Fivehundredmillionandone Tokens. Loading the AAC Container with Text Resources for Text Studies.

Hanno Biber, Evelyn Breiteneder
Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, 1010 Wien
E-mail: hanno.biber@oeaw.ac.at, evelyn.breiteneder@oeaw.ac.at

Abstract

The "AAC - Austrian Academy Corpus" is a diachronic German language digital text corpus of more than 500 million tokens. The text corpus has collected several thousands of texts representing a wide range of different text types. The primary research aim is to develop text language resources for the study of texts. For corpus linguistics and corpus based language research large text corpora need to be structured in a systematic way. For this structural purpose the AAC is making use of the notion of container. By container in the context of corpus research we understand a flexible system of pragmatic representation, manipulation, modification and structured storage of annotated items of text. The issue of representing a large corpus in formats that offer only limited space is paradigmatic for the general task of representing a language by just a small collection of text or a small sample of the language. Methods based upon structural normalization and standardization have to be developed in order to provide useful instruments for text studies.

Keywords: text corpora, literary studies, corpus linguistics

1. The AAC Container

Figure 1: Loading

The "AAC-Austrian Academy Corpus" is a large digital text corpus operated by the "Institute for Corpus Linguistics and Text Technology" of the "Austrian Academy of Sciences" in Vienna. The texts integrated into the AAC stemming from the last 150 years are predominantly German language texts and are of considerable historical and cultural significance. Thousands of language documents and literary objects by thousands of authors have been collected, representing an astonishing range of different text types from all over the German speaking areas.

Among the AAC's sources, which cover manifold domains and genres, there are literary journals, newspapers, novels, dramas, poems, advertisements, essays, travel accounts, cookbooks, pamphlets, political speeches as well as plenty of scientific, legal, religious texts. The historical period covered by the corpus is ranging from the 1848 revolution to the fall of the iron curtain in 1989. In this period significant historical changes with remarkable influences on the language and the language use can be observed. The AAC corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the linguistic and textual properties of these texts.

More than 500 million running words of text have already been scanned, converted into machine-readable text and carefully annotated and basic structural mark-up and selected thematic mark-up has been applied according to annotation and mark-up schemes based upon XML related standards. The AAC's text technology working group has been working on issues of digitization of historical language data, establishing efficient workflows as well as developing usable software.

While the original objectives of the build-up phase of the project were focused on issues of corpus creation, the application phase will see more work on analysis and exploitation of textual resources. The work will be to adapt existing resources, to develop new resources where necessary and to document and describe these resources in a manner that enables users from different backgrounds and disciplines to do research with textual resources. These endeavors can be regarded as basic research in humanities methodologies.

The AAC has followed text-oriented concepts attaching great importance to a perspective that does not allow the reduction of text resources to only collections of words or sentences. Without decent instruments that provide the user of large language resources and text corpora with structured access to and historically correct information about these text documents in which the language data is contained, knowledge and insights about these resources will be problematic.

Figure 2: Problems

Research will have to focus on methods and resources for making large amounts of texts accessible in a well-structured way. Efforts are made to develop usable tools, attempting to add to them wherever necessary, to provide and to promulgate relevant expertise while building up what we call the AAC Container, a systematic central and well-structured access point to the holdings of the entire corpus.

In contrast to other corpus-oriented projects, the AAC proceeds from literary and text lexicographic research. Corpus research and the creation of large electronic text collections has traditionally been the domain of linguists. Literary digitization initiatives are often restricted to particular writers and many of these projects did neither produce large amounts of data nor pursue research on methods of how to tackle the problems involved in working with such data. Being aware of the need of digital resources in many fields of the humanities, the AAC has started to work on applications, tools and methods geared towards a wider range of applications trying to pursue a path of text-oriented computing. While the needs of linguists have not been ignored, they tried to work towards applications that also offered access to coherent texts, convinced that for many applications it is indispensable for researchers to have access to the text as such.

From the start, the AAC relied on XML technologies as the foundation of their corpus build-up activities. The AAC tools support this technology which allows both the controlled application of markup as well as automated validation of large amounts of data. The AAC applies an encoding scheme characterized by a combined approach to capture both structural features of the texts as well as a certain amount of data describing the physical appearance of the original texts. This approach has led to a hybrid system of markup not only representing the basic semantic structures of the texts but also a certain amount of layout information. In digitizing historical data, semantic and presentational data will remain intertwined. When working on large amounts of such texts, it is considerably easier to capture formal data than to translate typographic idiosyncrasies into consistent structural markup.

The AAC's digitizing activities are characterized by a strong connection to the physical objects of digitization. This may be seen as one of the motives behind the one-page-one-file principle, which implies that each page of a printed publication is stored as a separate digital object in the digital medium, which form larger digital objects achieved by means of implicit and explicit linking that in turn represent coherent texts. By doing so, the AAC attaches importance to the semantic structures of the text as well as to the physical appearance of the text. The output is visualized via XSLT in browsers and encoding tools. It contains precise specifications and explanations of the elements and attributes that make up the system, and furnishes numerous examples intended to help users to correctly apply mark-up.

