Originally published in: Fišer, Darja/Witt, Andreas (Eds.): CLARIN. The infrastructure for language resources.
– Berlin/Boston: de Gruyter, 2022. Pp. 275-306. (Digital Linguistics 1)
DOI: *https://doi.org/10.1515/9783110767377*

Christoph Draxler, Alexander Geyken, Erhard Hinrichs,
Annette Klosa-Kückelhaus, Elke Teich, and Thorsten Trippel

How to Connect Language Resources, Infrastructures, and Communities

Abstract: This chapter will present lessons learned from CLARIN-D, the German CLARIN national consortium. Members of the CLARIN-D communities and of the CLARIN-D consortium have been engaged in innovative, data-driven, and community-based research, using language resources and tools in the humanities and neighbouring disciplines. We will present different use cases and users' stories that demonstrate the innovative research potential of large digital corpora and lexical resources for the study of language change and variation, for language documentation, for literary studies, and for the social sciences. We will emphasize the added value of making language resources and tools available in the CLARIN distributed research infrastructure and will discuss legal and ethical issues that need to be addressed in the use of such an infrastructure. Innovative technical solutions for accessing digital materials still under copyright and for data mining such materials will be presented. We will outline the need for close interaction with communities of interest in the areas of curriculum development, data management, and training the next generation of

Acknowledgments: The work reported here was funded by the Federal Ministry of Education and Research, Germany (BMBF) and the home institutions of the authors within various projects and funding programmes, especially in the CLARIN-D and CLARIAH-DE contexts. It also involves work within affiliated projects supported by other funders such as the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK), and the German Research Foudation (DFG), and projects funded by the European Commission involving the institutions of the authors.

Christoph Draxler, Ludwig-Maximilians-Universität München, Institut für Phonetik, Munich, Germany, e-mail: draxler@phonetik.uni-muenchen.de
Alexander Geyken, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, Germany, e-mail: geyken@bbaw.de
Erhard Hinrichs, University of Tübingen, Tübingen, Germany, e-mail: erhard.hinrichs@uni-tuebingen.de
Annette Klosa-Kückelhaus, Leibniz Institut für Deutsche Sprache, Mannheim, Germany, e-mail: klosa@ids-mannheim.de
Elke Teich, Saarland University, Saarbrücken, Germany, e-mail: e.teich@mx.uni-saarland.de
Thorsten Trippel, University of Tübingen, Tübingen, Germany; Leibniz Institut für Deutsche Sprache, Mannheim, Germany, e-mail: thorsten.trippel@uni-tuebingen.de

https://doi.org/10.1515/9783110767377-011

digital humanities scholars. The importance of community-supported standards for encoding language resources and the practice of community-based quality control for digital research data will be presented as a crucial step toward the provisioning of high quality research data. The chapter will conclude with a discussion of important directions for innovative research and for supporting infrastructure development over the next decade and beyond.

Keywords: CLARIN-D, research infrastructure, humanities, user communities, use cases

1 Introduction

The availability of digital research data of various kinds has led to new research paradigms and innovative research results in many fields of science, including the humanities, the social sciences, and related disciplines. Findability of research data, easy access to such data, interoperability among research data, and the reuse of data have become important desiderata. These requirements have been summarized in the FAIR principles (see Wilkinson et al. 2016) and more recently in the additional CARE principles (see Carroll et al. 2021).

Language data play a key role in this digital turn since unstructured textual data account for up to 80% of all digital data (see ESFRI Roadmap 2018: 108). Given the enormous and ever-increasing volume of digital data, text and data mining techniques in combination with sophisticated data analysis and data visualization tools have become an indispensable part of data-driven research. More generally, these demands have led to the development of research data infrastructures that couple data resources with such analysis tools and a rich portfolio of other services that facilitate uptake of digital research methods by a growing number of researchers.

1.1 The digital turn

In the humanities and social sciences, research is increasingly based on empirically collected data, especially in the Digital Humanities (DH), sometimes also referred to as eHumanities. While in early DH projects, a main concern was the (retro-) digitization of data (see, e.g., Presner 2010), more recent work is based on large stocks of digitized material feeding into working environments to create, manage, and deal with digital knowledge. This new way of dealing with data

results in innovative questions that lead to prototypical computer-assisted approaches to analysis in the digital humanities (see also Schaal and Kath 2014).

The development of legacy – "born analogue" – and data already digitally archived from the beginning – "born digital" – does not constitute a discrete differentiation, but rather forms a continuum. The extremes here are analogue data at one end of the spectrum, with data that is available on paper accessible only in restricted facilities, and fully interoperable, interlinked, and reusable data at the other end. The latter is often referred to as FAIR, as mentioned above, indicating data that is Findable, Accessible, Interoperable, and Reusable.

The availability of data on the continuum between legacy and born digital data opens up new methodological approaches or entirely new scientific questions. These questions go hand in hand with discussions on the "digital turn" (see, e.g., Berry 2011; Baum and Stäcker 2015). Domain-related research infrastructures take up these methods and have the task of supporting research in all phases – in data research, the digital provision of data, the linking of data to form virtual collections, the analysis of data with the aid of interoperable software tools. It also includes the storage and archiving of the resulting research data. Originally often installed on the personal computing devices of researchers, more and more tools are becoming available with web interfaces (see, e.g., Gomes et al. 2022). The web based infrastructure allows complex querying of data, including data that is distributed at different institutions. Users apply the tools without having to install them, work collaboratively, and unknowingly benefit from hardware and service scalability due to the operation of service providers.

To achieve transparency in an opaque server-side processing of research data, the interaction and discussion between applying researchers and service providers becomes an indispensable requirement. This interaction is needed on both sides, on the side of the users and on the side of the research infrastructures. The researchers need the interaction to understand the limits and capabilities of the infrastructure to assess the results provided in the process. The infrastructure providers on the other side need the discourse to understand the requirements and research questions to adjust the services as needed. The discussion requires Research Data Management (RDM) services and consultation. Here the users receive the support they need to efficiently provide their own results according to the FAIR principles without the overhead costs of having to provide the services themselves. A helpdesk for specific questions helps researchers to work with the tools and find the expert knowledge they might require for their specific questions. This process may result in consulting requirements to adjust the infrastructure services or to find appropriate methods available for a given research question. Often the first point of contact between young researchers and services provided by the infrastructures is in the context of academic teaching

or at conferences and workshops run by professional associations. This contact allows the infrastructures to connect to the researchers who will use the services and create new datasets that may be made available for reuse by other researchers with the help of the infrastructure providers.

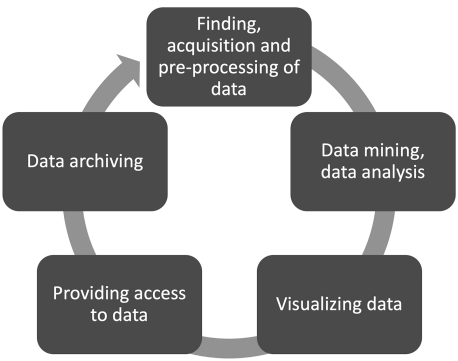

Figure 1: Illustration of the Research Data Lifecycle as a continuous process of data reuse and re-analysis, as often practised in the humanities.

1.2 Research Data Infrastructures (RDI) by researchers for researchers

Data is created at all stages of the research data lifecycle, from (1) finding, acquisition, and pre-processing data for reuse or creating new primary data, to (2) analysing data, including through data mining, (3) visualizing data, (4) making data available for review, and finally (5) archiving data.

Figure 1 illustrates the research data life cycle, which is used in different variants in many disciplines. What they have in common is that the entire process is viewed from the research perspective. However, some of the phases require cooperation with research infrastructure providers, for example, long-term archiving, which individual researchers can hardly be expected to do. Infrastructures are also useful for other tasks along the research data lifecycle, be it the provision of inventory data, tools for converting or searching and analysing data, virtual research environments, computing capacity, or the like. The research-driven data processing using infrastructures requires a continuous dialogue between data providers and data users – which can in some cases hardly be distinguished – and research infrastructures. In the case of CLARIN, a research infrastructure initiative was even created by and for researchers, along with national

nodes. The present chapter originates from participants in the German part of CLARIN.[1] Researchers joined forces to provide a sustainable infrastructure for their reference data and tools. Through sharing and collaboration they started to provide their data and services according to FAIR principles even before this term was coined, opening their data and services also for researchers, with an initiative that is open to new contributions and developments.

1.3 Use cases to extend the portfolio of Research Data Infrastructures

The dialogue between users of a research infrastructure (RI) and researchers providing the RI is needed to extend the portfolio of services and data. Though the researchers providing the RI also contribute to new developments based on their own research interests, new impulses can efficiently result from researchers not originally part of this development. This dialogue becomes transparent by providing uses cases.

Via use cases, the infrastructures demonstrate their existing abilities and options with data that is provided. Users of the infrastructure, on the other hand, also describe additional functionalities and required datasets by drafting a use case that fits their research interest. Hence, use cases are an effective means of extending the portfolio of research data and associated tools and for improving the usability of research data infrastructures. These use cases allow us to describe how research infrastructures can be used for new research topics, for new datasets, with new technologies, and for illustrating research questions in academic education.[2]

In the next section we will provide examples of the continuous enhancement and use of the research infrastructure, as illustrated by use cases.

