AP5 Second Year Activity Report

April 2013
Contents

1 Introduction 4

2 Adaptation and conversion of language resources for use with the CLARIN-D infrastructure 5
  2.1 Text corpora and archives ............................................. 5
  2.2 Speech and multimodal corpora ................................... 8
  2.3 Lexical resources ....................................................... 10
  2.4 Other resources ......................................................... 13

3 Integration of existing tools into the CLARIN-D infrastructure 14
  3.1 CLARIN-D infrastructure development ............................. 14
  3.2 Morphological analyzers ......................................... 18
  3.3 Taggers ................................................................. 18
  3.4 Named entity recognizers (NER) ................................ 18
  3.5 Spoken language tools ............................................. 18
  3.6 Other tools ............................................................ 19
1 Introduction

This report documents the work and achievements of the CLARIN-D consortium with respect to the tasks defined for work package 5 “Dienste und Ressourcen” (tools and resources) during the second year of CLARIN-D (June 2012 – Mai 2013).

The work described in this report mostly builds upon the curation efforts for linguistic resources and tools (LRT) reported on in the first year activity report.

During the second project year CLARIN-D member institutions focused on extending the available set of LRT for the CLARIN-D infrastructure by preparing and integrating existing LRT of the individual partner institutes that have been created in the context of other projects (see sections 2 and 3).

In December 2012 the “CLARIN-D User Guide” – a detailed handbook describing the CLARIN-D infrastructure and illustrating its use and application was published on the CLARIN-D website (http://www.clarin-d.de/en/language-resources/userguide). It was collaboratively written by members of all CLARIN-D centers and is freely available under a CC-BY-ND license for online reading and for download. The userguide will be constantly revised and further refined throughout the third year.

AP5 organized a workshop at the conference “Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzerperspektiven” (“Historical corpora for the humanities and social sciences. Use issues and perspectives,” 18–19 February 2013, Berlin) which was held in cooperation with the Deutsches Textarchiv project (German Text Archive). An introduction to CLARIN-D and its infrastructure was presented together with hands-on demonstrations of technical aspects of integrating tools and resources into the WebLicht system. Over 60 researchers from various humanities disciplines attended the event.

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

- BAS Bayerisches Archiv für Sprachsignale, LMU München
- BBAW Berlin-Bandenburgische Akademie der Wissenschaften
- IDS Institut für deutsche Sprache Mannheim
- IMS Institut für maschinielle Sprachverarbeitung, Universität Stuttgart
- MPI MPI for Psycholinguistics Nijmegen
- UdS Universität des Saarlandes
- UHH Universität Hamburg
- ULei Universität Leipzig
- UTüb Universität Tübingen
2 Adaptation and conversion of language resources for use with the CLARIN-D infrastructure

In this section we summarize all LRT that were made available to CLARIN-D during the second year of the project. We also included LRT for which the adaptation and/or conversion is not yet completely finished. These LRT will become available at the beginning of the third year.

A complete tabular overview of CLARIN-D LRT and their adaptation and availability status is provided on the CLARIN-D wiki (login required) at http://de.clarin.eu/mwiki/index.php/AP5:Ressourcenuebersicht (resources) and http://de.clarin.eu/mwiki/index.php/AP5:Tooluebersicht (tools). This overview will eventually comprise both D-SPIN and CLARIN-D LRT and will be frequently updated.

2.1 Text corpora and archives

A TEI-P5-compliant corpus encoding format that unifies the DTA base format and IDS-XCES, the two formats currently employed for the BBAW and IDS corpora, respectively, is currently being developed to foster the creation of a common text corpus representation format for CLARIN-D. This work will extend into the third year.

2.1.1 DaSciTeX (UdS)

The Darmstadt Corpus of Scientific Texts (DaSciTeX) contains full English scientific journal articles compiled from 23 sources covering nine scientific domains (see first year activity report for more details). The DaSciTeX corpus is now integrated in a CQPweb installation at UdS. Metadata for this corpus are available in Dublin Core and CMDI format.

2.1.2 DeReKo (IDS)

The German Reference Corpus (DeReKo) is the largest collection of electronic corpora of contemporary written German:

- DeReKo was expanded by approx. 0.7 billion words to now more than 6.1 billion words.

