BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 4
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0040

Text+ – Concept and Benefits for Empirical Researchers

Erhard Hinrichs, Thorsten Trippel

University of Tübingen, Seminar für Sprachwissenschaft, Keplerstr. 2, D-72074 Tübingen, Germany
and Leibniz-Institut für Deutsche Sprache (IDS) R 5, 6-13 D-68161 Mannheim, Germany
E-mails: erhard.hinrichs@uni-tuebingen.de trippel@ids-mannheim.de

***Abstract**: In this contribution, we report on ongoing efforts in the German national research infrastructure consortium Text+ to make research data and services for text- and language-oriented disciplines FAIR, that is findable, accessible, interoperable, and reusable, as well as compliant with the CARE principles for language resources.*

***Keywords**: Distributed research data infrastructure Text+, Text+, Language- and text-based research data, German national research infrastructure NFDI, NFDI, Text as data, Text analytics.*

1. Introduction

In this contribution, we report on ongoing efforts in the German national research infrastructure consortium Text+ to make research data and services for text- and language-oriented disciplines FAIR [1], i.e., findable, accessible, interoperable, and reusable, as well as compliant with the CARE principles (see CARE Principles for Indigenous Data Governance, **https://www.gida-global.org/care**) for language resources. The paper was originally presented during CLaDA-BG 2023 Conference: Language Technologies and Digital Humanities: Resources and Applications (LTaDH-RA), Sofia, Bulgaria, 10-12 May 2023.

Text+ is a collaborative effort by more than thirty research institutions across Germany with a total of well over 100 researchers. While the co-authors of this paper have prepared the contents and take responsibility for any errors that may have occurred, the work reported here is carried out by a large group of colleagues in Text+ and predecessor projects CLARIN-D, DARIAH-DE, and CLARIAH-DE. Their contributions are hereby acknowledged with deep appreciation and collegial gratitude.

This paper is structured as follows: In Section 2, we will briefly describe the goals of the German National Research Data Infrastructure (NFDI), its current state, and the division of labor between the four NFDI consortia that focus on research data in the Humanities and Cultural Studies: NFDI4Culture, NFDI4Memory, NFDI4Objects, and Text+. Section 3 contains illustrative examples of the breadth and depth of the Text+ portfolio of research data and presents the rationale for implementing Text+ as a federated and geographically distributed research infrastructure. Section 4 introduces the federated content search and the Text+ registry, which provide easy access to the Text+ portfolio of research data. The suite of tools for data analytics that Text+ currently offers is presented in Section 5. The ongoing efforts by Text+ to make research data available that are protected in one form or another, e.g., that are still under copyright, are described in Section 6. The paper concludes with an outlook toward future work and a short summary (Section 7). Appendix A offers a list of online resources that have been mentioned and referenced in the present paper.

Given space limitations, we can only include a small subset of the current Text+ portfolio of research data, tools, and services. We refer all interested parties to the Text+ webpage (**https://www.text-plus.org**), where a comprehensive overview of Text+ offerings is available and is continuously updated.

2. Text+ as Part of the German National Research Infrastructure (NFDI)

Research data management and research data infrastructures have gained considerable momentum over the past twenty years, with European initiatives such as CLARIN [2, 3] and DARIAH [4] as part of the landscape that has been organized in the context of ESFRI [5].

On the national scale in Germany, the government – in our federal system this includes the federal government and the 16 individual states – established the National Research Data Infrastructure (NFDI) (*Nationale Forschungsgs-dateninfrastruktur* in German). This decision was taken in 2018, and funding began in 2020 in three consecutive funding rounds. The goal is to ensure easy access on a long-term basis to research data across a wide range of different scientific disciplines by implementing the FAIR and CARE principles and facilitating long-term archiving of research data.

Currently, the NFDI consists of 25 consortia, each with a disciplinary specialization; this structure is complemented by one consortium for the implementation of ‘basic services’ across consortia, and an NFDI directorate responsible for coordination, governance, and communication for the entire NFDI. The NFDI is – as a research-driven initiative – organized according to disciplinary needs, while also encouraging exchange across disciplines and sharing basic services, technological know-how, and common infrastructure solutions, whenever feasible.

The vision of the NFDI is to provide data as a common asset for excellent research, organized by researchers in Germany. To achieve this vision, NFDI has the mission to create an organization for research data management and a legally compliant, interoperable, and sustainable data infrastructure. In its efforts to achieve these goals, the NFDI collaborates with national and international partners.

For the areas of Social Sciences and Humanities, six consortia are being funded in the NFDI. Four of them come from the Humanities. These are NFDI4Objects, NFDI4Memory, NFDI4Culture and Text+, all cooperating closely where applicable [6].

NFDI4Memory [7, 8] brings together researchers, memory institutions and information infrastructure facilities in a digital research infrastructure for researchers in disciplines that focus on historical data.

NFDI4Objects [9] aims to meet the infrastructure needs of researchers and practitioners who want to contribute to the material heritage of around three million years of human and environmental history, including Archaeology.

NFDI4Culture [10] is the consortium that deals with research data on tangible and intangible cultural assets. It includes Art History, Musicology, Theater and Media Studies, and Architecture.

The Text+ consortium [11, 12] will establish a text- and language-based research data infrastructure with a focus on three data domains: collections of language data, lexical resources, and editions. These types of resources are highly relevant for all language- and text-based disciplines, including Linguistics, Literary Studies, Philosophy, Classical Philology, Anthropology, Non-European Cultures and Languages, as well as language- and text-based research in the Social, Economic, Political and Historical Sciences.

The consortium Text+ has been initiated by the Leibniz-Institute for the German Language in Mannheim (IDS) (*Leibniz-Institut für Deutsche Sprache* in German), the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW) (*Berlin-Brandenburgische Akademie der Wissenschaften* in German), the German National Library in Frankfurt (DNB) (*Deutsche Nationalbibliothek* in German), the North Rhine-Westphalian Academy of Sciences and Arts (NRWAW) (*Nordrhein-Westfälische Akademie der Wissenschaften und Künste* in German), and the Göttingen State and University Library (SUB) (*Niedersächsische Staats- und Universitätbibliothek Göttingen* in German). Additionally, 26 other institutions participate in Text+, covering a wide range of the academies in the arts and sciences, universities, libraries, research foundations and academic centers.

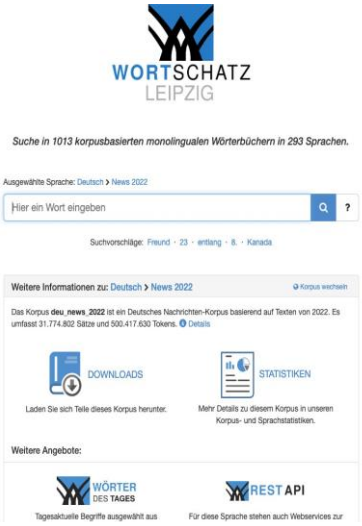
The mission of Text+ is to assist the text- and language-oriented Humanities and Social Sciences in the use of digital data and accompanying methods in their research and in developing a common data culture. In support of this mission and in support of promoting digital literacy, Text+ contributes to the FAIRness of text and language data, with the aim that access to digital resources becomes the common practice. Text+ is part of a diverse research community. It promotes interdisciplinary collaboration and scientific innovation through the integration of infrastructure and research. For this purpose, Text+ also closely observes the emergence of new research paradigms for the Humanities and Cultural Studies and promising developments in information and language technology.

3. Data in a distributed infrastructure for language resources

With more than 30 initial partners contributing their data and resources, Text+ has built on an extensive portfolio of legacy research data for language and text. This portfolio of language data is highly diverse along various dimensions: it contains data for written and spoken language, multimodal data; and data for contemporary language as well as diachronic data. In close consultation with the Text+ communities of interest, this initial portfolio is continuously extended.

3.1. Starting from existing datasets within Text+

While there is a strong focus on German data, Text+ also offers language data for a wide range of other languages. The Leipzig Wortschatz (Fig. 1) collection is a good example of a multi-lingual dataset. It consists of crawled web data for 293 languages. The Wortschatz collection also exemplifies the heterogeneous sources of research data in Text+; the Wortschatz data are automatically harvested web data and thus born digital data, while the majority of data sources in Text+ are curated collections of born-analogue data.



Suche in 1013 korpusbasierten monolingualen Wörterbüchern in 293 Sprachen.

Ausgewählte Sprache: Deutsch > News 2022

Hier ein Wort eingeben

Suchvorschläge: Freund · 23 · entlang · 8. · Kanada

Weitere Informationen zu: Deutsch > News 2022 Korpus wechseln

Das Korpus **deu_news_2022** ist ein Deutsches Nachrichten-Korpus basierend auf Texten von 2022. Es umfasst 31.774.802 Sätze und 500.417.630 Tokens. Details

DOWNLOADS

Laden Sie sich Teile dieses Korpus herunter.

STATISTIKEN

Mehr Details zu diesem Korpus in unseren Korpus- und Sprachstatistiken.

Weitere Angebote:

WÖRTER DES TAGES

Tagesaktuelle Begriffe ausgewählt aus

REST API

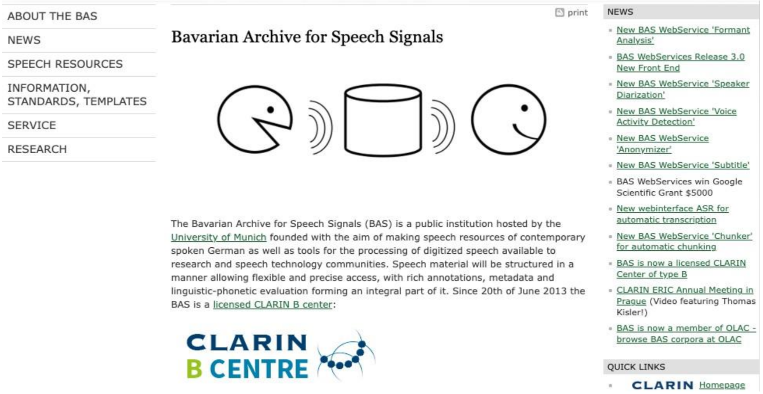
Für diese Sprache stehen auch Webservices zur

Fig. 1. The Leipzig Wortschatz portal for searching a total of 1.013 monolingual dictionaries which have been compiled from web-harvested corpora for 293 languages

For contemporary German, the German Reference Corpus DeReKo [13, 14, 15] is one of the resources offered by Text+ partners. It is the largest linguistically motivated collection of electronic corpora for German, with 55 billion words (as of March 2023), annotated at the token level. The data are contributed and licensed by publishers and other content providers and are freely available for academic, non-commercial use.

