thomisticus treebank (IT-TB) has been made available by its creators. This treebank is an extract of the works of Thomas Aquinas in Latin, manually annotated for syntactic dependencies, lemmatization, morphology and part-of-speech information. Presently, the IT-TB consists of 15,295 sentences. It is available to the public under a Creative Commons license and can be downloaded from the Tübingen repository, with a harvestable CMDI and a registered PID.

2.1.10 KoLaS (HZSK)

KoLaS (Kommentiertes Lernendenkorpus akademisches Schreiben, Commented Learner Corpus Academic Writing) contains texts in German (mainly seminar papers) from students, most of which use German as an L2, at the University of Hamburg. The texts were commented and revised during the peer tutoring for academic writing at the University of Hamburg (Schreibwerkstatt Mehrsprachigkeit) and most texts thus exist in several consecutive versions.

2.1.11 Old Bailey Corpus (UdS)

The Old Bailey Corpus is a socio-biographically, pragmatically, and textually annotated selection of the Proceedings of the Old Bailey. These speech-related texts document Late Modern English in London’s Central Criminal Court from
1720 to 1913. The corpus is the largest diachronic collection of spoken English with this detail of utterance-level sociolinguistic annotation. Almost 200 years of spoken Early Present Day English are annotated for the following socio-biographic, pragmatic and textual parameters:

**Socio-biographical speaker information:** gender, age, occupation (according to the Historical International Standard Classification of Occupations, HISCO), social-class (according to HISCLASS, a social-class scheme based on HISCO)

**Pragmatic information:** speaker role in the courtroom, e.g., defendant, judge, victim, witness, ...

**Textual information:** scribe, printer and publisher of the proceedings

The time span covered by the Old Bailey Corpus and the available socio-biographical speaker information are ideally suited for fine-tuned studies, including historical sociolinguistic approaches.

In addition, because of sheer size the Proceedings of the Old Bailey are a valuable textual source for the analysis of low-frequency features. For more information on the corpus see [http://www1.uni-giessen.de/oldbaileycorpus/](http://www1.uni-giessen.de/oldbaileycorpus/).

The Old Bailey Corpus was integrated into the CLARIN infrastructure (including FCS and a CQPweb installation at [https://corpora.clarin-d.uni-saarland.de/cqpweb/obc/](https://corpora.clarin-d.uni-saarland.de/cqpweb/obc/)) in March 2016.

### 2.1.12 Perseus Ancient Greek Dependency Treebank (UTüb)

UTüb has arranged with the creators of the Perseus Ancient Greek Dependency Treebank to host version 2.0 of the treebank in TüNDRA (see section 3.1.1). It includes texts attributed to Aesop, Aeschylus, Athenaeus, Herodotus, Hesiod, Homer, Lysias, Plato, Plutarch, Polybius, and Sophocles and currently contains 33,555 sentences annotated with syntactic dependencies, lemmatization, morphology and part-of-speech information. The treebank is available to the public under a Creative Commons license at [https://github.com/PerseusDL/treebank_data](https://github.com/PerseusDL/treebank_data).

### 2.1.13 Perseus Latin Dependency Treebank (UTüb)

UTüb has arranged with the creators of the Perseus Latin Dependency Treebank to host version 2.0 of the treebank in TüNDRA (see section 3.1.1). It includes texts attributed to Caesar, Cicero, Jerome, Vergil, Ovid, Petronius, Phaedrus, Sallust, Suetonius and currently contains 4,790 sentences annotated with syntactic dependencies, lemmatization, morphology and part-of-speech information. The treebank is available to the public under a Creative Commons license at [https://github.com/PerseusDL/treebank_data](https://github.com/PerseusDL/treebank_data).
2.1.14 Pros/Cons Sentences Automatically Annotated with Polarity (IMS)

This corpus comprises a set of 58,503 sentences annotated with polarity by taking advantage of semistructured reviews from epinions.com collected by Branavan et al. (2009) (camera and cellphone data sets, 17,442 reviews in total). For more information see http://wiltrud.hwro.de/research/data/proscons.html. A CMDI metadata description for the corpus was created and the respective PIDs were registered.

