Proceedings of he 8th Workshop on Challenges in the Management of Large Corpora, pages 47–51
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Challenges for Making Use of a Large Text Corpus such as the 'AAC – Austrian Academy Corpus' for Digital Literary Studies

Hanno Biber
Austrian Academy of Sciences,
Austrian Centre for Digital Humanities and Cultural Heritage,
Austrian Corpora and Digital Editions
1090 Vienna,
Sonnenfelsgasse 19, 3rd floor
Hanno.Biber@oeaw.ac.at

Abstract

The challenges for making use of a large text corpus such as the AAC-Austrian Academy Corpus for the purposes of digital literary studies will be addressed in this presentation. The research question of how to use a digital text corpus of considerable size for such a specific research purpose is of interest for corpus research in general as it is of interest for digital literary text studies which rely to a large extent on large digital text corpora. The observations of the usage of lexical entities such as words, word forms, multi word units and many other linguistic units determine the way in which texts are being studied and explored. Larger entities have to be taken into account as well, which is why questions of semantic analysis and larger structures come into play. The texts of the AAC-Austrian Academy Corpus which was founded in 2001 are German language texts of historical and cultural significance from the time between 1848 and 1989. The aim of this study is to present possible research questions for corpus-based methodological approaches for the digital study of literary texts and to give examples of early experiments and experiences with making use of a large text corpus for these research purposes.

Keywords: corpus research, corpus-based literary studies, computational philology

1. Introduction to a Challenging Research Question

In this presentation the challenges for making use of a large digital text corpus such as the AAC-Austrian Academy Corpus for the purpose of digital literary studies will be addressed and a brief introduction into possible ways to achieve that will be given.

The research question of how to use a complex digital text corpus of considerable size for such a particular research purpose is of considerable interest for corpus linguistics and for corpus research in general as it is for the fields of digital literary studies and digital philology alike. Text studies and in particular digital literary text studies rely to a very large extent more and more on the existence, the availability and the specific functionalities of large digital text corpora. Being able to investigate – just to mention a very common feature of such electronic resources – lexical units of various kinds and thereby to search for words, word forms, multi word units, collocations, lexical patterns, named entities and many other or similar linguistic structures in various digitalized texts within a corpus framework has considerably changed and determined the way in which texts are being studied and explored, not only by language scholars and literary historians. For particular questions of literary studies also larger linguistic units of such texts have to be taken into account here as well, which is why corpus-related questions of narrative studies and also to some extent of discourse studies could also come into play, but only very few examples will be given here to show the potential of a possible research agenda following the principles of corpus-based digital literary studies.

The AAC-Austrian Academy Corpus was founded in 2001 and the texts of the AAC are German language texts of historical and cultural significance from the period between 1848 and 1989. The time frame and the text frame of these highly valuable digital collections of German language texts from all over the German speaking areas constitute the first two important dimensions of the text corpus and its research approaches which are based upon a variety of different parameters. The language use and the considerable variety of the text production at the times of the historical periods in focus of this text corpus immediately raise many questions of how to build such a representative texts corpus, in which ways it would be related to other similar endeavours of creating linguistic text corpora with lexicographic objectives or questions

regarding a comparison to more balanced corpora of rather basic linguistic objectives.

At the centre of the considerations for the selection process of the texts to be integrated into the Austrian Academy Corpus however, stands the question of cultural and historical significance, which has led to the construction of a text corpus founded upon principles of specific parameters guided by critical historical, linguistic, literary, cultural as well as empirical principles.

Selected literary texts from the AAC are to be used as examples from this text corpus in order to demonstrate the corpus-linguistic possibilities of lexicographic studies of literature. The aim of the presentation is to present the potential of corpus-based methodological approaches for the study of literary texts, of a whole range of literary production of a certain historical period, and in particular for the study of lexical items and linguistic structures in literary texts. The framework of the AAC-Austrian Academy Corpus offers research options for such corpus-based literary studies. Examples of this research and its potential for corpus-based text studies will be given.

2. The AAC-Austrian Academy Corpus as a Large Text Corpus

Figure 1: AAC-Poster (copyright H. Biber).

