The AAC [Austrian Academy Corpus]
An Enterprise to Develop Large Electronic Text Corpora

Dr Hanno Biber, Dr Evelyn Breiteneder

AAC - Austrian Academy Corpus, Austrian Academy of Sciences

Sonnenfelsgasse 19/8, 1010 Vienna, Austria

Hanno.Biber@oeaw.ac.at, Evelyn.Breiteneder@oeaw.ac.at

Abstract

The AAC [Austrian Academy Corpus] is a corpus research institution based at the Austrian Academy of Sciences in Vienna. The AAC is a very large and complex electronic text collection. Its aims are to create an innovative text corpus and to conduct scholarly and scientific research in the field of electronic text corpora. In the first phase of the corpus build up the AAC is committed to have at least 100 million running words of carefully selected and scholarly annotated significant texts. The corpus approach of the AAC will allow a variety of investigations into the linguistic properties, the textual structures and the historical and literary significance of the selected texts. In the second phase of application development the size of the AAC will increase to around one billion running words. In this phase selected subcorpora will be annotated in greater detail following the AAC schemes for annotation and according to its editorial principles. The AAC working group is endeavouring to establish a corpus that meets the needs of textual studies and conveys essential information about the German language as well as about the history of the time in focus as a history of texts and of language.

1. The AAC [Austrian Academy Corpus]

The AAC is a corpus research institution based at the Austrian Academy of Sciences in Vienna. The AAC is a very large and complex electronic text collection. The primary aims of the AAC are to create an innovative and experimental text corpus of significant texts as well as to conduct scholarly and scientific research in the field of electronic text corpora. The texts selected for inclusion into the corpus date from the period between the 1848 Revolution and the fall of the Berlin Wall in 1989. The texts will be predominantly German language ones, but other specific parallel corpora and multimedia collections will be included.

Figure 1: AAC

AAC Corpus Build Up

In the first phase of the corpus build up, which will be completed by the end of the year 2005, the AAC is committed to have at least one hundred million running words of carefully selected and scholarly annotated significant texts ready in digital format. At present, however, we have already approximately two hundred million running words to hand. The corpus approach of the AAC will allow a variety of investigations to be carried out into the linguistic properties, the textual structures and the historical and literary significance of these texts. In the second phase of application development the size of the AAC will increase ten-fold to around one billion running words at the end of the year 2010. In this phase selected subcorpora will be carefully annotated in greater detail. The annotation process will be following the AAC schemes for annotation and mark up, and it will be done according to the AAC's editorial policies and principles. The AAC working group, who have had expertise in linguistics and literary studies, in computer supported lexicographic

research and other related fields, are endeavouring to establish a corpus that meets the needs of textual studies. At the same time the corpus will convey essential information about the German language used in the selected texts as well as about the history of the time in focus, which is to be regarded as a history of language and a history of texts.

AAC Selection of Texts

The sources of the AAC will stem from a variety of different fields and domains which reflect not only linguistic and literary but also historical and cultural processes. The AAC functions as a text research institution and as an example of an experimental corpus that is designed for use in scholarly textual studies. The AAC will provide a highly developed computational infrastructure in order to discover, structure and deliver information about the texts themselves as well as about the processes and phenomena to be observed in these sources. For this reason the AAC aims to include a wide range of text types from various cultural domains. All these texts will be carefully selected as being of key historical significance and as highly culturally relevant. The selected texts will represent a variety of genre and different text types, such as articles, essays, letters, poems, novels, anecdotes, funeral sermons and electoral speeches, propaganda slogans and advertising slogans, pop song lyrics and political speeches, comic books, instructions, travel guides, programmes, mailing catalogues and many other text types. The intention is to digitally present a wide selection of different sources of scholarly, literary, journalistic, scientific, political texts which exercised considerable influence over the last one hundred and fifty years. Texts will be taken into account that were read, published or generally regarded as being of importance at the time. This does not imply that the texts must be of Austrian origin or must only be original German language texts. The AAC will take into consideration texts that were in some way or another culturally relevant at the time and are at the same time corresponding to the thematic selection principles defined of the AAC. The texts are being digitized and the electronic text will be annotated and made accessible and searchable by means of the Extensible Markup Language XML. Digital images of the individual pages of the source texts will be included in the corpus. This will be done so because the original graphical and typographical information fixed on the paper of the original publication is of considerable importance for conveying the meaning and for allowing adequate interpretations of the texts. Therefore the AAC is also concerned with the representation, description and annotation of the graphic, typographic and specific conceptual properties of the texts in question.

