AP5 Fourth Year Activity Report

May 2015
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 25

3 Integration of language tools 27
   3.1 CLARIN-D infrastructure development 27
   3.2 Morphological analysers 28
   3.3 Taggers 28
   3.4 Named entity recognisers (NER) 28
   3.5 Parsers 29
   3.6 Spoken language tools 29
   3.7 Other tools 31

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

BAS Bayerisches Archiv für Sprachsignale, LMU München
BBAW Berlin-Bandenburgische Akademie der Wissenschaften
HZSK Hamburger Zentrum für Sprachkorpora, Universität Hamburg
IDS Institut für deutsche Sprache, Mannheim
IMS Institut für maschinelle Sprachverarbeitung, Universität Stuttgart
MPI MPI 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 fourth year of CLARIN-D (from June 2014 to May 2015). The work described in this report generally builds upon the curation efforts for linguistic resources and tools (LRT) reported in the preceding three yearly activity reports R5.1–R5.3.

During the fourth project year, CLARIN-D member institutions focussed on the curation and integration of LRT that originate from different research institutions outside the consortium. However, 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 three 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

As a curation effort specifically targeting metadata curation with respect to its application in the Virtual Language Observatory (VLO) AP5 members formed a dedicated VLO taskforce (VLO-TF). During the report period, BBAW was responsible for coordinating the taskforce work. The VLO-TF aims at providing AP3 and specifically the VLO development group with informed feedback and suggestions to further improve the VLO as the central entry point for many prospective CLARIN-D users. During regular (virtual) taskforce meetings, detailed tasks were formulated and directed to AP3 by means of TRAC tickets (see https://trac.clarin.eu/report/10, login required) and white papers.

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

CLARIN Concept Registry Coordinators: After the discontinuation of the ISOcat data categories registry that is now replaced by the CLARIN Concept Registry (CCR) as a central building block of the infrastructure, the members of the former national ISOcat coordinators group are now appointed CCR coordinators. Facet definitions in the VLO rely primarily on CCR entries for semantic interoperability among metadata fields. Therefore, close cooperation with the CCR coordinator group is mandatory for the work of the VLO-TF. On the first CCR coordinators meeting on March 23, 2015 in Amsterdam (NL) the current status and future plans for the VLO facet definitions were
presented. The CCR coordinators will base their initial curation work on those former ISOcat data categories (now imported as concepts into the CCR) that will serve as facet definitions provided by the VLO-TF.

**Metadata Curation Taskforce:** This taskforce aims at designing an evaluation schema for metadata quality among all European CLARIN centres. As the VLO-TF will ultimately provide a best practice for metadata descriptions for the VLO, it will define the standard for this specialised subset of metadata descriptions.

A joint (virtual) meeting of both taskforces was held on February 18, 2015.

**CLARIN Standards Committee:** On April 10, 2015, the annual CSC meeting took place in Vienna (AT). The work of the VLO-TF was presented and implications for the CSC were discussed. Though only little progress was achieved with respect to the ISO standardisation of the CMDI model and format, CMDI continues to be the mandatory format for full CLARIN support for metadata. The CSC will revise the list of CLARIN recommended and (at least partly) supported standards for LRT. This revision can then be exploited as controlled vocabulary in metadata instances and thus foster LRT discovery.

**CMDI-1.2 Taskforce:** Many members of the VLO-TF also participated in the development of the CMDI-1.2 specification and framework. Pending issues with respect to the VLO development could be addressed successfully, such as the provision of a component versioning schema and a best practice suggestion for modeling relations among metadata instances and the LRT they describe.

The work of the VLO-TF was presented at the CLARIN Annual Conference (CAC 2015), October 23–25, in Soesterberg (NL). For more information on the VLO-TF see [https://trac.clarin.eu/wiki/VLO-Taskforce](https://trac.clarin.eu/wiki/VLO-Taskforce) (login required).
2 Curation and adaptation of language resources

In this section we summarise all LRT that were made available to CLARIN-D during the fourth 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 Berliner Wendekorpus (BBAW, IDS)

The Berliner Wendekorpus was developed within the project Kollektives Gedächtnis (collective memory) at the Institut für Deutsche und Niederländische Philologie of the Freie Universität Berlin. In this project, the changes in society after the fall of the Berlin wall in 1989 were documented as a collection of group-specific experiences.

The corpus describes discursive contrasts of how two groups with social conflicts (based on stereotypes among Eastern and Western inhabitants of Berlin) perceive the time of crisis. It contains 50 recordings from 1992 to 1996 with a total duration of 26 hours and 15 minutes. The interviews were carried out with 30 East and 26 West Berlin people of ages between 19 and 55. The recordings were digitised by AGD (Archiv für gesprochenes Deutsch, Archive for Spoken German). The transcripts and New High German normalisations thereof were provided by BBAW and aligned with the audio recordings by AGD. The corpus is available in DGD2 (see section 2.2.2). CMDI metadata, transcription and normalisation data have been converted to standardised, CLARIN-D compliant formats (CMDI and TEI/XML for metadata, TEI/XML for text data, TCF for linguistic corpus annotation) by BBAW and are available at the CLARIN-D repository at BBAW (see http://clarin.bbaw.de/objects/dwds:5/). The CMDI metadata of the corpus is available in the CLARIN-D repository at IDS.

2.1.2 BulTreeBank-DP (UTüb)

The BulTreeBank-DP contains dependency information encoded in the BulTreeBank (a HPSG-based Treebank of Bulgarian). It consists of 11,990 sentences (approx. 196,000 tokens) and contains sentences from Bulgarian grammar textbooks, newspapers, literature, and other sources of texts. The treebank was developed by the BulTreeBank Group led by Kiril Simov (Linguistic Modelling Laboratory, IICT, Bulgarian Academy of Sciences) and Petya Osenova (Sofia University and Linguistic Modelling Laboratory, IICT, Bulgarian Academy of Sciences). The BulTreeBank-DP has been made available by the CLARIN-D centre at UTüb within the Tündra tool for querying treebanks (see section 3.1.3).
2.1.3 Chess Dataset (IMS)

This corpus consists of annotated chess games that were posted on http://www.chess.com/. A CMDI metadata description for the Chess Dataset was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on the Chess Dataset see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/chess/index.en.html.

2.1.4 DaSciTeX2014 (UdS)

The Darmstadt Corpus of Scientific Texts 2014 (DASCITEM2014) is a newly encoded version of the DaSciTex (see R5.3 report). Improvements have been made with regard to the distinction between document parts (abstract, introduction, main part, and conclusion).


2.1.5 DeReKo (IDS)

The German Reference Corpus (DeReKo) is the largest collection of electronic corpora of contemporary written German. In 2014, it comprises more than 25 billion words (as of 15 September 2014) and has a growth rate of approximately 1.7 billion words per year.

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 started to develop a new corpus analysis platform KorAP in 2011 (see section 3.7.7).

2.1.6 Deutscher Wortschatz (ULei)

For many years the project Deutscher Wortschatz has offered large corpora and statistical data on its Corpus-Portal. The statistical analysis of these corpora is based on language independent automatic and semi-automatic algorithms.

