IDS

Korpuslinguistik und interdisziplinäre
Perspektiven auf Sprache

Band 5

Jost Gippert / Ralf Gehrke (eds.)

# Historical Corpora

## Challenges and Perspectives

narr
VERLAG

Jost Gippert / Ralf Gehrke (eds.)

# Historical Corpora

## Challenges and Perspectives

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

# Contents

ALEXANDER GEYKEN / THOMAS GLONING

# A living text archive of 15th-19th-century German

## Corpus strategies, technology, organization

## Abstract

The corpus situation for the study of Early New High German (ENHG) and the older stages of New High German (oNHG) is far from satisfying. Indeed, there is no integrated and balanced corpus of a sufficient size that is publicly available. In this paper we give an outline of a living text archive of ENHG and oNHG in a distributed infrastructure that is supposed to fill this gap in the coming years. Such an archive is designed as a collaborative, sustainable and interoperable platform where historical texts are integrated together with relevant metadata and expert computational technology. We briefly mention examples of research questions and of corpus building scenarios and comment on the organizational aspects of such an archive including its potential as a basis for reference corpora for specific purposes.

## 1. Introduction

The notion of a corpus is intimately linked to three basic ideas: first, the idea of a corpus as a balanced or even representative sample of a language, a language stage, a variety of a language or a specific form of language use; second, the idea that a corpus is designed in order to fulfill a specific research question on an empirical basis; third, the idea that different research questions, purposes and layers of investigation (like semantics, morphology, textual organization) may require very different types of corpora in respect of size and composition. When Hoffmann (1998) gave his overview on historical corpora of German, there were not many electronic items available. Today, historical language corpora are usually digitized collections and historical corpus linguistics is closely connected with the use of computational tools.

However, the corpus situation for the study of Early New High German (roughly 15th-17th cc.; ENHG) and the older stages of New High German (roughly 17th-19th cc.; oNHG) is far from satisfying. There is no integrated and balanced corpus of a sufficient size that is publicly available.[1] Small or middle

[1] The Bonn Corpus of Early New High German (www.korpora.org/Fnhd/) is balanced as to parameters of time, region and text types; however, it is relatively small and was built and encoded specifically for the analysis of grammatical aspects.

sized subcorpora of research projects are usually not made available to other researchers. Up to now, there is no platform and no established culture of publicly sharing ENHG and oNHG resources among historians of the German language. Moreover, the electronic data that are produced for the publication of printed editions are not normally further processed for corpus usage, and in many cases, digital rights are given away to publishing houses without negotiation of moving wall solutions for the later use of the text in core corpora of the German language. On the other hand, there *are* many electronic texts on the hard drives of individual scholars and out there on the web, but they come in a huge diversity of technical formats and transcription schemas, so that it is hardly possible to work with these resources in a systematic way.

In addition to ongoing corpus projects funded by the Deutsche Forschungsgemeinschaft (German Research Council),[2] a curation project within the CLARIN-D infrastructure project[3] funded by the Bundesministerium für Bildung und Forschung (Federal Ministery for Education and Research) aimed to establish a distributed infrastructure for what we call a *living text archive* of ENHG and oNHG. Such an archive is designed as a collaborative, sustainable and interoperable platform where historical texts are integrated together with relevant metadata and expert computational technology. While the profile of such an archive will be 'opportunistic', it will be possible to choose specific subcorpora in a systematic way by using metadata on linguistic variation parameters. In a first round, an inventory of suitable legacy resources was compiled and, as a feasibility study, more than 78,000 pages were curated and integrated into the infrastructure, covering a wide range of text types, topics, dates of publication etc. This forms the basis for a long term curation initiative in a distributed platform that supports corpus based research on historical texts in general.

[2] Referenzkorpus Frühnheuhochdeutsch (Reference Corpus of Early New High German), see: http://gepris.dfg.de/gepris/projekt/200609649 (10/05/2014); Deutsches Textarchiv (German Text Archive), see: www.deutschestextarchiv.de. All URLs quoted were last accessed on May 5, 2014.

[3] On CLARIN-D see www.clarin-d.de. On the curation project see www.deutschestextarchiv.de/clarin_kupro and http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.

