Resources, Tools, and Applications at the CLARIN Center Stuttgart

Cerstin Mahlow Kerstin Eckart Jens Stegmann André Blessing
Gregor Thiele Markus Gärtner Jonas Kuhn
Institute for Natural Language Processing (IMS)
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart
firstname.lastname@ims.uni-stuttgart.de

Abstract

This NECTAR track paper (NECTAR: new scientific and technical advances in research) summarizes recent research and curation activities at the CLARIN center Stuttgart. CLARIN is a European initiative to advance research in humanities and social sciences by providing language-based resources via a shared distributed infrastructure. We provide an overview of the resources (i.e., corpora, lexical resources, and tools) hosted at the IMS Stuttgart that are available through CLARIN and show how to access them. For illustration, we present two examples of the integration of various resources into Digital Humanities projects. We conclude with a brief outlook on the future challenges in the Digital Humanities.[1]

1 Introduction

CLARIN-D[2] is the German branch of the European CLARIN initiative[3]. The overall goal is to implement a web-based and center-based infrastructure to facilitate research in the social sciences and humanities. This is achieved by providing linguistic data, tools, and services in an integrated, interoperable, and scalable infrastructure. CLARIN-D is funded by the German Federal Ministry for Education and Research (BMBF).

The Institute for Natural Language Processing (IMS) at the University of Stuttgart is one of currently nine German centers. CLARIN centers undergo thorough external and internal evaluation regarding mostly technical requirements—e.g., metadata, repository system, documentation, legal issues, authentication, and authorization. The IMS was awarded the *Data Seal of Approval* in March 2013 and gained the status of an official CLARIN center in June 2013.[4]

The integration of existing linguistic resources and tools includes efforts towards availability of resources as well as towards the creation and publication of metadata to enable the discovery of resources. All German centers closely collaborate on technical aspects and issues in the curation of language resources. Exchange on the European level is facilitated via the annual CLARIN ERIC conference and specific task forces.

The IMS provides a number of well-established as well as some recently created lexical and corpus resources; it also offers various tools in order to process linguistic data. They are usually made available both as a download package (to be installed and executed locally by the user) and as a web service. The latter is clearly in line with the general CLARIN philosophy of seamless access and usability of resources via the WWW. One particular interest is domain adaptation, resulting in

[2]http://www.clarin-d.de.

[3]http://www.clarin.eu.

[4]http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-95

many follow-up questions—e.g., related to the extendability of resources, the feedback of expert users, and the design of pertinent user interfaces. The IMS is involved in the development of several applications and showcases that demonstrate their potential for enabling Digital Humanities research.

In the rest of this NECTAR paper we first provide an overview of the resources developed and hosted at the IMS Stuttgart that are available through CLARIN-D (section 2). In section 3 we present two examples of how to use those resources in actual Digital Humanities projects. Section 4 concludes with a summary and a brief outlook on the future challenges in the Digital Humanities.

2 Resources

We use the term *linguistic resource* in a broad sense. Resources can be text, speech, multimodal corpora, and lexical knowledge bases, but also the tools utilized to create, annotate, and query linguistic information and data collected within experiments or studies. This also includes web services, i.e., tools that can be applied via a web browser and run on servers of the providing organizations. Similarly, important parts of these tools, such as grammars or statistically trained language models are also resources on their own.

The objective of CLARIN, however, is not only to provide resources, but to set up an infrastructure to support the applicability and interaction of these resources. Important aspects are (a) the possibility to find existing resources and to determine whether they fit one's own needs, (b) the possibility to store, access, execute, process, and cite linguistic resources, and (c) the possibility to reproduce experiments or studies based on specific versions of resources. All aspects contribute to the sustainability of the respective resources.

To be able to search for linguistic resources the *Virtual Language Observatory* (VLO)[5] has been set up (van Uytvanck et al., 2012). The faceted browser allows for a search based on free text, but also provides facets which allow users to filter resources by specific features, e.g., by language or resource type. A large number of resources are already listed in the VLO. Current development focuses on improvement of user interaction.