Currently, these corpora are available in more than 200 different languages. The data is accessible via a webportal (http://corpora.informatik.uni-leipzig.de/), SOAP webservices (http://wortschatz.uni-leipzig.de/Webservices/), and some of the corpora are available for download (http://corpora.informatik. 7
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 the R5.2 report.

The software tool that allows the automatic deployment of Wortschatz corpora (including supporting webservices and metadata) into the CLARIN-D infrastructure was further developed. By the end of year three, corpora in a variety of European (and other) languages were made available by the means of this system. By the end of the fourth year, the number of available datasets was further increased. Missing European languages were added and corpora for other important languages of the world were checked for quality and added to the system.

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 approx. 1600 to 1900. Due to the DTA’s primary focus on the history of the German language, the earliest editions accessible are digitised, and the full text transcriptions meticulously document the original historic text.

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 report (R5.3) and http://www.deutschestextarchiv.de/.

As of April 2015, the DTA core corpus comprises digitised and structurally as well as linguistically annotated full texts of 1129 volumes with 421,353 pages (approx. 101 million tokens). During the report period, the DTA core corpus was further extended on the basis of cooperations with partner institutions outside of CLARIN-D:

Hamburgischer Correspondent: In the course of a cooperation with the University of Paderborn (project management: Prof. Dr. Britt-Marie Schuster) 180 issues (1,072 pages) of the journal Staats- und Gelehrte Zeitung des Hamburgischen Unpartheyischen Correspondenten and some of its precedents have been digitised and structurally as well as linguistically annotated. All texts have been integrated in the corpus of the DTA. Thus, they are freely available in different CLARIN-D compliant formats (CMDI metadata, TEI, TCF) via the DTA website. Perspectively, they will be integrated into the CLARIN-D repository.

Alexander von Humboldt’s Kosmos-Lectures: The project Hidden Kosmos – Reconstructing Alexander von Humboldt’s Kosmos-Lectures at Humboldt University Berlin (project management: Prof. Dr. C. Kassung) is editing
auditor’s notes of Alexander von Humboldt’s Kosmos-Lectures and publishing these texts within the DTA. Perspectively, 11 issues (approx. 3500 pages), created in 1827/1828, will be integrated into the CLARIN-D infrastructure. Currently, 4 volumes (1489 pages) are available after login via the DTA collaborative quality assurance platform DTAQ (in TEI and TCF formats; CMDI metadata are also available) and will – after a quality assurance period – be made available via the CLARIN-D repository at the BBAW.

The project applies the extensively documented digitisation and annotation guidelines of the DTA (see R5.3 for details) as far as possible. It is thus a good example for the early involvement of a CLARIN-D centre in a digitisation project, a constellation which significantly facilitates the integration of the newly created resources into the CLARIN-D infrastructure.

**Grenzboten:** In the course of a project conducted by the SuUB Bremen and the DTA the digitised journal Die Grenzboten (from 1841 to 1922, 270 volumes, 180,000 pages) is substantially enhanced in terms of structural annotation and transcription quality. These works are conducted based on the DTA transcription and annotation guidelines resulting in CLARIN-D compliant texts.

Currently, work on the semi-automatic TEI annotation according to the DTA Base Format (DTABf, the CLARIN-D recommendation for the transcription of historical texts, see R5.3) has been finished. The next steps include the merging of automatically corrected texts (Bremen) with corresponding TEI-annotations (BBAW) and the publication of the results via – among others – the CLARIN-D repository at the BBAW.

In total, DTA and its extensions currently comprise approx. 1.5 billion characters (approx. 200 million tokens).

### 2.1.8 GECCo (UdS)

GECCo is a bilingual corpus containing approx. 1.44 million tokens of parallel as well as comparable English and German subcorpora. GECCo offers a continuum of different text types (registers) from written to spoken discourse. More precisely, it includes English and German texts of 14 registers, eight of which represent written discourse and include:

- fictional texts (FICTION)
- political essays (ESSAY)
- instruction manuals (INSTR)
• popular-scientific texts (POPSCI)
• letters to shareholders (SHARE)
• prepared political speeches (SPEECH)
• tourism leaflets (TOU)
• corporate websites (WEB).

This part was imported from the existing corpus CroCo.

The written texts are saved in two subcorpora according to the language: English written texts (EO), German written texts (GO). The further registers are of spoken discourse and include recorded and transcribed interviews and academic speeches. The spoken texts are also stored in two further subcorpora classified according to the language of origin: English spoken texts (EO-SPOKEN) and German spoken texts (GO-SPOKEN). The corpus comprises approx. 1.4 million tokens. It is pre-annotated on several levels which include information on tokens, lemmas, morpho-syntactic features (e.g. case, number), parts-of-speech, phrase chunks, and their grammatical functions, as well as sentence boundaries. The annotations of the written section were partly imported from CroCo, whereas for the spoken section, we used the Stanford part-of-speech tagger and the Stanford parser.

GeCCo was also semi-automatically annotated for information on cohesion (devices and relations), including coreference, conjunctive relations, substitution, and lexical cohesion. Ellipsis was manually annotated. The corpus is encoded in the CWB format and can be queried with the Corpus Query Processor (CQP).

For more information on GECCo see http://www.gecco.uni-saarland.de/GECCo/.

2.1.9 Hamburg Dependency Treebank (HZSK)

The corpus was compiled at the Natural Language Systems Division of the Department of Informatics at University of Hamburg. The primary data (plain text, 261,830 sentences) was acquired from the online newscasts of the Heise news service (http://www.heise.de/) from 1996 to 2001. All data was annotated automatically with dependency structures, 80% of the data has been manually corrected, and 40% has been additionally checked for consistency. The corpus will be made fully available and searchable by the CLARIN-D federated content search.

2.1.10 PolDiLemma (UdS)

The PolDiLemma corpus is a diachronic corpus made of political, religious, scientific, and historical texts from different authors of the Middle Polish period (16th – 18th century).
Characteristic for this period is the slow development of a supra-regional standard language, a process of standardisation on the basis of the variety of the Polish nobility, under the influence of Latin and other foreign languages as well as different social or regional varieties.

All texts (free licenses) are gathered from Federacja Bibliotek Cyfrowych (Digital Library Federation). The Middle Polish texts illustrate the history of the language and give the opportunity to explore some first-hand evidence of the development of Polish in its historical context.

Studying the history of the language is a way to familiarise oneself with aspects of the history of Poland in general. It also helps to build up valuable methodological knowledge in diachronic linguistics and philology.

The corpus includes lemma annotation created with the PolDiLemma-Tool (see 3.7.9).

PolDiLemma has been integrated into the CLARIN-D federated content search since June 2014. For more information on PolDiLemma see http://fedora.clarin-d.uni-saarland.de/poldilemma/index.html.

2.1.11 Sentiment Relevance Corpus (IMS)
The Sentiment Relevance Corpus contains 3,847 sentences, taken from 125 documents annotated for sentiment relevance. The data is a subset of the 2.0 Movie Polarity Dataset.

A CMDI metadata description for the corpus was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on the corpus see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sentimentrelevance/index.en.html.

2.1.12 SFB 632 corpora, project B4 (HZSK)
The following resources created by the project B4 (The role of information structure in the development of word order regularities in Germanic) which is part of the special research centre 632 (SFB 632, Information Structure) in Potsdam have been integrated and made fully available for academic use.