2.1.15 Royal Society Corpus (UdS)

The Royal Society Corpus (RSC) is a diachronic corpus of written scientific English. It contains articles from the Philosophical Transactions and Proceedings of the Royal Society of London from the first two centuries of the journal (1665–1869). As a historical corpus it represents a middle ground between big and “poor” data and small and “rich” data comprising approximately 35 million tokens.

The RSC contains metadata for each article such as author, title, year and type of publication. Besides, the corpus is annotated with linguistic annotation on token level, such as word, lemma, part-of-speech and normalized (modernized) word forms. The RSC is searchable via the CQPweb interface of UdS and will be fully integrated into the CLARIN infrastructure including FCS by May 2016.

2.1.16 SFB 632 datasets (UTüb)

In cooperation with information infrastructure projects of special research centers (SFBs), CLARIN-D offers sustainable solutions for storing research data. As an external data provider, UTüb accepted to host data of the SFB 632 (“Information structure: The linguistic means for structuring utterances, sentences and texts”, Potsdam) which came to an end in 2015, in addition to the Tübingen SFBs 441 and 833, with which the CLARIN center at UTüb cooperated before.

A visualization tool for the SFB 632 dataset was developed by IMS (see section 3.7.1).

2.1.17 SFB 632 corpora of the projects A5 and B7 (HZSK)

The following resources created within the special research center (SFB) 632 on information structure have been integrated:

- From the project A5 “Focus realization, focus interpretation, and focus use from a cross-linguistic perspective”, an annotated Hausa news corpus.
• From the project B7 “Predicate-centered focus types: A sample-based typological study in African languages”, two annotated Wolof corpora from Wikipedia and web forum sources.

2.1.18 SubCo (UdS)

SubCo (SUBtitle COrpus) is a corpus comprising original English subtitles (SRC), both human (HT) and machine translations (MT) into German, as well as post-editions (PE) of the MT output. HT and MT are annotated with errors. Moreover, human evaluation is included for HT, MT, and PE.

The source text used for this corpus is the documentary film “Joining the Dots” by Pablo Romero Fresco. The source text contained 132 subtitles amounting to 1557 words.

The translations were produced by human translators and by SUMAT, an MT system developed and trained specifically for the translation of subtitles.

The corpus was compiled as part of a course on subtitling targeted at students enrolled in the BA Translation Studies programme at UdS. The students carried out the human translations and the error and evaluation annotation.

Both HT and MT were annotated with labels for the identification of the errors and were assigned evaluation scores at sentence level.

2.1.19 TüBa-D/Z (UTüb)

Release 10.0 of TüBa-DZ has been made available for academic use in CLARIN. The TüBa-D/Z treebank is a syntactically annotated German newspaper corpus based on data taken from the daily issues of “die tageszeitung” (taz). The treebank currently comprises 95 595 sentences (1 787 801 tokens; 3 644 newspaper articles).

2.2 Speech and multimodal corpora

2.2.1 Speech corpora at the BAS

BAS now supports FCS across most of the speech corpora in its repository (except for pure video resources).

Throughout the fifth year, the following resources have been added to the BAS repository:

CI_1: The CI_1 corpus contains speech of patients with Cochlear implants.

HOESI: The HOESI corpus comprises speech of elderly speakers wearing hearing aids and communicating under the influence of noise.
**SC10:** The SC10 corpus contains read and non-prompted German and mother tongue speech of 70 different speakers from 17 mother tongues (L1) in a variety of speaking styles, e.g., reading, retelling, free talk.