In this paper we give an outline of such a living text archive of ENHG and of oNHG in a distributed infrastructure (section 2), we briefly mention examples of research questions and of corpus building scenarios (section 3), and comment on the technical (section 4) and the organizational aspects (section 5) of such an archive including its potential as a basis for reference corpora for specific purposes. Technical aspects and aspects of workflow are described in more detail in the paper by Thomas and Wiegand (this volume).

## 2. The idea of a living text archive of Early New High German and older New High German

The basic idea of the CLARIN-D curation project (Kurationsprojekt 2012; Thomas/Wiegand, this volume) was to 'feed' the existing infrastructures at the Deutsches Textarchiv (DTA) of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)[4] and at the Digital Library of the Herzog August Bibliothek (Wolfenbüttel)[5] with historical texts from the 15th to the 19th centuries that are already available: (i) in collections like wikisource, (ii) in the legacies of research projects, (iii) from individual scholars and (iv) scattered across different places on the web. Such texts had first to be evaluated with respect to their quality. The ones to be integrated into the repository had to be enriched by metadata (e.g. on the transcription schema, on the degree of accuracy, and on linguistic variation parameters), and finally they had to be brought into a TEI compliant format. The texts are now available in a distributed infrastructure at the BBAW, the HAB and in the near future at the Institute for the German Language (IDS, Mannheim), both for federated search and for download under a creative commons license.

The aim of the curation project was threefold: (i) to produce an inventory of available electronic texts in our time frame; (ii) to provide an evaluation and integration infrastructure that can be used on a long term basis; (iii) to curate and to integrate 35,000 pages of historical corpus texts into the respective infrastructures in a 'first round'. At the end of the project, more than 78,000 pages were integrated. The result of this integration of digital texts from the 15th to the 19th centuries is not yet a balanced reference corpus but rather an opportunistic repository. However, the use of metadata on linguistic variation parameters (e.g. date, place of publication, subject field, text type) will allow

[4] www.deutschestextarchiv.de.

[5] www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html.

the building of specific subcorpora according to criteria of selection and their combination (see section 4.3, sampling procedure).

In the medium term the infrastructures both at the BBAW and the HAB will be used for ongoing work on the *living archive* of historical German from the 15th to the 19th centuries. Basically, work on the enhancement of the archive will consist in a number of different kinds of activities:

- adding new texts and new text groups according to criteria like time zone (decades), subject field (e.g. balneology), text type (e.g. newspaper reports), gender, the manuscript/print ratio, and others;
- adding new and/or more fine-grained metadata, e.g. with respect to subject fields or historical text types (cf. Budin/Kabas/Mörth 2012);
- improving aspects of balance/representativeness in a systematic way, e.g. in respect of the time zones (decades, centuries), the text types, the topics and subject fields, the gender ratio and other criteria of corpus balance;
- improving the quality of texts: while there are already numerous high quality texts in the Deutsches Textarchiv, we intend to include in the archive a number of working transcriptions as well. This means that together with image data of the underlying prints or manuscripts it will be possible to do serious work with the texts even if there may be a few remaining transcription errors that will be corrected in the course of the work.

Apart from these dynamic activities on core aspects of the texts and of the text collection as a whole we provide the opportunity of adding or connecting new (layers of) linguistic information and the results of other kinds of scientific investigation with the texts. This possibly includes introducing new (types of) metadata categories according to specific research questions and according to the results of specific research projects. E.g., cataloguing texts that belong to a certain historical controversy has not been one of the core aspects of corpus architecture; it is, however, an important aspect of Early Modern intellectual history. Therefore, such kinds of further information from specific research projects should be integrated as well.

## 3. Research questions and corpus building scenarios

Empirical work on specific research questions in the history of the German language must be based on suitable subcorpora. The design of specific subcorpora from a large textual archive for specific purposes requires both fine-grained metadata and knowledge about the history of the German language, its text types in a historical perspective, a broad overview on the historical development of the German language and sound knowledge of factors of language change and linguistic evolution. We will now illustrate this point with some examples.