2. Examples of the AAC Collections

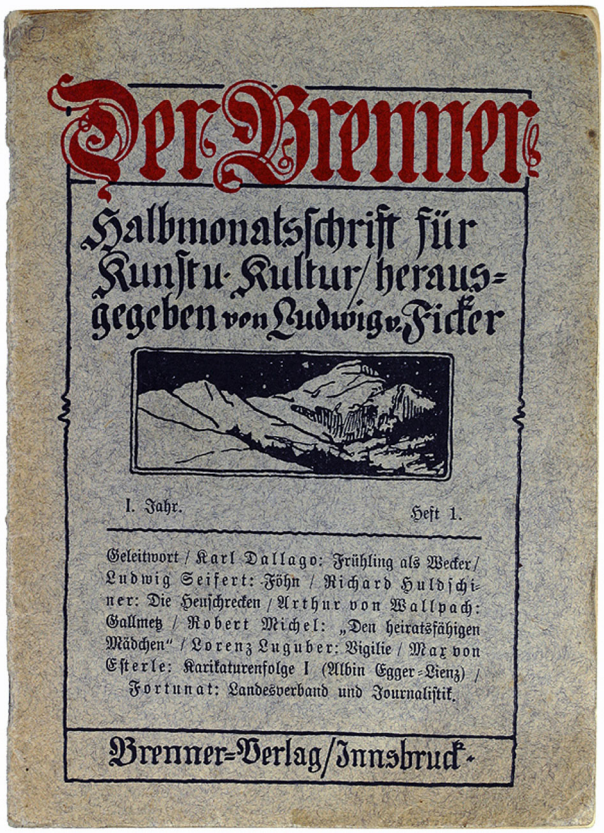
Example 1: AAC Literary Journals Subcorpora

The AAC will consist of a variety of different sources. The question of the graphic qualities, of contexts and arrangements as well as the compositional aspects of the texts is particularly relevant and interesting in the case of complex text structures such as newspapers or literary journals. Literary journals and in a similar way newspapers comprise a whole variety of functionally different text types within their specific structures. The AAC has, among many other sources, fully digitized and structurally integrated several influential and notable literary and political journals of the last one hundred and fifty years.

Five examples of the AAC's literary journals subcorpora holdings will suffice: A journal of the theatrical and political worlds of the first decades of the 20th century has been selected, "Die Schaubühne" (1905-1918), which was later called "Die Weltbühne" (1918-1933), and was published in Berlin. It calls to mind three names, Siegfried Jacobson, the founder of the journal and respected theatre critic, the author and short-time editor Kurt Tucholsky and finally Carl von Ossietzky, the famous pacifist and political journalist. Of the many journals included, the important expressionist journal "Der Sturm" (1910-1930), published by Herwarth Walden, and Franz Pfemfert's radical political magazine

"Die Aktion" (1911-1918), which were likewise published in Berlin, have to be mentioned here.

The Austrian journal "Der Brenner" (1910-1954), published in Innsbruck by Ludwig von Ficker roughly around the same time, will be included. The electronic edition of this literary journal will be the first example of a full-text online publication based upon the philological editorial and design principles of the AAC.


Der Brenner

Halbmonatsschrift für Kunst u. Kultur / herausgegeben von Ludwig v. Ficker

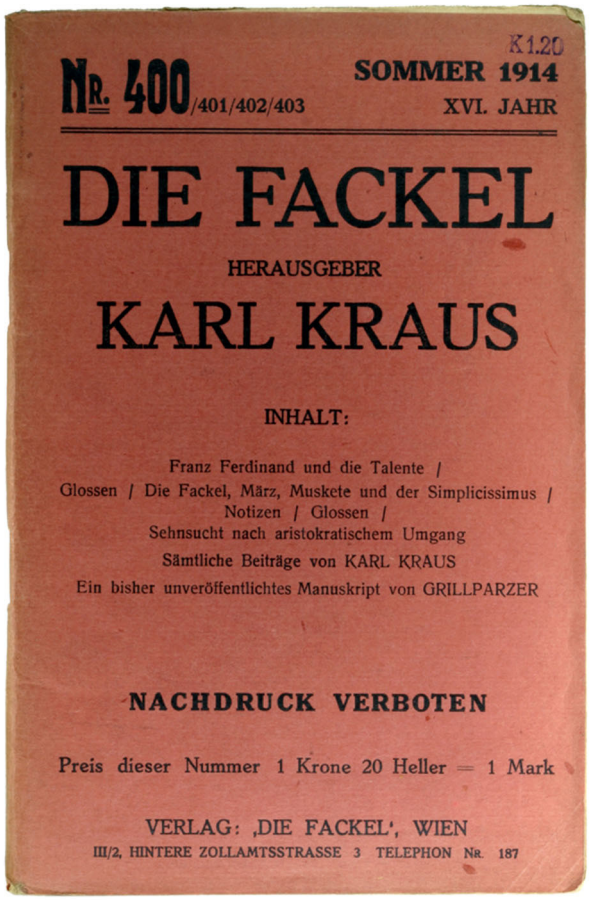
I. Jahr. Heft 1.