Metadata describing production processes and details about the physical sources of the digitized object are the backbone of a digital data collection's usability.The AAC collects two types of data that fall into these categories. The first consists in descriptive metadata concerning the digitized objects. This information has been drawn-up on a regular basis when the physical objects were scanned. It is stored in a relational database containing around 6000 records holding all relevant information about the physical objects so far ingested into the corpus. In many cases this data is much more detailed than regular library records. The record fields in this database were designed in a way that they can be easily mapped onto the fields of TEI headers. Mechanisms exists that allow to associate these with the relevant documents, either by storing them as stand-alone documents, or inserting them as TEI headers into the respective XML documents. To access and display the underlying data in a comfortable manner, it was necessary to find an adequate display mode. Having established a working infrastructure for the digital texts available, the AAC has been developing more sophisticated methods of utilizing large scale corpora on the basis of various database systems as well as XML-aware indexing tools to establish standard procedures of accessing large XML text repositories.

Two AAC projects have to be highlighted in connection with our aims to develop the AAC container infrastructure for text studies, the digital editions of the literary journals "Die Fackel" and "Der Brenner", the AAC-FACKEL and BRENNER ONLINE. How can a valuable historical text source based upon the principles of corpus research be used for research in the field of literary studies and linguistics?

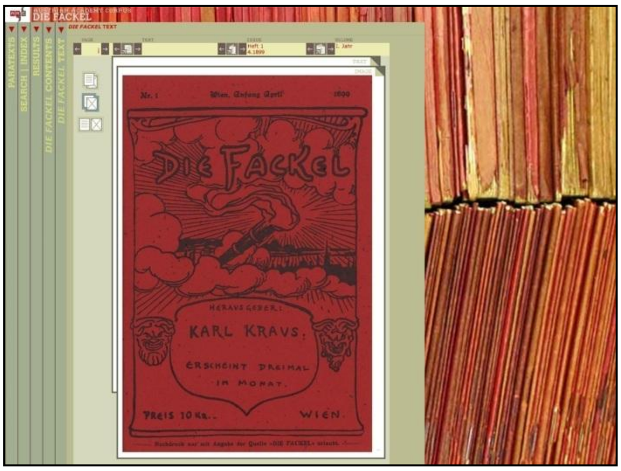
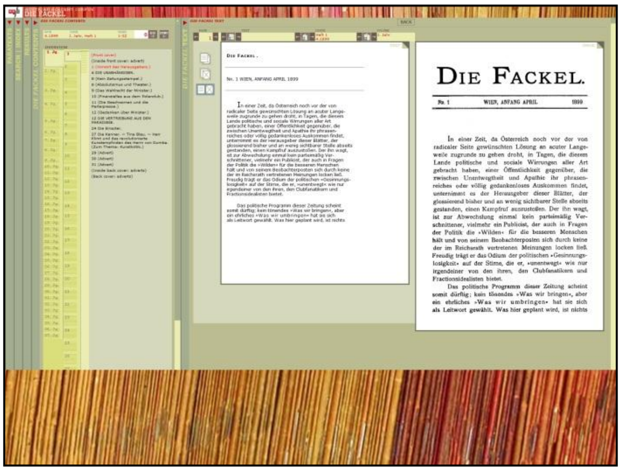


Figure 3: AAC Fackel - Cover Page



DIE FACKEL.

Nr. 1 WIEN, ANFANG APRIL 1899

In einer Zeit, da Österreich noch vor der von radicaler Seite gewünschten Lösung an acuter Langeweile zugrunde zu gehen droht, in Tagen, die diesem Lande politische und sociale Wirrungen aller Art gebracht haben, einer Öffentlichkeit gegenüber, die zwischen Unentwegtheit und Apathie ihr phrasenreiches oder völlig gedankenloses Auskommen findet, unternimmt es der Herausgeber dieser Blätter, der glossierend bisher und an wenig sichtbarer Stelle abseits gestanden, einen Kampfruf auszustoßen. Der ihn wagt, ist zur Abwechslung einmal kein parteimäßig Verschnittener, vielmehr ein Publicist, der auch in Fragen der Politik die »Wilden« für die besseren Menschen hält und von seinem Beobachterposten sich durch keine der im Reichsrath vertretenen Meinungen locken ließ. Freudig trägt er das Odium der politischen »Gesinnungslosigkeit« auf der Stirne, die er, »unentwegt« wie nur irgendeiner von den ihren, den Clubfanatikern und Fractionsidealisten bietet.

Das politische Programm dieser Zeitung scheint somit dürftig; kein tönendes »Was wir bringen«, aber ein ehrliches »Was wir umbringen« hat sie sich als Leitwort gewählt. Was hier geplant wird, ist nichts

Figure 4: AAC Fackel - Text Page

The edition interfaces, which have been designed by Anne Burdick, have five frames synchronized within one single window. The frames can be opened and closed as required. The paratext section situated within the first frame provides additional information about the background of the edition and scholarly essays about the journal. The contents section provides access to the whole run of issues and all of the contents of the journal in chronological order. The text section has a complex and powerful navigational bar at the top so that the reader can easily navigate and read within the journal either in text-mode or in image-mode from page to page, from text to text, from issue to issue, and with the help of hyperlinks.