1 Other national partners in CLARIN contributing to this volume are South Africa with Hennelly et al. 2022, Portugal with Silva et al. 2022, Czech Republic with Hajič et al. 2022, Lithuania with Petrauskaitė et al. 2022, and Austria with Trognitz, Ďurčo, and Mörth 2022.

2 More use cases for data and services are also included in this volume, for example in Silva et al. 2022 on diachronic Portuguese corpora; Lindahl and Rødven-Eide 2022 on Swedish corpora; Hoeksema, de Glopper, and van Noord 2022 on investigating secondary school writing; Pozzo et al. 2022 on aligning Chinese translations of Kant; Kučera 2022 on using NLP tools in psychological research; Fridlund et al. 2022 on cross-lingual text mining.

2 Development of the infrastructure by user-driven use cases

We established the need for infrastructures and user communities to interact. This interaction makes sure that new developments in the infrastructure are made available to the communities and the communities provide impulses for new developments as needed. To illustrate the interaction we draw on a number of use cases. We distinguish three classes of uses cases:

- addressing emerging research topics driven by public discourse;
- application of new technologies;
- new, faster more precise answers to established research questions;
- integration of new research data and developing community-methods for maintaining and improving research data quality.

In the remainder of this section, we will provide examples for each. These examples illustrate the interaction of user communities and infrastructure providers, which was key to answering the research questions.

2.1 Addressing emerging research topics driven by public discourse

Public discourse can lead to new research questions, for which an answer should be provided as part of this discussion. These questions may result from natural phenomena or long-term developments in society.

An example of natural phenomena influencing public discourse is the Covid-19 pandemic, which appeared in public media in 2020. Besides research in the life sciences and so forth, it also initiated research with regards to language change and language use, addressing the pandemic from a lexicographic perspective. An almost real-time investigation requires the availability of data and tools provided by a research infrastructure.

An example of long-term developments in society driving research topics is based on an intensive and long-lasting discussion rooted in emancipation and striving for non-discriminatory communication strategies. Here, the investigation of gender-neutral forms and their influence on pronunciation provides a new perspective on research on language change. Again, the tools and data for investigating such a research question can reuse data and tools developed for other purposes, provided by research infrastructures.

Interactions between infrastructures and scholarly users dealing with this type of question are characterized by their embeddedness in current events, but not necessarily by new methods and technologies. Though some new resources may be added, these questions are addressed with existing tools and often with existing resources.

2.1.1 Addressing the pandemic with lexicography

In 2020, the Covid-19 pandemic changed the world on a large scale; it also affected lexicographic work, as new words and phrases or new meanings of established words emerged on a daily basis and medical as well as epidemiological terminology became part of the general language. This is why early on in the pandemic, the *Digital Dictionary of the German Language* (DWDS)[3] compiled a thematic glossary with approximately 300 entries, containing medical terms (e.g., *Triage* 'triage', *Tröpfcheninfektion* 'droplet infection'), older lexemes with high current relevance in the public discourse on the pandemic (e.g., *Mundschutz* 'face mask', *Kontaktsperre* 'contact ban'), and neologisms (e.g., *Coronaparty* 'party during the Covid-19 pandemic defying rules of social distancing').[4] Existing entries were updated and new entries compiled based on corpus evidence to document the current changes in the German lexicon promptly. The thematic glossary presents the entries in an alphabetical list with (mostly) only the definition(s), but links them to the complete entry for each lexeme in the dictionary itself (with corpus citations, information on frequency, etc.).

The *Neologismenwörterbuch*[5] chose a different approach. This dictionary focuses on German neologisms from the three decades 1991–2000, 2001–2010, and 2011–2020. Starting in April 2020, it presents Covid-19 neologisms (new words, phrases, and meanings) in a continually updated list containing (as of March 2021) roughly 1,300 entries.[6] The meaning of each word or phrase is explained and at least one corpus citation is given. Not all words and phrases have yet been lexicographically described

3 See https://www.dwds.de/.

4 See https://www.dwds.de/themenglossar/Corona. Later in 2020, a thematic glossary on the US election campaign and one with Christmas words were published, see https://www.dwds.de/themenglossar/US-Wahl-2020 and https://www.dwds.de/themenglossar/Weihnachten, respectively.

5 See https://www.owid.de/docs/neo/start.jsp; more information on this portal can be found in Engelberg, Klosa-Kückelhaus, and Müller-Spitzer 2020; for dictionary portals in general see Engelberg and Müller-Spitzer 2013.

6 See https://www.owid.de/docs/neo/listen/corona.jsp.

in full, as neologisms are usually monitored for some years[7] before being accepted into the dictionary as part of the general language. Many of the Covid-19 neologisms will probably disappear at some point (e.g., many synonyms for grown-out haircuts due to periods of lock-down throughout the pandemic, such as *Coronafrisur*, *Coronamatte*, *Coronamähne*, *Lockdownfrisur*, *Lockdownlocken*, etc.). Thus, the *Neologismenwörterbuch* list is a snapshot of the current extension of the lexicon, based on evidence from online press and social media. One corpus-linguistic tool used to find candidates for the list is the *cOWIDplus Viewer* (cf. Section 2); information in the entries is also based on data from *Deutsches Referenzkorpus – DeReKo*.[8]

2.1.2 Exploring language change: The pronunciation of gender-neutral forms

With the ongoing debate about equal rights, non-discriminatory communication is part of public discourse. Currently, there is an ongoing discussion about gender neutrality in language in many countries. In German, the traditional male form to collectively refer to groups of people, for example, *Bäcker*, includes both male and female members, but with a perceived bias towards the male sex/gender.[9] New forms to express gender neutrality in German first appeared in written language in newspapers, social media, and job descriptions, for instance, a capital 'I' inside a word as in *BäckerIn*, or an asterisk or an underscore, as in *Bäcker*in* or *Bäcker_in*.

In a class for students in the master's programme, three students (two studying phonetics, one studying Ancient Greek) decided to explore how these gender neutral forms are spoken. Their hypothesis: gender-neutral forms are spoken with a perceivably lengthened final /I/ vowel. A quick check via general-purpose search engines and in CLARIN's virtual language observatory did not return matching resources. Thus, the students decided to record their own corpus, compute a phonetic segmentation which contains both the sound label (in the IPA alphabet) and their duration, run their analyses, and create a speech database to be added to a CLARIN-D repository for others to work with.[10]

---

7 A list of all words or phrases currently monitored by the dictionary project has been published online since 2019, see https://www.owid.de/docs/neo/listen/monitor.jsp.

8 For the use of a diachronic corpus to detect language change (such as lexical or semantic change) see Pettersson and Borin 2022.

9 For some interesting thoughts on the topic see, for example, https://www.nzz.ch/feuilleton/gendern-genus-und-sexus-sind-eng-miteinander-verbunden-ld.1578299.

10 This database is now available in the BAS repository: http://hdl.handle.net/11022/1009-0000-0003-FF39-F.

For the recordings, the students collected sentences from the German newspaper *taz, die tageszeitung*, edited them for readability, and generated both a gender-neutral and a non-gendered version of each sentence. The sentences were then read by 18 speakers in the studio of the phonetics institute in Munich, using the SpeechRecorder software (Draxler and Jänsch 2004).

The orthographic transcription was generated by listening to the audio files and manual modification of the prompt text to create a verbatim transcript, including filled pauses, self-repairs, and deviations from the prompt text. The result of these steps was a collection of more than 600 pairs of audio and text files.

A first look at the transcripts showed an unexpected phenomenon: for the production of gender neutral forms, speakers deviate from the given prompt in roughly 26% of the recordings. They expand the given form by adding its complement, either with or without a junctor, or they substitute the given form with the male form (see Table 1). Apparently, some speakers try to *avoid* the gender-neutral forms – the reasons are unknown.

Table 1: Avoidance strategies for sentences with a gender neutral form, for instance, *BäckerInnen*.

Type	Example	%
elliptical expansion	*Bäcker Bäckerinnen*	15.9%
complete expansion	*Bäcker und Bäckerinnen*	6.8%
substitution by male form	*Bäcker*	4.3%
other		2.0%

To generate a phonetic segmentation, that is a time-aligned annotation with the duration of words and individual speech sounds, the WebMAUS service (Kisler, Schiel, and Sloetjes 2012) was used. The students uploaded the file pairs to the CLARIN-D server via the graphical interface, selected standard German as the input language, IPA symbols as the output character set, and the Praat TextGrid file format (Boersma 2001). After a few minutes, the service displayed the segmentation in the Emu WebApp viewer (Winkelmann and Raess 2014), and the resulting files were downloaded to the local computer – this would have taken weeks if done manually.

The TextGrid files were converted to a tabular format and imported into a relational database system to be accessed from the statistics package R.

A statistical analysis of the duration of the final /I/ vowel showed both that the median duration was higher and the variation greater for gender neutral forms (see Figure 2 (b)). A plausible interpretation is that speakers produce gender neutral forms differently, but that there is not yet a consensus on how they should be produced.

The paper was successfully submitted to a phonetics conference (Slavik, Cronenberg, and Draxler 2018). A reviewer noted that this work describes one of the rare cases where orthography leads the way in sound change – a thrilling experience for students, made possible by CLARIN tools.