In the VLO, resources are described by their metadata. Since relevant metadata aspects are not easy to be defined a priori for all resource types, CLARIN proposed the flexible *Component MetaData Infrastructure* (CMDI, (Broeder et al., 2012)). In CMDI, metadata schemes reflecting the specific needs of the different resource types can be created by common means. This way CLARIN also helps to improve the documentation of resources, since metadata are one prerequisite for a resource to become part of the CLARIN infrastructure. For all resources we present in this paper, CMDI descriptions have been created or enriched.

The metadata and also the resource itself can be stored and made available via data repositories. Such repositories are hosted at the CLARIN centers. The metadata stored can be automatically harvested via the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). The term *harvesting* means that automatic collector services—e.g., from the VLO—regularly access a pertinent service exposed by the repository and copy all the disseminated metadata. Therefore it is not necessary to explicitly register resources at the VLO or to commit changes there. Since metadata do not contain any part of the resource itself and usually do not contain sensitive information, the CLARIN requirements stipulate that they have to be free to read and free to harvest via the web. The metadata of all the resources presented here are freely harvestable via the OAI-PMH service of the IMS repository.[6]

While metadata are freely available, this is often not the case for the resources themselves. We usually find a legal limbo with respect to language resources. However, we can distinguish

[5] http://catalog.clarin.eu/vlo/.

[6] http://clarin04.ims.uni-stuttgart.de/oaiprovider/oai?verb=ListRecords&metadataPrefix=cmdi.

(a) resources which can be freely distributed, (b) resources which are restricted to research purposes, and (c) resources with additional restrictions. Free resources, for example, can provide a download link in the VLO. Restricted resources, however, have to be adressed via (a) specific legal licensing schemes and (b) an authentication and authorization strategy that respects given restrictions. CLARIN adresses the former by providing licensing templates for resource creators corresponding to the respective categories mentioned above.[7]

The solution to the latter is the implementation of web-based single-sign on via authentication and authorization infrastructures (AAI) using the Shibboleth[8] technology. For example, the University of Stuttgart is a member of the DFN-AAI[9] federation (identity providers), so all researchers and students can login to CLARIN-D web applications using their University of Stuttgart credentials[10]. Additionally, the IMS CLARIN center has registered as a service provider in the DFN-AAI federation. We are currently in the process of reorganizing the mode of access via the IMS repository to make use of the federated Shibboleth-based approach.

Due to fast and constant development of resources, it is necessary to not only cite publications about resources, but also the datasets or resources themselves. This allows for a more precise citation and also supports the reproducibility of findings, when the exact version of the applied resources can be identified. Within CLARIN *persistent identifiers* (PIDs) are registered for data and metadata alike. These PIDs act as links to the sets of metadata, the download of the resource, and a landing page of the resource. They can be resolved in the address line of a browser, similar to DOIs. The advantage of PIDs is that they are not supposed to change. If a website moves, and thus the previous URL becomes invalid, it is not possible to find all places on the web that provided a link to the original website. If, however, all references are made to the PID, only the PID needs to be realigned with the new address of the web page, and the page and the resource remain accessible. It is a prerequisite for a resource in the CLARIN infrastructure to be identifiable by a PID. The PIDs have to be part of the CMDI metadata provided and can be registered via services provided by members of the EPIC consortium[11]. The IMS uses the service offered by GWDG.[12]

We now present resources hosted or created at the CLARIN-D center Stuttgart. For all resources metadata have been created and PIDs have been registered[13]. Most of the provided web services are also accessible via WebLicht, the CLARIN-D Web-based Linguistic Chaining Tool[14].

2.1 Corpora

The *Huge German Corpus* (HGC)[15] is a collection of German texts (newspaper and 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, and lemmatized and part-of-speech tagged by the TreeTagger (Schmid, 1994) using the STTS tagset. The corpus is partly based on data taken from the European Corpus Initiative Multilingual Corpus I (EMI/MCI). This corpus is now also maintained by the IMS.

SdeWaC[16] is based on the deWaC web corpus of the WaCky-Initative[17]. It contains parsable sentences from deWaC documents of the .de domain. (Faaß and Eckart, 2013) SdeWaC is limited to the sentence context. The sentences were

[7] https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiLicenceCategories.

[8] https://shibboleth.net/.

[9] https://www.aai.dfn.de/.

[10] This also works on the European level to access services provided by members of the CLARIN Service Provider Federation.

[11] http://www.pidconsortium.eu/.

[12] http://handle.gwdg.de:8080/pidservice/.