Other partners provide data on spoken German, such as the Bavarian Archive of Speech Signals (BAS, Fig. 2) and the Archive for Spoken Language at the IDS (Fig. 3). They offer recordings of various types: dialogues, prompted speech, and Oral History data. For many recordings, the speech signal is aligned with a transcript, which facilitates search and data mining.

For historical data, Text+ gives open-access to a wide range of editions and text collections. To name two examples: The TextGrid repository (see link in Appendix A and Fig. 4) housed at the SUB Göttingen, and the German Text Archive [16, 17] hosted at the Berlin-Brandenburg Academy of Sciences and the Humanities BBAW. These archives consist of retro-digitized books and documents in a consistent XML format. An illustration of the visual alignment of facsimile and the transcription in digital form is illustrated in Fig. 5. This example from the DTA shows the facsimile of a book published 1835. This book contains letters written by Johann Wolfgang von Goethe to Bettina von Arnim from 1809, displayed in the DTA web application as the facsimile and its aligned transcription.

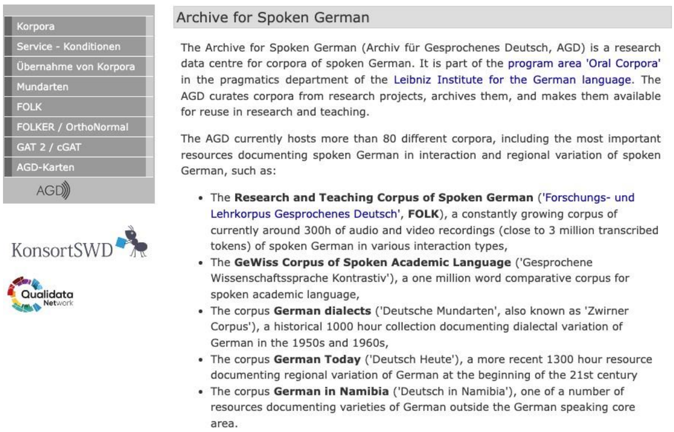


Bavarian Archive for Speech Signals

The Bavarian Archive for Speech Signals (BAS) is a public institution hosted by the University of Munich founded with the aim of making speech resources of contemporary spoken German as well as tools for the processing of digitized speech available to research and speech technology communities. Speech material will be structured in a manner allowing flexible and precise access, with rich annotations, metadata and linguistic-phonetic evaluation forming an integral part of it. Since 20th of June 2013 the BAS is a licensed CLARIN B center:

- New BAS WebService 'Formant Analysis'
- BAS WebServices Release 3.0 New Front End
- New BAS WebService 'Speaker Diarization'
- New BAS WebService 'Voice Activity Detection'
- New BAS WebService 'Anonymizer'
- New BAS WebService 'Subtitle'
- BAS WebServices win Google Scientific Grant $5000
- New webinterface ASR for automatic transcription
- New BAS WebService 'Chunker' for automatic chunking
- BAS is now a licensed CLARIN Center of type B
- CLARIN ERIC Annual Meeting in Prague (Video featuring Thomas Kisler!)
- BAS is now a member of OLAC - browse BAS corpora at OLAC

Fig. 2 Speech tools portal at the Bavarian Archive of speech signals



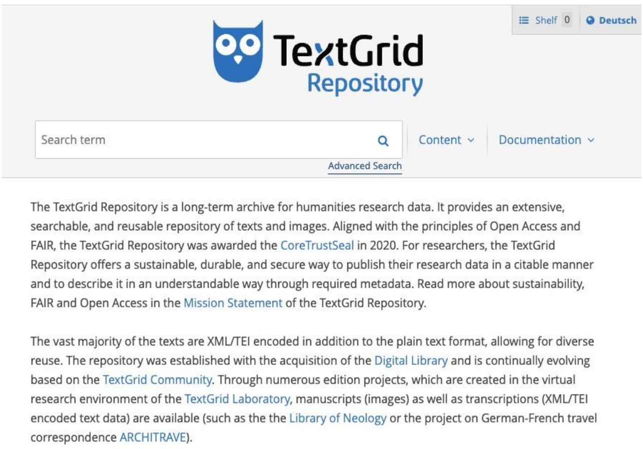
Archive for Spoken German

The Archive for Spoken German (Archiv für Gesprochenes Deutsch, AGD) is a research data centre for corpora of spoken German. It is part of the program area 'Oral Corpora' in the pragmatics department of the Leibniz Institute for the German language. The AGD curates corpora from research projects, archives them, and makes them available for reuse in research and teaching.

The AGD currently hosts more than 80 different corpora, including the most important resources documenting spoken German in interaction and regional variation of spoken German, such as:

- The **Research and Teaching Corpus of Spoken German** ('Forschungs- und Lehrkorpus Gesprochenes Deutsch', **FOLK**), a constantly growing corpus of currently around 300h of audio and video recordings (close to 3 million transcribed tokens) of spoken German in various interaction types,
- The **GeWiss Corpus of Spoken Academic Language** ('Gesprochene Wissenschaftssprache Kontrastiv'), a one million word comparative corpus for spoken academic language,
- The corpus **German dialects** ('Deutsche Mundarten', also known as 'Zwirner Corpus'), a historical 1000 hour collection documenting dialectal variation of German in the 1950s and 1960s,
- The corpus **German Today** ('Deutsch Heute'), a more recent 1300 hour resource documenting regional variation of German at the beginning of the 21st century
- The corpus **German in Namibia** ('Deutsch in Namibia'), one of a number of resources documenting varieties of German outside the German speaking core area.

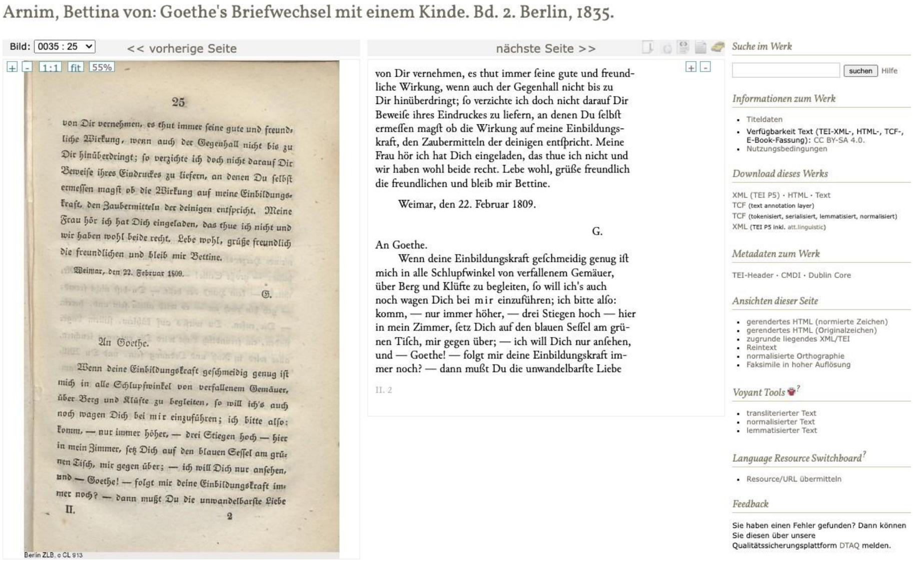
Fig. 3. Oral corpora are the data domain of the Archive for Spoken German (Archiv für Gesprochenes Deutsch, AGD) at the IDS, which also provides a portal for more than eighty corpora



The TextGrid Repository is a long-term archive for humanities research data. It provides an extensive, searchable, and reusable repository of texts and images. Aligned with the principles of Open Access and FAIR, the TextGrid Repository was awarded the CoreTrustSeal in 2020. For researchers, the TextGrid Repository offers a sustainable, durable, and secure way to publish their research data in a citable manner and to describe it in an understandable way through required metadata. Read more about sustainability, FAIR and Open Access in the Mission Statement of the TextGrid Repository.

The vast majority of the texts are XML/TEI encoded in addition to the plain text format, allowing for diverse reuse. The repository was established with the acquisition of the Digital Library and is continually evolving based on the TextGrid Community. Through numerous edition projects, which are created in the virtual research environment of the TextGrid Laboratory, manuscripts (images) as well as transcriptions (XML/TEI encoded text data) are available (such as the the Library of Neology or the project on German-French travel correspondence ARCHITRAVE).

Fig. 4. The portal of the TextGrid repository



Arnim, Bettina von: Goethe's Briefwechsel mit einem Kinde. Bd. 2. Berlin, 1835.

von Dir vernehmen, es thut immer ſeine gute und freundliche Wirkung, wenn auch der Gegenhall nicht bis zu Dir hinüberdringt; ſo verzichte ich doch nicht darauf Dir Beweiſe ihres Eindruckes zu liefern, an denen Du ſelbſt ermeſſen magſt ob die Wirkung auf meine Einbildungskraft, den Zaubermitteln der deinigen entſpricht. Meine Frau hör ich hat Dich eingeladen, das thue ich nicht und wir haben wohl beide recht. Lebe wohl, grüße freundlich die freundlichen und bleib mir Bettine.

Weimar, den 22. Februar 1809.

G.

An Goethe.