(i) *Word usage and historical semantics.* Assume that a historical lexicographer is working on an article, say on the history of the foreign word *Influenz* (Engl. *influence*) or on the adjective *billig* (Engl. *cheap*). The word *Influenz* is interesting for its early usage in a cosmological, astrological sense. The word *billig* is particularly interesting for its semantic development with its shifts from the senses 'fair, proper' to 'cheap, not expensive' and then to 'of poor quality'. Here the interesting question is in which contexts these shifts took place and how common knowledge (a fair price is a low price; a cheap product is usually of poor quality) fostered specific understandings that eventually were generalized and conventionalized. In this situation a lexicographer might be interested to see all the quotations and in further steps to filter the quotations according to centuries or textual groups. In this use case, the production of KWIC concordances based on the whole archive, on temporal subcorpora (for *billig*) or on subcorpora for specific subject fields like alchemy, astronomy and cosmology (for *Influenz*) allows to describe the historical semantics of these words more precisely along important evolutionary parameters of word usage. The example shows that metadata (time, subject fields) are technical instruments that have to be 'geared' by knowledge or by hypotheses about linguistic dynamics.

(ii) *Linguistic profiles of historical text types in a synchronic and diachronic perspective.* How did newspaper reports, culinary recipes, printed sermons, medical case studies, technical descriptions of machines, literary translations and many other text types look like around 1600, around 1700, around 1800, around 1900? How did they evolve over time? What are the (changing) principles of textual organization? Are there typical syntactic patterns? What are the basic aspects of organization of their lexicon? In order to treat these types of questions, the metadata of the living archive will enable users to choose subcorpora for specific text types. They allow for straightforward searches for those aspects that can be 'translated' to purely form-based queries (e.g., show all quotations for *hat man*[6] in 17th-century newspapers in order to find a news-specific type of construction; or: show all word formations in *-ung* together with their spelling variants that come from texts marked for the subject field 'politics'). For other questions, the annotation of subcorpora will be the way to go.

(iii) *The search for and automatic retrieval of specific constructions and grammatical phenomena* is not a trivial matter. Depending on the complexity of the research question, word based searches (e.g., *werden* for the German *werden* passive) will be helpful. In other cases (e.g., the structure of the NP in different historical text types of German), suitable subcorpora will facilitate further research on algorithms for automatic retrieval and analysis.

(iv) *The modelling of linguistic change and of the processes of linguistic dissemination and evolution* requires large historical corpora. Past experiences, e.g. in the history of modal verbs, show that word senses are often specific to certain domains and that the modelling of paths of change requires very large and richly diversified corpora.

(v) *Other use cases* include research on historical terminology development, on the evolution of spelling variants, on the history of morphological forms and systems, on the language use of certain groups of persons (e.g. women), on the connection of language use and the history of ideas (e.g. pietism, rise and fall of phlogiston theory), the language of famous authors, and many others.

It is one of our aims in the future not only to contribute further texts together with their metadata to the distributed CLARIN-D infrastructure, but also to delineate a number of typical use cases that show how the texts, the metadata and forms of annotation can be put to use in order to work on specific research questions in the history of the German language.

[6] A literal english translation für *hat man* is *has one/one has*. Formulae of the type *Aus Prag hat man, dass* ... were common in Early New High German newspapers to indicate that news were based on some kind of source.

## 4. Technical aspects

As stated above a distributed infrastructure for what we call a *living text archive* of ENHG and oNHG should be designed as a collaborative, interoperable and sustainable platform where historical texts are integrated together with relevant metadata and expert computational technology.

Since it is beyond the scope of this paper to discuss all technical aspects in detail, we will focus on those aspects that seem crucial for the collaborative aspect of the envisioned infrastructure and mention other aspects only briefly. In the following, we discuss three central technical components: an agreed standard for corpus encoding (4.1), a web-based infrastructure for ensuring quality assurance, for the integration of legacy data and for further annotation of data (4.2), and a sampling procedure that extracts a reference corpus from the entire document collection (4.3). Furthermore, historical texts pose specific problems to full text retrieval. This is due to the absence of consistent orthographic conventions in historical texts, which presents difficulties for any system requiring reference to a fixed lexicon. These issues as well as software to overcome these difficulties are addressed, e.g., in Gotscharek et al. (2009) and Jurish (2010).