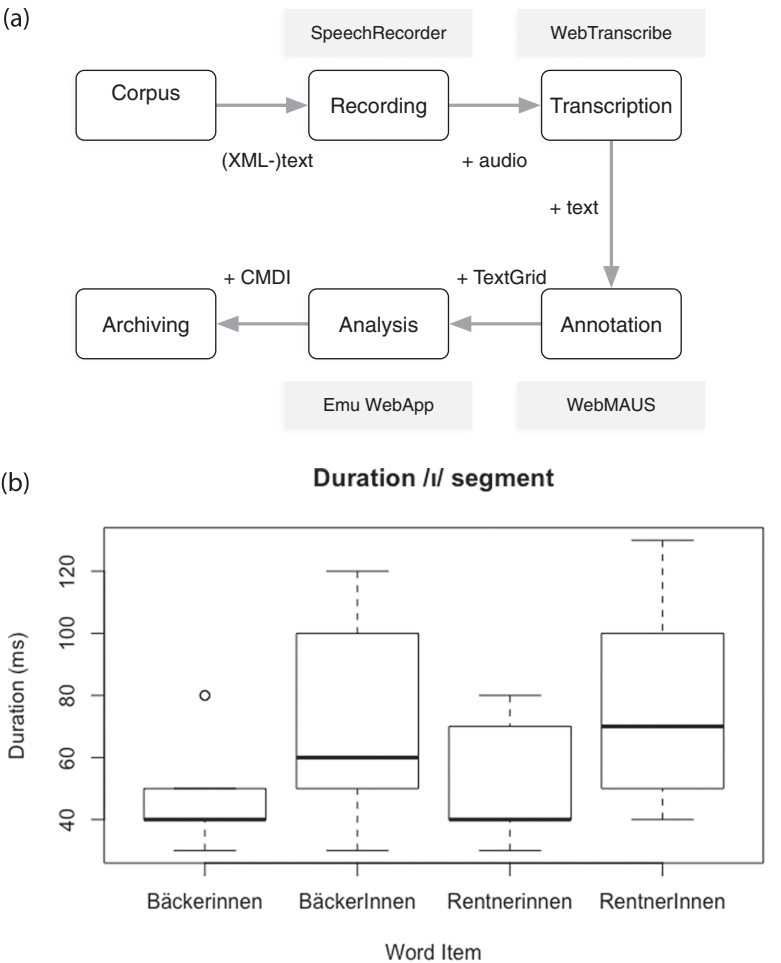

Figure 2: (a) Block diagram of the workflow with CLARIN tools (grey) and data types. (b) Duration of the final /I/ vowel in the gender-neutral and non-gendered forms of *Bäckerinnen* and *Rentnerinnen*.

Within a semester, the students were thus able to record, annotate, analyse, and publish a speech database, and to present their findings at an international pho-

netics conference. Since then, similar student projects, where all aspects of data collection, curation, and analysis are performed, have run every year, with topics as varied as the analysis of what makes a voice agreeable or interviews with immigrants on their life in Germany.

2.2 Application of new technologies

Research infrastructures need to constantly monitor the emergence of new research paradigms, research methods, innovative technologies, and new types of research data, in order to be able to serve the research needs of their community of interest well. Such responsiveness among research infrastructures is crucial for junior researchers and for more senior researchers who have progressed further in their careers. Doctoral and postdoctoral researchers are often major contributors to paradigm shifts and benefit directly from research infrastructures that offer novel research data and tools that directly serve their research goals. More advanced researchers can also benefit from such research data and tools – not only for their own research, but as extremely valuable resources for their teaching duties.

With the exponential growth in the availability of digital data (see Section 1.1), many scientific disciplines have experienced an empirical turn in their research paradigms and methods. Consequently, machine learning and other data-driven techniques now play a major role not only in computer science and in fields such as computational linguistics, but in a broader range of disciplines, including the (digital) humanities, which are based on data exploration and data analysis. In recent years, neural methods of machine learning have become particularly influential. These methods rely heavily on the distributional profiles of words that can be induced from very large corpora and that can be embedded into high-dimensional vector spaces. The resulting representations are therefore commonly referred to as *word embeddings*.

2.2.1 Advancing interoperability and reusability of word embeddings

With the support of CLARIN-D and under the guidance of Daniël de Kok, researchers at various stages of their careers at Tübingen University have advanced the interoperability of data formats for word embeddings, integrated neural tools into the annotation tool WebLicht, and developed an evaluation environment for assessing the data quality provided by deep learning tools for NLP. The Finalfusion tool which allows the use of a common data format for different word embeddings is described in de

Kok et al. (2020). Since the literature on deep learning implies that the amount of data is growing fast, it is timely and significant to offer a common data format that supports the interoperability and reuse of these formats. Finalfusion offers a data format that subsumes embeddings with character n-grams, quantized embedding storage, and memory mapping. Finalfusion also includes tools for training new embeddings, conversion tools (that map legacy formats into the final fusion format), and a code base for different programming languages, including Rust, Python, C, and C++. It is distributed with a set of new annotation tools and tool pipelines for Dutch and German, which are collectively referred to as the *sticker-2* tools. These tools provide high-quality annotations for both languages: lemmatization, part-of-speech tagging, and morphology at word level, and syntactic dependencies at sentence level. These tools can be used from within virtual research environments (VRE).

A Virtual Research Environment that integrates web services for processing language is provided by CLARIN-D. The Web-Based Linguistic Chaining Tool (WebLicht, M. Hinrichs, Zastrow, and E. Hinrichs 2010) provides a number of different tools for various languages for automatically annotating and analysing texts. WebLicht is productively used in academic education and research.

2.2.2 Enhancing virtual research environments

With the technical options of virtual research environments, tool suites for natural language processing (NLP) can be made available via web interfaces in a Service Oriented Architecture (SOA). These technologies enable scholars from various disciplines to utilize such tools for their own research without having to install them on their own computers or without requiring prior knowledge in programming. With WebLicht (M. Hinrichs, Zastrow, and E. Hinrichs 2010; Dima et al. 2012) such a research environment has been developed in CLARIN-D and has been widely used by humanities scholars in Germany and other CLARIN countries. WebLicht helps users to automatically annotate their reseach data. For this purpose, WebLicht provides a user with a selection of available NLP tools appropriate for a given language and a specific annotation task. Novice users can apply predefined tool chains, while experienced users can customize their own annotation workflows and select from a suite of available tools.

The WebLicht architecture has been designed with an open and scalable system architecture that allows for easy integration of additional annotation tools, as they become available. Given the fast-moving developments in deep learning and the improvements to be gained in annotation quality, researchers in CLARIN-D started to investigate how such neural annotation tools could be made available in WebLicht. In a disciplinary working group of CLARIN-D, they discussed

options and developed a neural part-of-speech tagger, to increase the performance of existing taggers. The result was sticker2, which is a sequence labeller. Trained for German and Dutch and capable of outperforming state of the art HMM taggers, sticker2 is a production ready multi-task sequence-labeller, lemmatizer, and dependency parser (de Kok, Falk, and Pütz 2020) which is itself used for further research (de Kok and Pütz 2020).

Another recent enhancement of the CLARIN infrastructure is offered by the virtual language environment *Language Resource Switchboard* (Zinn 2018). This tool suggests suitable tools available for a given dataset that a user wants to reuse for their own research. From these suggestions, the user can start the process directly with the data they have provided, including WebLicht, but also other tools such as Voyant (Sinclair and Rockwell 2016). With such a low, data-based entry threshold to the virtual language environments, the infrastructures provide easy access for all users, independently of their technical background.

Figure 3 illustrates the result of uploading an English-language PDF file to the Language Resource Switchboard. For this example, we uploaded an earlier version of this article as a PDF into the Switchboard. By dropping the file onto the

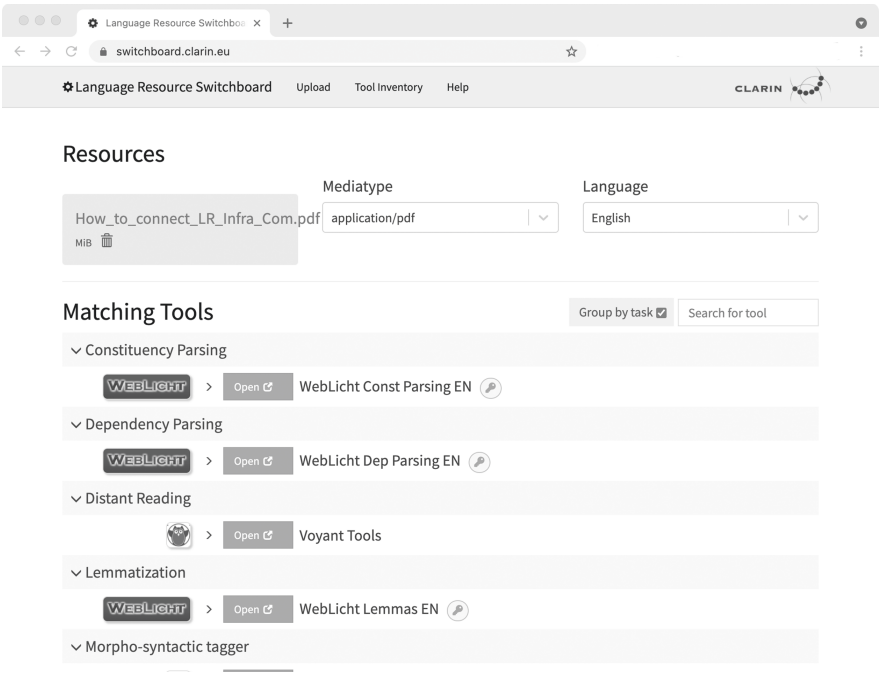

Figure 3: Result of uploading an English-language PDF file to the Language Resource Switchboard.

web page of the Switchboard, the Switchboard identifies the media type (here: PDF), and the language. Both can be adjusted manually if needed. The Switchboard then shows applicable tools: the first tools shown are various parsers and a tool for distant reading. Other tools are also presented but not shown due to the size of the browser window. By clicking on the respective "Open" button, the user directly invokes the tool. A dedicated section on the Switchboard is included in Zinn and Dima (2022).