[13] We thus give the respective PIDs for each resource

[14] http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

[15] http://hdl.handle.net/11858/00-247C-0000-0022-F7B4-4.

[16] http://hdl.handle.net/11858/00-247C-0000-0022-F7BA-7.

[17] http://wacky.sslmit.unibo.it

sorted and duplicate sentences within the same domain name were removed. In addition, some heuristics based on Quasthoff et al. (2006) have been applied. SdeWaC-v3 comes in two formats: (a) one sentence per line and (b) one token per line including part-of-speech and lemma annotation.

The *TIGER corpus*[18] is a German newspaper corpus enriched with part-of-speech annotation, morphological and lemma information and syntactic structure (Brants et al., 2004). Versioning is an important aspect of the proper modelling of linguistic resources via metadata. We use the TIGER corpus as testbed for exploring different possibilities in this respect. Questions related to versioning highlight aspects of the more general question of how to deal with relations among resources.

The *Discourse Information Radio News Database for Linguistic Analysis* (DIRNDL)[19] is a corpus resource based on hourly broadcast German radio news (Eckart et al., 2012). The textual version of the news is annotated with syntactic information. Syntactic phrases are labeled with information status categories (given vs. new information). The speech version is prosodically annotated, i.e., with pitch accents and prosodic phrase boundaries. 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.

GECO[20] has been created to investigate phonetic convergence in German spontaneous speech (Schweitzer and Lewandowski, 2013). The database consists of 46 dialogs of approximately 25 minutes length each, between previously unacquainted female subjects. Of these 46 dialogs, 22 dialogs were in a unimodal setting, where participants could not see each other, while the remaining 24 dialogs were recorded with subjects facing each other. The database was automatically annotated on the segment, syllable, and word levels using forced alignment with manually generated orthographic transcriptions.

2.2 Lexical Resources

The *German Logical Metonymy Database*[21] is the result of a corpus study for German verbs (*anfangen (mit)* ('to start (with)'), *aufhören (mit) ('to stop')*, *beenden* ('to end'), *beginnen (mit)* ('to begin (with)'), *genießen* ('to enjoy')), based on data obtained from the deWaC corpus. (Zarcone and Rued, 2012) The database contains 2'661 metonymies and 1'886 long forms with two expert annotations.

The *IMSLex dictionary database*[22] covers information on inflection, word formation, and valence for several ten thousand German base forms. (Fitschen, 2004)

The *German Verb Subcategorization Database*[23] contains verb subcategorization information from German MATE dependency parses of SdeWaC. The subcategorization 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. The SdeWaC subcategorization database comprises 73'745'759 lines (representing the number of extracted target verb clauses), resulting in 6.3 GB in compressed format.

2.3 Tools

For all tools we have CMDI data for a downloadable local executable version and for the webser-

[18] http://hdl.handle.net/11858/00-247C-0000-000D-FFB5-1.
[19] http://hdl.handle.net/11858/00-247C-0000-0022-F7B2-8.
[20] http://hdl.handle.net/11858/00-247C-0000-0023-5137-2.
[21] http://hdl.handle.net/11858/00-247C-0000-0023-5147-D.
[22] http://hdl.handle.net/11858/00-247C-0000-0022-F7B8-B.
[23] http://hdl.handle.net/11858/00-247C-0000-0023-8BCD-01.

vice version we provide for the CLARIN-D infrastrucure.

SMOR[24] is a German finite-state morphology implemented in the SFST programming language (Schmid et al., 2004). It is integrated in the CLARIN-D infrastructure by means of a web service, there is also an SMOR download tool.

We deployed a new morphology web service called *Stuttgart Morphology* for German which derives the morphological analysis from *RFTagger's* (see below) internal analysis.

The *TreeTagger*[25] is a tool for annotating text with part-of-speech and lemma information (Schmid, 1994). We deployed a new version (i.e., TreeTagger2013) of TreeTagger as web service implemented in Java. The new release achieves better performance.

RFTagger[26] is a part-of-speech tagger providing also morphological information and makes use of fine-grained tagsets (Schmid and Laws, 2008). The RFTagger web service is implemented in Java.