### 4.1 An agreed standard for corpus encoding

The set of annotation schemes developed by the Text Encoding Initiative (TEI; Burnard/Bauman 2012) is more and more wide-spread in current corpus projects and is going to become a de facto standard of corpus encoding for historical texts. Large infrastructure projects such as CLARIN-D and DARIAH recommend this de-facto standard. In its current version, TEI-P5 is a very flexible scheme that is adoptable for a large variety of text types. Due to this flexibility, TEI-P5 compliant corpora are generally not interoperable per se.[7] However, interoperability of corpora is one major backbone of a collaborative infrastructure on several levels. On the level of metadata, interoperability assures that texts can be uniformly processed and stored in central databases, thus providing a larger visibility of the corpus data. Standards such as OAI-PMH[8] are available for metadata exchange and have been adopted by many infrastructure projects. For object data, a common encoding facilitates the

[7] For a discussion of some of the problems arising from this fact cf. Unsworth (2011).

[8] Open Archives Initiative Protocol for Metadata Harvesting, a standard for the interchange of metadata.

exploitation of the text for further computational processing such as text mining or the annotation of the text with syntactic information. It also facilitates the common indexing of texts: texts encoded in a common way are searchable via search engines or federated search interfaces without further conversion work. Last but not least, texts with a common metadata and object data encoding can directly be stored in the repository of a larger corpus infrastructure, thus making corpus data sustainable.

How can the interoperability of corpus data be assured? First of all, it is important to state that the willingness to share data, to provide it with a license that enables the reuse of corpus data, is a prerequisite to the technical notion of interoperability. Second, interoperability of corpus data is comparatively new to corpus encoding. Up to the end of the 1990s, corpus compilation on the basis of the TEI was mainly a project-specific activity. Corpus documents were validated against a project-specific document grammar, possibly private character encodings were used, and the documents were transformed into proprietary formats in order to be indexed for full text retrieval. In that era of project-specific encoding, exchange of documents across projects was no goal per se and, in general, character encoding problems as well as differences in the document type grammar (DTD) were obstacles to a broader exchange of data. With the advent of XML and Unicode, documents encoded according to the recommendations of the TEI became interchangeable, but the problem of a lack of interoperability persisted due to the complexity and flexibility of TEI-P5. More recently, several attempts were made to increase the interoperability among different document collections by creating common formats. Subsets of TEI-P5 were created in such a way that the number of elements was largely reduced with respect to the full set of elements of the TEI Guidelines.[9] Also, the number of attributes and their corresponding values were restricted in order to obtain a better control of documents encoded in that format. Such formats – technically expressed as XML schemas – should allow for a basic structuring of all written texts and therefore serve as a starting point from which more detailed, possibly project-specific text structuring could start.

[9] The TEI recommends the definition of a subset of TEI elements appropriate to the anticipated needs of the project rather than to base the annotation of a corpus on the whole TEI tagset (Burnard/Baumann 2012: ch. 15.5) and promotes formats like TEI Tite (Trolard 2011), TEI Lite (Burnard/Sperberg-McQueen 2012) or the Best Practices for TEI in Libraries (TEI SIG on Libraries 2011). Other formats, such as TEI Analytics (Unsworth 2011, Pytlik-Zillig 2009), IDS-XCES (Institute for the German Language, Mannheim) and Textgrid's Baseline Encoding for Text Data in TEI P5 (Textgrid 2007-2009) were created.

The base format of the DTA project (henceforth DTABf) is a TEI-P5 subset that ensures the interoperability of corpus texts (cf. Geyken et al. 2012b). DTABf draws on the experiences of the DTA where a selection of 1,300 important texts of different text types (fictional, functional, and scientific texts), originating from the 17th to the 19th century, is currently being digitized and annotated. Linguistic analyses are added to the digitized text sources in a stand-off format for further corpus research. The tagset of the DTABf is a strict subset to the TEI-P5 guidelines, i.e., no new elements or attributes were added to the TEI-P5 tagset. It consists of about 80 TEI-P5 elements needed for the basic formal and semantic structuring of the DTA reference corpus. The purpose of the DTABf is to provide a faithful page per page presentation of the entire works and to maintain coherence on the annotation level (i.e., similar structural phenomena should be annotated similarly).