Wenn deine Einbildungskraft geſchmeidig genug iſt mich in alle Schlupfwinkel von verfallenem Gemäuer, über Berg und Klüfte zu begleiten, ſo will ich's auch noch wagen Dich bei mir einzuführen; ich bitte alſo: komm, — nur immer höher, — drei Stiegen hoch — hier in mein Zimmer, ſetz Dich auf den blauen Seſſel am grünen Tiſch, mir gegen über; — ich will Dich nur anſehen, und — Goethe! — folgt mir deine Einbildungskraft immer noch? — dann mußt Du die unwandelbarſte Liebe

II. 2

Suche im Werk

suchen Hilfe

Informationen zum Werk

- Titeldaten
- Verfügbarkeit Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung): CC BY-SA 4.0.
- Nutzungsbedingungen

Download dieses Werks

XML (TEI P5) · HTML · Text
TCF (text annotation layer)
TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
XML (TEI P5 inkl. att.linguistic)

Metadaten zum Werk

TEI-Header · CMDI · Dublin Core

Ansichten dieser Seite

- gerendertes HTML (normierte Zeichen)
- gerendertes HTML (Originalzeichen)
- zugrunde liegendes XML/TEI
- Reintext
- normalisierte Orthographie
- Faksimile in hoher Auflösung

Voyant Tools ?

- transliterierter Text
- normalisierter Text
- lemmatisierter Text

Language Resource Switchboard ?

- Resource/URL übermitteln

Feedback

Sie haben einen Fehler gefunden? Dann können Sie diesen über unsere Qualitätssicherungsplattform DTAQ melden.

Fig. 5. Rendering of a DTA resource on the website in three column format: Facsimile, transcription, and associated functionalities

The orthography and typography are kept as in the source document. The rightmost column in Fig. 5 provides metadata and additional functionality, including: download of various formats; rich metadata; different views. Links to post-processing tools are also provided, including: distant reading tools (Voyant Tools, see [18]); tools provided via the Language Resources Switchboard [19, 20]; and the quality assurance tool DTAQ, which allows users to suggest corrections to the transcription. Searching within the resources relies on normalized spellings and query expansion into all attested historical spelling variants, as described in [21].

NFDI funding does not support digitization via Optical Character Recognition (OCR) since there are separate funding opportunities dedicated to OCR. However, Text+ wants to integrate the results of OCR initiatives, wherever possible. For example, the Göttingen State and University Library will merge extensive multilingual collections, consisting of more than 13 million digitized pages in image format obtained from significant digitization initiatives that offer digital facsimiles of early German prints from the 16th to the 18th century, like those listed in the VD16, VD17 and VD18 catalogues (see online resources in Appendix A). These collections serve as the basis for automated text recognition in the OCR-D project. In the future, the outcomes of these efforts, along with additional resources such as scientific journals spanning from the 17th to the 21st century, will be integrated into the Text+ infrastructure and made accessible in accordance with legal regulations.

The data center with the largest collection of language data in Text+ is undoubtedly the German National Library DNB, with digital data resulting from an OCR workflow as well as born-digital data. However, the data are typically unstructured in the sense that they do not have XML or comparable markup and require some amount of pre-processing for data mining purposes.

When creating a research data infrastructure with these and many more existing resources, the first idea might be to create one large database to house them all centrally, possibly in the form of a duplicate of the original material from the data provider. However, this is neither possible nor desirable as the next section will discuss.

3.2. Motivation for a distributed infrastructure

Text+ is establishing a federated research data infrastructure as a network of data and competence centers. There are various reasons for such a federated approach. For legal reasons, it is often not possible to copy datasets from one institution to another. This is particularly the case for legacy datasets obtained from third parties under bilateral license agreements between the data providers and the hosting institutions of the data. Existing contracts and licenses are usually not transferable, i. e., they cannot be used by a different legal entity. Hence it is impossible to create an umbrella legal entity that would have the necessary legal authority to make the federated data available to research communities and/or the general public. And for the German National Library there is even a specific law [22] regulating their rights and obligations with regards to the data in their archives. The Text+ portfolio contains many such legacy data, which have been in high demand by scholarly communities of Text+, and which can be made available in a distributed research data infrastructure.

Datasets provided by third parties – and with text and language data this applies to almost all datasets – touches on the rights of individuals and legal entities. There are privacy restrictions imposed by GDPR [23], and intellectual property rights of publishers and authors must be respected. According to German law [24, paragraph 65], written works are copyright protected for 70 years after the author's death.

These are the main reasons why a consolidation into one database – including a search index – is not feasible. But of course, there are other reasons as well. Different data centers specialize in different types of research data with different data models and with established workflows for the creation, maintenance and analysis of data.

Based on our previous arguments, we find that due to the inclusion of legacy data from the Text+ partners, a decentralized data infrastructure becomes the only viable approach. The abundance of data within these institutions further motivates us to explore opportunities to provide access to these datasets beyond their original housing institutions.

4. Federated search in a distributed infrastructure

In a distributed infrastructure that provides research data according to FAIR principles, the tools that make data discoverable, enable access to the data, enable usability in different environments, and ensure reusability are of particular importance. We distinguish two different types of search: The first type is based on the metadata for archived objects, the second type searches within the object data. The metadata search allows users to search for datasets according to their properties, just like a search in a library catalog or in an online shop. The content search does not query the metadata but the content of the datasets, for example specific words in a text corpus.

The Text+ Registry allows searching on the metadata level and facilitates findability of research data. The registry provides information about where (i.e. at which institution) the data can be found and how to get access to it (i.e. access restrictions, access possibilities, license conditions, contact possibilities).

For metadata ingest, the Text+ Registry uses standard protocols of metadata harvesting such as OAI-PMH [25]. In this respect it uses the same underlying technology as the Virtual Language Observatory (VLO) [26], developed in the context of CLARIN. However, the Text+ Registry supports a broader range of metadata standards that are in use by the various partners, including (but not limited to) ISO 24622-1/24622-2 [27, 28], MARC21 [29], DCAT [30]. By utilizing these standard protocols and formats, the infrastructure remains open to new partners. Their data can be integrated by harvesting from their standard interfaces. The Text+ Registry is currently under development. A first version is expected to become available in the fourth quarter of 2023.

A second type of search functionality operates on the object level. As explained in the previous section, Text+ establishes a federated research data infrastructure as a network of data and competence centers. This has direct implications for the way users of the infrastructure can access research data. Depending on the research question at hand, users may want to search only Text+ data resources hosted at a particular Text+ data center. For such a scenario, users are best advised to consult the search portal for that particular data resource. But for other purposes, especially if a researcher is interested in all research data distributed over all or many of the Text+ data centers, then a federated search portal for aggregating the query results from data at different data centers is needed [31].

The challenge for a distributed infrastructure is to provide users with such a federated search portal. The portal should offer a unified query front end that interfaces with research data at different data centers at different locations, with diverse data formats in the back end.

Taking the Text+ domain of lexical resources as an example, Text+ partners provide a wide variety of research data. They include digital dictionaries such as the *Digitale Wörterbuch der Deutschen Sprache* [32], retro digitized encyclopedias such as the historic *Meyers Konversationslexikon* [33], and born-digital lexical resources such as the German wordnet GermaNet [34, 35], or the *Online-Wortschatz-Informationssystem Deutsch* [36], which is a portal to various resources of the IDS Mannheim. These lexical resources each have their data format and come with their query engine, tailored to the data type and structure of the particular resource. Fig. 6 illustrates the present situation for the data domain of lexical resources. These resources are encoded in a heterogeneous set of data formats. These specialized formats prohibit converting these data into a unified format without a major loss of information. On the other hand, users should be able to query the data in a unified query language. This suggests a federated search architecture with a unified frontend for querying the diverse data formats in the backend. In a federated search, the backend transparently maps multiple autonomous database systems into a single federated database – without physical data integration into one system. The constituent database systems remain independent and autonomous, they can also be used independently from each other.

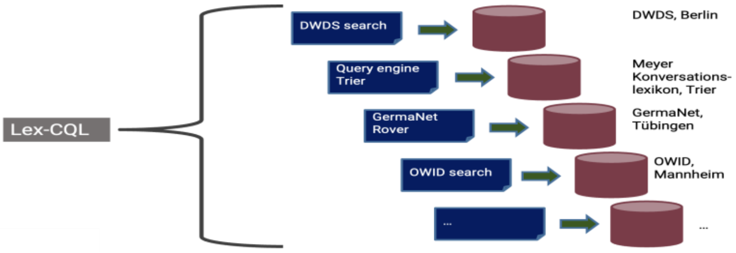


Fig. 6. Federated search exemplified with lexical resources; the individual lexical resources are left with the hosting institution; users are presented with a central search interface that distributes the queries to the federated locations and gathers the results for display

The federated search in Text+ utilizes the Search/Retrieve via URL (SRU) protocol and the *Contextual Query Language* (CQL), a standard query syntax for representing queries [37]. From the unified query interface, LexCQL-expressions (Lex-CQL is an extension of CQL for querying lexical resources.) are sent to all connected data centers in Text+. Each center translates the incoming query into the appropriate local query. The query results are then converted to the SRU format before being sent back to the portal front end. This way, the front end receives consistently formatted results from each center and can present the unified results to the user.

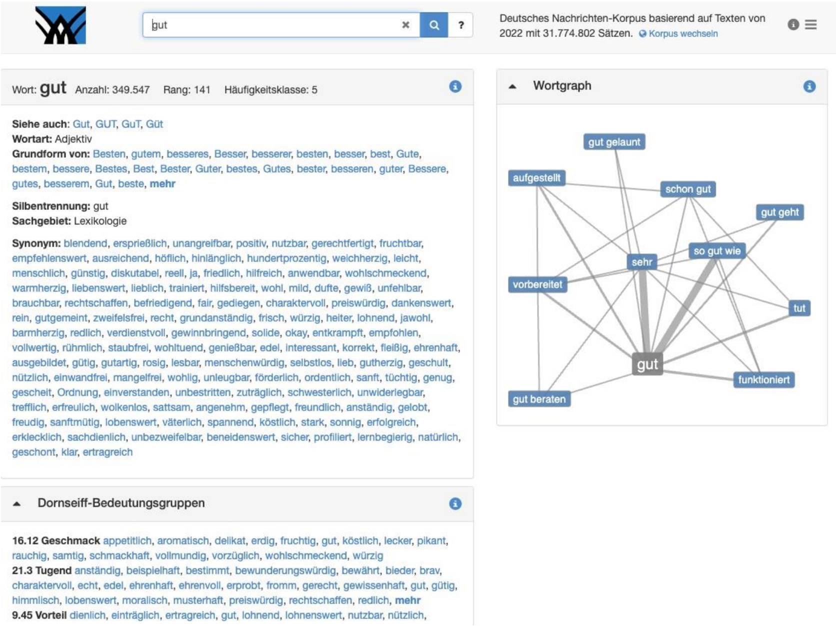
By this procedure, users can query multiple resources in parallel without having to learn the query language for each participating institution. This is already being implemented.