**SC2:** The SC2 corpus contains read speech of ten different speakers with screen prompted “automobile diagnosis phrases” recorded under real conditions in two different car maintenance halls. The language is German. All speakers are male native Germans and have never participated in such a task before. They are all experts in the field of car diagnosis. Each speaker produced 800 utterances consisting of three to seven words derived from 100 different sentences, resulting in a total of 8000 utterances.

### 2.2.2 Speech corpora at the HZSK

**EuroWiss:** The EuroWiss corpus contains authentic oral university communication including various communication types on several subjects from German and Italian universities. In addition to audio and transcription data, the corpus includes metadata on speakers and communications and relevant accompanying written materials.

**Wahrnehmungsdialektologie:** The corpus from the research project “Der deutsche Sprachraum aus der Sicht linguistischer Laien” on perceptual dialectology covers regional varieties from the entire Northern Germany area. In addition to audio and transcription data, the corpus includes detailed speaker metadata, metadata on communication settings, and partly further materials such as mental maps.

**SFB 632 corpora:** The following resources created within the special research center 632 on information structure (SFB 632 “Informationsstruktur”) have been integrated:

- from the project A5 “Focus realization, focus interpretation, and focus use from a cross-linguistic perspective”, an annotated Hausa corpus based on the film “Umarmin Uwa”
- from the project B1 “Focus in Gur and Kwa languages”, four annotated corpora on the languages Aja, Fon, Foodo, and Yom, respectively
- from the project B2 “Focus in Chadic Languages”, three annotated corpora on the languages Guruntum, Marghi, Tangale, respectively.

### 2.2.3 Speech corpora at the IDS

The “Datenbank für Gesprochenes Deutsch” (DGD2, German Spoken Language Database) comprises all speech corpora at IDS that are publically available.
metadata descriptions for DGD2 are integrated in the IDS repository.

During the report period, the following resources were integrated into DGD2:

**Deutsche Mundarten, DDR:** The Deutsche Mundarten, DDR corpus (DR) originates from the Akademie der Wissenschaften der DDR (Academy of Science of the German Democratic Republic) and was compiled by IDS. Following the recording scenario of the former Deutsches Spracharchiv (DSAv), samples of dialects spoken in East Germany were collected. Comparable material should be made from the collected samples according to uniform criteria.

The corpus contains recordings of stories, conversations and standard texts (comparative texts, word lists) from 1,582 speakers from the DDR and the former east German territories. Wenkersätze were used as the text material for the dialect comparison. The speakers were asked to translate the text and thereby freely use any word choice and word order. The recordings have been digitized in AGD (“Archiv für Gesprochenes Deutsch”, Archive for Spoken German), but the transcripts are not archived. The metadata of the corpus in CMDI are available in the IDS repository.

**Jugendkommunikation:** The Jugendkommunikation corpus (JK) was compiled within a project of the Goethe University Frankfurt in the Rhine-Main region. The project goal was to investigate everyday communication culture of adolescents in an ethnographic conversion study. The main concerns are linguistically-interactive ethno methods used by adolescents to socialize in a group, and the forms of social categorization used to interpret themselves and their social environment.

For external users, some parts of the corpus from 1996 to 1999 are available. Those parts contain six audio recordings with a total duration of four hours and 42 minutes. The speakers were teenagers (later adults) living in a small rural town in the Rhine-Main region and regularly visited the local youth club. The recordings were digitized by AGD and this corpus part is available in DGD. The corpus metadata in CMDI will be available in the IDS repository.

### 2.2.4 Speech corpora at the MPI

The following new corpora have been added to the archive:

**Preference and Timing:** This corpus contains raw EEG recordings of participants while listening to questions, offers, requests, proposals, and invitations from telephone conversations from the “Corpus Gesproken Nederlands” (Corpus of Spoken Dutch), followed by a silence of 300 or 1,000 ms and a “yes” or a “no” from elsewhere in the corpus. For more information on Preference and Timing see [https://hdl.handle.net/1839/00-0000-0000-0021-7E25-C@view](https://hdl.handle.net/1839/00-0000-0000-0021-7E25-C@view).
Breathing: This corpus consists of recordings of spontaneous dyadic conversations between friends including audio, video, and inductive plethysmography signals, as well as orthographic transcriptions and English translations. For more information on Breathing see https://hdl.handle.net/1839/00-0000-0000-0021-6297-B@view.