We deployed a new *NER web service for German*[27] based on the Conditional Random Field-based Stanford Named Entity Recognizer by Finkel and Manning (2009) which includes semantic generalization information from large untagged German corpora. (Faruqui and Padó, 2010)

BitPar[28] is a parser for highly ambiguous probabilistic context-free grammars (such as treebank grammars). BitPar uses bit-vector operations to speed up the basic parsing operations by parallelization (Schmid, 2004). It is integrated in the CLARIN-D infrastructure by means of a web service.

The *Bohnet Toolchain*[29] includes a lemmatizer, a part-of-speech tagger, a morphological tagger, and a state-of-the-art dependency parser for German (Bohnet, 2010). We deployed a new version of the Bohnet Toolchain web service. The new release includes some bugfixes and performance improvement. The Bohnet Toolchain is available as *MATE Tools* for download[30]; additionally, it is deployed at the High Performance Computing Center Garching as web service.

The *Interactive Text Analysis Tool* is a prototype system based on RESTful web services implementing an interactive relation extraction system (Blessing et al., 2012). It comprises a retrainable web service on top of a web service processing chain (tokenizer, tagger, parser) which merges automatic linguistic annotation on several levels. The system aims to demonstrate the dynamic interaction between such software and human users from the Digital Humanities.

The *TIGERSearch*[31] software helps to explore linguistically annotated texts. It is a specialized search engine for retrieving information from a database of graph structures (treebank) (Lezius, 2002). The text corpus to be searched by TIGERSearch must have been annotated beforehand, e.g., with grammatical analyses (syntax trees).

3 Case Studies

3.1 ICARUS

ICARUS[32] is a search and visualization tool that primarily targets dependency trees (Gärtner et al., 2013). It allows users to search dependency treebanks given a variety of constraints, including

[24] http://hdl.handle.net/11858/00-247C-0000-0022-F7BC-3.
[25] http://hdl.handle.net/11858/00-247Z-0000-0007-5EC0-4.
[26] http://hdl.handle.net/11858/00-247C-0000-000D-FFB4-3.
[27] http://www.nlpado.de/~sebastian/software/ner_german.shtml.
[28] http://hdl.handle.net/11858/00-247C-0000-0022-F7B0-C.
[29] http://hdl.handle.net/11858/00-247Z-0000-0007-6A0D-E.
[30] http://code.google.com/p/mate-tools
[31] http://hdl.handle.net/11858/00-247C-0000-0022-F7BE-0.
[32] http://hdl.handle.net/11858/00-247C-0000-0022-F7B6-F.

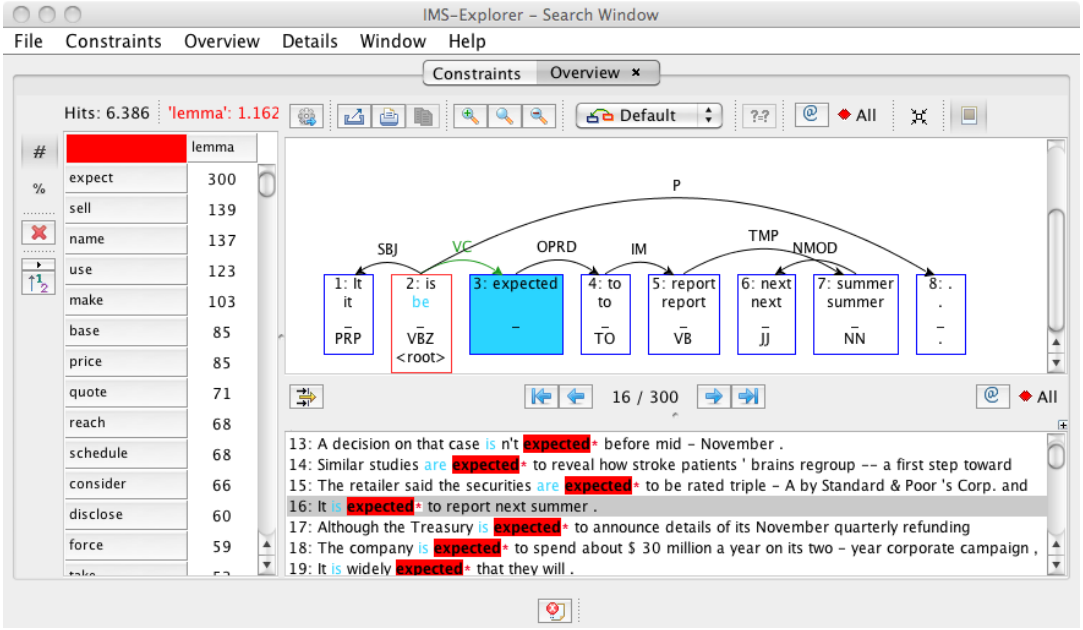

Figure 1: Passive constructions in the treebank grouped by lemma and sorted by frequency.