The web application for the LEX-CQL federated search also provides seemless access to the full lexical entry stored at the hosting institution. Fig. 7 shows the full entry found in the Leipzig Wortschatz resource and also provides a graphical presentation of typical collocates for the lemma gut. Finally, it displays the subject domains in the *Dornseiff* thesaurus [38] that the word gut is associated with.

Lex-CQL supports filtering results on part-of-speech, so that the results only include instances where the query term appears, for instance, as a noun or an adjective, as shown in Fig. 8. POS filtering is an extension of the latest version of the LEX-CQL query language, which is not yet supported by all of the data center endpoints. This is the reason why only 4 matching resources are found.

To show some other features of LEX-CQL, we will use a lemma search for the noun *Qualität* "quality". This word has a more complex morphology, with the highly productive suffix *-tät*. This suffix can be searched for, using the Kleene Star, resulting in all lemmas ending in this string of characters. It is possible to limit the number of hits per endpoint and/or to further restrict the search to specific properties. For the latter, one can use the def attribute, which will result in a full-text search of the word definition for a particular set of lemmas. One example for such a query is a search for lemmas with the suffix *-tät* and where the word definition of the lexical entry contains the string Eigenschaft "property". This CQL query can be written as:

`lemma="*tät" and def="Eigenschaft"`. This query will rule out a word such as *Aktivität* ‘activity’, which refers to an event rather than a property.



gut

Deutsches Nachrichten-Korpus basierend auf Texten von 2022 mit 31.774.802 Sätzen. Korpus wechseln

Wort: **gut** Anzahl: 349.547 Rang: 141 Häufigkeitsklasse: 5

Siehe auch: Gut, GUT, GuT, Güt
Wortart: Adjektiv
Grundform von: Besten, gutem, besseres, Besser, besserer, besten, besser, best, Gute, bestem, bessere, Bestes, Best, Bester, Guter, bestes, Gutes, bester, besseren, guter, Bessere, gutes, besserem, Gut, beste, **mehr**

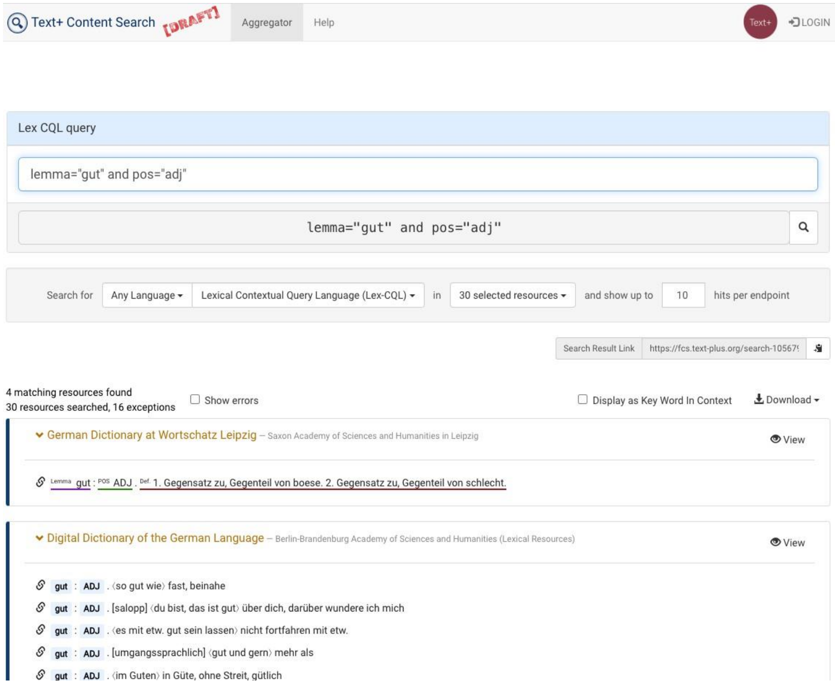
Silbentrennung: gut
Sachgebiet: Lexikologie

Synonym: blendend, ersprießlich, unangreifbar, positiv, nutzbar, gerechtfertigt, fruchtbar, empfehlenswert, ausreichend, höflich, hinlänglich, hundertprozentig, weichherzig, leicht, menschlich, günstig, diskutabel, reell, ja, friedlich, hilfreich, anwendbar, wohlschmeckend, warmherzig, liebenswert, lieblich, trainiert, hilfsbereit, wohl, mild, dufte, gewiß, unfehlbar, brauchbar, rechtschaffen, befriedigend, fair, gediegen, charaktervoll, preiswürdig, dankenswert, rein, gutgemeint, zweifelsfrei, recht, grundanständig, frisch, würzig, heiter, lohnend, jawohl, barmherzig, redlich, verdienstvoll, gewinnbringend, solide, okay, entkrampft, empfohlen, vollwertig, rühmlich, staubfrei, wohltuend, genießbar, edel, interessant, korrekt, fleißig, ehrenhaft, ausgebildet, gütig, gutartig, rosig, lesbar, menschenwürdig, selbstlos, lieb, gutherzig, geschult, nützlich, einwandfrei, mangelfrei, wohlig, unleugbar, förderlich, ordentlich, sanft, tüchtig, genug, gescheit, Ordnung, einverstanden, unbestritten, zuträglich, schwesterlich, unwiderlegbar, trefflich, erfreulich, wolkenlos, sattsam, angenehm, gepflegt, freundlich, anständig, gelobt, freudig, sanftmütig, lobenswert, väterlich, spannend, köstlich, stark, sonnig, erfolgreich, erklecklich, sachdienlich, unbezweifelbar, beneidenswert, sicher, profiliert, lernbegierig, natürlich, geschont, klar, ertragreich

▲ Dornseiff-Bedeutungsgruppen

16.12 Geschmack appetitlich, aromatisch, delikat, erdig, fruchtig, gut, köstlich, lecker, pikant, rauchig, samtig, schmackhaft, vollmundig, vorzüglich, wohlschmeckend, würzig
21.3 Tugend anständig, beispielhaft, bestimmt, bewunderungswürdig, bewährt, bieder, brav, charaktervoll, echt, edel, ehrenhaft, ehrenvoll, erprobt, fromm, gerecht, gewissenhaft, gut, gütig, himmlisch, lobenswert, moralisch, musterhaft, preiswürdig, rechtschaffen, redlich, **mehr**
9.45 Vorteil dienlich, einträglich, ertragreich, gut, lohnend, lohnenswert, nutzbar, nützlich,

Fig. 7. Referenced from the LexFCS output, the full entry of the source dictionary is linked, here from the Leipzig Wortschatz



Lex CQL query

lemma="gut" and pos="adj"

```
lemma="gut" and pos="adj"
```

Search for Any Language ▾ Lexical Contextual Query Language (Lex-CQL) ▾ in 30 selected resources ▾ and show up to 10 hits per endpoint

Search Result Link https://fcs.text-plus.org/search-10567!

4 matching resources found
30 resources searched, 16 exceptions

☐ Show errors ☐ Display as Key Word In Context Download ▾

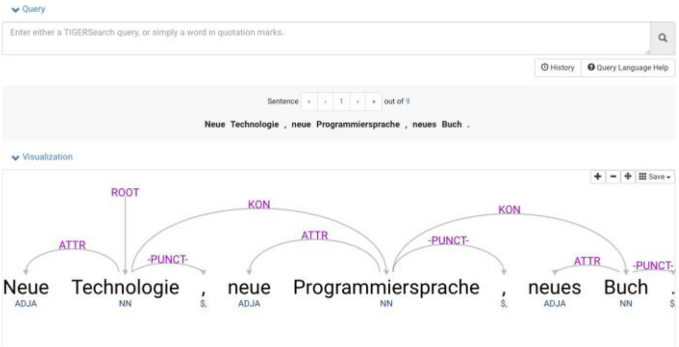
German Dictionary at Wortschatz Leipzig – Saxon Academy of Sciences and Humanities in Leipzig View

Lemma gut : POS ADJ . Def. 1. Gegensatz zu, Gegenteil von boese. 2. Gegensatz zu, Gegenteil von schlecht.

Digital Dictionary of the German Language – Berlin-Brandenburg Academy of Sciences and Humanities (Lexical Resources) View

gut : ADJ . ⟨so gut wie⟩ fast, beinahe
gut : ADJ . [salopp] ⟨du bist, das ist gut⟩ über dich, darüber wundere ich mich
gut : ADJ . ⟨es mit etw. gut sein lassen⟩ nicht fortfahren mit etw.
gut : ADJ . [umgangssprachlich] ⟨gut und gern⟩ mehr als
gut : ADJ . ⟨im Guten⟩ in Güte, ohne Streit, gütlich

Fig. 8. LexFCS output with a query for the lemma gut including the part of speech adj in the query

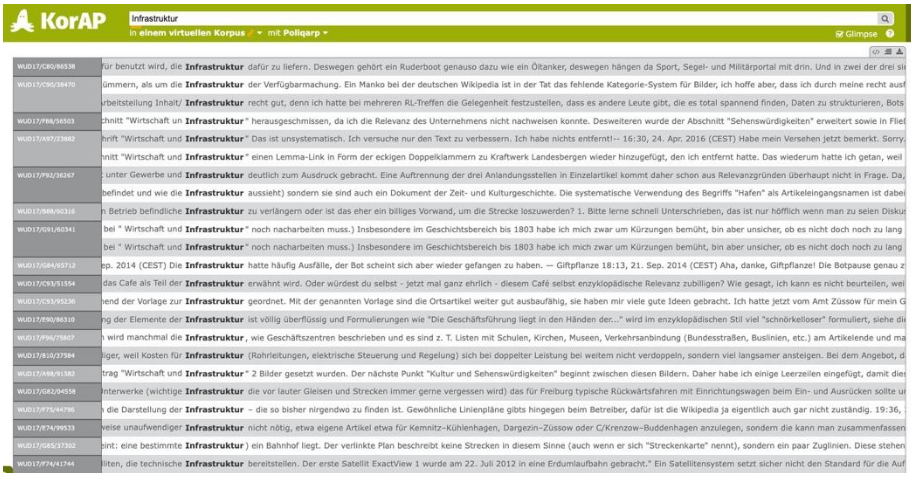


Query
Enter either a TIGERSearch query, or simply a word in quotation marks.
History
Query Language Help
Sentence
1
out of 9

Neue Technologie , neue Programmiersprache , neues Buch .