Modelling Noise-Robustness: Modelling the noise-robustness of infants’ word representations: the impact of previous experience. For more information on this corpus see https://hdl.handle.net/1839/00-0000-0000-0021-5234-D@view.

XTYP Lab: This set comprises various language corpora collected by the XTYP Lab group at the University of Bielefeld. For more information on XTYP Lab see https://hdl.handle.net/1839/00-0000-0000-0021-4DA2-2@view.

Iwaidja and Ikaan: The project investigates the interplay of gestures and prosodic events in two languages with very different prosody: Iwaidja, a stress-accent language spoken in Northern Territory, Australia, and Ikaan, a tone language spoken in Nigeria. Iwaidja speakers are bilingual with English as their dominant language. Ikaan speakers are multilingual, speaking Ikaan, Yoruba, further indigenous languages and English and Pidgin English to varying degrees. The project compares the way in which gestures and speech are coordinated in the speakers’ two languages, e.g. Iwaidja and English. Integration of this resource was possible through cooperation with DoBeS. For more information on Iwaidja and Ikaan see https://hdl.handle.net/1839/00-0000-0000-0021-6DD4-3@view.

Douala: The project on the language of perception in Douala (with special attention to smell and taste) is part of a larger project on olfactory language and cognition of the research group Meaning, Culture and Cognition, Center for Language Studies, Radboud University Nijmegen. For more information on Douala see https://hdl.handle.net/1839/00-0000-0000-0021-632D-D@view.

Psychology of Language: This set comprises various corpora from the MPI’s Psychology of Language department, e.g. Mental Math, Predicting Language in Context, Rhythmic Attention, Sustained Attention, and Incrementality and Flexibility in Sentence Production. For more information on these projects see https://hdl.handle.net/1839/00-0000-0000-001A-A607-3@view.

Neurobiology of Language: This set comprises various corpora from the MPI’s Neurobiology of Language department, e.g. Searching High and Low, Discourse Coherence EEG, Frequency Tagging EEG, RC Comprehension MEG,
Cognition 2015, and Personal Pronouns Affect Literary Reading. See https://hdl.handle.net/1839/00-0000-0000-000F-12AE-0@view for more information on these projects.

Several corpora have also been greatly amended, in particular:

**Evolution of Semantic Systems:** The project investigates how meanings vary over space and change over time. It focuses on different kinds of categories: containers (kinds of objects), colour (attributes of objects), body parts (parts of objects), and spatial relations (how objects are related to one another). For more background information see http://www.mpi.nl/departments/other-research/research-consortia/eoss.

The archive consists of data from 62 languages of the Indo-European language family, two languages of the Uralic language family and data from the European Isolate. It contains the audio recordings and coded data from around 20 participants of each language. For more information on this archive see https://hdl.handle.net/1839/00-0000-0000-0016-3009-3@view.

**Vodun Stories:** The corpus comprises Vodun stories recorded in 1976 in Benin. For more information on this resource see https://hdl.handle.net/1839/00-0000-0000-001E-9C2D-F@view.

### 2.3 Lexical resources

#### 2.3.1 DWDSWB (BBAW)


The dictionary was actively edited and considerably extended throughout the report period. Approx. 70,000 entries from the “Großes Wörterbuch der deutschen Sprache” (1999) were integrated, partly substituting outdated or biased entries from WDG. Additionally, approximately 45,000 basic articles (without semantic descriptions) were added, on the basis of the DWDS corpora. These basic articles will serve as the core of the manual lexicographical work currently undertaken at the BBAW. Based on those sources, the DWDSWB currently comprises approximately 210,000 entries. For more information on DWDSWB see http://zwei.dwds.de/.

