TITLE: **Austrian Academy Corpus Digital Resources and Textual Studies**
AUTHORS: **Hanno Biber, Dr Evelyn Breiteneder, Dr Karlheinz Moerth**

ABSTRACT:
The Austrian Academy Corpus (AAC) is a newly founded institution based at the Austrian Academy of Sciences in Vienna. It was designed to set up a text corpus and to conduct research in the field of electronic text corpora. The electronic text collections established at the AAC so far and its future projects will mainly focus on electronic representations not only of literary texts, literary magazines, journals and newspapers but also on a carefully considered selection of texts from many cultural and social domains. The aim of the proposed paper is to investigate the potential of digital resources for textual studies in various fields of the humanities. The paper will consider the advance of new systems of digital representation and its implications for the study of language, literature and cultural history. The paper will show the range of interests pursued in the AAC research group. It will be concerned with the general organisational structures of the AAC, the specific selection criteria for the great variety of texts which will form the AAC, and finally, examine practical issues in digitising the magazine “Die Weltbühne”, giving special attention to the applicability of XML Schemas in literary computing.
BIOGRAPHY:
AAC- working group editors: Hanno Biber, Evelyn Breiteneder, Karlheinz Moerth
The AAC working group has had expertise in digital text studies and in lexicography for more than ten years. The "Dictionary of Idioms" (Wörterbuch der Redensarten) was compiled by the editors, edited by Werner Welzig and published by the Austrian Academy Sciences Press in 1999.

PAPER PROPOSAL SUBMISSION:
TITLE: Austrian Academy Corpus - Digital Resources and Textual Studies
NAMES OF AUTHORS: Hanno Biber, Dr Evelyn Breiteneder, Dr Karlheinz Moerth
AFFILIATION: AAC (Austrian Academy Corpus), Austrian Academy of Sciences
WWW URL: http://www.oeaw.ac.at/litgeb/aac
CONTACT ADDRESS: Sonnenfelsgasse 19/8, A-1010 Vienna, Austria
PHONE NUMBER: +43151581333
FAX NUMBER: +43151581339
E-MAIL: hanno.biber@oeaw.ac.at, evelyn.breiteneder@oeaw.ac.at, karlheinz.moerth@oeaw.ac.at
KEYWORDS: corpus, text encoding, XML, german language, literature

PAPER:

Austrian Academy Corpus - Digital Resources and Textual Studies

Hanno Biber, Evelyn Breiteneder, Karlheinz Moerth

The Austrian Academy Corpus (AAC) is a newly founded institution based at the Austrian Academy of Sciences in Vienna. It has been designed to set up a text corpus and to conduct research in the field of electronic text corpora. For a long time electronic text collections in general were primarily focused on linguistic studies and on lexicography. They were designed and set up for these language orientated purposes. Only recently has the perspective changed towards providing material for scholars interested in texts from various fields of the humanities and in particular in the field of textual studies. Texts are the prime sources of knowledge and of historical investigations. The AAC has been trying to find solutions that meet the needs of textual studies and convey essential information about language and history. The AAC functions as an example of an experimental corpus that is predominantly designed for the study of texts and has been planned as an attempt to build up a complexly structured text collection in which sources from a variety of fields are to be included. And the AAC aims to present and represent a whole range of significant texts which have been and will be carefully selected as being of historical and cultural significance and relevance. The texts sampled, analysed and structured will be from various cultural domains.

The Austrian Academy Corpus will provide the first extensive and intricately structured collection of electronic texts with a systematic and scholarly basis in the German speaking area. The plan is to build up a corpus of electronic texts of the language and literature of the 19th and 20th centuries that will gradually be available on-line. The AAC will be a new institution that creates, structures, provides and analyses selected text sources from the last two centuries, thereby making use of the latest standards and techniques in electronic text. The AAC intends to digitally present a wide selection of different sources of scholarly, journalistic and political texts which were of considerable influence between 1848 and 1989 and continues with its digitisation and structured integration of texts, among which are several influential and notable literary and political journals, such as “Die Weltbühne” or “Die Aktion” published in Berlin in the first decades of the last century -, as well as many other sources, of which the most famous satirical magazine “Die Fackel” published in Vienna - will constitute the core and a starting point for future selections of texts. The corpus size of the AAC will be considerably large, a couple of hundred million words eventually. Images and manuscripts will be included, where necessary, because the graphical and typographical information is important for meaning and interpretation. This is particularly the case for complex text structures like newspapers or literary journals which comprise a whole variety of functionally different text types.