Fig. 9. Searching in treebanks and visualizing results in Tündra; integrated in WebLicht, Tündra can also visualize syntactic structures of user generated texts

In addition to the federated search for lexical resources and for corpora, researchers also have access to specialized search tools for annotated data – for example the search tool TüNDRA [39] for corpora that are annotated with syntactic constructions or semantic information (Fig. 9). Another of such tools is KorAP [40, 41, 42, 43], a tool for querying the German Reference Corpus (see Fig. 10 and Section 3.1 of this contribution). All of these tools also offer statistical information about the search results obtained.



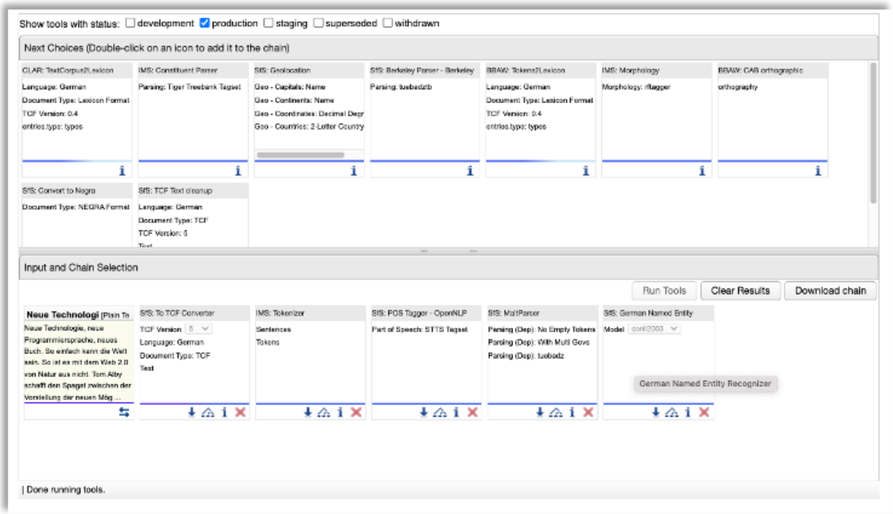
WUD17/C80/86538	für benutzt wird, die **Infrastruktur** dafür zu liefern. Deswegen gehört ein Ruderboot genauso dazu wie ein Öltanker, deswegen hängen da Sport, Segel- und Militärportal mit drin. Und in zwei der drei si
WUD17/C90/38470	ümmern, als um die **Infrastruktur** der Verfügbarmachung. Ein Manko bei der deutschen Wikipedia ist in der Tat das fehlende Kategorie-System für Bilder, ich hoffe aber, dass ich durch meine recht ausf
	rbeitsteilung Inhalt/ **Infrastruktur** recht gut, denn ich hatte bei mehreren RL-Treffen die Gelegenheit festzustellen, dass es andere Leute gibt, die es total spannend finden, Daten zu strukturieren, Bots
WUD17/F88/56503	chnitt "Wirtschaft un **Infrastruktur**" herausgeschmissen, da ich die Relevanz des Unternehmens nicht nachweisen konnte. Desweiteren wurde der Abschnitt "Sehenswürdigkeiten" erweitert sowie in Flie
WUD17/A97/23882	hrift "Wirtschaft und **Infrastruktur**" Das ist unsystematisch. Ich versuche nur den Text zu verbessern. Ich habe nichts entfernt!-- 16:30, 24. Apr. 2016 (CEST) Habe mein Versehen jetzt bemerkt. Sorry.
	nitt "Wirtschaft und **Infrastruktur**" einen Lemma-Link in Form der eckigen Doppelklammern zu Kraftwerk Landesbergen wieder hinzugefügt, den ich entfernt hatte. Das wiederum hatte ich getan, weil
WUD17/F92/36267	unter Gewerbe und **Infrastruktur** deutlich zum Ausdruck gebracht. Eine Auftrennung der drei Anlandungsstellen in Einzelartikel kommt daher schon aus Relevanzgründen überhaupt nicht in Frage. Da,
	befindet und wie die **Infrastruktur** aussieht) sondern sie sind auch ein Dokument der Zeit- und Kulturgeschichte. Die systematische Verwendung des Begriffs "Hafen" als Artikeleingangsnamen ist dabei
WUD17/B88/60316	n Betrieb befindliche **Infrastruktur** zu verlängern oder ist das eher ein billiges Vorwand, um die Strecke loszuwerden? 1. Bitte lerne schnell Unterschrieben, das ist nur höfflich wenn man zu seien Disku
WUD17/G91/60341	bei " Wirtschaft und **Infrastruktur**" noch nacharbeiten muss.) Insbesondere im Geschichtsbereich bis 1803 habe ich mich zwar um Kürzungen bemüht, bin aber unsicher, ob es nicht doch noch zu lang
	bei " Wirtschaft und **Infrastruktur**" noch nacharbeiten muss.) Insbesondere im Geschichtsbereich bis 1803 habe ich mich zwar um Kürzungen bemüht, bin aber unsicher, ob es nicht doch noch zu lang
WUD17/G84/65712	ep. 2014 (CEST) Die **Infrastruktur** hatte häufig Ausfälle, der Bot scheint sich aber wieder gefangen zu haben. — Giftpflanze 18:13, 21. Sep. 2014 (CEST) Aha, danke, Giftpflanze! Die Botpause genau z
WUD17/C93/51554	das Cafe als Teil der **Infrastruktur** erwähnt wird. Oder würdest du selbst - jetzt mal ganz ehrlich - diesem Café selbst enzyklopädische Relevanz zubilligen? Wie gesagt, ich kann es nicht beurteilen, wei
WUD17/C93/95236	hend der Vorlage zur **Infrastruktur** geordnet. Mit der genannten Vorlage sind die Ortsartikel weiter gut ausbaufähig, sie haben mir viele gute Ideen gebracht. Ich hatte jetzt vom Amt Züssow für mein G
WUD17/E90/86310	ng der Elemente der **Infrastruktur** ist völlig überflüssig und Formulierungen wie "Die Geschäftsführung liegt in den Händen der..." wird im enzyklopädischen Stil viel "schnörkelloser" formuliert, siehe di
WUD17/F96/75807	wird manchmal die **Infrastruktur**, wie Geschäftszentren beschrieben und es sind z. T. Listen mit Schulen, Kirchen, Museen, Verkehrsanbindung (Bundesstraßen, Buslinien, etc.) am Artikelende und ma
WUD17/B10/37584	liger, weil Kosten für **Infrastruktur** (Rohrleitungen, elektrische Steuerung und Regelung) sich bei doppelter Leistung bei weitem nicht verdoppeln, sondern viel langsamer ansteigen. Bei dem Angebot, d
WUD17/A98/91582	trag "Wirtschaft und **Infrastruktur**" 2 Bilder gesetzt wurden. Der nächste Punkt "Kultur und Sehenswürdigkeiten" beginnt zwischen diesen Bildern. Daher habe ich einige Leerzeilen eingefügt, damit dies
WUD17/G82/04558	Interwerke (wichtige **Infrastruktur** die vor lauter Gleisen und Strecken immer gerne vergessen wird) das für Freiburg typische Rückwärtsfahren mit Einrichtungswagen beim Ein- und Ausrücken sollte u
WUD17/F75/44796	die Darstellung der **Infrastruktur** – die so bisher nirgendwo zu finden ist. Gewöhnliche Linienpläne gibts hingegen beim Betreiber, dafür ist die Wikipedia ja eigentlich auch gar nicht zuständig. 19:36,
WUD17/E74/99533	weise unaufwendiger **Infrastruktur** nicht nötig, etwa eigene Artikel etwa für Kemnitz–Kühlenhagen, Dargezin–Züssow oder C/Krenzow–Buddenhagen anzulegen, sondern die kann man zusammenfassen
WUD17/G85/37302	eint: eine bestimmte **Infrastruktur**) ein Bahnhof liegt. Der verlinkte Plan beschreibt keine Strecken in diesem Sinne (auch wenn er sich "Streckenkarte" nennt), sondern ein paar Zuglinien. Diese stehen
WUD17/F74/41744	lliten, die technische **Infrastruktur** bereitstellen. Der erste Satellit ExactView 1 wurde am 22. Juli 2012 in eine Erdumlaufbahn gebracht." Ein Satellitensystem setzt sicher nicht den Standard für die Auf

Fig. 10. For searching the German Reference Corpus DeReKo, one of the tools is KorAP

5. Data analysis tools

In addition to easily accessible and easy-to-use tools for metadata and for search in the digital data themselves, there is an ever-increasing need for sophisticated tools for data analytics among Humanities scholars. This demand is, of course, triggered by the dramatic increase in digital language and text data in recent years and by the digital turn that research in the Humanities and beyond has witnessed in the process. For the purpose of data analytics, researchers need easy access to automatic or semi-automatic annotation tools, which can robustly process large datasets by enriching them with linguistic information at various levels of analysis. These analysis levels include, for example, annotations at the level of individual word tokens and lemmata, at the phrase and sentence level, or the clustering of texts by topic, author, sentiment. The use of such annotation tools can significantly enhance recall and precision for data analysis and information retrieval.