- In cooperation with the PolMine project (http://polmine.sowi.uni-due.de/) at the University of Duisburg-Essen, protocols of all German parliaments were digitized, converted to XCES/TEI P5, and, as of release 2013-I, integrated into DeReKo.

- The end-user license of the DeReKo-sub-corpus “Reden und Interviews” (“Talks and Interviews”) was changed to CC-BY-SA (thanks to Adrien Barbaresi for a large contribution).
2.1.3 Deutscher Wortschatz (ULei)

Since many years the Deutscher Wortschatz project offers large corpora and statistical data on its “Corpus-Portal”. The statistical analysis of these corpora is based on language independent, automatic and semiautomatic algorithms. Currently these corpora are available in more than 130 different languages. The data is accessible via webportal (http://corpora.informatik.uni-leipzig.de/), SOAP Webservices (http://wortschatz.uni-leipzig.de/Webservices/) and some of the corpora by 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 CLARIN-D, metadata of ten German and Nepali newspaper corpora has been made available for piloting purposes. These corpora include material published on newspaper websites in 2008 and 2010 in sizes 10k, 100k and 1M sentences in the Wortschatz Corpus format. These resources are now fully integrated into the CLARIN-D infrastructure:

- metadata is available in CLARIN-EU compatible way: (updated) CMDI components and profiles are registered to the CLARIN-EU component registry
- metadata is available for harvesting via OAI-PMH
- each corpus is identified by a PID
- each corpus is exposed to the CLARIN-D Federated Content Search

The metadata was also updated according to the new standards supported by the Virtual Language Observatory (VLO).

2.1.4 DTA (BBAW)

The Deutsches Textarchiv project (“German text archive”, DTA, available at http://www.deutsches-textarchiv.de/) aims at compiling a core set of German texts that illustrate language use from 1650 to 1900. It includes first editions of canonical as well as lesser known literary texts and scientific and other non-literary texts. The list of texts was edited with support from a board of experts representing a broad range of scientific domains. Due to their dates of publication none of the historical texts are bound to copyright restrictions. The full-text versions of the texts are provided under a CreativeCommons CC-BY-NC license (for more details see the first year activity report).

During the second year a detailed CMDI profile for historical texts was developed and allows for the representation of all metadata that is available within the DTA.

As DTA tries to be a “living archive”, it also comprises supplemental texts that are provided by external groups or projects (DTA-E), e.g. the “Neue Rheinische Zeitung” (a historical newspaper) provided by the project Marx-Engels-Gesamtausgabe (“complete works of Marx and Engels”) or the Polytechnisches Journal (a technical journal published 1820–1931) in collaboration with the project Digitizing the “Polytechnische Journal”.
The DTA corpora are constantly being increased and made publically available together with their CMDI metadata. Currently (April 2013) the DTA corpora contain 1,259 books (81 million tokens).

2.1.5 GRUG (UdS)

The Georgian Russian Ukainian German parallel treebank (GRUG) 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 languages compiled for the GRUG NLP lexicon project.

The languages (except German) involved in the project are considered “lesser-resourced” languages from the computational viewpoint.

The German subcorpus of GRUG is now integrated in a CQPweb installation at UdS. Metadata for this corpus are available in Dublin Core and CMDI format.

For the German subcorpus of GRUG, federated content search using the SRU protocol is available.

2.1.6 HGC (IMS)

The Huge German Corpus (HGC) is a collection of German texts (newspaper, law texts) of about 204 million tokens including punctuation in 12.2 million sentences (about 180 million “real” words). The corpus was automatically segmented into sentences. Furthermore, it was lemmatized and part-of-speech tagged by the TreeTagger using the STTS tagset. The corpus is partly based on data taken from the European Corpus Initiative Multilingual Corpus I (EMI/MCI). This already existing corpus is now also taken care of by IMS. For more details on HGC see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.en.html.

A CMDI description for HGC is currently being prepared. The respective PIDs will be registered after its completion and the corpus will be made available for academic use in CLARIN-D.