The background of the AAC and its overall research guidelines are reflected in the history of the project which has been determined by the compilation of a phraseological

dictionary, a lexicographic project undertaken at the Austrian Academy of Sciences. The “Wörterbuch der Redensarten” is a selective text-dictionary about "Die Fackel", the famous satirical magazine published by Karl Kraus from April 1899 until February 1936. The computer based working procedures and corpus based research techniques within this lexicographic project focused our attention towards the possibilities and needs of digital resources in this field. The idea of compiling a text-dictionary derives from our interest in language and how it is used in "Die Fackel". “Die Fackel" can be regarded as an ideal text basis for such a dictionary in that it has no equal in the German literature of the twentieth century either in terms of form and content or in the use of language. The first volume, the Dictionary of Idioms, was published in 1999, the 100th anniversary of "Die Fackel". This text dictionary can be regarded as an example and an application of corpus based textual studies.

The AAC’s affiliation with similar undertakings in the field of computing in the humanities should be briefly presented here. The Austrian Academy of Sciences has signed a trilateral corpus agreement with two sister organisations in the German speaking countries, with the 'Berlin-Brandenburg Academy of Sciences' and the 'Swiss Academy of Humanities and Social Sciences', in Vienna on the 12th of March 2001. This trilateral corpus agreement will lead to a joint effort in the three German speaking countries to develop and maintain text corpora for various purposes, among others for building a digital dictionary, the project undertaken in Berlin and to which the Swiss have joined in. The Austrian Academy Corpus however is more concerned with the establishment of a large and well selected corpus of significant texts and focuses its interests and research efforts to the very concept of a corpus itself, which in its electronic form and with the possibilities offered by digital means has become an increasingly interesting objective. The declaration signed by the three Academies emphasizes particular aspects of the value of digital resources in the humanities. “Being convinced of the need to research the cultural development of the German language in its different national contexts as well as to use modern techniques to preserve its diverse lexical and textual substance for posterity and desiring to exploit cross-border co-operation to promote and support interest and scholarship in the German language worldwide”, the three academies have declared their intention to collaborate in the construction of electronic text corpora.

In such a wider perspective, the common and individual settings and conditions of the German language in Austria, Germany, and Switzerland will have to be taken into consideration, as will their historical and contemporary literatures of various kinds. The constant cultural exchange between the countries opens particular research areas and fields of study for the linguists and scholars engaged in the establishment of digital resources and in computing activities in the humanities in the three countries. Setting up such a corpus scheme will require strategic planning and structured procedures in order to account for the diversity and differentiation involved in describing and exploring the language and literature of the German speaking countries. Whereas the efforts undertaken in Berlin and Berne are predominantly concerned with providing selected data and texts of the 20th century mainly for lexicographic purposes, the Austrian Academy of Sciences intends to tackle the problem of digital representation of literary, scholarly, journalistic and political texts which were of considerable influence between 1848 and 1989, and

these texts need not necessarily be original German texts.

Research projects in the field of humanities computing rely heavily on cooperation, collaboration and the constant exchange of knowledge and expertise. The success of undertakings in the domain of linguistic and literary computing will to a large extent depend on dialogue and joint efforts. This is the case in projects that are faced with large quantities of electronic texts, particularly in the areas of corpus linguistics and literary corpus research. Being a research unit within the Austrian Academy of Sciences, the AAC is set within a wider framework, in which the academic institutions in Austria, Germany, Switzerland as well as from other countries, the Czech Republic, the Russian Federation and the United Kingdom are participating. Efforts to find concerted means to maintain common interests and standards in the subject should be made. The Austrian Academy Corpus places special emphasis on international cooperation.