For the automatic annotation of textual data, we have developed a tool called WebLicht [44, 45]. WebLicht offers processing pipelines for a wide range of annotation tasks (e.g., tokenization, lemmatization, parsing, morphology, named-entity recognition) for text in a number of languages.



Show tools with status: □ development ☑ production □ staging □ superseded □ withdrawn

Next Choices (Double-click on an icon to add it to the chain)

CLAR: TextCorpus2Lexicon
Language: German
Document Type: Lexicon Format
TCF Version: 0.4
entries.type: types

IMS: Constituent Parser
Parsing: Tiger Treebank Tagset

SfS: Geolocation
Geo - Capitals: Name
Geo - Continents: Name
Geo - Coordinates: Decimal Degr
Geo - Countries: 2-Letter Country

SfS: Berkeley Parser - Berkeley
Parsing: tuebadztb

BBAW: Tokens2Lexicon
Language: German
Document Type: Lexicon Format
TCF Version: 0.4
entries.type: types

IMS: Morphology
Morphology: rftagger

BBAW: CAB orthographic
orthography

SfS: Convert to Negra
Document Type: NEGRA Format

SfS: TCF Text cleanup
Language: German
Document Type: TCF
TCF Version: 5
Text

Input and Chain Selection

Neue Technologi [Plain Te
Neue Technologie, neue Programmiersprache, neues Buch. So einfach kann die Welt sein. So ist es mit dem Web 2.0 von Natur aus nicht. Tom Alby schafft den Spagat zwischen der Vorstellung der neuen Mög ...

SfS: To TCF Converter
TCF Version 5
Language: German
Document Type: TCF
Text

IMS: Tokenizer
Sentences
Tokens

SfS: POS Tagger - OpenNLP
Part of Speech: STTS Tagset

SfS: MaltParser
Parsing (Dep): No Empty Tokens
Parsing (Dep): With Multi Govs
Parsing (Dep): tuebadz

SfS: German Named Entity
Model conll2003

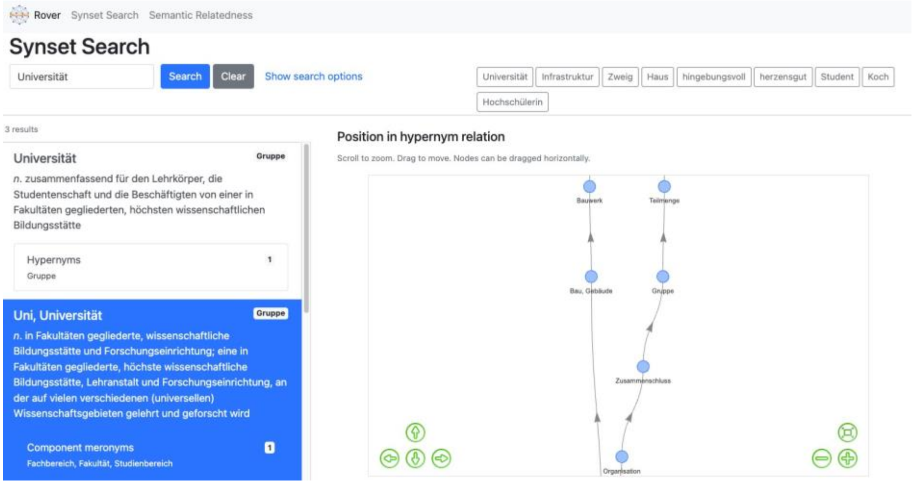
| Done running tools.

Fig. 11. WebLicht web application for configuring customized or pre-defined processing chains of annotation tools, which are applied to input data supplied by the user

The annotation tools for WebLicht (Fig. 11) have been contributed either by Text+ partners or by external providers. These include statistics-based tools, as well as tools trained with neural networks. In the future, we plan to integrate additional state-of-the-art services based on deep learning such as spaCy (See the spaCy website for details). TüNDRA, described above, is integrated into WebLicht, so that the annotated data is immediately visualized, and can be queried. For the annotation of very large datasets, WaaS (WebLicht as a Service) provides efficient processing from the command line.

In support of semantic analysis, Text+ offers ontology services such as wordnets and is developing linked-data resources that can aid in data analysis and in data linking. Fig. 12 shows the GermaNet Rover, which is a tool for querying GermaNet and visualizing semantic relations.

In addition to data analysis tools, Text+ also supports researchers with a tool switchboard, which matches data with appropriate data analysis or visualization tools. The Language Resource Switchboard [19, 20] is a very useful tool for researchers who are aware of data resources that are relevant for their research projects, but who may not be fully aware of the data analysis tools that can be used on these datasets.



Universität

3 results

Universität Gruppe

n. zusammenfassend für den Lehrkörper, die Studentenschaft und die Beschäftigten von einer in Fakultäten gegliederten, höchsten wissenschaftlichen Bildungsstätte

Hypernyms 1
Gruppe

Uni, Universität Gruppe

n. in Fakultäten gegliederte, wissenschaftliche Bildungsstätte und Forschungseinrichtung; eine in Fakultäten gegliederte, höchste wissenschaftliche Bildungsstätte, Lehranstalt und Forschungseinrichtung, an der auf vielen verschiedenen (universellen) Wissenschaftsgebieten gelehrt und geforscht wird

Component meronyms 1
Fachbereich, Fakultät, Studienbereich

Position in hypernym relation

Scroll to zoom. Drag to move. Nodes can be dragged horizontally.

Fig. 12. Visualizing semantic relations in Rover for GermaNet

6. Access to protected materials

With the large amount of legacy data and data with other legal and ethical constraints, Text+ is exploring ways for researchers to get access to textual materials that are still under copyright. To be on the safe side, it is still common practice in Germany to apply this protection to most resources published since the year 1920. The German National Library (DNB) has the mandate to collect all works published in Germany or about Germany, starting with the beginning of the 20th century. So in principle, the Text+ data center at the DNB could make all these materials available for non-commercial research purposes, if it were not for the fact that most of these materials are still under copyright. In view of the growing number of digital data, the European Union and its member states have reviewed and revised their copyright laws. For Germany, there is a new law that permits access to copyrighted materials for the purposes of text and data mining. But this does not give permission to individual researchers to distribute such materials. Text+ is currently investigating how copyrighted materials can be aggregated in digital form and made available for research in such a way that it does not violate copyright law. Obviously, this is a major issue for a research data infrastructure that focuses on textual data and language data more generally.

The strategy that Text+ is currently considering for this purpose is to utilize a range of derived text formats for textual sources. The German copyright law [24] does not provide clear criteria for evaluating whether a given derived text format infringes copyright or not. However, current legal practice recognizes three criteria: recognizability, reproducibility, and enjoyment of the work (See, for example, [46, 47] for a detailed discussion of the issues involved).

What is still under investigation and a topic of active research in Text+ is to what extent different derived text formats comply with the current copyright legislation – at least on a national scale. [48] identify six different text formats of this kind and investigate the usefulness of each derived text format for the following research areas in Digital Humanities: stylometry, the extraction of distinguishing features from a text, topic modelling, sentiment analysis, network analysis, text re-use, and language modelling. At the same time, [48] assess the legal criteria of recognizability, reproducibility, and enjoyment of the work. The authors of that study summarize their assessment in an overview table in their paper. Table 1 is an English translation of their original table ([48] use a three-valued classification scheme green, red, and yellow, which correspond to "yes", "no", and "maybe").

Table 1. Overview of derived textformats, their usefulness and a first legal assessment, adapted from [48]

Derived text format	Usefulness for research							Legal assessment		
	Stylometrics	Distinctiveness	Topic modeling	Sentiment analysis	Network analysis	Text Re-use	Language models	Protection from recognizability	Protection from reconstructability	Impossibility to enjoy the work
Simple TermDocument matrix	+	+	🟧	-	-	-	-	+	+	+
Scrambling of token sequences	+	+	+	o	o	-	o	o	o	+
Selective masking of tokens	-	o	+	o	o	+	o	o	o	o
N-grams	-	o	-	-	-	-	+	o	o	+
Static word embeddings	+	-	-	-	-	-	+	+	+	+
Contextual word embeddings	o	o	o	o	o	o	+	+	+	+

The six different derived text formats which [48] investigate include four existing representational formats: term-document matrices, which were first proposed for information retrieval purposes; n-gram representations, which were first proposed for language modelling purposes in Computational Linguistics; static word embeddings such as word2vec [49] and GloVe [50]; and contextual word embeddings such as BERT [51]. Each of these four representations can be used as the basis of a derived text format. A term-document representation of a text or a text collection abstracts away from the linear order of tokens in each document. Thus, it offers protection from the ability to reconstruct the original text and from being able to recognize the original text. Static or contextual word embeddings can be used in a derived text format where the word tokens in the original texts are substituted by their corresponding word embeddings. Static word embeddings offer a single vector representation per word or subword, depending on the tokenization that was used in training the embedding model. Contextual word embeddings offer more fine-grained context-dependent representations of word tokens, rather than a single vector per word. As long as there is no explicit link between the words or word tokens of the original text and the corresponding embeddings, it is not possible to reconstruct or to recognize the original text from the derived word embedding representations.

However, as [48] point out, numerical representations of word embeddings cannot be directly interpreted by humans.

An n-gram representation of a text will directly preserve sequence information of the original text only up to the level of the longest n-grams. However, it may be possible to interpolate sequences of tokens that exceed the length of the longest n-grams from combining n-grams of different lengths whose boundaries overlap. Thus, it is an open question under what conditions it is possible to reconstruct the original text from an n-gram characterization or whether it is possible to recognize the original text. The same uncertainty holds for the two derived formats that [48] propose in addition to these four existing representational formats discussed already: scrambling of word sequences from the original text and selective masking of individual tokens in the original text by part-of-speech labels (These two formats can be parameterized in the following way: for the scrambling approach, the length of the word sequences can be adjusted; for the selective masking approach, the number of word classes that are used for masking can be fine-tuned).