2.1.7 SdeWaC (IMS)

SdeWaC is based on the deWaC web corpus of the WaCky-Initiative (http://wacky.sslmit.unibo.it/). It contains parsable sentences from deWaC documents of the .de domain. SdeWaC is limited to the sentence context. The sentences were sorted and sentence duplicates within the same domain name were removed. In addition, some heuristics based on Quasthoff et al. 2006: “Corpus Portal for Search in Monolingual Corpora” have been applied. SdeWaC-v3 comes in two formats:

1. one sentence per line and

2. one token per line including part-of-speech and lemma annotation.
A CMDI description for SdeWaC-v3 is currently being prepared. The respective PIDs will be registered after its completion.

For more information on SdeWaC see http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html.

2.1.8 TIGER (IMS)

The TIGER corpus is a German newspaper corpus enriched with part-of-speech annotation, morphological and lemma information and syntactic structure. For version 2.1 of this text corpus resource, CMDI metadata have been created and the respective PIDs have been registered, therefore the metadata is now harvestable via the OAI-PMH protocol.

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

2.1.9 TLA (MPI)

The Language Archive (TLA, http://tla.mpi.nl/) at the MPI has taken further steps to curate its services and resources and to integrate them more tightly into the CLARIN-D infrastructure.

For the IMDI metadata files, a CMDI translation service was deployed, in order to generate CLARIN compatible metadata on the fly.

2.2 Speech and multimodal corpora

2.2.1 DGD (IDS)

The Datenbank für Gesprochenes Deutsch (DGD2, “German Spoken Language Database”) is a corpus management system with an interactive web-based interface to a part of the data managed by the Archiv für gesprochenes Deutsch, AGD at IDS (see first year activity report for details). The first official release (December 2012) provides browsing and query access to metadata, transcriptions and aligned audio recordings, as well as the possibility to fully download selected datasets. The DGD2 currently comprises 18 corpora, totalling about 9,000 recordings, 2,500 hours of audio, 4,000 transcripts or 7.5 million transcribed word tokens.

2.2.2 DIRNDL (IMS)

The Discourse Information Radio News Database for Linguistic Analysis (DIRNDL, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/dirndl.en.html) is a corpus resource based on hourly broadcast German radio news. The textual version of the news is annotated with syntactic information. On top of this, the syntactic phrases are labeled with information status categories (given-new information). The speech version is prosodically annotated, i.e. with pitch accents and prosodic phrase boundaries. As the textual and the speech version slightly deviate from each other due to slips of the tongue, fillers and minor modifications. A (semi-automatic) linking of the two versions
was carried out and the results were stored inside the database. With the help of these newly established links, all annotation layers can be accessed for exploring the relations between prosody, syntax and information status.

Due to the two deviating sets of primary data, we are currently defining a CMDI profile based on existing components for text as well as speech corpora. The respective PIDs will be registered after creating the CMDI metadata.

2.2.3 EXMARaLDA Demokorpus (UHH)

A selection of short audio and video recordings in various languages to be used for instruction and demonstration of the EXMARaLDA system was compiled by UHH. The corpus contains 21 recordings. The transcriptions are based on the HIAT transcription conventions.

2.2.4 FOLK (IDS)

The Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK, German Speech Corpus for Research and Teaching) strives to document the German language as it is spoken today by giving insight into the reality of social communication in present-day Germany in different German-speaking regions. It consists of a wide-ranged yet balanced collection of spontaneous speech data (recordings and transcriptions) from different areas of social life such as work, leisure, education and media. Corpus construction is ongoing. The current version published in the DGD2 (section 2.2.1) comprises 95 recorded speech events, corresponding to about 70h of audio recordings or 700,000 transcribed tokens.

2.2.5 Hamburg Modern Times Corpus (UHH)

The Hamburg Modern Times Corpus consists of audio recordings of film retelling tasks (“Modern Times” by Charles Chaplin) with adult L2 users of German. The speakers’ L1 vary, but for most L1s there is more than one speaker. The speakers’ L2 proficiencies vary, too, but extensive metadata to their language learning biographies (free text in summarizing transcription) is available.

The corpus contains of 26 recordings (+1 by a L1 German speaker) between 5 and 15 minutes length. The transcription is based on the HIAT transcription conventions and is completed for 14 recordings.

Metadata for the Hamburg Modern Times Corpus is available in CMDI. The corpus can be used by academic users in CLARIN-D and will also be made accessible for the federated content search (see 3.1.1) in year three.