Digital resources in the form of electronic text corpora should be regarded as structures for representing complex information. The electronic text collections established at the AAC so far and its future projects will focus on electronic representations not only of literary texts, literary magazines, journals and newspapers but also on a carefully considered selection of texts from many cultural and social domains. Text corpora should provide electronically available data for scholarly research in the fields of language, literature, and cultural history. We should consider the advance of new systems of digital representation and its implications for the study of language, literature and cultural history. Special emphasis will be placed on areas that have been rather neglected in humanities computing so far. Journals and newspapers pose an especially difficult task when it comes to representation in digital form. An equally difficult task is the analysis and description of the media's decisive historical influences and contexts. The study and detailed investigation of texts has always been crucial for our understanding of historical processes. The knowledge of texts and the accessibility of textual knowledge can be furthered by means of large text corpora like the AAC.

To set up a text corpus several conditions and considerations are required. In the past twenty years, electronic text corpora have been built up in academic institutions of many European countries, such as France, Norway, Sweden, Slovenia, Spain, the Czech Republic, and the UK. The setting up of these corpora is motivated by the state's will to document the national language in a comprehensive manner and to make the corpora available for scientific, especially linguistic, application. The AAC has a different starting point. For the construction of an Austrian corpus one must consider complicated issues relating to the history of the past two hundred years on the one hand and to our own specific interests on the other. The text selection for the AAC, which will take place at the same time as the corpus work, will be guided by thematic and empirical criteria, as well as factors specifically related to the type of text. The specificity of text type is therefore a factor for the choice of texts, but also for their categorisation in a corpus: letters by Oskar Kokoschka, anecdotes by Max Liebermann, writings of Adolf Loos, narrations by Adalbert Stifter, feature articles by Daniel Spitzer, funeral sermons and electoral speeches, propaganda slogans and advertising slogans, pop song lyrics and political speeches, comic books, instructions, travel guides, TV programmes, mailing

catalogues. These and other text types as well as the various kinds of text ‘carrier’ are important for the choice of text.

In recent years, the establishment of large German language corpora has been restricted to the field of linguistic and lexicographic studies. So far, there have not been any large-scale endeavours in the area of text-centred studies. Although more and more literary texts are becoming available, many of these came into existence as by-products of efforts to amass data for lexicographic research. Generally speaking, the historical period on which the Austrian Academy Corpus is working is poorly documented in terms of digital literary texts. This applies even more when it comes to collective text ‘carriers’ such as magazines, papers, year-books, commemorative volumes and similar materials. To our knowledge there do not exist any large amounts of digitised historical magazines or papers in the German language. The sources being digitised for the AAC at the moment are historical literary magazines of major importance. In the first instance there is “Der Brenner”, which was published by Ludwig Ficker in Innsbruck from 1910 until 1934. Among the contributors to the “Brenner” are figures as renowned as Carl Dallago, Theodor Haecker, Else Lasker-Schüler, Adolf Loos and Georg Trakl. The other two magazines on which the AAC is working were both published in Berlin. The journal “Die Aktion” (1911 - 1932) was edited by Franz Pfemfert. Among its contributors were Peter Altenberg, Hermann Bahr, Walter Benjamin, Max Brod, Richard Dehmel, Salomo Friedlaender, Georg Heym, Kurt Hiller, Max Oppenheimer, Egon Schiele and August Strindberg. The last journal to be mentioned here and perhaps the most important one of those being worked on at the moment is the weekly Berlin journal ‘Die Schaubühne’ (1905 - 1918), later renamed ‘Die Weltbühne’ (1918 - 1933,) which was edited by Siegfried Jacobsohn, Kurt Tucholsky and Carl von Ossietzky. Among the writers who contributed to the ‘Weltbühne’ were Henry Barbusse, Bertolt Brecht, Alfred Döblin, Lion Feuchtwanger, Arthur Koestler, Heinrich Mann, Alfred Polgar, Romain Rolland and Leon Trotsky.