While the study by [48] offers a valuable first point of orientation, further research on derived text formats is still necessary. As Table 1 shows, a positive legal assessment of a derived data format does not correlate well with a positive assessment of the usefulness for research. [48] also point out that the set of possible derived text formats is by no means restricted to the six formats that they considered. Moreover, [48] only consider derived formats in isolation. However, what if a Text+ data center wants to release more than one derived data format for a source text (collection) that is still under access protection? Is it possible to enjoy the original work, recognize the original text(s), or reconstruct them from a combination of derived data formats? These are all open questions that Text+ still needs to investigate, before it can distribute derived data formats to its communities.

7. Summary and outlook

In addition to the legal aspects outlined in the previous section, Text+ is also addressing ethical and social issues of research data provenance. Such questions arise, inter alia, in the area of language documentation for minority languages and endangered languages, two data domains that participating institutions of Text+ are collecting research data on and are sharing with their research communities. Here the rights of minority populations or indigenous people are at issue.

Text+ is by no means the only NFDI consortium that is concerned with legal and ethical issues. NFDI consortia - in the Life Sciences and in Medicine face ethical and legal issues with respect to patient data or data on the human genome. For a cross-disciplinary research data infrastructure such as the NFDI, it is therefore imperative to develop common guidelines for legal, social, and ethical aspects of research data management and data sharing.

In order to facilitate cross-disciplinary dialogue and the development of common guidelines, the NFDI has formed five special sections for topics that cut across different disciplinary NFDI consortia. In addition to the section on Ethical and Legal issues [52], there are sections on the cross-cutting topics of (Meta-)data, Terminologies, Provenance [53]; of Common Infrastructures [54]; of Training & Education [55]; and of Industry Engagement [56]. Each of these sections has written a section concept that is published on zenodo. Text+ is contributing to all of these cross-disciplinary sections.

Due to space restrictions, we will mention only two additional areas of active involvement of Text+ in cross-cutting sections of the NFDI. The issue of language and text collections as data resources with commercial value has recently become a major topic in the context of large language model applications such as conversational AI. It is therefore necessary for Text+ to identify common interests and seek collaborations with industrial partners. What Text+ can offer in such collaborations is access to high-quality data that have been carefully curated and can be used under transparent license conditions. These activities are currently coordinated through the NFDI section on Industrial Relations. In the NFDI section on Education and Training, Text+ is promoting the concept of data literacy among researchers in the Humanities and Cultural Studies.

Support for the Text+ user communities is, of course, not restricted to the conceptual work in the NFDI section on education and training. Text+ is organizing training events for researchers at different stages of their careers, but with special emphasis on doctoral and post-doctoral researchers. Such training events are organized on a continuous basis, often in conjunction with annual conferences of participating Text+ disciplines. Support for individual researchers is channeled via the Text+ helpdesk. The helpdesk offers the possibility for researchers to field specific questions to the participating institutions of Text+. Such questions often concern topics such as the drafting of research data management plans, the hosting of research data by Text+ data centers, the use of specific collections of research data, or the use of specific software tools offered by Text+.

Ultimately, the success of Text+ will depend on the active involvement of its user communities. In order to facilitate this involvement, Text+ has formed four coordination committees with its communities. These committees are addressing the future development of the Text+ portfolio of research data and infrastructure services. The members of these committees are representatives from the relevant professional organizations in the scientific disciplines that participate in Text+. The coordination committees meet on a regular basis to give feedback on the progress of Text+ and to select cooperation projects proposed by institutions that are not already part of the Text+ consortium. Such proposals are solicited via an open call that is addressed to universities and other research institutions in the fields of Humanities and Cultural Studies that are interested in integrating their research data and services into the Text+ research infrastructure and are prepared to share their data via Text+.

Needless to say, the construction of a national research data infrastructure will require considerable time and concerted effort by many stakeholders. It is encouraging to see that the German government has identified this effort as a matter of national priority. Currently, the funding of the NFDI is guaranteed until 2026. However, talks are already under way to extend the funding for the NFDI with the ultimate goal of funding a national research infrastructure on a permanent basis.

Text+ is a collaborative effort by more than thirty research institutions across Germany with a total of well over 100 researchers. While the co-authors of this paper have prepared its contents and take responsibility for any errors that may have occurred, the work reported here is carried out by a large group of colleagues in Text+ and predecessor projects CLARIN-D, DARIAH-DE, and CLARIAH-DE. Their contributions are hereby acknowledged with deep appreciation and collegial gratitude. In addition, we thank all colleagues from the other NFDI consortia and from the NFDI directorate for their collaboration on cross-cutting NFDI issues.

***Acknowledgments*:** This publication was created in the context of the German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states through the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). Text+ funding is identified by the DFG Project No 460033370.

References

1. Wilkinson, M., D. Dumontier, I. J. Aalbersberg et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. – Scientific Data, Vol. **3**, 2016. DOI: 10.1038/sdata.2016.18.
2. D. Fišer, A. Witt, Eds. CLARIN. The Infrastructure for Language Resources. Berlin, Boston, De Gruyter, 2022. DOI:10.1515/9783110767377.
3. Hinrichs, E., S. Krauwer. The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. – In: Proc. of 11th International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association, Reykjavik, Iceland, 2014, pp. 1525-1531. **http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.**
4. Hedges, M., H. Neuroth, K. M. Smith, T. Blanke, L. Romary, M. Küster, M. Illingworth. TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship. – Journal of the Text Encoding Initiative, 2013. DOI:10.4000/jtei.774.
5. European Strategy Forum on Research Infrastructures (ESFRI). Strategy Report on Research Infrastructures: Roadmap 2018, Report 2018. **http://roadmap2018.esfri.eu/media/1060/ esfri-roadmap-2018.pdf.**
6. Brünger-Weilandt, S., K.-C. Bruhn, A. W. Busch et al. Memorandum of Understanding by NFDI Initiatives from the Humanities and Cultural Studies. 2020. DOI:10.5281/zenodo.4045000.
7. Paulmann, J., C. Wood, K. Ceynowa et al. NFDI4Memory. Consortium for the Historically Oriented Humanities. Proposal for the National Research Data Infrastructure (NFDI), 2022. DOI:10.5281/zenodo.7428489.
8. Paulmann, J., C. Wood, F. Cremer. Linkage – Digitale Gegenwart und Zukunft Historischer Forschung. Die Ziele der Konsortialinitiative 4Memory. – VHD Journal, 2020, pp. 26-34.
9. Bibby, D., K.-C. Bruhn, F. Dührkopp et al. Digitales Forschungsdatenmanagement in der Archäologie und die Initiative NFDI4Objects. – BliCKpunkt Archäologie, 2021, 2021, pp. 150-164.
10. Altenhöner, R., I. Blümel, F. Boehm et al. NFDI4Culture – Consortium for Research Data on Material and Immaterial Cultural Heritage. – Research Ideas and Outcomes, Vol. **6**, 2020. DOI:10.3897/rio.6.e57036.
11. Hinrichs, E., A. Geyken, P. Leinen et al. Text+: Language- and Text-Based Research Data Infrastructure, 2022. DOI:10.5281/zenodo.6452002.
12. Trippel, T. Mit Text+ Forschungsdaten digital vernetzen – ein Fall für die Sprachwissenschaft? – Sprachreport, Vol. **38**, 2022, pp. 1-7. DOI:10.14618/sr-1-2022_trip.

13. Kupietz, M., H. Lüngen, P. Kamocki, A. Witt. The German Reference Corpus DeReKo: New Developments – New Opportunities. – In: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, Eds. Proc. of 11th International Conference on Language Resources and Evaluation (LREC'2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 4353-4360.

14. Kupietz, M., C. Belica, H. Keibel, A. Witt. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. – In: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, D. Tapias, Eds. Proc. of 7th Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA'10), 2010, pp. 1848-1854.

15. Kupietz, M., H. Keibel. The Mannheim German Reference Corpus (DeReKo) as a Basis for Empirical Linguistic Research. – In: M. Minegishi, Ed. Workings Papers in Corpus-Based Linguistics and Language Education, Vol. **3**, Tokyo University of Foreign Studies, 2009, Tokyo, 2009, pp. 53-59.

16. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache, 2023. **https://www.deutschestextarchiv.de/**

17. Geyken, A., S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas, F. Wiegand. Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. – In: S. Schomburg, C. Leggewie, H. Lobin, C. Puschmann, Eds. Digitale Wissenschaft – Stand und Entwicklung digital vernetzter Forschung in Deutschland, Köln, 2010, pp. 157-161. **https://hbz.opus.hbz-nrw.de/frontdoor/index/index/docId/206**

18. Sinclair, S., G. Rockwell. Voyant Tools. **http://voyant-tools.org, 2016**

19. Zinn, C. The Language Resource Switchboard. – Computational Linguistics, Vol. **44**, 2018, pp. 631-639. DOI:10.1162/coli_a_00329.

20. Zinn, C., E. Dima. The CLARIN Language Resource Switchboard: Current State, Impact, and Future Roadmap. – In: D. Fišer, A. Witt, Eds. CLARIN. The Infrastructure for Language Resources, deGruyter, Berlin, 2022, pp. 83-106.

21. Jurish, B. Finite-State Canonicalization Techniques for Historical German. Ph.D. Thesis, Universität Potsdam, 2012. **http://opus.kobv.de/ubp/volltexte/2012/5578/, (completed 2011, published 2012)**

22. Deutscher Bundestag, Gesetz über die Deutsche Nationalbibliothek (DNBG), Bundesgesetzblatt Jahrgang, 2006. **https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html**

23. European Commission, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance), 2016. **https://eur-lex.europa.eu/eli/reg/2016/679/oj**

24. Deutscher Bundestag, Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtgesetz). Bundesgesetzblatt Jahrgang, 1965. **https://www.gesetze-im-internet.de/urhg/index.htm**

25. OAI-PMH, The Open Archives Initiative Protocol for Metadata Harvesting, Technical Report. – The Open Archives Initiative, 2015. **https://www.openarchives.org/pmh/**

26. Van Uytvanck, D., C. Zinn, D. Broeder, P. Wittenburg, M. Gardelleni. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. – In: Proc. of 7th Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association, 2010, pp. 900-903. **http://www.lrec-conf. org/proceedings/lrec2010/pdf/273_Paper.pdf**

27. ISO 24622-1, Language Resource Management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model, International Standard, International Organization for Standardization (ISO). Geneva, 2015.