The lemma list of DWDSWB is available through the CLARIN repository at the BBAW including grammatical properties of the lemmas, detailed CMDI metadata, and PIDs.
2.3.2 EtymWB (BBAW)

The “Etymologisches Wörterbuch des Deutschen” (Etymological Dictionary of the German Language, EtymWB) by Wolfgang Pfeifer is based on the retro-digitized print edition published in 1989. It was actively edited and extended throughout the report period. EtymWB currently comprises approximately 9500 morphologically simple headwords together with approximately 14000 subordinated headwords. For more information on EtymWB see previous annual reports.

The lemma lists of all versions of EtymWB are made available through the CLARIN repository at the BBAW including grammatical properties of the lemmas, detailed CMDI metadata, and PIDs.

2.3.3 GermaNet (UTüb)

During the reporting period, GermaNet, a lexical-semantic web, was further extended. GermaNet relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets. GermaNet has much in common with the English WordNet and can be viewed as an on-line thesaurus, or a light-weight ontology.

Release 11.0 contains 10000 new entries (7870 nouns, 1420 verbs, and 710 adjectives), with a total of 108920 synsets and 141810 lexical units.

2.3.4 Polarity Reversing Constructions (IMS)

Polarity reversing constructions are syntactic constructions, which reverse the polarity of a sentiment word. This resource is a list of such constructions for English. The constructions were annotated manually by an annotator. For more information see http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/sentiment_prcs.html. A CMDI metadata description for the resource was created and the respective PIDs were registered.

2.4 Other resources

2.4.1 LAUDATIO repository (IMS)

The LAUDATIO (Long-term Access and Usage of Deeply Annotated Information) team has built an open access research data repository at Humboldt University, Berlin, for historical linguistic data. For access and (re)use of historical linguistic data, the LAUDATIO repository uses a flexible and appropriate documentation schema with a subset of TEI customized by TEI ODD. The extensive metadata schema contains information about the preparation and checking methods applied
to the data, tools, formats and annotation guidelines used in the project, as well as bibliographic metadata, and information on the research context (e.g., the research project). For more information see http://www.laudatio-repository.org/repository/. A CMDI metadata description for this repository was provided by the resource creators. The respective PIDs were registered.

2.4.2 Semantic norms (IMS)

The psychological community frequently investigates semantic norms of properties produced by native speakers after being presented concept words, and these norms are of great value for a wide variety of psychological experiments. This resource was developed at the Center for Mind/Brain Sciences, University of Trento, and presents a set of norms that includes a collection of properties from a production experiment for the German and the Italian languages. Stimuli consisted of 50 concrete objects taken from 10 different concept classes. The data comprise annotations of semantic relation types and several statistical measures, which facilitates the comparison of the two target languages. For more information see //dx.doi.org/10.3758/s13428-010-0028-x. A CMDI metadata description for the data set was created and the respective PIDs were registered.

3 Integration of language tools

3.1 CLARIN-D infrastructure development

Most of the development work of the core infrastructure for CLARIN-D is reported by AP3 (see the R3.5 report). In this section, we shortly describe the efforts that have been contributed by AP5 members to foster LRT harmonization and compatibility. Much of the effort reported in this section concerns the WebLicht web-service orchestrator. For more details on WebLicht see previous reports for both AP3 and AP5.