2.3 New, faster, more precise answers to established research questions

Independent of benefits for the infrastructure and for new research, another aspect of this cooperation is in working with established research questions. These are typical questions that are used in teaching but also occur in other research processes. One example is the variation in translations, which is explored in translation studies. With detailed analysis of translated works researchers are able – for example, with pen and paper – to explore this variation and prove hypotheses. Assisted by data and tools from within research infrastructures, this process can be sped up considerably. Another example presented here is access to lexicographic information in dictionaries. Scholars can access dictionaries on their shelves or, more recently, via affiliated websites, but with the help of research infrastructures they can access lexicographic information from multiple sources in parallel. Again, the same information may be gathered by other means, but the process is accelerated considerably if the desired resources are accessible.

2.3.1 Exploring variation in translation

Translation studies have a long-standing tradition in the humanities, often resulting in collections of texts and translations. At the heart of the language and text-based disciplines are *corpora* and *comparative methods*. Adequate technology must thus offer support for comparing texts and languages from the socio-cultural and cognitive perspectives. There are two immediate implications for tools supporting the comparison of text and language data. First, tools should help users explore corpora with regard to relevant variables in order to find linguistic features in which variation becomes manifest. For example, if we observe that sermons in the 17th century tend to use a lot of 1st person plural pronouns, is that a *distinctive* feature of sermons in that time period? Second, tools should enable

users to extract linguistic features from corpora that then undergo quantitative and qualitative analysis. For example, is the use of passive constructions in academic text a *significant* feature? What are the *contexts of use* of the passive in academic text?

We illustrate here how we use a combination of two tools that are part of the CLARIN-D portfolio to show students how to find distinctive features by comparing two corpora and further exploring their usage context, both quantitatively and qualitatively. For the first step, we implemented a dedicated visualization tool that highlights distinctive words in two (or more) corpora under comparison (Fankhauser, Knappen, and Teich 2014). For the second step, a sophisticated concordance tool provides the means for quantitative and qualitative analysis (Evert and the CWB Development Team 2019).

This concrete example is taken from translation studies, where we are interested in the linguistic differences between (simultaneous) interpreting and translation (see example in Figure 4), but any question of intralingual variation can be approached in the same way. The underlying corpora are the EuroParl-UdS (translation) and the EPIC-UdS (interpreting) (Karakanta, Vela, and Teich 2018), both of which are available at the Saarbrücken CLARIN-D centre.[11]

The underlying models are uni-gram models. The word cloud not only encodes relative frequency (item colour) but also distinctivness of words (item size). The measure underlying distinctivness is relative entropy (here, Kullback-Leibler Divergence [KLD]). KLD measures the number of bits needed for encoding when the underlying model is non-optimal. In the example shown in Figure 4, we model interpreting based on translation and vice versa. The items with the greatest distinctiveness for interpreting are the hesitation markers 'euh' and 'hum' (which are also high frequency items) and the 1st person plural 'we'. These clearly mark online, spoken production. For translation, by contrast, the most distinctive items are 'this', 'we' and 'that' ('that' is the most frequent among the three but slightly less distinctive). Note that it is not surprising that the most distinctive items are grammatical words (pronouns, deictic elements), since grammatical use is a marker of mode and style. In contrast, lexical items (mostly in blue shades) are in lower frequency bands and are not very distinctive.

From the visual representation (shown in Figure 4) of corpora under comparison, we can enter the Corpus Query Processor (CQP; Evert and the CWB Development Team 2019), simply by clicking on a word. CQP runs as a web application at the Saarbrücken CLARIN-D centre and is accessible upon registration.[12] For

11 http://hdl.handle.net/21.11119/0000-0000-D5EE-4.
12 http://corpora.clarin-d.uni-saarland.de/cqpweb/.

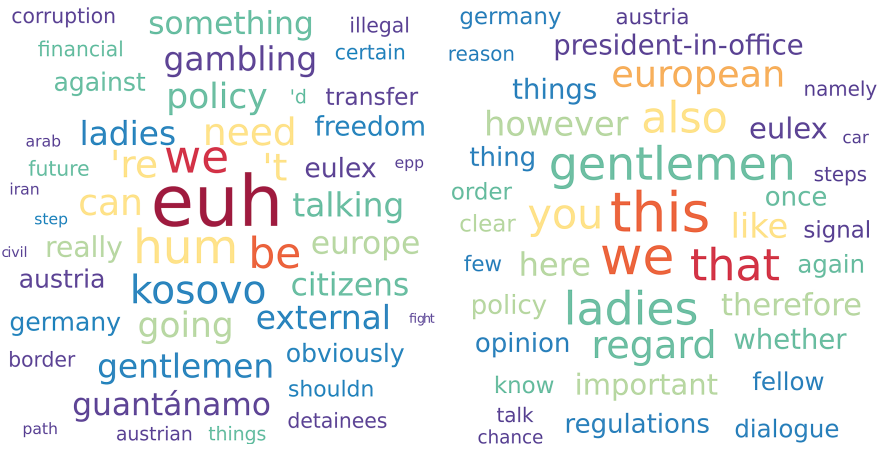

Figure 4: Variation in translation mode in target language English from source language German: interpreting (left), translation (right). Item colour: relative frequency (red=high, blue=low), item size: degree of distinctiveness, $p < 0.5$.

the given example, we are now interested in the context of the hesitation marker 'euh' as a highly distinctive item for simultaneous interpreting. Querying 'euh' in CQP provides detailed information on its surrounding context as well as number of occurrences and distribution in the corpus (see Figure 5). If the corpus is part-of-speech (POS) annotated, for better generalization we can inspect the context at POS level. Interestingly, for 'euh' we observe that it primarily occurs in the context of proper nouns as well as common nouns. This is an interesting descriptive result that provides a good basis for hypothesis building regarding the specific processing difficulties in simultaneous interpreting: nouns, and proper nouns in particular, are generally considered high entropy items and can therefore be expected to incur a high processing cost. This cost may be particularly high in interpreting. To test this, further analysis would be needed.

The kind of exploration and analysis shown in this example provides a typical agenda for a one-week course at a summer school, and we taught it many times at the European Summer University (ESU).[13] In our experience, students at all levels are extremely grateful to be offered an exploratory perspective on corpus comparison that can be easily combined with familiar tools such as concordances for more hypothesis-driven analysis. Exploration of potentially interesting and relevant features prior to qualitative and quantitative analysis lowers the initial

13 https://esu.culintec.de/, https://esu.fdhl.info/

Your query "[word="euh"]" returned 5,536 matches in 421 different texts (in 180,659 words [463 texts]; frequency: 30,643.37 instances per million words) [0.002 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Solution 1 to 50		Page 1 / 111
1	ORG_SP_EN_003	hands Firstly let us be clear this Parliament today will support the	**euh**	UN Security Council Resolution eighteen sixty and it should be implemented withc
2	ORG_SP_EN_003	in and to distribute aid and it is an issue of proportionality	**euh**	Save the Children say that the killing of a hundred and thirty
3	ORG_SP_EN_003	the killing of a hundred and thirty nine children since the conflict	**euh**	began euh and one thousand two hundred seventy one injured it can
4	ORG_SP_EN_003	of a hundred and thirty nine children since the conflict euh began	**euh**	and one thousand two hundred seventy one injured it can not be
5	ORG_SP_EN_004	the right to its own nationals and not allowed British nationals to	**euh**	to work that would be a violation of European Union law discrimination
6	ORG_SP_EN_004	of nationality as it would be if the company was undermining Eur	**euh**	British legal requirements as it is required to observe under the euh
7	ORG_SP_EN_004	euh British legal requirements as it is required to observe under the	**euh**	Posted Workers Directive if however the protesters are saying that only British
8	ORG_SP_EN_005	any other Member State of the European Union Thank you Madam President	**euh**	Commissioner I welcome you your predecessor Mister Mandelson was well known
9	ORG_SP_EN_005	Ireland for reasons which I 'm sure you 're very familiar with	**euh**	the issue of the Doha euh round is not being talked about
10	ORG_SP_EN_005	sure you 're very familiar with euh the issue of the Doha	**euh**	round is not being talked about amongst the people of Europe it
11	ORG_SP_EN_005	being a great thing yet in the financial sector it has not	**euh**	been thus in relation to agriculture that the other speakers just prior
12	ORG_SP_EN_005	I regard agriculture as rather important because it produces food and therefore	**euh**	higher up the the scale tha than he placed it euh and
13	ORG_SP_EN_005	therefore euh higher up the the scale tha than he placed it	**euh**	and I think we should remember that we voted in this House
14	ORG_SP_EN_005	it should be an issue that is discussed at the Doha level	**euh**	the issue of how European producers farmers can be competitive when in
15	ORG_SP_EN_005	than there is now so I would ask you perhaps in your	**euh**	c concluding comments if you could address some of those very real

Figure 5: CQP concordance for 'euh' in interpreting corpus EPIC-UdS.

threshold for coming up with an original topic for a BA, MA, or even PhD thesis, which may feasibly be carried out technically at the same time.