Each resource consists of one single medieval text (monologue) that has been transcribed from manuscript, translated to Modern German (or English), manually annotated with part of speech information, syntactic categories, grammatical functions, clause status, numbers of syllables (per constituent), alliteration, information status, topic / comment, position of phrase in sentence, definiteness, focus / background, focus marker, and comments on context:

Heliand: Old Saxon historic manuscript (9th century epic poem)
Muspilli: Old High German historic manuscript (epic poem)

Otfrid: Old High German historic manuscript

Tatian: Old High German historic manuscript (2nd century epic poem, translation of Latin original)

2.1.13 SFB 833 datasets (UTüb)

During the period of this report, the descriptions of the primary research data created in the special research centre 833 (SFB 833, The Construction of Meaning) became available through the CLARIN-D centre at UTüb. The primary data in that repository currently consists of 117 sets of data of the types experiment, text corpus, lexical resource and tool. Most of the resources are made available by the individual researchers upon request. With this primary research data close to ongoing research projects UTüb shows that though the data is open in principle, there are privacy constraints by the researcher for first publication of research results based on the data and possibly also additional copyright and privacy issues.

2.1.14 TüBa-D/W (UTüb)

In December 2014, the CLARIN-D centre at UTüb released TüBa-D/W, a large automatically annotated treebank of modern written German. TüBa-D/W follows common annotations standards and has the following annotation layers:

- part-of-speech tags (STTS),
- lemmas (TüBa-D/Z),
- morphology (TIGER)
- dependency structure (TüBa-D/Z)

The treebank is based on Wikipedia text and consists of 36.1 million sentences (615 million tokens) in CONLL-X format.

2.1.15 TüBa-D/Z (UTüb)

Release 9.1 of TüBa-D/Z has been made available for academic use in CLARIN-D. 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 85,358 sentences (1,569,916 tokens; 3,444 newspaper articles). Release 9.1 includes 17,910 manual annotations of a selected set of lemmas (30 nouns, 79 verbs) with their corresponding senses in the German wordnet GermaNet
(see section 2.3.3) with the goal of providing a gold standard for word sense disambiguation.

2.2 Speech and multimodal corpora

2.2.1 Speech corpora at the BAS

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

**SaGA:** The primary data of the SaGA corpus are made up of 25 dialogs of 50 interlocutors, who engage in a spatial communication task combining direction-giving and sight description. Six of those dialogues with data only from the direction giver are available including audio (WAV) and video (MP4) data.

The secondary data consists of annotations (EAF) of gestures and speech-gesture referents, which have been completely and systematically annotated based on an annotation grid.

The SaGA corpus is comprised of 9,881 isolated words and 1,764 isolated gestures. The stimulus is a model of a town presented in a virtual reality environment. Upon finishing a bus ride through the VR town along five landmarks, a router explained the route as well as the wayside landmarks to an unknown and naive follower.

The SaGA Corpus was curated 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.

**NM-MoCap:** The Natural Media Motion Capture Corpus (NM-MoCap-Corpus) originates from a case study recorded in Aachen, Germany in 2011 for a theory of gesture form analysis, which aimed at eliciting object descriptions containing gestural information about stimulus objects. Gesture form analysis bases on differentiating between the physical configuration of the articulating body part (articulator form) and the spatial information that an observer abstracts from articulator form (gesture form), for example by profiling specific parts of an articulator. Of particular interest were differences in the participants’ depiction of size information versus size and shape information.

The corpus consists of data from 18 participants, whose task was to describe nine objects each to an experimenter, without using everyday vocabulary about forms, sizes, or objects. The participants were recorded on audio and several video cameras, and their hand movements were recorded using an optical VICON motion capture system. ELAN annotations for gestural holds
displaying size or shape information were generated semi-automatically from the motion capture data.

Each participant’s sessions contains ten combined motion capture and video recordings (nine object descriptions and one calibration task). Each motion capture recording consists of three video files and two data files. For each object description one ELAN annotation file was produced. Furthermore, each participant’s data contain one HD video of the entire session.

In total the corpus consists of: 557 video files (MP4, one file missing), 360 annotation files (EAF), 162 motion capture data files (CSV).

The NM-MoCap-Corpus was curated 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.

**Ph@ttSessionz**: Ph@ttSessionz is a speech database created by BAS. It comprises 1,019 recording sessions of adolescent speakers recorded via the web in German secondary schools.

**AsiCa**: AsiCa is a speech database of Calabrese Italian created by the department of Romanistic Languages at LMU, München. In extension of the annotations which were already available in the BAS repository, academic users now have access to the signal data as well.

**CL-I**: This corpus contains speech recordings of normal hearing speakers and speakers equipped with Cochlear Implants (CI). Speech data were collected with the software SpeechRecorder, for each recording a BPF file was generated (PAR), on which the MAUS segmentation was based (TextGrid); The formants 1-4 (fms), fundamental frequency (f0), short-time energy (rms), and zero-crossing rate (zcr) were calculated in Emu for 23 selected speakers. The formant values were corrected manually and analysed through the Emu-R interface.

**RVG1-CLARIN**: The corpus is a collection of more than 500 speakers of different dialect regions of Germany. The recordings were made using four different microphones (two in low and two in high quality) and consist of single digits, connected digits, phone numbers, phonetically balanced sentences, computer command phrases prompted on a screen, and one minute spontaneous speech (monologue). The speakers were recorded in normal office environments. The background noise was limited to the usual noise in office environment, eg. door slam, backround crosstalk, phone ringing, paper rustle, PC noise, etc.

---

1 as used for analysis in the Master thesis of Veronika Neumeyer (2009, LMU München)
SC1: The corpus contains speech of 88 different speakers, reading the German story “Der Nordwind und die Sonne.” Subcorpus T contains the recordings of 16 native Germans (L1). The other 72 speakers which were born and educated in other countries (L2) are pooled in subcorpus C. Every speaker has a distinct accent.

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 etc.

SC2: The corpus SC2 contains read speech of ten different speakers with screen prompted “automobil 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 has spoken 800 utterances of three to seven words derived from 100 different sentences, resulting in a total of 8 000 utterances.

2.2.2 Speech corpora at the IDS

As of April 2015, the Datenbank für Gesprochenes Deutsch (DGD2, German Spoken Language Database) comprises 18 corpora, totalling about 9 000 recordings, 2 500 hours of audio, 4 000 transcripts, or 7.5 million transcribed word tokens. CMDI metadata descriptions for DGD2 were created and integrated in the IDS repository.

Most of the following resources are part of DGD2:

Australiendeutsch: Australiendeutsch (AD) is a corpus of spoken German containing narratives, interviews, and descriptions of pictures from 81 elderly men and women, where partially three generations of their family have been living in South Australia. The recordings were carried out in 1967 within a project directed by Michael Clyne from Monash University in Melbourne. In 2014, they were revised and transcribed by AGD (Archiv für gesprochenes Deutsch). There are 46 recordings with a total duration of 13 hours and 30 minutes, and 33 recordings have been transcribed. The CMDI metadata of the corpus is available in the IDS repository.