Geleitwort / Karl Dallago: Frühling als Wecker / Ludwig Seifert: Föhn / Richard Huldschiner: Die Heuschrecken / Arthur von Wallpach: Gallmetz / Robert Michel: „Den heiratsfähigen Mädchen" / Lorenz Luguber: Vigilie / Max von Esterle: Karikaturenfolge I (Albin Egger-Lienz) / Fortunat: Landesverband und Journalistik.

Brenner-Verlag / Innsbruck

Figure 2: "Der Brenner"

The famous satirical journal published by Karl Kraus in Vienna, has had considerable influence not only upon all these journals, but upon many authors of the time in Austria and Germany. On more than twenty two thousand pages in his journal Karl Kraus criticised, analysed, commented, parodied, quoted or polemicized against texts by others, texts of journalistic, political or literary origin, of newspapers or other publications. In regard of its satirical and critical potential "Die Fackel" (1899-1936) will constitute the core of the AAC holdings and will be a starting point for future selections of texts.


K 1.20

Nr. 400/401/402/403 SOMMER 1914 XVI. JAHR

DIE FACKEL

HERAUSGEBER

KARL KRAUS

INHALT:

Franz Ferdinand und die Talente / Glossen / Die Fackel, März, Muskete und der Simplicissimus / Notizen / Glossen / Sehnsucht nach aristokratischem Umgang

Sämtliche Beiträge von KARL KRAUS

Ein bisher unveröffentlichtes Manuskript von GRILLPARZER

NACHDRUCK VERBOTEN

Preis dieser Nummer 1 Krone 20 Heller = 1 Mark

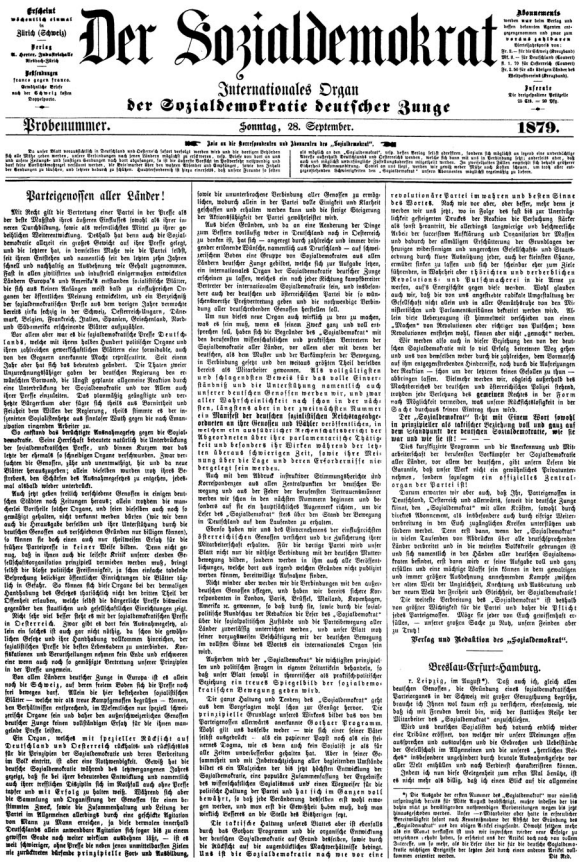
VERLAG: ‚DIE FACKEL', WIEN

III/2, HINTERE ZOLLAMTSSTRASSE 3 TELEPHON Nr. 187

Figure 3: "Die Fackel"

Example 2: AAC Newspaper Subcorpora

The AAC will include several runs of influential German language newspapers, for example the newspapers published in Vienna such as "Neue Freie Presse", "Arbeiter-Zeitung" or "Reichspost", as well as the German newspaper "Berliner Tageblatt" and the Bohemian German language newspaper "Prager Tagblatt". The historical newspaper subcorpora constitute only a segment of the AAC. We have chosen to integrate selected issues and runs of newspapers that are acknowledged to show and have been analysed as of being in thematic connection with other texts present in the AAC. "Die Fackel" for example frequently and satirically comments upon the press of the day and the journalistic discourse of the time. The politically highly influential paper for the social democratic movement of the late 19th century in the German speaking countries, "Der Sozialdemokrat" (1879-1890), is being prepared as another complete electronic edition of the AAC.



Der Sozialdemokrat

Internationales Organ der Sozialdemokratie deutscher Zunge

Probenummer. Sonntag, 28. September. 1879.

Parteigenossen aller Länder!

The AAC has developed a policy of selection which is based upon thematic interests that are pursued by the working group and their scholarly and scientific fields of research, by empirical and by technological parameters. At the AAC, all texts are digitized and thoroughly annotated by means of XML in order to facilitate systematic investigations and efficient research into the textual qualities of the corpus holdings. How modern technology and standards can be integrated in traditionally-oriented fields of research such as linguistics and literary and cultural studies will be exemplified through the digitization projects and digital editions of the AAC.

http://www.aac.ac.at

Figure 4: "Der Sozialdemokrat"