2.3.2 Access to lexicographic information: The German lexicographic-lexicological portals *OWID* and *ZDL*

The CLARIN infrastructure offers access to 95 dictionaries, most of them monolingual, others bi- or multilingual, accounting for 14 languages, German being one. In the vast majority of cases, the dictionaries can be directly downloaded from the national repositories or queried through an easy-to-use online search.[14] While dictionaries "were primarily created for human use (e.g., language learning/teaching, translation, lexicology) and are typically semasiological", the data collected in dictionaries is now used for the development of language tools and technology of all kinds, for example, speech recognition or word processing tools. Thus, CLARIN offers one of the oldest and most cherished ways of conveying the meaning and usage of words to scholars, researchers, and citizen-scientists from very different backgrounds, linking a large variety of dictionaries, exemplified here by language resources covering – to some extent – the German lexis: *Low German Loanwords in the Estonian Language*,[15] *Digital Dictionary of the German Language (DWDS)*,[16] *Rendering Dictionary of Personal Names*,[17] *Slovenian–German Dictionary of Maks Pleteršnik (1894–1895)*,[18] and others.

All online reference works can (theoretically) be updated continually. But those dictionaries that are officially completed also profit from their integration into lexicographic-lexicological portals, as users can easily find more and potentially more recent information on their search items from (a) other sources and (b) from corpus data.[19]As shown above, some German dictionaries (in the OWID and ZDL portals) are indeed "works in progress".

In this chapter, we describe cross-linking of different lexical resources in dictionary portals and how they may be connected to other data, such as corpora. We discuss the challenges of keeping information in online dictionaries (such as a dictionary of neologisms) up-to-date and we present some ideas on lexical resources as connections between (the academic discipline of) linguistics and

14 https://www.clarin.eu/resource-families/dictionaries
15 See http://www.eki.ee/dict/asl/.
16 See https://www.dwds.de/.
17 See https://www.letonika.lv/groups/default.aspx?g=2&r=1109.
18 See https://www.fran.si/136/maks-pletersnik-slovensko-nemski-slovar.
19 For one example in the Norwegian context, see Rauset et al. 2022.

the language community. As an example, the *Online-Wortschatz-Informationssystem Deutsch (OWID)*,[20] a dictionary portal developed at the Leibniz-Institute for the German Language (IDS), Mannheim,[21] one of the CLARIN-D centres,[22] is introduced. One of the dictionaries in OWID is the *Neologismenwörterbuch*. This dictionary is also one of the online resources presented in a dictionary portal of the *Zentrum für digitale Lexikographie der deutschen Sprache (ZDL)*,[23] containing information on the German lexicon from its beginnings to the present day, hosted at the Berlin-Brandenburg Academy of Sciences and Humanities, another of the CLARIN-D centres.

The OWID dictionary portal offers (as of March 2021) access to 10 different lexicographic resources comprising, for example, a paronym dictionary[24] documenting easily confusable expressions in their current public usage, a dictionary on German proverbs and slogans, the revised edition of *Deutsches Fremdwörterbuch*[25] explaining the origin and meaning of today's learned everyday language, the *Neologismenwörterbuch* and others. OWID contains retro-digitized online dictionaries as well as dictionaries that were developed directly for online publication. Besides completed dictionaries, there are some that are constantly worked on and are published dynamically (e.g., the *Paronymwörterbuch*), and there are diachronic (e.g., *Deutsches Fremdwörterbuch*) as well as synchronic dictionaries (e.g., *Neologismenwörterbuch*). All dictionary content can be accessed by search functions on two levels: the level of the portal and the level of an individual dictionary, thus addressing two different user needs (searching for one word in any dictionary, cf. Figure 6, or restricting the search to one specific dictionary).

In addition, appropriate advanced searches for each dictionary in the portal are developed using diverse technologies. All dictionaries in OWID are based on extensive empirical, mostly corpus-derived, linguistic data and are products of scholarly lexicography resulting from lexicological-lexicographic and metalexicographic research. They are not only innovative in choosing specific parts of German vocabulary as dictionary matter, but also in developing new types of lexicographic information by consistently linking between lexicographic information and corpus data, and in presenting information to users in new ways that have been adapted to each dictionary type. Although most of them focus on specific areas of vocabulary and not the general language, exploring them in the OWID

20 See https://www.owid.de/.
21 See https://www1.ids-mannheim.de/.
22 See https://www.clarin-d.net/de/aufbereiten/clarin-zentrum-finden.
23 See https://www.zdl.org/.
24 See https://www.owid.de/parowb/.
25 See https://www.owid.de/wb/dfwb/start.html.

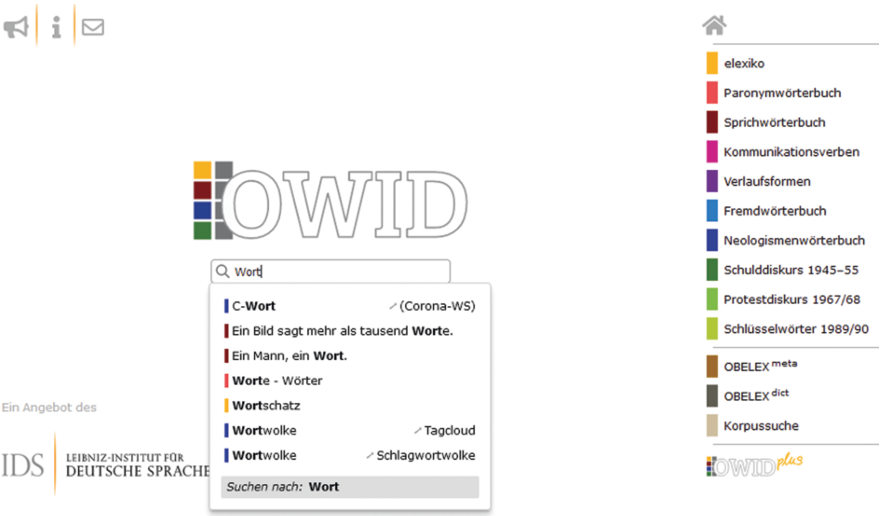

Figure 6: Search for *Wort* in OWID with results from five different dictionaries.

portal offers end users fascinating insights into the German vocabulary. In addition, the experimental platform *OWIDplus*[26] was established at IDS, containing a variety of lexicological-lexicographical data in mono- and multilingual interactive applications, for example, a *Lexical Explorer*[27] for corpus data on spoken German, browsable log file statistics of six Wiktionary language editions,[28] or the *cOWIDplus Viewer*[29] in which frequency curves of the use of word forms during the Covid-19 pandemic in 13 German online media are visualized. As of 2021, work is being done on a common faceted search option that will connect the resources in OWID and OWIDplus. In addition, OWID offers an easy-to-use corpus query interface with *DeReKo – Deutsches Referenzkorpus* of IDS.[30]

The dictionary portal of ZDL gives access to six dictionaries: the first and second edition of the diachronic general language dictionary *Deutsches Wörterbuch*.[31], the diachronic general language dictionary of Swiss German *Schweiz-*

26 See https://www.owid.de/plus/index.html.
27 See https://www.owid.de/lexex/.
28 See https://www.owid.de/plus/wikivi2015/index.html.
29 See https://www.owid.de/plus/cowidplusviewer2020/.
30 See https://www1.ids-mannheim.de/kl/projekte/korpora.html.
31 See information on https://www.dwds.de/d/wb-1dwb and https://www.dwds.de/d/wb-2dwb. *Deutsches Wörterbuch* in both editions was retro-digitized in collaboration with the Trier Centre for Digital Humanities and the Göttingen Academy of Sciences and Humanities. The Trier Centre for Digital Humanities is part of the CLARIAH-DE initiative, where CLARIN-D and DARIAH-DE are

erisches Idiotikon,[32] the new diachronic dictionary focused on central lexemes of politics and society *Wortgeschichte digital*,[33] the synchronic general language dictionary *Digitales Wörterbuch der deutschen Sprache (DWDS)*,[34] and the synchronic *Neologismenwörterbuch* of IDS. *Schweizerisches Idiotikon*, *DWDS*, *Wortgeschichte digital* and *Neologismenwörterbuch* are continually updated, while work on the first as well as the second edition of *Deutsches Wörterbuch* is now completed. Any search in ZDL generates a search result page where extracts from each dictionary containing the lemma are shown (cf. Figure 7). When clicking on the links "Vollständigen Artikel im . . . lesen" ('Read full entry in . . .') or "Detailansicht . . ." ('Detailed view of . . .'), users leave the ZDL portal and access the lexicographic or lexicological content of separate web pages.

In addition, users are shown a word frequency curve created from corpus queries in *Deutsches Textarchiv*[35] and the DWDS corpora[36] and a word cloud with typical collocates of the lemma generated from the DWDS corpora, thus cross-linking content from dictionaries and corpora successfully. ZDL also offers access to DeReKo – Deutsches Referenzkorpus at IDS as well as the diachronic language tool DiaCollo,[37] where information on the diachronic development of collocational behaviour can be obtained. Overall, both portals presented here facilitate the search for information on meaning and usage of words and phrases, as they offer easy access to different sources (dictionaries, lexicological interactive applications, visualizations of corpus data and corpora).