2.2.6 Speech corpora at BAS

The speech corpora at BAS are currently distributed on traditional media (CD-R, DVD-R, harddisk) via BAS and ELDA. After the ingestion into the CLARIN-D repository these corpora will be available online to CLARIN members free of charge. However, it is still possible to order these corpora on media via BAS or ELDA.
The BAS has now set up a CLARIN-D repository which can be accessed using the PID http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E or directly via the URL https://clarin.phonetik.uni-muenchen.de/BASRepository/.

Access to the metadata in the repository is unrestricted. Access to the primary data is restricted to registered CLARIN users via Shibboleth authentication. The metadata in the repository can be searched via the OAI-PMH endpoint at https://clarin.phonetik.uni-muenchen.de/cgi-bin/BASRepository/oaipmh/oai.pl?verb=Identify. Federated content search accesses the content data via an SRU endpoint at https://clarin.phonetik.uni-muenchen.de/BASSRU/.

In the BAS repository, corpora and sessions within the corpora are identified via PIDs; updated versions of corpora and sessions receive new PIDs so that the earlier versions remain accessible via their original PIDs.

Available resources (for detailed descriptions see first year activity report):

- SmartWeb SHC SMC (multimodal human machine interaction)
- HEMPEL (spontaneous telephone speech)
- PD2 (read speech)
- Ph@ttSessionz (read and spontaneous speech of adolescents)
- ALC (Alcohol Language Corpus)
- VM I (controlled dialogue)
- VM II (free dialogue)
- SMARTKOM AUDIO (men-machine dialogue)
- SMARTKOM PUBLIC (men-machine dialogue)
- SMARTKOM HOME (men-machine dialogue)
- SMARTKOM MOBIL (men-machine dialogue)
- ZIPTEL (numbers, street/city names)
- TAXI (bilingual dialog (German, English)
- VOYS (adolescents, read/spontaneous, Scottish English)

2.3 Lexical resources

2.3.1 dlexDB (BBAW)

dlexDB (http://dlexdb.de) is a cooperative project of the departments of cognitive psychology and theoretical computational linguistics at the University of Potsdam and the project Digital Dictionary of the German Language (DWDS, http://www.dwds.de/)
at BBAW. Goal of the project is to establish a lexical database for psychological and linguistic research.

dlexDB comprises frequency information on a broad range of linguistic and graphematic levels (e.g. frequencies and n-gram frequencies for tokens, morphems, characters or word association measures). During the second year a WebLicht compatible interface was developed for dlexDB and proposals for the adaptation of WebLicht’s TCF format were discussed to enable the representation at least of parts of the available (lexico)statistical information. This work will extend into the third year.

2.3.2 1DWB (BBAW)

1DWB is the retro-digitized version of the first edition of the Grimm brother’s “Deutsches Wörterbuch” which was compiled and printed from 1854 to 1961. This 33-volume dictionary contains about 300,000 main headwords and a yet not fully known number of related minor headwords.

Detailed CMDI metadata has been recorded for this resource and will be made available together with a webservice providing the list of headwords and grammatical information (e.g. part-of-speech, gender) by the end of the second year.

2.3.3 Deutscher Wortschatz (ULei)

The corpora of the Deutscher Wortschatz project (see section 2.1.3) can also be described and accessed as a lexical resource, since not only sentences, but also automatically computed data is available for each wordform:

- frequency and frequency-class,
- sentence-based cooccurrences,
- right and left neighbours and
- example sentences (including source reference).

The following webservises are available for the seven corpora described in section 2.1.3:

- words cooccurring to the left / right of target word,
- words cooccurring with the target word on sentence level,
- frequency of the target word,
- sentences containing the target word,
- sentences contained in target source, and
- sentence for a given sentence ID.

Most of these webservises are compatible to WebLicht (TCF 0.4). Metadata for these webservises is available in CMDI (according to the WebLicht webservises CMDI profile) via the OAI-PMH interface of the CLARIN-D resource center Leipzig. Additionally a PID for each of these services was registered.
2.3.4 EtymWB (BBAW)

A CMDI metadata description for EtymWB was revised and BBAW plans to make parts of the dictionary available through a web service (headwords, related headwords, grammatical information, definitions).