3.1.1 TüNDRA treebank search and visualization (UTüb)

TüNDRA is a web application for searching treebanks. During the fifth year, work was started on a major rewrite of the source code. The refactoring includes a better separation of the frontend and backend, using current rendering libraries for the visualizations, and incorporating full visualization and search of TCF data. The import of TCF data will allow WebLicht to share TüNDRA’s frontend, which is now only possible for TCF data that contains parse trees. This code sharing will have the advantage that all TCF data will be searchable, and a more unified user experience can be achieved.
Several treebanks have been added to TüNDRA:

- release 10.0 of the TüBa-D/Z (see section 2.1.19)
- release 10.0 of the TüBa-D/Z with automatically constructed dependencies
- the new release of the Index Thomisticus (dependency, see section 2.1.9)
- Perseus Ancient Greek v2.0 (dependency, see section 2.1.12)
- Perseus Latin v2.0 (dependency, see section 2.1.13)

3.2 Tokenizers and sentence splitters

3.2.1 Turkish sentence splitter and tokenizer (UTüb)

This service makes use of an existing OpenNLP-based sentence and word tokenizer. The sentence splitter is trained on the Turkish part of the Leipzig corpora collection (see section 2.1.5). The word tokenizer is trained on the METU-Sabancı treebank.

3.3 Morphological analysers

3.3.1 MorphAdorner (UTüb)

MorphAdorner (http://morphadorner.northwestern.edu/), a Java-based tool for performing a number of linguistic annotation tasks on English texts, has been integrated into WebLicht in form of two services: lemmatization and morphological feature annotation.

3.3.2 Turkish morphological processing (UTüb)

This service performs morphological analysis and disambiguation for Turkish. The morphological analysis is done using TRmorph. The disambiguation system for this service is developed in-house. Since Turkish dependency parsing uses sub-word syntactic units, this service accepts plain text input, and produces a tokenization where words may be tokenized into multiple syntactic units. Each token is associated with the most likely POS tag and morphological features.

3.4 Named entity recognizers

3.4.1 Illinois named entity recognizer (UTüb)

This is a state of the art named entity tagger for English that tags plain text with named entities. This service tags entities with the “classic” four label type set
3.5 Parsers

3.5.1 Charniak (UTüb)

This service is based on the well-known Charniak parser for English constituency parsing. It requires tokenized text in the input and produces constituency parses along with POS tags.

3.5.2 TurboParser (UTüb)

TurboParser is a multi-lingual statistical dependency parser. This web service uses TurboParser to annotate POS-tagged input with dependency parses.

3.5.3 Turkish dependency parsing (UTüb)

This service makes use of an existing web service based on the Malt Parser. The model used for parsing Turkish was trained on the METU-Sabancı treebank. The service requires POS tags and morphological features assigned to sub-word units as determined by the morphological analysis and disambiguation service described in section 3.3.2. The output of the service are labeled dependencies using the METU-Sabancı tagset.

3.6 Spoken language tools

3.6.1 Chunk preparation service (BAS)

This is a new service that allows users to process manually created so-called “chunk segmentations” in Praat, TextGrid or ELAN EAF files to be processed by MAUS.

3.6.2 ELAN (MPI)

Throughout the reporting period, CLARIN’s annotation tool ELAN was continuously developed and improved further. This includes:

- a first implementation of an inter-rater agreement calculation, which also supports the Staccato algorithms (developed by F-AG 6, Bielefeld),
- further improved integration in WebLicht,
- a new VLC-based video player for Linux.
From June 2015, the work on ELAN was mainly centered on bug fixes, but also included minor changes and improvements, such as the preliminary development of a beta version for Mac OS X, which uses a JavaFX player (instead of an outdated version), adding Korean as a language to the user-interface, among others. For more detailed information see https://tla.mpi.nl/tools/tla-tools/elan/release-notes/. Currently, the development of ELAN includes building an extensional lexicon component.

3.6.3 G2P (BAS)

Grapheme-to-phoneme conversion (G2P) is a necessary preprocessing step for the automatic segmentation, but it is also a useful service in its own right. G2P has been extended to currently 13 languages. Several input and output formats have been added to the service.

3.6.4 TextAlign (BAS)

This new service allows the optimized alignment of symbolic strings, e.g., two different pronunciations of the same utterance, based on dynamic programming and decision trees.