Binnen- und auslandsdeutsche Mundarten: Varia The Binnen- und auslandsdeutsche Mundarten: Varia (MV) corpus consists of selected DGD2 archived audio recordings from 1965 to 1974. The corpus contains 183 recordings with a total duration of 69 hours and 25 minutes. The 184 speakers, women and men of different ages, are from Australia, Germany, Canada, Mexico, Austria,
Switzerland, and the USA. The corpus documents are narratives, dialogues, some texts, and lists of words. The recordings were digitised by AGD. 14 transcripts from Australia, Canada, and the USA were published in Phonai Series (Volume 6, 10, 18, and 21), however, they are not digitally archived. The recordings are available in DGD2. CMDI metadata of the corpus is available in the IDS repository.

**Biographische und Reiseerzählungen:** The Biographische und Reiseerzählungen (BR) corpus consists of 7 recordings from 1985 to 1990 with a total duration of 5 hours and 30 minutes. They are narratives and interviews from 24 young men and women from East Germany, Poland, and the Czech Republic. The recordings were carried out under the supervision of Katharina Meng (Zentralinstitut für Sprachwissenschaft der Akademie der Wissenschaften der DDR) and digitised by AGD. The transcripts are available in DGD2. The metadata of the corpus in CMDI is available in the IDS repository.

**Deutsche Mundarten, ehemalige deutsche Ostgebiete:** The corpus Deutsche Mundarten: ehemalige deutsche Ostgebiete (OS) was developed by DSAv in cooperation with the research centre Deutscher Sprachatlas. The corpus was built as an addition to the corpus Deutsche Mundarten: Zwirner-Korpus (ZW) and aimed to collect audio data of German dialects in former eastern territories of Germany (before World War II) and different parts of Eastern and south-eastern Europe.

The corpus consists of 981 audio recordings from 1962 to 1965 with a total duration of 462 hours and 5 minutes of data, e.g. interviewer talks, telling, reading, conversations, dialogues. The speakers were primarily older people (men and women) speaking East and South-East German dialects, which represent the historical state of the language before 1945.

The recordings were digitised by AGD. 280 of them are aligned with their transcripts and archived. All recordings are available in DGD2. The CMDI metadata of the corpus is integrated into the IDS repository.

**Deutsche Mundarten, Kreis Böblingen:** The data for the Deutsche Mundarten, Kreis Böblingen (BB) corpus was collected by Ulrich Engel from 1965 to 1967. It consists of 73 recordings with a total duration of 42 hours and 28 minutes. It contains narratives, dialogues, and some texts that were read out by dialect speakers, namely women and men of different ages in the Böblingen area. The recordings were digitised by AGD and the metadata of the corpus in CMDI has been published in the IDS repository.

**Deutsche Mundarten, Schwarzwald:** The Deutsche Mundarten: Schwarzwald (SW) corpus was developed by the permanent working group Sprache in
Südwestdeutschland in Tübingen with the purpose of enhancement of the corpus Deutsche Mundarten: Zwirner-Korpus (ZW) with the documentation of the dialect spoken in the region of Black Forest. Interest was primarily focused on the dialects from the hamlets of Schönminz, Romishorn, and St. Roman.

The corpus contains 126 recordings from 1964 to 1974 with a total duration of 37 hours and 31 minutes. The 126 speakers were from places nearby the cities Freudenstadt and Wolfach. The recordings document the reading, telling stories, and standard texts.

The recordings were digitised by AGD. The transcripts are aligned with the audio recordings and available in DGD2. The CMDI metadata of the corpus is available in the IDS repository.

Deutsche Mundarten, Südwestdeutschland und Vorarlberg: The Deutsche Mundarten: Südwestdeutschland und Vorarlberg (SV) corpus was developed by the permanent working group Sprache in Südwestdeutschland in Tübingen with the aim to document the spoken language of the South-West Germany dialects. It contains 242 recordings from 1966 to 1970 with a total duration of 72 hours and 6 minutes. The main genres of the recordings are narratives and standard texts (week names and numbers). The transcription of the audio data is not available.

All recordings were digitised by AGD and are available in DGD2. The CMDI metadata of the corpus is available in the IDS repository.

Deutsche Mundarten, Zwirner-Korpus: The corpus Deutsche Mundarten: Zwirner-Korpus (ZW) is a large dialect collection for spoken German. It was built by a project at DSAv. The aim of the project was to collect data of the German dialects as complete as possible. The places for recordings and the number of interviewed people at each place have been carefully selected in accordance to the number of different dialects.

ZW was initiated in 1955. The 5796 recordings were made from 1955 to 1972 with a total duration of 1077 hours and 15 minutes. The 5888 speakers, women and men, were from the former Western federal states of Germany, Vorarlberg, Liechtenstein, Alsace, and the Netherlands. The recordings document different types of texts such reading, telling stories, standard texts.

The recordings were digitised by AGD. The recordings and the 2310 transcriptions, which are aligned with the audio recordings, and metadata are available in DGD2. The CMDI metadata of the corpus is available in the IDS repository.
Deutsche Umgangssprachen, Pfeffer-Korpus: The corpus Deutsche Umgangssprachen: Pfeffer-Korpus (PF) was built within a project at the Institute for Basic German of Stanford University in cooperation with DSAv, Institut für deutsche Sprache und Literatur der Akademie der Wissenschaften der DDR and some other institutions.

The archived version of the corpus comprises 398 audio recordings from 1961, with a total duration of 79 hours and 15 minutes. The audio data includes talks, reports, stories, and other reading materials. The speakers were 402 women and men of different ages, all levels of qualification, with different professional backgrounds from Germany, Austria, and Switzerland.

The recordings were digitised by AGD. All recordings are transcribed, aligned with the audio data and archived. The text data are lemmatised and annotated with POS tags. All recordings with transcription and other data materials are available in DGD2. The CMDI metadata of the corpus is available in the IDS repository.

Dialogstrukturen: The Dialogstrukturen (DS) corpus was developed in a collaborative project by IDS-Forschungsstelle Freiburg, Deutsches Seminar of Universität Freiburg, and Lehrstuhl für Psychologie of Universität Gießen. In the project, the corpus was used to analyse the discourse organisation of natural language dialogues.

The corpus consists of 70 recordings taken from 1960 to 1977 with a total duration of 15 hours and 18 minutes. 51 recordings were taken from Korpus Grundstrukturen: Freiburger Korpus and re-transcribed according to the project conventions. There were 152 speakers of standard German including standard-like daily language in official and informal communications. It contains various kinds of conversation themes such as registration, consultation, discussion, and exam.

The recordings were digitised by AGD and the transcripts have been archived. The corpus is available in DGD2 and its CMDI metadata are available in the IDS repository.

Elizitierte Konfliktgespräche zwischen Müttern und jugendlichen Töchtern: The corpus was developed by the project C2 (Arguments in conflict conversations between parents and adolescents in the research area 245: communication and language understanding in social contexts) in Mannheim and Heidelberg. Particularly, it describes arguments between mothers and their young daughters. Based on the corpus, an integrated language psychology and conversation category system was built.
The corpus contains 138 recordings with 214 speakers and its total duration is 12 hours and 23 minutes. The recordings were carried out from 1988 to 1990. The mothers were between 31 and 58, and the daughters between 12 and 24 years old.