Dictionaries and lexicographic-lexicological portals address primarily human users. They serve as a link between research on words and its documentation and speakers of natural language. Data in dictionaries or lexicological information systems is based on corpus evidence and utilizes what corpus linguistics and language technologies have to offer. Users contribute to the compilation of language resources as well, either directly (e.g., by filling out feedback forms, such as the form to suggest a new word to the editors in the *Neologismenwörterbuch*[38]) or indirectly (e.g., when dictionaries use log-file analysis to find out which words are looked up most often; see de Schryver, Wolfer, and Lew 2019 and Wolfer et al. 2014).

---

combined in one network for research infrastructure: see https://dig-hum.de/forschung/projekt/clariah-de.

32 See https://www.idiotikon.ch/.

33 See information on https://adw-goe.de/forschung/weitere-forschungsprojekte/wortgeschichte-digital-teilprojekt-im-zdl/.

34 See https://www.dwds.de/.

35 See https://www.deutschestextarchiv.de/.

36 See https://www.dwds.de/r.

37 See https://clarin-d.net/de/kollokationsanalyse-in-diachroner-perspektive.

38 See https://www.owid.de/wb/neo/mail.html.

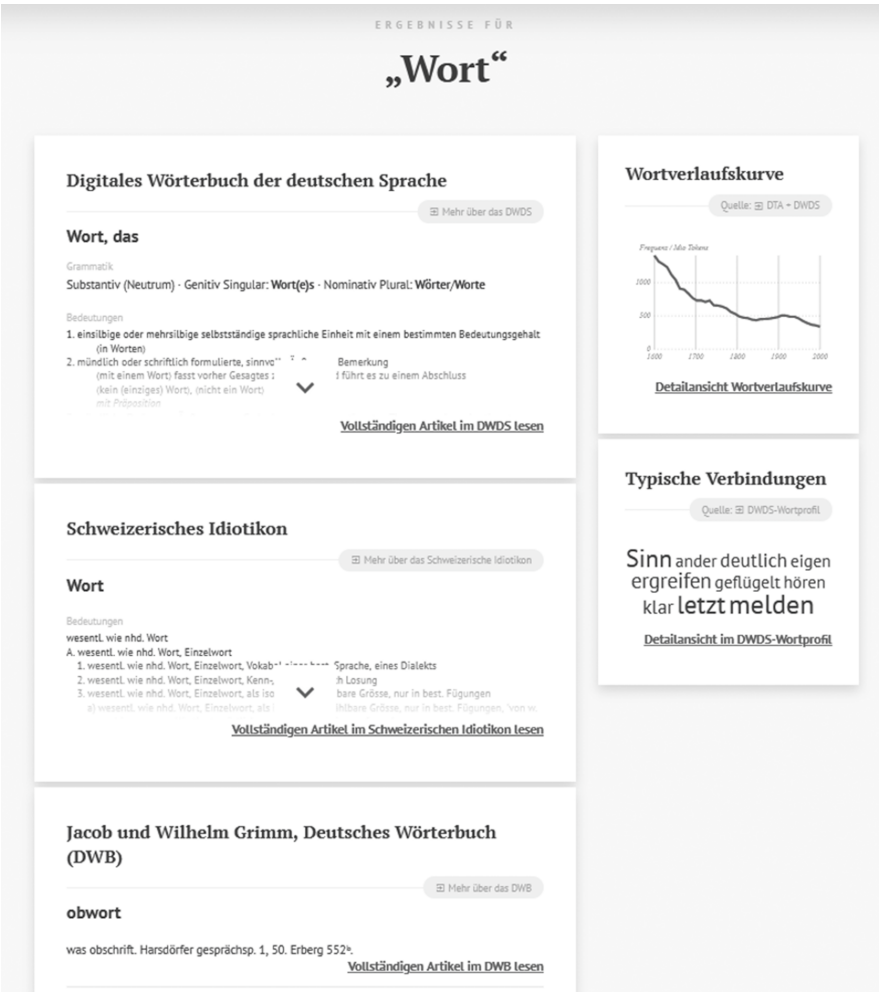

Figure 7: Result page of search for *Wort* in ZDL with results in three dictionaries and corpus-based additional information.

Lexicography and lexicological research are a perfect example for illustrating manifold connections: between different dictionaries and other lexical sources in portals, using infrastructure such as provided in the CLARIN framework; between lexicographers or lexicological researchers on one side and corpus linguists and language technology on the other, such as found in the CLARIN network; and finally between linguistic research (in its widest sense) and the language community.

2.3.3 The German Text Archive: An active archive for historical data in CLARIN

The *German Text Archive* (DTA), located at the Centre for Language at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), was funded by the German Research Foundation (DFG) from 2007 to 2016 and now forms an essential component of the research data infrastructure of the German part of CLARIN. In this section, the DTA is presented as a web-based research platform for the creation and curation of corpus texts as well as for corpus analysis.

The aim of the DTA has been to create a basic stock of German-language texts spanning all disciplines and genres for the period ca. 1600–1900. The text selection was based on an extensive bibliography, annotated and supplemented by members of the BBAW Academy. From this, the DTA project group compiled a text corpus balanced according to text types and disciplines, which serves as the basis for a reference corpus on the development of New High German. In order to reflect the historical state of the language as accurately as possible, the first editions of the works were generally used as a basis for digitization. The DTA core corpus compiled according to these criteria is continuously being expanded. It currently comprises about 1,500 works with a volume of about 120 million words. In addition, there are another nearly 4,000 works (about 100 million tokens) that have been curated together with external projects for the DTA platform (as of April 2021); most of them via the DTAQ quality assurance platform (see below).

The basis for the DTA is a structured format that was developed from the multitude of different texts it contains in order to be usable for as many contexts as possible. This so-called DTA Base Format (DTABf), in addition to serving as an interchange format for different corpora, ensures interoperability for use cases as diverse as corpus display, full-text search, and text mining. The DTABf is a true subset of the TEI's text document encoding guidelines: the TEI's tagset has been reduced in terms of available elements and attributes and specified in terms of attribute values (Haaf, Geyken, and Wiegand 2015; Geyken, Haaf, and Wiegand 2012). The DTABf annotation scheme for historical prints (and other document classes such as newspapers and manuscripts, cf. Haaf and Thomas 2015), together with extensive documentation and a Schematron rule set, forms the basis for XML markup of all works in the DTA. With the help of conversion tools, numerous other formats can be automatically generated from DTABf documents for further processing with linguistic tools, for search engine indexes, for presentation of the texts (e.g., reading versions for various media), and for export (e.g., to citation

environments, graph databases, or in the CLARIN context for WebLicht). The further development of the DTABf guidelines[39] is ensured by a steering group.[40]

For the quality assurance of the full texts and the structural data, a web-based platform was developed (DTA Quality Assurance, DTAQ[41]), which allows the distributed proofreading and correction of texts. For this purpose, flexible options for text import from different formats and a text-image view were created, and an editor was integrated into the platform, with which texts can be edited without the need for additional software to be installed. At the end of the correction process, the work is published on the DTA website, where it is accessible via a text-image view and linked to various analysis tools (see below). DTAQ includes a user management system that provides multiple levels of access and annotation options for different user groups through roles and permissions. Users of DTAQ register with a personalized account on the platform and can specify various types of expertise (expertise in literary or linguistic history, knowledge of foreign languages, expertise in transcribing mathematical formulas, etc.). This makes it possible to specifically address other users with the help of the ticket system when in doubt or when using difficult text passages, and thus to work collaboratively on the documents. In addition, this makes it easy to work in a team, as certain types of errors can be specifically assigned to individual users. Personalization also makes it possible to save the user's own preferences with regard to the DTAQ display for each account, including the optimal text and image width or the preferred text view, among others. As of June 2021, more than 2,000 users have been active on DTAQ; some have commented on text errors and others have curated entire works via the platform.

Another key element of DTA is its collection of analysis tools. CAB (Cascaded Analysis Broker; cf. Jurish 2012), a tool for normalizing historical spellings, provides a spelling-tolerant full-text search across all texts in the DTA. In addition, with the integration of GermaNet (Hamp and Feldweg 1997; Henrich and E. Hinrichs 2010), a lexical resource that groups nouns, verbs, and adjectives into SynSets according to similarity of meaning, full-text search by semantic categories is also made possible. Furthermore, a number of lexicometric analysis tools are available, including the visualization of diachronic collocations (Jurish and Nieländer 2019), and a quantitative text analysis based on the Voyant tools.[42]

39 See https://www.deutschestextarchiv.de/doku/basisformat/leitlinien.html.
40 See https://www.deutschestextarchiv.de/doku/basisformat/steuerungsgruppe.html.
41 See https://www.deutschestextarchiv.de/dtaq/.
42 See https://voyant-tools.org/.