3.6.5 WebMAUS (BAS)

MAUS is a tool for the automatic segmentation of speech signals based on forced alignment and WebMAUS is its web-based graphical frontend. The automatic segmentation system was first installed in 2012. In the reporting period, new languages were added (e.g. French, Russian, Georgian, Dutch), and the output now also allows IPA manner or place. The code of the web application has been restructured, and the web layout has been adapted to the CLARIN-D design guidelines. For further information on the WebMAUS tool see https://clarin.phonetik.uni-muenchen.de/BASWebServices/.

3.7 Other tools

3.7.1 ANNIS Search and Visualization System for Corpora (IMS)

ANNIS (ANNotation of Information Structure) is an open source, cross platform (Linux, Mac, Windows), web browser-based search and visualization architecture for complex multilayer linguistic corpora with diverse types of annotation. The system was originally designed to provide access to the data of the SFB 632 (Information Structure: The Linguistic Means for Structuring Utterances, Sentences and Texts). It has since then been extended to a large number of projects annotating a variety
of phenomena. Since complex linguistic phenomena such as information structure interact on many levels, ANNIS addresses the need to concurrently annotate, query and visualize data from such varied areas as syntax, semantics, morphology, prosody, referentiality, lexis, and more. For projects working with spoken language, support for audio and video annotations is also required. For more information see http://corpus-tools.org/annis. A CMDI metadata description for ANNIS was provided by the Corpus Linguistics and Morphology group at Humboldt-Universität zu Berlin. The respective PIDs were registered.

The SFB 632 data is hosted independently by UTüb (see section 2.1.16).

3.7.2 ASV Online Toolbox (ULei)

The ASV (“Abteilung Automatische Sprachverarbeitung”, Department for Automatic Language Processing) Toolbox has been available for many years as an offline stand-alone tool. It is a modular collection of tools for the exploration of written language data. The topics covered contain language detection, POS-tagging, base form reduction, named entity recognition, and terminology extraction. For more information on the ASV Toolbox see http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/.

ULei continued work on transforming these tools into a collection of REST web services that are CLARIN compatible (e.g., PID, CMDI metadata available via OAI-PMH) and that are accessible by human users without installation by using a simple web application integrated into the CLARIN infrastructure. As mentioned in the previous reports, the graphical user interface was already created and the first web services were integrated. Throughout the current reporting period, existing tools were stabilized and enhanced (e.g., with further import and export options, or analysis methods).

3.7.3 Atomic (IMS)

Atomic is a multilevel annotation tool and platform for linguistic corpora. It was developed at Friedrich-Schiller University, Jena, and is implemented as an OS-independent desktop software in Java. It is easily extensible via its plugin framework based on the Eclipse Rich Client Platform. Possible extensions include new editor types, NLP components, corpus views, among others. Atomic’s data model is Salt, a generic graph-based model for linguistic data, devoid of semantics and therefore usable for all kinds of annotation types, annotation schemes, tagsets, theories, and the like. Atomic also encapsulates the Pepper converter framework for linguistic data (see 3.7.10), hence it is compatible with a wide range of corpus formats. For more information see http://corpus-tools.org/atomic/. A CMDI metadata description for Atomic was provided by the Corpus Linguistics and
Morphology group at Humboldt University, Berlin. The respective PIDs were registered.

3.7.4 COALA (BAS)

The manual creation of CMDI compliant metadata is very tedious. COALA simplifies this task by reading input from plain text tables. COALA template table files with sample data (in English and German) have been improved based on user feedback. COALA is now installed as a web service and a web interface on the public server.