28. ISO 24622-2, Language Resource Management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language, International Standard, International Organization for Standardization (ISO). Geneva, 2019.

29. MARC21, MARC 21 Format for Bibliographic Data, Technical Report, 1999-2023. **http://www.loc.search-gov/marc/bibliographic/**

30. Data Catalog Vocabulary (DCAT). Version 3. Technical Report. World Wide Web Consortium, 2023.
https://www.w3.org/TR/vocab-dcat-3/

31. Stehouwer, H., M. Durco, E. Auer, D. Broeder. Federated Search: Towards a Common Search Infrastructure. – In: N. Calzolari, N., K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, Eds. Proc. of 8ght International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3255-3259.
http://www.lrec-conf. org/proceedings/lrec2012/pdf/524_Paper.pdf

32. Klein, W., A. Geyken. Das Digitale Wörterbuch der Deutschen Sprache (DWDS). – Lexikographica, 2010, pp. 79-93.

33. Meyers Großes Konversationslexikon (6. Auflage, 1905-1909), digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Wörterbuchnetz des Trier Center for Digital Humanities, 2023.
https://www.woerterbuchnetz. de/Meyers

34. Henrich, V., E. Hinrichs. GernEdiT – the GermaNet Editing Tool. – In: Proc. of International Conference on Language Resources and Evaluation (LREC) 7, European Language Resources Association (ELRA), Valletta, 2010, pp. 2228-2235.
http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf

35. Hamp, B., H. Feldweg. GermaNet – A Lexical- Semantic Net for German. – In: P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, Y. Wilks, Eds. Proc. of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Association for Computational Linguistics, Somerset, NJ, 1997, pp. 9-15.

36. OWID – Online-Wortschatz-Informationssystem Deutsch, Since 2008.
http://www.owid.de

37. SearchRetrieve. Version 1.0. OASIS Standard. OASIS, 2013.
https://docs.oasis-open.org/ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part0-overview.html

38. Dornseiff, F. Der deutsche Wortschatz nach Sachgruppen, Berlin, Boston, De Gruyter, 2020. DOI:10.1515/9783110457742.

39. Martens, S. TüNDRA: A Web Application for Treebank Search and Visualization. – In: Proc. of 12th Workshop on Treebanks and Linguistic Theories (TLT'12), 2013, pp. 133-144.

40. Bański, P., J. Bingel, N. Diewald, E. Frick, M. Hanl, M. Kupietz, P. Pęzik, C. Schnober, A. Witt. KorAP: The New Corpus Analysis Platform at IDS Mannheim. – In: Z. Ventulani, H. Uszkoreit, Eds. Proc. of 6th Language and Technology Conference (LTC'13), Poznań: Fundacja Uniwersytetu im. A. Mickiewicza, 2013, pp. 586-587.
https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3261/file/Banski_KorAP_2013.pdf

41. Kupietz, M., N. Diewald, M. Hanl, E. Margaretha. Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. – In: M. Konopka, Ed. Grammatische Variation. Empirische Zugänge und theoretische Modellierung. Jahrbuch des Instituts für Deutsche Sprache, 2016, de Gruyter, Berlin/Boston, 2017, pp. 319-329.

42. Diewald, N., E. Margaretha. Krill: KorAP Search and Analysis Engine, Corpus Linguistic Software Tools. – Journal for Language Technology and Computational Linguistics (JLCL), Vol. **31**, 2016, pp. 73-90.

43. Diewald, N., M. Hanl, E. Margaretha, J. Bingel, M. Kupietz, P. Bański, A. Witt. KorAP Architecture – Diving in the Deep Sea of Corpus Data. – In: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, Eds. Proc. of 10th International Conference on Language Resources and Evaluation (LREC'16), Paris, European Language Resources Association (ELRA'16), Portoroz, Slovenia, 2016, pp. 3586-3591.

44. Hinrichs, M., T. Zastrow, E. Hinrichs. WebLicht: Web-Based LRT Services in a Distributed eScience Infrastructure. – In: N. Calzolari, Ed. Proc. of International Conference on Language Resources and Evaluation (LREC), Vol. **7**, 2010, pp. 489-493.

45. Hinrichs, E., M. Hinrichs, T. Zastrow. WebLicht: Web-Based LRT Services for German. – In: Proc of ACL 2010 System Demonstrations, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 25-29.
https://aclanthology.org/P10-4005
46. Glisse, K. Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte: Rechtliche Bedingungen und Möglichkeiten, RuZ – Recht und Zugang, 2020, pp. 143-159. DOI:10.5771/2699-1284-2020-2-143.
47. Jotzo, F. Der Schutz großer Textbestände nach dem UrhG– Die Nutzbarmachung fremder Textbestände für die Forschung, RuZ – Recht und Zugang, 2020, pp. 128-142. DOI:10.5771/2699-1284-2020-2-128.
48. Schöch, C., F. Döhl, A. Rettinger, E. Gius, P. Trilcke, P. Leinen, F. Jannidis, M. H., Jörg Röpke. Abgeleitete Textformate: Prinzip und Beispiele, RuZ – Recht und Zugang, 2020, pp. 160-194. DOI:10.5771/2699-1284-2020-2-160.
49. Mikolov, T., K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. – In: Y. Bengio, Y. LeCun, Eds. Proc. of 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2-4 May 2013, Workshop Track Proceedings, 2013.
http://arxiv.org/abs/1301.3781
50. Pennington, J., R. Socher, C. Manning. GloVe: Global Vectors for Word Representation. – In: Proc. of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532-1543. DOI:10.3115/v1/D14-1162.
51. Devlin, J., M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. – In: Proc. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. **1** (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171-4186. DOI:10.18653/v1/N19-1423.
52. Boehm, R., B. Buchner, D.-K. Kipker, A. Kuntz, G. Petri, U. Sax, K. Schaar, D. von Suchodoletz, O. Vettermann. Sektionskonzept Ethical, Legal & Social Aspects (Section-ELSA), 2021. DOI:10.5281/zenodo.5675972.
53. Koepler, O., T. Schrade, S. Neumann, R. Stotzka, C. Wiljes, I. Blümel, C. Bracht, T. Hamann, S. Arndt, J. Hunold. Sektionskonzept Meta(daten), Terminologien und Provenienz zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2021. DOI:10.5281/zenodo.5619089.
54. Diepenbroek, M., S. Schimmler, B. Ebert. Sektionskonzept Common Infrastructures zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2021. DOI:10.5281/zenodo.5607490.
55. Herres-Pawlis, S., P. Pelz, N. Kockmann, et al. Sektionskonzept Training & Education zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2022. DOI:10.5281/zenodo.6475541.
56. Stahl, F., A. Hamann. Sektionskonzept Industry Engagement zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2023. DOI:10.5281/zenodo.7900079.

Appendix A. Online Resources FIRST APPENDIX

The following online resources have been mentioned and referenced in the text. All links were last visited in July 2023, but each of them is expected to be persistent.

- Archiv für Gesprochenes Deutsch, **https://agd.ids-mannheim.de/**
- BAS Webservices, **https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface**
- CLARIAH-DE, project to merge CLARIN-D and DARIAH-DE, **https://www.clariah.de**
- CLARIN, European Research Infrastructure Consortium (ERIC), **https://www.clarin.eu**

- CLARIN-D, German chapter of the European CLARIN infrastructure network, **https://www.clarin-d.net**
- Deutsches Referenzkorpus (DeReKo), **https://www.ids-mannheim.de/digspra/kl/projekte/korpora/**
- Deutsches Textarchiv (DTA), **https://www.deutschestextarchiv.de/**
- Text+ Federated Content Search (FCS), **https://fcs.text-plus.org/**
- GermaNet Rover, **https://weblicht.sfs.uni-tuebingen.de/rover/**
- KorAP, **https://korap.ids-mannheim.de/**
- Language Resource Switchboard, **https://switchboard.clarin.eu/**
- German national research data infrastructure NFDI, **https://www.nfdi.de/**
- NFDI4Culture, **https://nfdi4culture.de/**
- NFDI4Memory, **https://4memory.de/**
- NFDI4Objects, **https://www.nfdi4objects.net/**
- NFDI-Section: Ethical and Legal issues, **https://www.nfdi.de/section-elsa/**
- NFDI-Section: (Meta-)data, Terminologies, Provenance, **https://www.nfdi.de/section-meta/**
- NFDI-Section: Common Infrastructures, **https://www.nfdi.de/section-infra/**
- NFDI-Section: Training & Education, **https://www.nfdi.de/section-edutrain/**
- NFDI-Section: Industry Engagement, **https://www.nfdi.de/section-industry-engagement/**
- Online Wortschatz-Informationssystem OWID, **http://www.owid.de**
- spaCy, **https://spacy.io/**
- Text+ webpage, **https://www.text-plus.org**
- TextGrid Repository, **https://textgridrep.org/**
- TüNDRA, **https://weblicht.sfs.uni-tuebingen.de/Tundra**
- Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (VD 16), English: Register of printed works of the 16th century published in German-speaking countries (VD 16), **http://www.vd16.de**
- Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts (VD 17), English: Union Catalogue of Books Printed in German Speaking Countries in the 17th Century (VD 17), **http://www.vd17.de**
- Verzeichnis Deutscher Drucke des 18. Jahrhunderts (VD 18), English: The Index of 18th century German prints (VD 18), **http://www.vd18.de**
- Voyant tools, **http://voyant-tools.org/**
- WebLicht, **https://weblicht.sfs.uni-tuebingen.de/**
- Wortschatz Leipzig, **https://wortschatz.uni-leipzig.de/de**

Received: 12.11.2024; Accepted: 26.11.2024. Fast-track.