searching for particular subtrees. Emphasis has been placed on functionality that makes it possible for users to switch back and forth between a high-level, aggregated view of the search results and browsing of particular corpus instances. Users can create queries graphically and results will be returned as frequency lists and tables (i.e., quantitatively) as well as qualitatively by connecting the statistics to the matching sentences and allowing the user to browse them graphically. The first application using ICARUS is a search engine to explore dependency trees in treebanks as shown in Figure 1.

ICARUS provides plugins for the integration of existing tools or pipelines like the Bohnet Toolchain. So far, two additional applications have been developed: ICE, the *ICARUS Coreference Explorer* (Gärtner et al., 2014), and a graphical interface for automatic error mining of annotation in corpora (Thiele et al., 2014). Both applications use annotated corpora and make use of the general ICARUS features.

ICE is an interactive tool to browse and search coreference annotation. The annotation can be displayed as tree, as entity grid, or as text. Figure 2 shows the entity grid with the predicted annotations and the complete text. Different annotations of the same text can be compared, thus facilitating evaluation. Two usergroups are in focus: NLP developers designing coreference resolution systems—here ICE serves as interactive diagnosis and evaluation tool towards a gold standard—and corpus linguists—here ICE serves as research instrument. The built-in search engine of ICARUS is adapted to allow queries over sets of documents to actually allow searching a corpus. ICE is the first graphical coreference exploration tool offering three different visualizations and thus supporting various user needs.

The ICARUS error mining extension is a tool for finding annotation errors and inconsistencies in large annotated corpora. It implements the automatic error mining algorithms proposed in (Dickinson and Meurers, 2003) and (Boyd et al., 2008) for part-of-speech and dependency annotations, respectively. The tool allows the user to find potential annotation errors by presenting a list of candidates generated by the algorithm. It presents statistics on the label distribution of the candidate and connects the error candidate with the sentences in the corpus in which it occurs. The user can then decide if the annotation is indeed erroneous and needs to be corrected. Figure 3 illustrates the candidate list for the part-of-

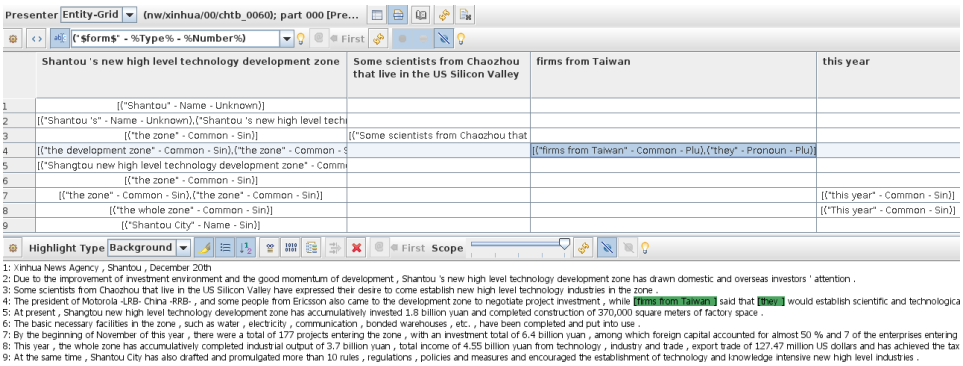

Figure 2: Entity grid over the predicted clustering in the example document.