The recordings were digitised by AGD and the transcripts have been archived. The corpus is available in DGD2 and its CMDI metadata is available in the IDS repository.

Deutsche Hochlautung: Deutsche Hochlautung (DH) is a corpus developed by DSAv within the project Hochlautung. 27 audio recordings of the corpus are publicly accessible. They were recorded from 1971 to 1975 and their total duration is 1 hour and 57 minutes. The speakers were 9 TV journalists, news presenters, and government spokesmen. The conversation themes were diverse such as news, interviews, comments, reports and statements in broadcasts and press conferences.

The recordings were digitised by AGD (before DSAv). The transcripts have been archived and aligned with the audio data. The publicly accessible part of the corpus is available in DGD2, and its CMDI metadata is available in the IDS repository.

Deutsche Standardsprache, König-Korpus: The corpus Deutsche Standardsprache: König-Korpus (KN) was created by Werner König (Augsburg University) and consists of recordings of interviews, conversations, and readings from 1975 to 1976 and was created for the investigation of pronunciation of written German in different places in the former FRG. For legal reasons, only 43 audio recordings with a total duration of 5 hours and 48 minutes from 1975 are available in DGD2. The speakers were 43 students and academic staff, men and women between the ages of 17 and 27 years.

The recordings were digitised by AGD. The transcripts aligned with the audio data are available. The audio data were transcribed literally, orthographically normalised (old German spelling rules), lemmatised, and annotated with POS tags. The CMDI metadata of the corpus is available in the IDS repository.

Emigrantendeutsch in Israel: The major part of the corpus Emigrantendeutsch in Israel (IS) was built within a DFG project about language and culture identity of the emigrants of age 50–60 after their immigration to Palestine or Israel. Since 2000, the first part of the collection has been supplemented by some video recordings such as extra recordings with the same interviewees 20 year after the first interviews, and some special interviews on selected themes, for instance the childhood of Michela Melz from Salzburg.
The archived version of the corpus in the IDS repository comprises 176 audio recordings and 4 audio copies of video recordings from 1989 to 2011, with a total duration of 284 hours and 40 minutes. It mainly contains argumentative and narrative autobiography interviews with many monological (but also strong dialogical) passages with 181 Jewish emigrants from German-speaking regions in middle Europe. The conversations were mostly carried out in private apartments of the interviewees.

The recordings were digitised and edited by AGD (before DSAv). 104 transcripts of various kinds have been archived. 16 of them have been aligned with the audio data. The recordings and 22 transcripts are available in DGD2. 82 transcripts which have not been edited, can be requested from AGD. The CMDI metadata of the corpus is available in the IDS repository.

**Emigrantendeutsch in Israel: Wiener in Jerusalem:** The major part of the corpus Emigrantendeutsch in Israel: Wiener in Jerusalem (ISW) was developed during an excursion to Israel of students and teachers of the department of German studies in Salzburg. The corpus is a supplement to the IS corpus. All of the interviews were done in December 1998, except one interview with Ari Rath, which was first started during the excursion and continued later in Salzburg and Jerusalem. In 2010 and 2011, the corpus was extended with three additional interviews with some interviewees, that have participated in the previous interviews in 1998.

The archived version of ISW in IDS consists of 28 audio recordings from 1998 to 2011. The speakers were 24 Jewish men and women of age 69–90, that were born and grew up in Austria (mostly in Vienna), and lived in Jerusalem. Similar to the recording procedures of the IS corpus, the recordings were carried out in the apartments of the interviewees.

The recordings were digitised and edited by AGD (before DSAv). 20 transcripts and four unedited ones have been archived. 16 of them have been aligned with the audio data. The recordings and the 20 transcripts are available in DGD2. The unedited transcripts can be requested from AGD. The CMDI metadata of the corpus is available in the IDS repository.

**Forschungs- und Lehrkorpus gesprochenes Deutsch:** The corpus Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK, German speech corpus for research and teaching) is a collection of spoken German covering a wide range of daily life conversations such as at a driving school, a theater, and an employment office. The conversations were recorded in different German-speaking regions and transcribed by using FOLKER.
The corpus contains recordings, documentations, transcripts, and extra materials, which can be accessed via DGD2. On March 5, 2015, FOLK comprises 169 recorded speech events, corresponding to about 135 hours of audio recordings and 1.3 million transcribed tokens. CMDI metadata for this resource has been created and integrated in the IDS repository.

**GeWiss:** GeWiss (Gesprochene Wissenschaftssprache, Spoken Academic Language) is a comparative corpus for spoken academic language which can be freely used for research and teaching purposes, as well as for academic studies. The corpus contains 400 events consisting of 400 recordings (in WAV and MP3 format) and 400 Transcriptions (in EXS and EXB format) involving 510 Speakers.

Since 2013 GeWiss has become a curation project of CLARIN-D carried out by the Herder-Institut of Universität Leipzig in cooperation with IDS, ULei, and HZSK. The objective of the curation project is to make the resources of the GeWiss project available and compatible with the CLARIN-D infrastructure. Therefore, for all GeWiss resources, 1710 CMDI metadata records have been created and are available in the IDS repository.

**Grundstrukturen: Freiburger Korpus:** The corpus was developed by the IDS Forschungsstelle Freiburg within the project Grundstrukturen der deutschen Sprache (Fundamental Structures of German). The project aims at describing grammatical and stylistic features of standard German. To achieve this, an extensive audio archive and transcripts of about 500,000 words were created.

The corpus consists of 222 audio recordings from 1960 to 1974 with a total duration of 68 hours and 6 minutes. The speakers were 812 men and women speaking standard German including standard-like daily language in official and non-official communications. The conversation themes include consultation, press conference, presentation, and reports in radio or TV broadcasts.

The recordings were digitised by AGD and 221 transcripts of the recordings have been archived. The transcripts are also aligned with the audio data. The corpus is available in DGD2 and its CMDI metadata is available in the IDS repository.

**Kindersprache, Saarbrücker Korpus:** The corpus Kindersprache: Saarbrücker Korpus (SA) was developed in a project at Saarland University with the aim to investigate the spontaneous successive language acquisition in children with Turkish and Italian as first language, and German as second language.

The corpus comprises 65 audio recordings from 1982 to 1984, with a total duration of 4 hours and 33 minutes. The recordings cover the interactions
of the children with adults and document the children speech during the tasks with telling and retelling of stories, conversations, descriptions, playing actions, etc. The probands were children between 9 and 13 years. The transcripts for all recordings were published in Phonai Series (Volume 32).

The recordings were digitised by AGD. All recordings with transcription and other data materials are available in DGD2. The CMDI metadata of the corpus is available in the IDS repository.

**Slawische Mundarten im Ruhrgebiet**: The Slawische Mundarten im Ruhrgebiet (SR) corpus was built within a cooperation project between the Seminar für Slawistik at the University of Bochum and DSAv.

The corpus comprises 23 audio recordings of narratives with a total duration of 6 hours and 40 minutes. The speakers were men and women between 17 and 78 years. The recordings were taken in 1969 in Bochum, Bottrop, Essen, Herne, and Recklinghausen. They cover the German dialect in the Ruhr District, but also speech in Slavic languages such as Polish, Slovenian, Ukrainian, and Russian.