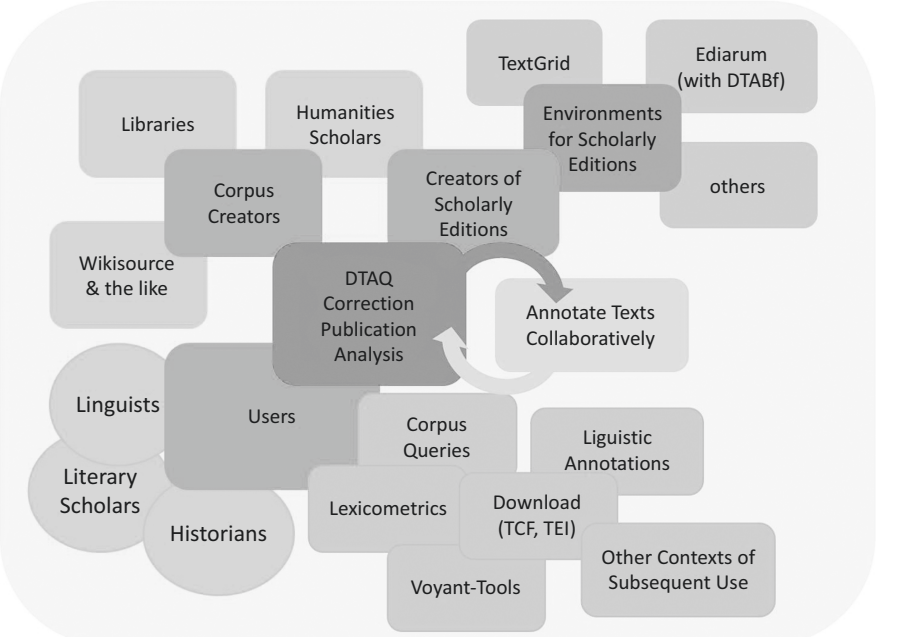

Figure 8: DTA as a research, publication, and analysis platform.

All texts of the DTA are under an open Creative Commons (CC)[43] license and can thus be easily reused as a complete set in scientific contexts.[44] Furthermore, due to the interoperability ensured by the encoding in DTABf, all texts of the DTA can be easily converted into different formats.

Figure 8 summarizes the various components, at the centre of which is DTAQ as a proofreading, publication, and analysis platform. On one side are the various corpus producers (humanities and social scientists, libraries, and non-academic initiatives such as Wikisource); on the other side are edition environments and producers of editions. The "classic" use of DTAQ consists of collaborative annotation of texts. All DTA texts can be corrected and annotated at any time, and the continually updated version can be exported from the platform. The fourth and final component is analysis, with the aforementioned CAB and GermaNet tools for linguistic annotation, the various analysis tools, and the export formats for flexible reuse in other contexts.

43 See https://creativecommons.org/.
44 See https://deutschestextarchiv.de/download.

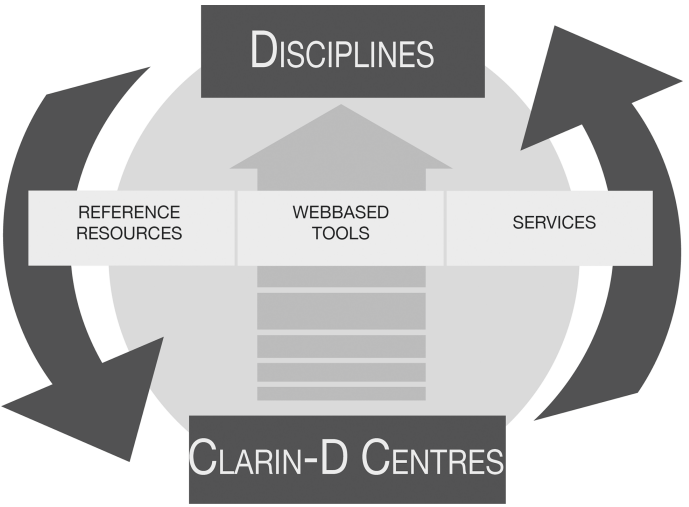

Figure 9: Disciplinary cooperation of users and infrastructure providers.

3 Instruments for supporting sustainable user involvement

In the development of the CLARIN infrastructure, we see large benefits on both sides from a strong cooperation between users of a research infrastructure and infrastructure providers, as illustrated by Figure 9. We have shown how the developments have already cross-fertilized in the past and have led to a significant improvement on the research side and to an enhancement of the offerings. In addition to the fact that the offerings were initially developed very much in line with the research background of those who also provided the offerings to others, further measures were established in CLARIN-D to ensure collaboration. In this section, we present some of the measures that we have taken to foster this collaboration, namely:

- discipline-specific working groups
- curation projects
- tools for collaboration on resources
- training[45] and consulting activities.

45 Hennelly et al. 2022 describe the motivation and processes for training in South Africa.

Discipline-specific working groups were created to integrate disciplinary needs in the infrastructure and to encourage feedback to service providers. For scholars using the infrastructure, the working groups also established a channel to spread information about the availability and use of the infrastructure. Chaired by distinguished scholars in the field, around 200 researchers from Germany, with varying backgrounds in the humanities and social sciences, met in eight discipline-specific working groups, supported by travel grants and with administrative support from CLARIN-D. These working groups included disciplines such as German philology, other philologies, linguistic fieldwork, anthropology, language typology, human speech processing (including psycho-linguistics, speech technology and other modalities), applied linguistics and computational linguistics, content analysis in social sciences, and history. The groups met on a regular basis, reviewing the infrastructure, services, and available datasets. They also devised application scenarios, projects, and uses of the infrastructure and presented at academic conferences and workshops. In the process of applying scenarios, they detected usability issues, gaps in the infrastructure, and valuable add-ons to the infrastructure. The discipline-specific working groups also establish a bridge to professional associations. With their publications, conferences, and workshops, these associations provide another point of contact between infrastructure providers and the research community.

Curation projects in CLARIN-D are measures within the infrastructure to help close the detected gaps and integrate valuable add-ons. Supported by the infrastructure, the discipline-specific working groups decided on priorities, such as the preparation and depositing of legacy data, or the development of new tools. For this, each discipline-specific working group received a budget and an infrastructural partner with which to work on curating data resources or tools.

The activities of the discipline-specific working groups, curation projects, and technical tools for collaboration are complemented by established outreach activities, including workshops and tutorials, summer school courses, consulting services, and a helpdesk. Each of these activities disseminates the infrastructure's resources and provides a low access threshold for scholars at all stages of their academic career.

One example for supporting training activities is the European Summer University in Leipzig, Germany. This established summer school is used by CLARIN-D to disseminate tools, services, and other resources by training individuals to use them. The classes are based on the requirements and feedback of participants. For example, users pointed to the need for training on low-level query methods for CLARIN data, the application of tools and services for specific research questions, applying and evaluating NLP technologies in the humanities, and analysing language data for humanities scholars. Together with other classes

on data management, legal and ethical questions, metadata modelling, and so on, CLARIN offered a wide spectrum of infrastructure-related training to young researchers.

4 Conclusion

In this chapter we have illustrated that the integration of language resources infrastructures and communities is beneficial both for the communities and for the services provided by the infrastructures. The German national project CLARIN-D established strong bonds with the research community through discipline-specific working groups, curation projects that were prioritized by the discipline-specific working groups, training, and dissemination activities. With this strong connection between the community and the infrastructure, researchers achieved results when addressing emerging research questions, confirmed research hypotheses faster and with more precision, and developed new methods, contributing to new research paradigms. The cooperation between users and infrastructure providers thus contributed to the success story of CLARIN in Germany.

Bibliography

Baum, Constanze & Thomas Stäcker. 2015. Methoden – Theorien – Projekte. In Constanze Baum & Thomas Stäcker (eds.), *Grenzen und Möglichkeiten der Digital Humanities: Sonderband der Zeitschrift für digitale Geisteswissenschaften*, 4–12. https://doi.org/DOI10.17175/sb001_023.

Berry, David M. 2011. The computational turn: Thinking about the digital humanities. *Cultural Machine* 12. 1–22.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10). 341–345.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell & Shelley Stall. 2021. Operationalizing the CARE and FAIR Principles for indigenous data futures. *Scientific Data* 8(1). 108. https://doi.org/10.1038/s41597-021-00892-0.

Dima, Emanuel, Erhard Hinrichs, Marie Hinrichs, Alexander Kislev, Thorsten Trippel & Thomas Zastrow. 2012. Integration of WebLicht into the CLARIN infrastructure. In *Service-oriented architectures (soas) for the humanities: Solutions and impacts joint clarin-d/dariah workshop at digital humanities conference 2012*, 17–23. http://clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf.

Draxler, Christoph & Klaus Jänsch. 2004. SpeechRecorder – a universal platform independent multi-channel audio recording software. In *International Conference on Language*

Resources and Evaluation (LREC) 4, 559–562. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/summaries/242.htm.

Engelberg, Stefan, Annette Klosa-Kückelhaus & Carolin Müller-Spitzer. 2020. Internet lexicography at the Leibniz-institute for the German language. *K Lexical News* 28. 54–77. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-99953.

Engelberg, Stefan & Carolin Müller-Spitzer. 2013. Dictionary portals. In Rufus Hjalmar Gouws, Ulrich Heid, Wolfgang Schweickard & Herbert Ernst Wiegand (eds.), *Dictionaries. An international encyclopedia of lexicography: Supplementary volume: Recent developments with focus on electronic and computational lexicography*, 1023–1035. Berlin: De Gruyter Mouton. https://doi.org/doi:10.1515/9783110238136.

European Strategy Forum on Research Infrastructures (ESFRI). 2018. *Strategy report on research infrastructures: Roadmap 2018*. Report. http://roadmap2018.esfri.eu/media/1060/esfri-roadmap-2018.pdf.