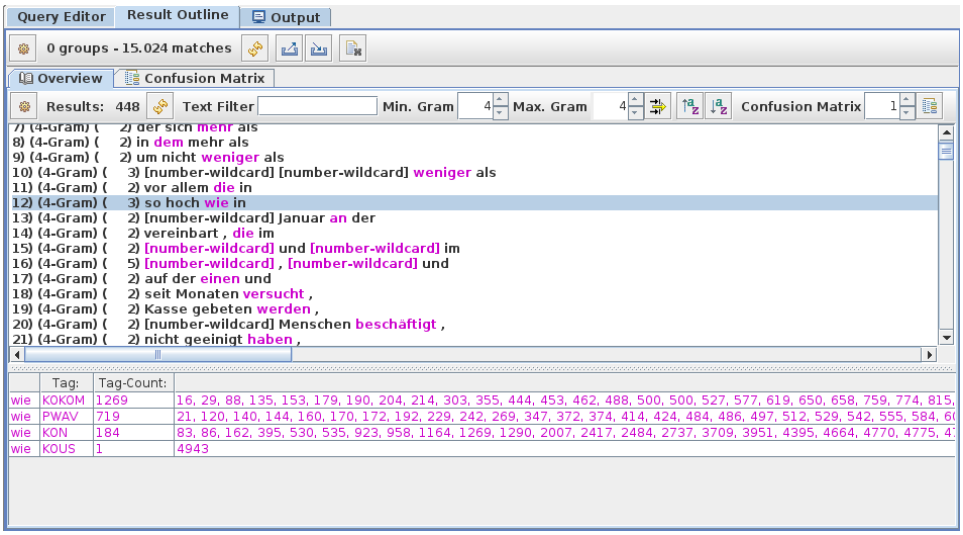

Figure 3: N-gram view of the error mining application based on ICARUS.

speech error mining algorithm. In the upper part, candidate tokens are shown with their surrounding context. In the lower part, a label distribution for the candidate token is shown. Clicking on the candidate or on one of the labels opens the corpus browser with the appropriate sentences. The user can inspect the relevant sentences and decide if there are erroneous annotations. This tool is thus intended to support corpus creation and curation, the processing step before corpus linguists may actually query the corpus to answer dedicated research questions. Annotations to be checked for errors and inconsistencies may stem from both manual or automatic processing.

3.2 Textual Emigration Analysis

Textual Emigration Analysis (TEA)[33] is a web-based application that transforms raw textual data into a graphical display of migration source and target countries. (Blessing and Kuhn, 2014) The tool serves as showcase demonstrating the use of language technology to support research in the humanities. It is used in ongoing research projects. For instance, from the sentence "*Erika Lust grew up in Kazakhstan and emigrated to Germany in 1989.*" we can extract the triple `emigrate(Erika Lust, Kazakhstan, Germany)` by using several web services (tokenizer, TreeTagger, Bohnet Toolchain, NER) provided by the CLARIN-D in-

[33] `http://clarin01.ims.uni-stuttgart.de/geovis/showcase.html`.

frastructure. Those triples are then visualized on a map.

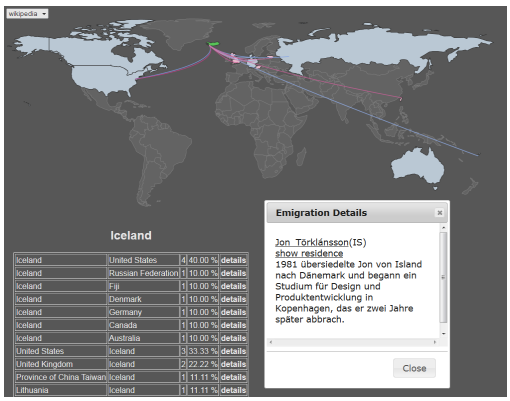

Figure 4: Screenshot of the user interface of the TEA web application showing emigration from and to *Iceland*.

TEA is intended to be used by humanities scholars; it offers a visual impression of the aggregated data as well as means for qualitative inspection of the underlying sources. Figure 4 shows a screenshot of the TEA-user-interface. The user selects a country (*Iceland* in the given example) to get the list of related migration events. The details of the row *Iceland-Denmark-1* are selected and the user sees the textual source which describes that Jon Törklánsson emigrated from Iceland to Denmark. This way, the graphical visualization is more transparent, which leads to a better acceptance of automatic tools in the humanities; users can always refer to the corresponding sources.

4 Summary and Perspectives

In this paper, we presented the resources (corpora, lexical resources, and tools) provided at the CLARIN center at the IMS Stuttgart. We created CMDI metadata and registered PIDs for all resources, so they can be discovered and accessed by users. As examples for the use of those resources in actual applications, we elaborated on two use cases, the ICARUS search and visualization tool and the Textual Emigration Analysis to be used in Digital Humanities research.