The recordings were digitised by AGD and are available in DGD2. The corpus does not include any transcriptions. The CMDI metadata of the corpus is available in the IDS repository.

**Zweite Generation deutschsprachiger Migranten in Israel**: The Zweite Generation deutschsprachiger Migranten in Israel (ISZ) corpus was developed as follow-up project to the IS und ISW corpora. The collection was created from 1999 to 2012. The corpus consists of 68 interviews in 70 audio recordings with a total duration of 125 hours and 27 minutes. The speakers were 63 children of German speaking Jews in Israel (Yekkes). Almost all of the speakers were born in Palestine or Israel. Only a few speakers came to the country in early childhood together with their parents. Although most of the interviews are in German, sometimes the interviewees switched between German and English.

The recordings were digitised and edited by AGD. In addition there are 64 non-edited transcripts, a table of contents, and a detailed list of the linguistic peculiarities of the speech. The recordings and the additional materials are available in DGD2. The transcripts are available via personal contact with AGD. The CMDI metadata of the corpus are available in the IDS repository.

### 2.2.3 Speech corpora at the MPI

**LAISEANG**: The geographical region of insular South East Asia and New Guinea is well-known as an area of mega-biodiversity. Less well-known is the extreme
linguistic diversity in this area: over a quarter of the world’s 6,000 languages are spoken here. As small minority languages, most of these will cease to be spoken in the coming few generations. The Language Archive of Insular South East Asia and West New Guinea (LAISEANG) will ensure the preservation of unique records of languages and the cultures encapsulated by them in the region. The language resources have been gathered by twenty linguists at, or in collaboration with Dutch universities over the last 40 years, and will be compiled and archived in collaboration with The Language Archive (TLA) at the MPI. The resulting archive will constitute an unrivaled collection of multimedia materials and written documents from over 50 languages in Insular South East Asia and West New Guinea. At the TLA, the data will be archived according to state-of-the-art standards (TLA holds the Data Seal of Approval): the component metadata infrastructure CMDI will be used; all metadata categories as well as relevant units of annotation will be linked to the CCR. This guarantees the proper integration of the language resources into the CLARIN-D framework.

Through the archive, future speaker communities and researchers will be able to exploit the materials for answers to their own questions, even if they do not themselves know the language, and even if the language dies.

For more information on LAISEANG see http://hdl.handle.net/1839/00-0000-0000-0018-CB72-4@view.

VALID: Research groups from the universities of Nijmegen, Amsterdam (UvA), and Utrecht have decided to prepare a nationwide, open access multimedia archive of language pathology data collected in the Netherlands, primarily on Dutch (Vulnerability in acquisition: language impairments in Dutch, VALID). In this enterprise, as many other research groups as possible, both from universities and care and educational institutes, will be involved. Although data sharing is becoming increasingly important to widen the scope and depth of empirical research in all kinds of areas, no tradition has been established yet in language and speech pathology research. This means that in various places a wealth of relevant and precious language data exists that cannot be easily and optimally accessed and exploited. The aim of the present project is to curate five existing, digital data sets, in order to make them available for scientific research in CLARIN-D compatible format. This is a first, major step in the development of a VALID data archive.

For further information on this project see http://hdl.handle.net/1839/00-8C315BC1-AD5E-4348-9A79-A41FE3DE1150@view.

Slips of the tongue data: This dataset is a collection of slips of the tongue in Dutch, British English, New Zealand English, and American English. The
data concerns slips of the tongue (not only blends) involving phrasal lexical items.

For more information on slips of the tongue data see http://hdl.handle.net/1839/00-0000-0000-001E-6DBE-9@view.

**Nijmegen Corpus of Spanish English:** The corpus was recorded in April and May 2012 in the laboratory of the Grupo de Tecnología del Habla at the Escuela Técnica Superior de Ingenieros de Telecomunicación of the Universidad Politécnica de Madrid.

Students of the Escuela Técnica Superior de Ingenieros de Telecomunicación and other Madrid universities were recruited via e-mail. This message was entirely in Spanish, as was all communication with all Spanish participants prior to their arrival at the recording day. Since the students were unaware that the recordings would be in English, self-selection was avoided as much as possible. Thirty-four Spanish students replied to our invitation to participate in a research project. The two Dutch speakers in the corpus were recruited at the Radboud University in Nijmegen, the Netherlands. They were both active in a student theater group and therefore confident in role playing.

In the informal recordings, casual speech was elicited between the Spanish student and the Dutch confederate of the same sex. The Spanish student was unaware that the Dutch speaker was a confederate. There were two parts in the informal recordings. In the first part the speakers were asked to get to know each other and talk about any topic they liked, while the researcher was out of the room to get the task they would be performing. Once the speakers ran out of conversation topics, the researcher went back into the recording room with a name-guessing game, which formed the second part of the informal recordings.

The formal speech was elicited during an interview about Spanish politics and the economic crisis in Spain. The Spanish speaker was interviewed by the Dutch confederate of the opposite sex, who was introduced as a colleague of the researcher in a journalism master’s thesis project. Both speakers wore formal clothing and a video camera was overtly present.

For more information on this corpus see http://hdl.handle.net/1839/00-0000-0000-0020-5D09-1@view.
2.3 Lexical resources

2.3.1 Database of paradigmatic semantic relation pairs (IMS)

The database of paradigmatic semantic relation pairs is a collection of semantically related word pairs in German which was compiled via human judgement experiments hosted on Amazon Mechanical Turk. Three paradigmatic relations are addressed: antonymy, hypernymy, and synonymy. The database consists of three parts:

- A representative selection of target lexical units drawn from GermaNet, using a principled sampling technique and taking into account the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency.

- A set of 8,910 human-generated semantically related word pairs, based on the target lexical units.

- A subset of 1,684 semantically related word pairs, rated for the strengths of relations.

The dataset fulfils the following criteria:

- It focuses on multiple paradigmatic relations.

- It systematically works across word classes.

- It explicitly balances the targets according to semantic category, polysemy, and type frequency.

- It explicitly provides positive and negative rating evidence.

A CMDI metadata description for the database was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on the database see http://www.ims.uni-stuttgart.de/forschung/ressourcen/experiment-daten/sem-rel-database.en.html.

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-digitised print edition published in 1989. The EtymWB was actively edited and extended throughout the report period. It currently comprises approx. 8,700 morphologically simple headwords together with approx. 14,000 subordinated headwords.
The lemma list of all versions of EtymWB is made available through the CLARIN-D repository at the BBAW together with grammatical properties of the lemmas, detailed CMDI metadata, and PIDs.

For more information on EtynWB see previous annual reports.

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 10.0 was prepared and made available. This release contains 10,000 new entries: 994 verbs, 7,945 nouns, and 1,061 adjectives. Therefore, GermaNet now consists of 101,366 synsets and 131,810 lexical units.

2.3.4 German Verb Subcategorisation Database (IMS)

Based on the SubCat-Extractor, verb subcategorisation information from German MATE dependency parses was induced. The subcategorisation database is represented in a compact but linguistically detailed and flexible format, comprising various aspects of verb information, complement information, and sentence information within a one-line-per-clause style.