2.3.5 Lehnwortportal (IDS)

The Lehnwortportal Deutsch (“loan word portal German”, http://lwp.ids-mannheim.de/) is a lexical information system developed at the IDS. It provides unified access to a growing number of manually interlinked dictionaries of German loanwords in other languages. The initial release (November 2012) hosts three loanword dictionaries:

- Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standard-sprache (http://lwp.ids-mannheim.de/dict/wdlp), a dictionary of German loanwords in Standard Polish,
- Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen (http://lwp.ids-mannheim.de/dict/wdlt), a dictionary of German loanwords in the Teschen dialect of Polish,
- Deutsche Lehnwörter im Slovenischen (http://lwp.ids-mannheim.de/dict/st), a dictionary of German loanwords in Slovene.

The two Polish dictionaries were previously published online; the Slovene dictionary is a partially retro-digitized version of a print monograph. Apart from conventional access to the individual works, the information system features advanced cross-dictionary search facilities and provides an “inverted loanword dictionary” (Wörterbuch der deutschen Herkunftswörter des Portals, http://lwp.ids-mannheim.de/dict/meta) that traces the way of German words into the different target languages. The web application operates on a database that represents pertinent lexicographical information as a cross-resource network (directed acyclic graph) of relations between words (etyma and loanwords with their respective variants, derivatives etc.).

Integration of the Lehnwortportal data in CLARIN-D is expected to begin in 2014.

2.3.6 OWID (IDS)

In the Lexis department at IDS various corpus-based dictionaries and lexical reference works are being compiled and continuously extended. The monolingual lexical resources are integrated into the dictionary portal Online-Wortschatz-Informationssystem Deutsch (OWID, Online Lexis Information System for German) and can be accessed from there. At the time of writing OWID contained the following dictionaries:

- elexiko (http://www.owid.de/wb/elexiko/start.html), a dictionary of contemporary German,
- Feste Wortverbindungen (http://www.owid.de/wb/uwv/start.html), a dictionary of fixed multi-word expressions,
• Kommunikationsverben (http://www.ovid.de/docs/komvb/start.jsp), a dictionary on German speech act verbs,

• Neologismenwörterbuch (http://www.ovid.de/wb/neo/start.html), a dictionary of neologisms and lexical innovations,

• Protestdiskurs 1967/68 (http://www.ovid.de/wb/disk68/start.html), a discourse dictionary covering the discourse around the topic “democracy” in the late 1960s,

• Schuldiskurs 1945–55 (http://www.ovid.de/wb/disk45/einleitung.html), a reference work covering the lexemes that establish the discourse about issues of German war guilt in the early post-war era from 1945 to 1955, and

• Sprichwörterbuch (http://www.ovid.de/wb/sprw/start.html), a dictionary of German proverbs, compiled within the EU-project “SprichWort. Eine Internetplattform für das Sprachenlernen”.

During the report period contact with CLARIN-D regarding possibilities for cooperation was ensured at all times.

2.4 Other resources

2.4.1 Information systems at IDS

IDS maintained and improved the following information systems during the second year:

• Gesprächenanalytisches Informationssystem (GAIS, http://prowiki.ids-mannheim.de/bin/view/GAIS/), an information system of conversation analysis,

• grammis (http://hypermedia.ids-mannheim.de/), a grammatical information system,

• bibliographic databases:
  – Konnektoren (http://www.ids-mannheim.de/gra/konnektoren/anfrage.html), an online bibliography on connectors,
  – Präpositionen (http://www.ids-mannheim.de/gra/konnektoren/p-anfrage.html), an online bibliography on prepositions,
  – Bibliografie zur deutschen Grammatik (BDG, http://hypermedia.ids-mannheim.de/call/public/bib.ansicht), an online bibliography on German grammar,
  – Gesprächsforschung (http://hypermedia.ids-mannheim/de/pragdb/bgf.html), an online bibliography on conversation analysis,
  – OBELEX, an online bibliography on electronic lexicography for dictionaries (http://www.ovid.de/obelex/dict) and research literatur (http://www.ovid.de/obelex/meta).
3 Integration of existing tools into the CLARIN-D infrastructure

3.1 CLARIN-D infrastructure development

3.1.1 CLARIN-D federated content search aggregator

The federated content search aggregator (FCS aggregator) provides a web-based interface for end-users to access the functionality of SRU/CQL conforming endpoints and their resources.