On a technical level, an important focus for future work at the Stuttgart CLARIN center is on metadata: Currently, relations between resources are not covered by the provided CMDI data. Similarly, there is no agreed upon standard to describe different versions of a resource due to improvements of tools, extension or extraction of corpora, or the like. CMDI in general offers to describe relations and versions, however, various possibilities could be used. The use in the VLO requires some information and sets some costraints, but consistent procedures are still missing. For example we can register a PID for a resource and a PID for the respective CMDI description, but we cannot define which is depending on which. As mentioned before, we use the TIGER corpus as testbed for versioning and the creation of corresponding metadata to hopefully develop a proposal for general use.

Taking a broader Digital Humanities perspective, experience shows that an operational technical infrastructure is an important ingredient for innovative avenues of research, but there are remaining methodological challenges that cannot be resolved on a purely technical level. It is very important to engage in open-minded interdisciplinary collaborations and learn to better understand each other's working assumption and methodological conventions. The IMS is involved in several such interdisciplinary projects using the CLARIN-D infrastructure and the resources provided, and contributing to the formation of a Digital Humanities methodology. In the BMBF-funded project "e-Identity", a large corpus of newspaper texts from Austria, Germany, Ireland, France, the UK, and the USA is investigated with respect to national identities in critical political situations after the Cold War (Kolb et al., 2009; Blessing et al., 2013; Kliche et al., 2014). In the BMBF-funded project "ePoetics",[34] hermeneutic and algorithmic methods are combined to investigating a corpus of German scholarly aesthetics and poetics from 1770 to 1960 (Richter, 2010). The CLARIN center also collaborates closely with the infrastructure project of SFB 732 "Incremental specification in con-

[34] http://www.epoetics.de.

text",[35] a joint effort of theoretical and computational linguistics in which corpus resources and analysis tools play a central role. In the third funding period, the SFB focuses on the generalization of models and theories to non-canonical data types and phenomena and aims to build up a large collection of annotated corpora, adopting a "silver standard" approach of transparent and quality-controlled automatic annotation.

With the recent advances in computational linguistics and language technology, including machine learning paradigms that can be easily extended beyond a linguistically oriented approach to large text collections, there is no doubt about the great potential lying in these techniques for the broader Digital Humanities. But to intergrate them effectively with the established body of knowledge in the humanities and social sciences, the field needs a more systematic methodology that breaks down analytical processes into building blocks whose "deeper" functionality is transparent to the users in the humanities, so they are in a position to make their own critical assessment of the reliability of a particular component or component chain—and arrange for adjustments as necessary. Crucially, the meta-architecture to be established should include best practices for non-computational intermediate steps too, which are required to bridge the methodological gap between data-based empirical results and higher-level disciplinary research questions. Ultimately, Digital Humanities scholars should feel fully competent to draw upon a flexible methodological toolbox so they can try backing up any partial results from one component with evidence obtained from other sources, make informed adjustments to the components, or attempt an entirely different way of approaching the available information sources.

In other words, the mid- to long-term goal should not have IT specialists optimize a tool chain for fully automatic analysis so as to achieve the best possible performance for some specified task—which is bound to be imperfect for any non-trivial question anyway, thus requiring a responsible integration into higher-level research questions. The Digital Humanities should rather aim to create transparency within a complex multi-purpose network of interacting information sources of variable quality or reliability—in plain extension of the classical competences humanities scholars have always had regarding approaches to their object of study. Contrary to the assumptions one can make about the typical users in a standard web-oriented application scenario of language technology and visual analytics (where users rarely have any philological or other meta-level attachment to the text basis from which they are seeking information), humanities scholars have far-reaching competences and intuitions about their objects of study and their sources. This makes the goal of developing an interactive framework for a network of knowledge sources a promising endeavor, drawing on techniques for aggregation, diagnostic and explorative visualization, quantitative analysis and linking back to data instances and (re-)annotation tools, but crucially also including "soft" non-technical components, i.e., theoretically informed steps of analysis and reflection.

[35] www.uni-stuttgart.de/linguistik/sfb732/.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) in the CLARIN-D project, and the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg.