A CMDI metadata description for the database was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on this resource see http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/subcat-database.en.html.

2.3.5 Merkmalsnormen für deutsche Nomenkomposita (IMS)

Feature norms are short descriptions of typical attributes for a set of objects. They often describe the visual appearance (a firetruck is red), function or purpose (a cup holds liquid), location (mushrooms grow in forests), and relationships between objects (a cheetah is a cat). The underlying features are usually elicited by asking a subject to carefully describe a cue object, and recording their responses. The quantification of the resulting object–feature pairs is called feature norms.

For the dataset Merkmalsnomen für deutsche Nomenkomposita, feature norms were collected for a set of 572 concrete, depictable German nouns, comprising 244 noun–noun compounds and their corresponding constituents (e.g. features for Schneeball (snowball), Schnee (snow), and Ball (ball)).
A CMDI metadata description for the feature norms was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on this dataset see http://www.ims.uni-stuttgart.de/forschung/ressourcen/experiment-daten/feature-norms.en.html.

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 R3.3 report). In this section, we shortly describe the efforts that have been contributed by AP5 members to foster LRT harmonisation and compatibility. Much of the effort reported in this section concerns the WebLicht web-service orchestrator. For more details on WebLicht see previous reports.

3.1.1 Monitoring (UTüb)

UTüb is continuously working on the improvement and maintenance of the west (WEbServiceTester) Java library. This library is used by the monitoring system to test both the availability and functionality of WebLicht web services. It generates sample test input for web services and must be extended when a web service is added that requires a type of input, which is not yet supported.

3.1.2 Bombard (UTüb)

The Bombard tool for stress testing services was made available to other CLARIN-D centres for testing their web services with regard to reliability and performance.

3.1.3 TüNDRA treebank search and visualisation (UTüb)

TüNDRA is a web application for searching treebanks. The work on TüNDRA continued during the fourth year. Most notably, the user interface was remodeled, providing increased space for viewing the trees and a more intuitive navigation.

After the successful tests with the Index Thomisticus Treebank (see R5.3), more dependency treebanks have been included in TüNDRA. These include:

- BulTreeBank-DP (see section 2.1.2)
- TüBa-D/W (see section 2.1.14)
- a version of TüBa-D/Z with automatically constructed dependencies (see section 2.1.15).
3.2 Morphological analysers

3.2.1 Spanish Malaga Morphology (IMS)

The Spanish Malaga Morphology (SMM) is an implementation of a morphological analyser for Spanish. SMM is based on the grammar development Malaga, which is based on the formalism of Left-Associative Grammar. SMM allows the analysis of Spanish word forms concerning inflection, derivation, and compounding. It is able to analyse verb forms with clitical pronouns. SMM analyses single word forms as well as large corpora. In order to use SMM, you need to install the current version of Malaga (available from Malaga Website), and the current version of SMM Rules and SMM Lexicon. SMM Rules and Lexicon are available as Open Source under GPL License and Creative Commons License.

A CMDI description for SMM is currently being prepared. The respective PIDs will be registered after the completion of the metadata description.

For more information on the Spanish Malaga Morphology see http://www.mahlow.ch/SM/i.

3.3 Taggers

3.3.1 Jitar Part-of-Speech Tagger (UTüb)

Jitar is an open source Hidden Markov Model part-of-speech tagger inspired by the non-redistributable TnT tagger. It was developed by Daniël de Kok at the University of Groningen. Within CLARIN-D, Jitar was improved to achieve higher accuracy for German and English. It was made available as a web service in WebLicht. For German it outputs tags in the STTS format. For English the tags are provided in Penn Treebank tag set format.

3.3.2 Stanford part-of-speech tagger (UTüb)

The Stanford part-of-speech tagger was included in WebLicht as a web service. This service includes the model provided by the Stanford Natural Language Processing Group for English. The tagger uses the Penn Treebank tag set.

3.4 Named entity recognisers (NER)

3.4.1 Stanford named entity recogniser (UTüb)

The Stanford named entity recognition labels sequences of words in a text, which are the names of things, such as person and company names. This service includes the models provided by Stanford: CoNLL 2003, MUC, and a model that has the
intersection of CoNLL and MUC labels, and the union of their data. The Stanford NER has been made available as a web service in WebLicht for English text.

3.5 Parsers

3.5.1 FSPar (IMS)

FSPar is a rule-based dependency parser. It is based on the approach of partial parsing by finite state cascades. FSPar also includes its own internal pipeline of lexically informed tokenizing and lemmatizing and applies the TreeTagger for part-of-speech tagging. It makes use of a large integrated lexical knowledge base and generates underspecified graphs with respect to head selection as well as dependency labels. Additional annotations include: sentence border, token, lemma, part-of-speech tag (STTS), morphological information, and some candidates for coreference resolution.

A CMDI metadata description for FSPar was created and the respective PIDs were registered. All metadata for this resource are harvestable via the OAI-PMH protocol.

For more information on the FSPar parser see http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/fspar.en.html.

3.6 Spoken language tools

3.6.1 Chunk preparation (BAS)

This service allows a pre-processing of Praat TextGrid annotation files containing a so called “chunk segmentation” data into WebMAUS processible BAS Partitur Format (BPF) files that contain a chunk segmentation layer TRN. That way users applying Praat for their orthographic transcription of larger video or speech recordings (often interviews) can obtain high quality word and phonetic segmentations.

3.6.2 G2P (BAS)

Grapheme-to-phoneme conversion (G2P) is a necessary preprocessing step for the automatic segmentation, but it is also quite a useful service in its own right. G2P has been extended to more input and output formats as well as new languages: American English, Estonian, Finnish, and Georgian.

3.6.3 IMS German Festival speech synthesis system (IMS)

The speech synthesis activities at the IMS (Experimental Phonetics) concentrates on various linguistic and application oriented aspects of speech synthesis. The
goal is to achieve naturally sounding and linguistically motivated speech synthesis. The speech synthesis system IMSGermanFestival is based on the original Festival speech synthesis framework developed at CSTR, University of Edinburgh. The current voice of the system uses diphones taken from the MBROLA project.

A CMDI description for the system is currently being prepared. The respective PIDs will be registered after the completion of the metadata description.

For more information IMSGermanFestival see http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/IMS_German_Festival.html.

3.6.4 Pho2Syl (BAS)

This new service provides an automatic syllabification of phonetic transcripts for the languages: German, Estonian, English (British, Australian, American, New Zealand), Italian, Hungarian, Dutch, Polish, Georgian, Finnish, and French. It supports canonical and spontaneous speech transcriptions and can be customised with respect to word boundary synchronisation.

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 webinterface of WebMAUS was re-designed within the reporting period to conform the CLARIN-D corporate identity design. Since the last reported version new languages were added: American English, Estonian, Finnish, and Georgian. The output now also allows phonetic feature categories such as IPA manner of articulation and IPA place of articulation.

The WebMAUS Multiple service now offers EMU legacy database and EMU-R database as output format. That way users can seamlessly apply highly sophisticated database queries to MAUS results. The webinterface has been informally evaluated by naive users and consequently improved; a FAQ and use case section has been added to the webinterface. Based on the number of backend calls per days we see an exponentially growing demand for this service since beginning of 2013.