Creating an intuitive navigable structure in retrodigitized historical journals poses a considerable number of problems on top of which is the compilation of contents lists. These contents lists were created by the literary scholars and based upon original material retrieved from library sources. The readers can navigate the texts in four ways in the digital editions: from page to page, from text to text, from issue to issue, from volume to volume. The readers can read the text page by page as digital text or as facsimile image. The search and index section gives access to a variety of indexes, databases and full-text search mechanisms. The results of these queries and lists are displayed in the adjacent results section. These digital editions function as models for similar applications for the access to corpus based text studies provided for scholars and the interested public alike. They provide well-structured and well-designed access to the sources. After a thorough survey of all periodicals mentioned in the journal, this data was disambiguated and enriched with external data. And linguistic data produced were restricted to lemmatization and the assignment of word class information and POS and lemma data was created by a standard tagger, as described in the paratexts.

Database records have been created for each token which in turn not only contain linguistic data but also information as to the position within the original document, whereby it is possible to refer from the search results back into the texts as well as to allow complex queries including range operators.

These tools as well as other ones developed for the AAC corpus have to be seen as work for the AAC Container, by which the working group understands the framework for the future presentation of the entire corpus. It should be capable of performing general purpose word form and searches and related searches and can at the same time manage any type of token attributes thereby making the instrument highly adjustable to various tasks. Style sheet transformations can be applied to particular parts of documents, such as a particular paragraph, which is a useful feature when working with larger text documents.

The requirement specification for these access tools determine an application which has several components, for input queries and to display the results in form of KWICs which also provide the essential metadata such as information on sources, text types, authors, date of publication, historical origin, modifications and so on. In order to display the related pages of the results as digital text, it should also allow access to the XML source of the texts and to define custom style sheets, to view facsimiles of the texts or to display metadata of the digital objects. A unit to query the metadata database is necessary as well as the navigational control, offering random access to the book, the journal issue, the collection subsection etc. To do literary and historical research it is indispensable to have access to the complete texts in their entirety. Therefore, access is offered to the corpus not only through query result lists but also through a kind of library unit which allows readers to navigate to any desired part of the corpus.

The text corpus has collected several thousands of texts representing a wide range of different text types. The issue of representing a large corpus in formats that offer only limited space is paradigmatic for the general task of representing a language by just a small collection of text or a small sample of the language. The primary research aim within the framework of the "AAC - Austrian Academy Corpus" is to develop text language resources for the study of texts. For corpus linguistics and corpus based language research large text corpora need to be structured in a useful and systematic way, in which methods based upon structural normalization and standardization have to be utilized in order to provide suitable instruments for text studies. For this structural purpose the AAC is making use of the conceptual notion of container, in the context of corpus research a flexible system of pragmatic representation, manipulation, modification and structured storage of annotated items of text.

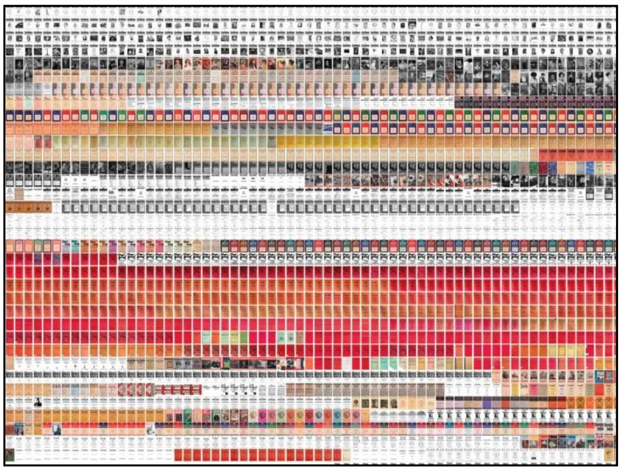


Figure 5: AAC Magazines

2. References

AAC-Austrian Academy Corpus: AAC-FACKEL. Online Version: "Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936". AAC Digital Edition No 1 (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörth), http://www.aac.ac.at/fackel

AAC-Austrian Academy Corpus and Brenner-Archiv: BRENNER ONLINE. Online Version: "Der Brenner. Herausgeber: Ludwig Ficker, Innsbruck 1910-1954", AAC Digital Edition No 2, (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörth), http://www.aac.ac.at/brenner

Mörth, Karlheinz (2002): The representation of literary texts by means of XML: some experiences of doing markup in historical magazines. In: Michael Fraser, Nigel Williamson and Marilyn Deegan (Eds.), *Digital Evidence. 2002. Selected papers from DRH 2000, Digital Resources for the Humanities Conference*. Office for Humanities Communication 14, pp. 17-32.