The authors would like to thank Heike Zinsmeister who made significant contributions to the Stuttgart CLARIN center during her time in Stuttgart, and Anders Björkelund and Wolfgang Seeker for their advice and contributions, in particular in the context of the ICARUS tool.

References

André Blessing and Jonas Kuhn. 2014. Textual emigration analysis. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan

Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2089–2093, Paris. European Language Resources Association (ELRA).

André Blessing, Jens Stegmann, and Jonas Kuhn. 2012. SOA meets relation extraction: Less may be more in interaction. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities*, pages 6–11.

André Blessing, Jonathan Sonntag, Fritz Kliche, Ulrich Heid, Jonas Kuhn, and Manfred Stede. 2013. Towards a tool for interactive concept building for large scale analysis in the humanities. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 55–64, Stroudsburg, PA. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. 2008. On Detecting Errors in Dependency Treebanks. *Research on Language and Computation*, 6(2):113–137.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a Component Metadata Infrastructure. In *Proceedings of the Workshop on Describing Language Resources with Metadata (LREC'12)*, Paris. European Language Resources Association (ELRA).

Markus Dickinson and W. Detmar Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114, Stroudsburg, PA. Association for Computational Linguistics.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, pages 65–75. Springer, Heidelberg.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC—A Corpus of Parsable Sentences from the Web. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer, Heidelberg.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Stroudsburg, PA. Association for Computational Linguistics.

Arne Fitschen. 2004. *Ein computerlinguistisches Lexikon als komplexes System*. AIMS // Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung, Lehrstuhl für Computerlinguistik, Universität Stuttgart. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Markus Gärtner, Gregor Thiele, Wolfgang Seeker, Anders Björkelund, and Jonas Kuhn. 2013. Icarus—an extensible graphical search tool for dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Stroudsburg, PA. Association for Computational Linguistics.

Markus Gärtner, Anders Björkelund, Gregor Thiele, Wolfgang Seeker, and Jonas Kuhn. 2014. Visualization, search, and error analysis for coreference annotations. In *Proceedings of the 52nd Conference of the Association for Computational Linguistic: System Demonstrations*, Stroudsburg, PA. Association for Computational Linguistics.

Fritz Kliche, André Blessing, Jonathan Sonntag, and Ulrich Heid. 2014. The e-identity exploration workbench. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 691–697, Paris. European Language Resources Association (ELRA).

Peter Kolb, Amelie Kutter, Cathleen Kantner, and Manfred Stede. 2009. Computer- und korpuslinguistische Verfahren für die Analyse massenmedialer politischer Kommunikation: Humanitäre und militärische Interventionen im Spiegel der Presse. In Wolfgang Hoeppner, editor, *Technischer Bericht Nr. 2009-01. GSCL-Symposium Sprachtechnologie*

und eHumanities, pages 62–71, Duisburg. Universität Duisburg-Essen.

Wolfgang Lezius. 2002. TIGERSearch – Ein Suchwerkzeug für Baumbanken. In Stephan Busemann, editor, *Proceedings der 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken.

Uwe Quasthoff, M. Richter, and C. Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1799–1802, Paris. European Language Resources Association (ELRA).

Sandra Richter. 2010. *A History of Poetics: German Scholarly Aesthetics and Poetics in International Context, 1770-1960*. De Gruyter, Berlin, New York.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 777–784, Stroudsburg, PA. Coling 2008 Organizing Committee.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Paris. LREC, European Language Resources Association (ELRA).

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing (NeMLaP 1994)*, pages 44–49, Manchester,UK.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Stroudsburg, PA.

Antje Schweitzer and Natalie Lewandowski. 2013. Convergence of articulation rate in spontaneous speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pages 525–529.

Gregor Thiele, Wolfgang Seeker, Markus Gärtner, Anders Björkelund, and Jonas Kuhn. 2014. A graphical interface for automatic error mining in corpora. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 57–60, Stroudsburg, PA. Association for Computational Linguistics.

Dieter van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the Virtual Language Observatory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Paris. European Language Resources Association (ELRA).

Alessandra Zarcone and Stefan Rued. 2012. Logical metonymies and qualia structures: an annotated database of logical metonymies for German. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Paris. European Language Resources Association (ELRA).