For further information on the WebMAUS tool see https://clarin.phonetik.uni-muenchen.de/BASWebServices/.

3.6.6 WebMINNI (BAS)

This new service allows the segmentation and labelling of speech without any orthographic transcript. A first set of languages were implemented as webservice and webinterface on the production server: German, Italian, American English, Australian English, Estonian, and Hungarian.
3.7 Other tools

3.7.1 ASV Online Toolbox (ULei)

The ASV Toolbox has been available for many years as an offline stand-alone tool. The ASV Toolbox is a modular collection of tools for the exploration of written language data. The tools work either on word lists or text and solve several linguistic classification and clustering tasks. 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 is working on transforming these tools into a collection of REST webservices that are CLARIN-D 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-D infrastructure (e.g. via workspaces). As mentioned in the previous reports, the graphical user interface was already created and the first webservices were integrated. After a successful testing phase, further tools were added in the recent months.

Among these new tools were services such as:

- tokeniser
- topic model creation
- spell checking (using Levenshtein distance)
- Zipfel tool for analyzing word frequency distributions
- Viterbi tagger

In the following year, the development of the ASV Online Toolbox will be continued and additional tools will be added.

3.7.2 B3 Database (IMS)

The B3-Database is a relational database schema which allows to store primary data, information on processing tools, analyses, and annotations as generic objects. Objects and relations are typed and divided into a macro layer (containing workflow information), and a micro layer (containing the analyses).

A CMDI description for this database is currently being prepared. The respective PIDs will be registered after the completion of the metadata description.
3.7.3 COALA (BAS)

The manual creation of CMDI compliant metadata is very tedious. COALA simplifies this task by reading input from plain text tables.

Automatic creation of CMDI-compliant metadata is now implemented as web-service and webinterface on the BAS production server.

3.7.4 Cosmas II (IDS)

Cosmas II (Corpus search, management, and analysis system) is a corpus search engine providing access and analysis on corpora of contemporary written German in DeReKo (see section 2.1.5). As one of the search engines of the CLARIN-D federated content search, Cosmas II enables search on IDS public resources such as Goethe literatures. Moreover, Cosmas II has been integrated in Weblicht as a web-service tool.

3.7.5 FOLKER (IDS)

FOLKER is a transcription editor developed in the AGD according to GAT2 (Gesprächsanalytisches Transkriptionssystem, conversation analytic transcription system). Primarily, FOLKER is used to transcribe and annotate the recordings of FOLK. A CMDI metadata description of FOLKER is available in the IDS repository.

3.7.6 Higher Order Tree Coreference (IMS)

The Higher Order Tree Coreference (HOTCoref) system is a data-driven coreference resolution system. It models coreference within a document as a directed rooted tree. For learning it adopts the idea of latent antecedents and exploits the tree structure for the purpose of non-local (with respect to a single pair of mentions) features. Higher order features is a term often used to describe non-local features in the context of dependency parsing. The system obtains the best results published to date on all languages from the CoNLL 2012 Shared Task. It is written entirely in Java and is thus platform independent. The download package includes binaries, sources, and a description how to replicate the experiments from the paper. The system is licensed under the GNU General Public License (GPL).

A CMDI description for HOTCoref is currently being prepared. The respective PIDs will be registered after the completion of the metadata description.

For more information on HOTCoref see http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/HOTCoref.en.html.
3.7.7 KorAP (IDS)

KorAP (Korpusanalyseplattform der nächsten Generation, next generation corpus analysis platform) is a web-platform for searching and analysing very large corpora. 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 2.1.5) and has been integrated into the CLARIN-D federated content search and WebLicht.

3.7.8 OrthoNormal (IDS)

OrthoNormal is a tool for normalizing the orthography of transcriptions. Results of automatic lemmatisation and POS-tagging can also be manually corrected. The tool’s metadata is described in CMDI and available in the IDS repository.

3.7.9 PolDiLemma-Tool (UdS)

Historical word forms may differ in orthography and/or grammatical categories. It is a linguistic challenge to make these old forms accessible. The PolDiLemma-Tool is a Python tool for generating possible middle Polish inflected word forms including the part of speech tags from a given (new) Polish word. It contains a middle Polish morphology in XFST-format (compiled and plain text) and a stemmer (Morfologik, a full form lexicon and finite state based stemmer).

3.7.10 Stanford Tokeniser (UTüb)

The Stanford Tokeniser is a fast, deterministic, rule-based tokeniser for English. It is designed to largely mimic Penn Treebank (PTB) tokenisation. Since annotation tools for English are commonly trained on the Penn Treebank, the use of a compatible tokeniser typically results in better annotations.

The Stanford tokeniser was integrated in the CLARIN-D infrastructure as a WebLicht service. It takes TCF as input and provides tokens and sentences as output.

3.7.11 TEI to TCF to TEI converter (BBAW, UTüb)

In order to accommodate the needs of a growing user-base, we have implemented tools to facilitate the processing of XML source texts encoded according to the Text Encoding Initiative (TEI) in WebLicht by a two-stage process.

First, the TEI data are “encoded” into CLARIN-D’s native Text Corpus Format (TCF) used by WebLicht by means of linguistically motivated heuristics to re-
serialise inline elements such as footnotes and marginalia. Arbitrary WebLicht toolchains may then be used to analyse the TCF-encoded TEI text data. Finally, the encoded and analysed TCF data is “decoded” into a TEI XML document, whereby the TCF `<tokens>` and `<sentences>` layers are realised as (possibly fragmented) `<w>` and `<s>` elements in the decoded TEI.

Sharing of token- and sentence-identifiers allows interpretation of the TCF analysis output as a form of stand-off markup for the decoded TEI data. Both encoder and decoder were implemented using the `DTA::TokWrap` module developed at the BBAW in the context of the Deutsches Textarchiv project (see section 2.1.7), and made accessible as WebLicht web-services. Inclusion of the source TEI document in the encoded TCF is required for stateless (RESTful) processing, to which end the TCF schema as well as the WebLicht orchestration infrastructure has been updated by UTüb to provide a `<textSource>` element, flexible enough to allow additional source types as the need arises.

The converter is available as a stand-alone service at [http://kaskade.dwds.de/tei-tcf/](http://kaskade.dwds.de/tei-tcf/), WebLicht web-services are registered under their PIDs [http://hdl.handle.net/11858/00-203C-0000-0024-758D-2](http://hdl.handle.net/11858/00-203C-0000-0024-758D-2) (the encoder) and [http://hdl.handle.net/11858/00-203C-0000-0024-758E-F](http://hdl.handle.net/11858/00-203C-0000-0024-758E-F) (the decoder). For more information on `DTA::TokWrap` consult [http://odo.dwds.de/~moocow/software/dta-tokwrap/](http://odo.dwds.de/~moocow/software/dta-tokwrap/).

### 3.7.12 Text cleanup (UTüb)

Text cleanup is a web service made available in WebLicht. The development of this service was motivated by the observation that some tools do not work well with non-ASCII Unicode punctuation characters that were not seen during training. This service replaces such non-ASCII characters by ASCII counterparts. This web service is language-neutral and can be used to clean up both plain text and TCF.