The user can specify a query and select endpoints and resources to query (figure 1). The aggregator web application dispatches the query to the target repositories, collects the partial results from keyword-in-context data view, and presents them to the user. The user can either download the search results, or save them to their CLARIN-D personal workspace (figure 2).

The aggregator supports endpoint selection by providing explicit `x-aggregation-context` parameters, making it possible to integrate it with other resource searching tools such as the Virtual Language Observatory.

3.1.2 Harvesting and monitoring

As part of an ongoing effort to integrate WebLicht into the wider CLARIN infrastructure, we have introduced support for distributed webservice registration. Making a webservice available to WebLicht users can be achieved in two steps:

1. A webservice description is created in a standardised format specifically devised for WebLicht. A special online editing tool is provided to simplify this process.

2. Once a description is written, it must be published in the repository of a participating CLARIN center. WebLicht periodically gathers these descriptions from all the participating centers, putting all the currently available webspes at user’s disposal.

Another requirement for the CLARIN infrastructure is to constantly monitor the availability of webspes. In case of a technical malfunction to a webservice, the responsible CLARIN center can take action to restore its full functionality. A special tool has been developed that tests the functionaity of a WebLicht webservice in addition to its availability. It automatically generates an appropriate sample input for a WebLicht webservice, then runs the webservice, and makes sure that the output conforms to the expectation.

3.1.3 WebLicht geolocation webservice (UTüb)

During the second year WebLicht has been substantially improved and extended. WebLicht’s geolocation webservice looks for location names in the named entities of TCF-annotated texts and adds map coordinates and geographic information to them. It inserts
Figure 1: FCS aggregator user interface: corpus selection.
into the annotation a specific longitude, latitude, and altitude, as well as the continent and country where it can be found, and capital cities where possible. Location data used by the service comes from GeoNames (http://www.geonames.org/), indexing all cities with a population over 1000.

Currently, this geolocation service supports names in English, German and Spanish, but can be extended to support more languages. It is accessible as an NLP service via WebLicht, taking TCF formatted input with tokens and named entity annotations, and producing a TCF document with geo annotations appended.

WebLicht also provides map visualization for geolocation output (see figure 3)

### 3.1.4 TüNDRA treebank search and visualization (UTüb)

The Tübingen aNnotated Data Retrieval Application (TüNDRA) provides search and tree visualization services for stored corpora and for TCF annotated treebanks using a query interface based on the TIGERSearch and CQP query languages. It supports dependency, constituency and hybrid annotations, and support for general segment-annotated data (even if not structured as a hierarchical tree) is in progress. Queries can range over the labels and features of tokens, edges and constituents or other abstract tree elements, including full support for existential negations (like requesting nodes that are not parents of nodes matching some query). Support for additional non-tree edges makes it possible to query non-hierarchal annotations outside the scope of simple trees (figure 4).

Users have a selection of tree drawing formats to choose from, and internal support for serving webfonts makes it possible to reliably display treebanks for any language with an available Unicode-compliant OpenType or TrueType font, including languages written from right to left, on any up-to-date browser with webfont support.
Figure 3: WebLicht’s geovisualization output.

Figure 4: TüNDRA treebank visualization.
3.2 Morphological analyzers

3.2.1 SMOR (IMS)

SMOR is a German finite-state morphology. It is integrated in the CLARIN-D infrastructure by means of a web service. A landing page has been created for the download of the tool. CMDI metadata for the download tool is currently being developed.

3.3 Taggers

3.3.1 RFTagger (IMS)

The RFTagger is a part-of-speech tagger for Hungarian, German, Slovenian, and Czech, and makes use of fine-grained tagsets. In addition to the web service, CMDI metadata for the download tool, along with the respective PIDs are provided by IMS.

3.3.2 TreeTagger (IMS)

The TreeTagger (http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html) is a tool for annotating text with part-of-speech and lemma information. There are parameter files currently for German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese, Swahili, Latin, Estonian and old French. However, the TreeTagger is also trainable for other languages. The TreeTagger for Italian, English, French and German is included in WebLicht. In addition to the web service, CMDI metadata for the download tool, along with the respective PIDs are provided by IMS.