Evert, Stefan & the CWB Development Team. 2019. The IMS Open Corpus Work Bench (cwb), CQP query language tutorial (CWB version 3.4.16). http://cwb.sourceforge.net/files/CQP_Tutorial.pdf.

Fankhauser, Peter, Jörg Knappen & Elke Teich. 2014. Exploring and visualizing variation in language resources. In *International Conference on Language Resources and Evaluation (LREC) 9*, 4125–4128. Reykjavik: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/185_Paper.pdf.

Fridlund, Mats, Daniel Brodén, Tommi Jauhiainen, Leena Malkki, Leif-Jöran Olsson & Lars Borin. 2022. Trawling and trolling for terrorists in the digital Gulf of Bothnia: Cross-lingual text mining for the emergence of terrorism in Swedish and Finnish newspapers, 1780—1926. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*, 1780–1926. Berlin: De Gruyter.

Geyken, Alexander, Susanne Haaf & Frank Wiegand. 2012. The DTA 'base format'. A TEI-subset for the compilation of interoperable corpora. In Jeremy Jancsary (ed.), *11th conference on natural language processing (KONVENS), Ithist 2012 workshop* (Scientific Series of the ÖGAI 4), 383–391. Vienna: Österreichische Gesellschaft für Artificial Intelligence.

Gomes, Luís, Ruben Branco, João Silva & António Branco. 2022. Open and inclusive language processing: Language processing services by PORTULAN to meet the widest needs of CLARIN users. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Haaf, Susanne, Alexander Geyken & Frank Wiegand. 2015. The DTA " Base Format": A TEI subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative* 8. https://doi.org/10.4000/jtei.1114.

Haaf, Susanne & Christian Thomas. 2015. Enabling the encoding of manuscripts within the DTABf: Extension and modularization of the format. *Journal of the Text Encoding Initiative* 10. https://doi.org/10.4000/jtei.1650.

Hajič, Jan, Eva Hajičová, Barbora Hladká, Jozef Mišutka, Ondřej Košarko & Pavel Straňák. 2022. LINDAT/CLARIAH-CZ: Where we are and where we go. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Hamp, Birgit & Helmut Feldweg. 1997. GermaNet – A lexical-semantic net for German. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo & Yorick Wilks (eds.), *Proceedings of the ACL workshop automatic information extraction and building of lexical semantic resources for NLP applications*, 9–15. Somerset, NJ: Association for Computational Linguistics.

Hennelly, Martin, Langa Khumalo, Juan Steyn & Menno van Zaanen. 2022. Training of digital language resources skills in South Africa. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Henrich, Verena & Erhard Hinrichs. 2010. GernEdiT – the GermaNet editing tool. In *International conference on language resources and evaluation (LREC) 7*, 2228–2235. Valletta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf.

Hinrichs, Marie, Thomas Zastrow & Erhard Hinrichs. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In N. Calzolari (ed.), *International Conference on Language Resources and Evaluation (LREC) 7*, 489–493.

Hoeksema, Jack, Kees de Glopper & Gertjan van Noord. 2022. Syntactic profiles in secondary school writing using PaQu and SPOD. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Jurish, Bryan. 2012. *Finite-state canonicalization techniques for historical German*. (completed 2011, published 2012). Universität Potsdam dissertation. http://opus.kobv.de/ubp/volltexte/2012/5578/.

Jurish, Bryan & Maret Nieländer. 2019. Using DiaCollo for historical research. In *CLARIN annual conference 2019 (Leipzig, Germany, 30 September – 2 October, 2019)*. https://www.clarin.eu/clarin-annual-conference-2019-abstracts#L.

Karakanta, Alina, Mihaela Vela & Elke Teich. 2018. Europarl-UdS: Preserving metadata from parliamentary debates. In Darja Fišer, Maria Eskevich & Franciska de Jong (eds.), *ParlaCLARIN@LREC2018, at International Conference on Language Resources and Evaluation (LREC) 11*. Miyazaki: European Language Resources Association (ELRA). https://www.clarin.eu/sites/default/files/ParlaCLARIN_Session2_2.2.EuroParl-UdS_Alina-Karakanta_LREC2018.pdf.

Kisler, Thomas, Florian Schiel & Han Sloetjes. 2012. Signal processing via web services: The use case WebMAUS. In Erhard Hinrichs, Heike Neuroth & Peter Wittenburg (eds.), *Workshop on service-oriented architectures (SOAs) for the huamnities: solutions and impacts at digital humanities 2012*, 30–34. Hamburg: Universität Hamburg. https://www.mpi.nl/publications/item_1850150.

Kok, Daniël de, Neele Falk & Tobias Pütz. 2020. Sticker2: A neural syntax annotator for Dutch and German. In Constanza Navarretta & Maria Eskevich (eds.), *Proceedings of the CLARIN annual conference 2020*, 27–31. https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.

Kok, Daniël de, Sebastian Pütz, Eric Schill & Erhard Hinrichs. 2020. Finalfusion: Fusing all your embeddings into one format. In *Knowledge, language, models: Volume in honor of Prof. Galia Angelova*, 57–73. Shoumen: INCOMA Ltd.

Kok, Daniël de & Tobias Pütz. 2020. Self-distillation for German and Dutch dependency parsing. *Computational Linguistics in the Netherlands Journal* 10. 91–107. https://www.clinjournal.org/clinj/article/view/106.

Kučera, Dalibor. 2022. Application of CLARIN linguistic tools in psychological research. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Lindahl, Anna & Stian Rødven-Eide. 2022. Argumentative language resources at Språkbanken text. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Petrauskaitė, Rūta, Darius Amilevičius, Virginijus Dadurkevičius, Tomas Krilavičius, Gailius Raškinis, Andrius Utka & Jurgita Vaičenoniene˙. 2022. CLARIN-LT: Home for Lithuanian language resources. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Pettersson, Eva & Lars Borin. 2022. Swedish Diachronic Corpus. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Pozzo, Riccardo, Timon Gatta, Hansmichael Hohenegger, Jonas Kuhn, Axel Pichler, Marco Turchi & Josef van Genabith. 2022. Aligning Immanuel Kant's work and its translations. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Presner, Todd. 2010. Digital humanities 2.0: A report on knowledge. In Melissa Bailar (ed.), *Emerging disciplines: Shaping new fields of scholarly inquiry in and beyond the humanities*. Online: OpenStax CNX. http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6.

Rauset, Margunn, Gyri Smørdal Losnegaard, Helge Dyvik, Paul Meurer, Rune Kyrkjebø & Koenraad De Smedt. 2022. Words, words! Resources and tools for lexicography at the CLARINO Bergen centre. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Schaal, Gary S. & Roxana Kath. 2014. Zeit für einen Paradigmenwechsel in der politischen Theorie? In André Brodocz, Dietrich Herrmann, Rainer Schmidt, Daniel Schulz & Julia Schulze Wessel (eds.), *Die Verfassung des Politischen: Festschrift für Hans Vorländer*, 331–350. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-04784-9_20.

Schryver, Gilles-Maurice de, Sascha Wolfer & Robert Lew. 2019. The relationship between dictionary look-up frequency and corpus frequency revisited: A log-file analysis of a decade of user interaction with a Swahili-English dictionary. *GEMA Online Journal of Language Studies* 19(4). 1–27. https://doi.org/10.17576/gema-2019-1904-01.

Silva, João, Sara Grilo, Márcia Bolrinha, Rodrigo Santos, Luís Gomes, António Branco & Rui Vaz. 2022. Where do I belong in six centuries of literature? Datasets and AI-based tools for Portuguese literary documents made possible and available by PORTULAN CLARIN. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Sinclair, Stéfan & Geoffrey Rockwell. 2016. *Voyant Tools*. http://voyant-tools.org/.

Slavik, Korbinian, Johanna Cronenberg & Christoph Draxler. 2018. A study on the pro-nunciation of gender-neutral nouns in German. In Malte Belz, Christine Mooshammer, Susanne Fuchs, Stefanie Jannedy, Oksana Rasskazova & Marzena Żygis (eds.), *Proceedings of the conference on phonetics & phonology in german-speaking countries (P&P 13)*, 185–188. Berlin: Leibniz-Zentrum Allgemeine Sprachwissenschaft & Humboldt-Universität.
https://doi.org/http://dx.doi.org/10.18452/18805.

Trognitz, Martina, Matej Ďurčo & Karlheinz Mörth. 2022. Text technology for the digital humanities: Maximizing impact in a diverse field of disciplines. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo,

Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific data* 3. https://doi.org/https://doi.org/10.1038/sdata.2016.18.

Winkelmann, Raphael & Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *International Conference on Language Resources and Evaluation (LREC) 9*, 4129–4133. Reykjavik: European Language Resources Association (ELRA).

Wolfer, Sascha, Alexander Koplenig, Peter Meyer & Carolin Müller-Spitzer. 2014. Dictionary users do look up frequent and socially relevant words. Two log file analyses. In Andrea Abel, Chiara Vettori & Natascia Ralli (eds.), *Proceedings of the XVI Euralex International Congress, Bolzano/Bozen, 15.-19.07.2014*, 281–290. Bozen: Institute for Specialised Communication & Multilingualism. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-31125.

Zinn, Claus. 2018. The language resource switchboard. *Computational Linguistics* 44(4). 631–639. https://doi.org/10.1162/coli_a_00329.

Zinn, Claus & Emanuel Dima. 2022. The CLARIN Language Resource Switchboard: Current state, impact, and future roadmap. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.