The DTABf attempts to meet the criteria of interoperability mentioned by Unsworth (2011) in that it "focuses on non-controversial structural aspects of the text and on establishing a high quality transcription of that text". Therefore, the goal of the DTABf is to provide as much expressiveness as necessary by being as precise as possible. For example, DTABf is restrictive not only considering the selection of TEI-elements but also with respect to attribute-value pairs, and allows only a limited set of values for a given attribute. Unlike initiatives such as TEI Analytics (as presented in Pytlik-Zillig 2009), the goal of DTABf is not to build a schema that validates as many cross-collections as possible but to convert resources from other corpora so as to keep the structural variation as small as possible.

The necessity of a common standardized format for the annotation of printed texts seems to be opposed to the fact that different projects usually have different needs as to how a corpus may be exploited. Therefore annotation practices vary according to the variable queries on a certain corpus. This problem may be addressed by defining different *levels of text annotation* that represent different text structuring depths. The TEI Recommendations for the Encoding of Large Corpora foresee four different levels of annotation defining required, recommended, optional, and proscribed elements.[10]

The DTABf consists of such annotation levels, which serve as classes subsuming and, by that, categorizing all available 'base format' elements:

[10] Cf. www.teic.org/release/doc/tei-p5-doc/en/html/CC.html.

- Level 1/required: elements that are mandatory for the basic semantic structuring of corpus texts.
- Level 2/recommended: elements that are recommended for the semantic structuring of corpus texts. These elements are systematically used in the DTA-corpus.
- Level 3/optional: elements that need not be considered for text annotation. Level 3 elements are not (yet) part of the DTA guidelines and are therefore not used systematically in the texts of the DTA corpus. They are, however, compatible with the DTA schema.
- Level 4/proscribed: elements that were explicitly excluded from the DTA guidelines. They should be avoided in favour of the solutions offered in the DTA guidelines.

## 4.2 An infrastructure for the lifecycle of a digital text

The second step towards an interoperable corpus platform is, from a technical point of view, to provide software enabling an easy integration of legacy data into the DTABf as well as to correct the texts in case their quality does not meet the criteria established by the community.

DTABf is a flexible format in the sense that it consists of mandatory, recommended, and optional criteria. Hence, software that validates legacy data against DTABf can do this on those three levels. Thus, legacy texts can be integrated in the platform even though they are possibly not DTABf compliant on all levels. Currently, two generic conversion tools are provided in the CLARIN-D context: a generic web-based software (TEI-Integrator, Th. Eckart, Univ. Leipzig[11]) that assists the user in the conversion and the upload process of legacy data, and a specific oXygen-Framework where any TEI-P5 compliant text can be validated against the DTABf and a GUI is used to assist the user in evaluating the amount of work needed to convert the legacy document into a valid DTABf document.[12]

Since it can be expected that legacy data from heterogeneous origins do not meet all the required criteria, a collaborative platform is needed where uploaded texts can be proofread and evaluated. Since generic collaborative proof-reading platforms did not exist for TEI-P5 texts, such a platform was

[11] http://clarin.informatik.uni-leipzig.de/program.

[12] http://lecture2go.uni-hamburg.de/konferenzen/-/k/13952.

implemented for the DTA (DTAQ).[13] Apart from the proof-reading facilities where the user has the possibility to check the text for errors against the image page per page, the user is also provided with several presentation formats of the text: the original XML/TEI format, a formatted HTML presentation, a pure text format, and a normalized view of the text after being processed by a lemmatizer of historical German word forms (Jurish 2010).

The third issue of the corpus management infrastructure concerns the fact that an electronic document is always 'living', i.e., it is always subject to further correction and structural or semantic annotation. From the point of view of the 'living archive' it is important to note here that additional annotations are not only carried out for a specific research but can also be valuable for other researchers. Therefore, these additional annotations (as far as the DTABf is concerned) should be carried out 'within' the archive, thus enabling other researchers to benefit from previous annotations. In addition, such a platform should integrate computational tools such as Named-Entity-Recognizers, tools for statistic computations, or tools for the automatic analysis of citations. Such tools are currently in preparation in the CLARIN-D infrastructure and will be accessible to the DTAQ platform as web-services. Finally, the versioning of the documents as well as the use of persistent identifiers play a role here since stable references to the documents are a necessary requirement for any sustainable infrastructure. Solutions for these problems exist and are currently implemented on a larger scale for infrastructure projects such as CLARIN[14].