3.7.5 ISO-TEI converters (HZSK)

To enhance the support for spoken language data within the CLARIN infrastructure, a number of WebLicht compatible web services for conversion into the new ISO-TEI format for transcriptions of spoken language (see http://www1.ids-mannheim.de/prag/muendlichekorpora/isodin.html) have been developed. Several commonly used formats can be converted into the standardized ISO-TEI format. There is also a service to convert data from the ISO-TEI format into the WebLicht internal TCF format to allow researchers working with spoken language data formats to benefit from existing and future WebLicht services.

3.7.6 KorAP (IDS)

KorAP (Korpusanalyseplattform der nächsten Generation, next generation corpus analysis platform) is a web platform for searching and analysing very large corpora with multiple annotations. It supports principally corpora of any size and without limitations as to the number of annotation layers. Unlike a typical corpus search engine, it is not restricted to work only on a specific corpus query language, but it supports several such languages as used in Cosmas II, Poliqarp, and ANNIS. Like Cosmas II, it provides access to DeReKo (see section 2.1.4) and has been integrated into the FCS and WebLicht.

3.7.7 Language Profiler (ULei)

Being faced with a high demand for OCR-related tools in the digital humanities community, the Language Profiler and the corresponding web services are now hosted by the CLARIN center at ULei. Researchers can use the PoCoTo (Post Correction Tool) client to connect with the server and send their data for post-correction to the profiler web service. Currently, there are three language profiles available (German, Greek, and Latin).
The Language Profiler is a tool that is developed by the CIS OCR Group at the University of Munich. It is targeted at OCR post-correction with a focus on historical texts. OCR on historical documents can be difficult due to bad quality of the scans or even the original documents. But also rare fonts and deprecated spelling variants can be challenging. Traditional OCR post-correction tools are often based on dictionaries which are rarely available for historical languages. The Language Profiler overcomes this disadvantage by using a dictionary of modern words, a set of recurring patterns describing the differences in spelling and a sample containing text in the historical variant of the specific language.

The Language Profiler not only detects errors but also suggests corrections for misspelled words by examining the Levenshtein distance between those words and the dictionary entries. Furthermore, it is able to identify unknown words by combining the dictionary information, historical spelling patterns, and hints from the OCR engine. The output is a list of the most common patterns encountered in the document and the input document enriched with correction suggestions.

3.7.8 LDA Topic Modelling (UTüb)

This service generates and provides visualization for a topic model, given a set of text documents. It was the result of incorporating the work done in the KobRA project (http://www.kobra.tu-dortmund.de/) into CLARIN.

3.7.9 OrthoNormal (IDS)

OrthoNormal is a tool for normalizing the orthography of transcriptions. Results of automatic lemmatization and POS-tagging can also be manually corrected. Within the reporting period, OrthoNormal was adapted to the needs of the curation project for the Dortmunder Chatkorpus (see section 2.1.6). The tool’s metadata is described in CMDI and available in the IDS repository.

3.7.10 Pepper converter framework (IMS)

Pepper is a pluggable, Java-based, open source converter framework for linguistic data. It was developed at Humboldt University, Berlin, to convert data across linguistic data formats. Supported formats include EXMARaLDA, Tiger XML, MMAX, ANNIS, Treetagger output, RST, PAULA, and others. For more information see http://corpus-tools.org/home/. A CMDI metadata description for Pepper was provided by the Corpus Linguistics and Morphology group at Humboldt-University, Berlin. The respective PIDs were registered.
3.7.11 TEA (IMS)

TEA (Textual Emigration Analysis) is a web-based application showcase that applies textual analysis in the humanities. The TEA tool is used to transform raw text input into a graphical display of emigration source and target countries (under a global or an individual perspective). It provides emigration-related frequency information and gives access to individual textual sources, which can be downloaded by the user. The application is built on top of the CLARIN infrastructure. It processes a large set of documents to extract information about people who emigrated. The current implementation integrates several data sets. For more information see http://clarin01.ims.uni-stuttgart.de/tutorial/tea.html. A CMDI metadata description for the most recent showcase of TEA was created and the respective PIDs were registered.