3.4 Named entity recognizers (NER)

3.4.1 SemiNER (UdS)

SemiNER is a semi-supervised NER tool with pre-trained models for German. In 2012 a new release of SemiNER was published. It is now fully integrated in the sequor package. UdS is currently preparing a web service to integrate SemiNER into WebLicht.

3.5 Spoken language tools

3.5.1 EXMARaLDA toolset (UHH)

The EXMARaLDA search and analysis tool was extended to be able to access the HZSK repository directly.

3.5.2 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 (see figure 5). WebMAUS now offers three interfaces: the basic MAUS with language selection only, general MAUS with access
to all parameters of MAUS, and multiple MAUS for drag and drop processing of many files at once.

WebMAUS can be found online at https://clarin.phonetik.uni-muenchen.de/BASWebServices/

3.6 Other tools

3.6.1 ASV Online Toolbox (ULEi)

The ASV Toolbox has been available for many years as an offline stand-alone tool. In short, the ASV Toolbox is “a modular collection of tools for the exploration of written language data. They 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” (http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/). Currently ULei is working on transforming these tools into a collection of REST webservices 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-D infrastructure (e.g. via workspaces).

The webservices and the graphical user interface are already implemented and are currently in a testing phase. Metadata and additional functionality will be added in the
near future.

3.6.2 Bohnet Toolchain (IMS)

The Bohnet toolchain (http://code.google.com/p/mate-tools/) includes a lemmatizer, a part-of-speech tagger and a state-of-the-art dependency parser for German. In addition to the existing web service specifications, CMDI metadata has also been created for the download tool and the respective PIDs have been registered.

3.6.3 COALA (BAS)

COALA is a script to generate valid CMDI meta data files for the profiles media-corpus and media-session from a set of five comma-separated text files. Such text files are easier to edit using standard word processing or spreadsheet applications, or they may be exported directly from databases. COALA is written in perl and has successfully been used to generate the CMDI metadata files for the ALC and the SmartKom corpora at BAS, the ASiCa corpus created by an external partner, and for preliminary versions of the multimodal corpora created by the F-AG 8 (Sprache und andere Modalitäten) in Bielefeld.

3.6.4 EXMARaLDA Basic Transcription Converter (UHH)

UHH developed a converter between the EXMARaLDA format and several other commonly used transcription formats used in CLARIN (Praat, ELAN, TEI). A REST-style webservice was created for integration in WebLicht TCF processing chains.

3.6.5 Interactive Text Analysis Tool (IMS)

The Interactive Text Analysis Tool is a prototype system which is based on RESTful web services and implements an interactive relation extraction system. It involves a (re-)trainable web service on top of a web service processing chain (tokenizer, tagger, parser) that brings about automatic linguistic annotation on several levels. The system aims to show how the dynamic interaction between such software and human users from the digital humanities can be brought about.

3.6.6 TLA tools (MPI)

The following tools at The Language Archhive (TLA) were significantly updated and extended in year two:

- The content-search engine TROVA was connected to the SRU/CQL example code provided by CLARIN-D. This enables users to search in the corpora hosted at the TLA. For the moment the following (open) collections are made SRU searchable: ESF, CGN, Childes, TalkBank, and IFA.
- ELAN has been extended to support web-service plugins, e.g. for the WebMAUS phonetic alignment service provided by BAS. Lexus was completely re-implemented and now features LMF compatible export functionality. The authentication of the TLA archive was connected to the CLARIN Service Provider Federation. This allows users from external identity providers to access the archive.

- The Annex annotation viewer was extended to support handles (persistent identifiers) as an input parameter.

- The CMDI support for Arbil was significantly improved. Support was given to researchers to use handles in their publications.

All webservice hosted by the TLA have been described with CMDI and made available via OAI-PMH and the Virtual Language Observatory (VLO).

### 3.6.7 Wortprofil 2012 (BBAW)

Wortprofil 2012 is a tool similar to the WordSketch engine but focusses on German (see first year activity report for details). The tool is already exploited at the DWDS website (http://www.dwds.de/) and is a valuable tool for current lexicographic work at BBAW. It supports TCF-0.4 input and output and was fully integrated into WebLicht in year two.