To produce a digital version of the magazine “Die Weltbühne”, the original text has to undergo the usual stages of electronic processing: After being scanned, the text is made readable by means of up-to-date OCR. Then pages, paragraphs and lines are identified by automatic routines. The application of markup is the last step in this process. Tags describing contents are carefully inserted by literary scholars especially trained for this job. This process, which takes several runs, is accompanied by proofreading against the original and constant checking and validating of the achieved results. Literary projects in the past used to employ SGML, very often in connection with the TEI guidelines. The AAC also makes extensive use of XML’s modular system of specifications. Aside from the basic XML specification, several other specifications exist, all of them having their more or less well-defined place within the overall framework. The exact nature of some of these sub-specifications is not yet clear (XLink, XML Query), as everything is very much in a state of flux at the moment. Those that are classified as recommendations are XSLT (Extensible Stylesheet Transformations) and XPath (a language for addressing parts of an XML document). The implications of others such as XLink (Extensible Linking Language), XPointer (an abstract language that specifies locations), and XQL (Extensible Query Language) for literary computing will have to be considered in due

course. As XML comes of age, the issue of a standard way of defining the structure of documents becomes more and more important. Both traditional DTDs (document type definitions) and XML Schemas are technologies that provide such descriptions of document structures. Whereas DTDs in the traditional sense have been around for some time and are widely accepted in the field of SGML-based text-encoding, XML Schemas must be regarded as a fledgling technology that still has to win its spurs.

XML Schemas are commonly regarded as an attempt at an XML answer to the problem of defining the structure, content and semantics of documents. There are several arguments in favour of XML Schemas, among which are XML syntax, object orientation, inheritance, polymorphism and datatyping. Firstly, XML Schemas follow XML syntax rules, which makes it possible to parse them with XML tools. Nowadays, authors of XML documents often regard traditional DTDs as unwieldy and inconsistent with the structure of the overall XML system. With XML Schemas, validating parsers can be built on the basis of XML syntax. Secondly, XML Schemas may include explicit restrictions on the data types an element may hold. They let the text programmer attribute data types such as strings, numbers (integer, floating point), date and time formats, boolean and others to elements constituting an XML document. In addition, XML Schemas are also supposed to allow the text worker to define new data types to refine the markup system being used.

Selected references:

Deak, Istvan: Weimar Germany’s Left-Wing Intellectuals. A Political History of the Weltbühne and Its Circle. Berkley and Los Angeles 1968.

Dietzel, Thomas / Hügel, Hans-Otto: Deutsche literarische Zeitschriften 1880 - 1945. Ein Repertorium. München, New York, London, Paris 1988.

Werner Welzig (ed.): Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«. Wien 1999.

Document Object Model (DOM) Level 3 Core Specification. Version 1.0. W3C Working Draft 01 September, 2000 (<http://www.w3.org/TR/DOM-Level-3-Core>http://www.w3.org/TR/DOM-Level-3-Core)

Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6 October 2000 (<http://www.w3.org/TR/REC-xml>http://www.w3.org/TR/REC-xml)

Extensible Stylesheet Language (XSL) Version 1.0. W3C Working Draft 27 March 2000 (<http://www.w3.org/TR/xsl>http://www.w3.org/TR/xsl)

XML Path Language (XPath) Version 1.0. W3C Recommendation 16 November 1999 (<http://www.w3.org/TR/xpath>http://www.w3.org/TR/xpath)

XML Schema Part 0: Primer. W3C Candidate Recommendation 24 October 2000

(<http://www.w3.org/TR/xmlschema-0>http://www.w3.org/TR/xmlschema-0)

XML Schema Part 1: Structures. W3C Candidate Recommendation 24 October 2000
(<http://www.w3.org/TR/xmlschema-1>http://www.w3.org/TR/xmlschema-1)

XML Schema Part 2: Datatypes. W3C Candidate Recommendation 24 October 2000
(<http://www.w3.org/TR/xmlschema-2>http://www.w3.org/TR/xmlschema-2)

XSL Transformations (XSLT) Version 1.0. W3C Recommendation 16 November 1999
(<http://www.w3.org/TR/xslt>http://www.w3.org/TR/xslt)