### 4.3 Corpus chooser

The result of the previous steps is an interoperable distributed corpus repository annotated with more or less finegrained metadata information. This repository is not per se a reference corpus. As stated above, there may be more than one reference corpus at hand depending on the research purpose. In order to establish a subset of the repository that qualifies as a reference corpus for a given research task, sampling procedures are needed that exploit the metadata in such a way that an optimal subcorpus of the repository can be extracted. There are at least two possible ways to deal with these sampling procedures: either as a corpus bitmap where any user can choose the appropriate "virtual" subcorpus for her or his research (Kupietz et al. 2010), or as a

[13] www.deutschestextarchiv.de/dtaq.

[14] www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf.

complex set of criteria chosen by the user beforehand, which is then computed by a sampling procedure trying to find an optimal subcorpus for these criteria (cf. Geyken 2007).

We illustrate this method with the example of building a maximally balanced corpus out of the entire text collection that is equally distributed over all the decades of the 15th to the 19th century. The balance of the selection follows a previously determined distribution of text types. In order to determine the maximally balanced corpus, we identify the decade with the least number of tokens. The sampling process then calculates the expectation value according to the text type distribution. For practical reasons, this algorithm cannot be applied strictly since the decades are not equal from the 15th to the 19th century; e.g., in the period of the Thirty Years War there was a much sparser text production than in the 1890s. Therefore we assume that the token number for some decades can deviate from the mean value. The deviation depends on the difference between the minimally balanced corpus and the corpus size to be attained. Moreover, the sampling procedure begins with a so-called 'initial corpus'. The initial corpus consists mainly of texts by major writers and scientists as well as texts that are considered to be of high interest. This guarantees that works by Goethe, Marx, Büchner, Boltzmann or Liebig will be present independently of the sampling procedure whereas, for example, serial corpus texts such as newspaper articles or exchange of letters are selected randomly. Of course, the initial corpus must not exceed the boundaries of the required text distribution, i.e., it will have to be ensured beforehand that no decade and no text class in the start corpus exceed the required token number. Further sampling strategies will be demonstrated in the future with respect of specific use cases.

## 5. Organizational aspects of collaboration on a living text archive

A living text archive basically needs (i) contributors who are willing to share, and to work on, textual resources, and (ii) an infrastructure with a core team whose members organize collaboration on a long term basis.

(i) From a contributor's point of view, the main question is why she or he should share textual resources. At present, there are three partial answers to this question.

- We aim to create a system of 'reputation' for making resources available, including personal ascription in the metadata of work that has been donated to the archive and the possibility to produce contributor profiles. The success of such mechanisms depends on the importance that is granted to these contributions to infrastructures in comparison to traditional publications.[15]
- We are able to display the texts that have been given to the archive *locally* via a JSON interface. In this way texts can be consulted both in the large archive *and* on the websites of research teams, universities etc. in a professional way, together with expert corpus technology.
- There is also a moral aspect to this issue: If we as researchers want *huge* corpora, we all have to contribute by integrating into a common infrastructure the data many of us have produced. On the other hand, sharing one's resources is not only altruistic but is also a way to make this work visible to others and to receive recognition for it. Thus, sharing produces a win/win-situation.

In addition, to share texts together with different annotations allows to combine research aspects beyond the interests of the original annotators. E.g., to project a syntactic annotation of text X by person A onto an annotation of textual structures of text X by person B allows to describe typical syntactic structures of textual elements in a given text of a certain text type. At present, this type of collaborative use of different annotations is rare in research on the history of German.

(ii) A new culture of sharing and collaborating in a living text archive needs infrastructure centers with core teams whose members organize collaboration on a long term basis. Among the ongoing tasks of such a core team are: to catalogue newly available digital resources in a canonical way; to evaluate and to acquire new digital texts; to support and advice in conceptual, technical, and legal matters; to adjust the principles of work to new developments; to take care of and to advance corpus technology for his-

[15] The German Council of Science and Humanities (Wissenschaftsrat) is encouraging the scientific community to count contributions to digital infractures as ordinary publications: "Auch um die Behebung des Reputationsdefizits ist der Wissenschaftsrat bemüht, wenn er anregt, Infrastrukturarbeit als eigenständige Forschungsleistungen nicht anders als Publikationen zu bewerten. Bei der Besetzung wichtiger Posten sollen technische und wissenschaftliche Qualifikation von gleichem Gewicht sein." (www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html).

torical texts; to provide best practices, use cases and strategies for the use of subcorpora for specific research questions; etc. We hope that the CLARIN-D curation project will demonstrate the fruitfulness of such a long term living archive of ENHG and oNHG.

## References

Budin, Gerhard/Kabas, Heinrich/Mörth, Karlheinz (2012): Towards finer granularity in metadata. Analysing the contents of digitised periodicals. In: Journal of the Text Encoding Initiative 2 (February 2012): 1-8. http://jtei.revues.org/416.

Burnard, Lou/Bauman, Syd (2012): P5: Guidelines for electronic text encoding and interchange, Version 2.1.0, June 17th, 2012. www.tei-c.org/release/doc/tei-p5-doc/en/html/.

Burnard, Lou/Sperberg-McQueen, Michael C. (2012): TEI Lite: Encoding for Interchange: an introduction to the TEI. Final revised edition for TEI P5, August 2012. www.tei-c.org/Guidelines/Customization/Lite/.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (ed.): Collocations and idioms: linguistic, lexicographic, and computational aspects. London: Continuum, 23-41.

Geyken, Alexander/Gloning, Thomas/Stäcker, Thomas (2012a): Compiling large historical reference corpora of German: quality assurance, interoperability and collaboration in the process of publication of digitized historical prints. Panel DH2012, Hamburg. [Abstract and video lecture].

Geyken, Alexander/Haaf, Susanne/Wiegand, Frank (2012b): The DTA 'base format': A TEI-subset for the compilation of interoperable corpora. In: Jancsary, Jeremy (ed.): 11th Conference on Natural Language Processing (KONVENS): Empirical Methods in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2012. (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). Wien: ÖGAI, 383-391. www.oegai.at/konvens2012/proceedings.pdf#page=383.

Geyken, Alexander/Gloning, Thomas/Kupietz, Marc/Stäcker, Thomas/Thomas, Christian/Witt, Andreas (2012c): Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur, Vorhabensbeschreibung für ein Kurationsprojekt der F-AG 1 Deutsche Philologie. www.deutschestextarchiv.de/doku/clarin_kupro_index.

Geyken, Alexander/Haaf, Susanne/Jurish, Bryan/Schulz, Matthias/Thomas, Christian/Wiegand, Frank (2009): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. In: Jahrbuch für Computerphilologie 2009. http://computerphilologie.tu-darmstadt.de/jg09/geykenetal.html.

Gotscharek, Annette/Neumann, Andreas/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus (2009): Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data. New York: ACM, 69-76.

Haaf, Susanne/Wiegand, Frank/Geyken, Alexander (2012): Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text. In: Journal of the Text Encoding Initiative 4. http://jtei.revues.org/pdf/739.

Hoffmann, Walter (1998): Probleme der Korpusbildung in der Sprachgeschichtsschreibung und Dokumentation vorhandener Korpora. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (eds.): Sprachgeschichte. Zweite, vollständig neu bearbeitete und erweiterte Auflage. Volume 1. Berlin/New York: de Gruyter, 875-889.

Jurish, Bryan (2012): Finite-state canonicalization techniques for Historical German. PhD thesis, Universität Potsdam. http://opus.kobv.de/ubp/volltexte/2012/5578/pdf/jurish_diss.pdf.

Jurish, Bryan (2010): More than words – using token context to improve canonicalization of Historical German. In: Journal for Language Technology and Computational Linguistics (JLCL) 25/1: 23-39.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010). Paris: ELRA, 1848-1854.

Pytlik-Zillig, Brian (2009): TEI Analytics: converting documents into a TEI format for cross-collection text analysis. In: Literary and Linguistic Computing 24(2): 187-192. doi:10.1093/llc/fqp005.

Sinclair, John (2005): Corpus and text – basic principles. In: Wynne, Martin (ed.): Developing linguistic corpora: a guide to good practice. Oxford: Oxbow Books: 1-16.

TextGrid (2007-2009): TextGrid's baseline encoding for text data in TEI P5. www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf.

Unsworth, John (2011): Computational work with very large text collections. Interoperability, sustainability, and the TEI. In: Journal of the Text Encoding Initiative 1. http://jtei.revues.org/pdf/215.