The AAC-Austrian Academy Corpus is a large text corpus of around 600 million tokens. It consists of a large variety of different texts predominantly in German language from the period between 1848 until 1989 with a strong emphasis on the first half of the 20th century. 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. It has been founded in 2001 and most of its textual resources were created in the first decade of the 21st century. The selection of texts to be part of this large text corpus has been based upon a variety of parameters. As the main purpose of the construction of this text corpus has been a primarily lexicographic, or to be more precise text-lexicographic one, the sources of the AAC stem from a variety of different sociological fields and linguistic domains thereby reflecting not only linguistic and literary but also historical and cultural processes. The AAC provides "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." (Biber and Breiteneder 2004). Among the sources are more than one hundred full runs of political, cultural or literary journals, such as "Die Schaubühne" and "Die Weltbühne" or "Die Aktion", published in Berlin in the early 20th century, as well as many other similar sources, of which the most famous satirical journal "Die Fackel", published in Vienna by Karl Kraus, constitutes "the core and starting point for future selections of texts". (Biber and Breiteneder 2002). 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." (Biber and Breiteneder 2004). The sub-corpora of magazines is contributing to a high variation of different text-types to be found in this large text corpus, because traditionally journals include also letters, notes, poems, advertisements, essays and so on. The AAC includes, apart from the magazines a large section of what is called collections, i.e. text books, almanacs, reading books etc., containing articles from various authors, and also a large number of books of fiction, poetry, popular science, essayistic literature or scholarly publications, and so on. Newspapers are also included to some extent, thereby putting an emphasis to publications or certain months of particular historical importance.

In almost all cases all texts had been scanned with industrial book scanners, so that both the actual images of all the digitalized texts are conserved and accessible when searching for words or annotated content. Then these images have been OCR-read with commercial and highly efficient software, before XML-mark-up has been applied in order to deal with the structural elements describing basic properties of the texts and consequently with the application of linguistic standard annotations (such as STTS PoS-tagging for example) as well as on top of that providing layers of more specific semantic or literary types of annotation, for example in the field of named entities like toponyms or personal names and the like, depending on the research efforts possible.

Several digital editions and text corpus tools have been developed within the AAC for the purpose of detailed

investigations of large amounts of literary texts. In order to be able to explore digital text corpora and to be able to conduct research in the fields of text analysis, several million pages of text are available in this form and have been converted into machine-readable text of more than six hundred million tokens of annotated text. As most of the work started some time ago, newer standards of annotation schemes might have to be applied, funding permitted.

Because the AAC represents such a wide range of different text types from many different domains and genres, it is particularly interesting for the corpus-based study of historical developments by means of looking into the lexicographic and lexical data provided by such a resource, as it can be followed over time. The text corpus includes apart from newspapers, literary journals, novels, dramas, poems, essays, advertisements etc. also travel accounts, cookbooks, pamphlets, political speeches as well as scientific, legal, and religious texts, to name more text types.

The AAC provides a great number of reliable resources for investigations into the linguistic and textual properties of these texts, of which not only the literary ones are well selected with consideration and do by no means follow an opportunistic pattern of selection. The intention has from the very beginning been "to digitally present a wide selection of different sources of scholarly, literary, journalistic, scientific, political texts which exercised considerable influence." (Biber and Breiteneder 2004). This text corpus and its methodological approach of text selection gives scholars who are making use of this text corpus a reliable resource to conduct their research. In the following some examples of first experiments and thorough explorations will be given. An overview of the AAC-Austrian Academy has been given in the second "CMLC" workshop (cf. Biber and Breiteneder 2014).

3. Examples, Experiments and Explorations of Digital Literary Text Studies performed within the AAC

The methodological approaches of the AAC-Austrian Academy Corpus are governed by principles of philological exactness, clear efforts in the structuring of texts, systematic and standardized annotation, specific editorial techniques, lexicographic indexing, scholarly commenting, and so on. Therefore the texts are to be made accessible for research efforts in corpus linguistics and digital philology alike. Examples of experimental explorations into the potential of such a text corpus approach can help to describe the scope and possible directions, leading to digital editions, corpus-based dictionaries, digital libraries or data collections, and corpus research in a broader sense.

The "AAC-Fackel" (Biber et al. 2007a), the first AAC digital edition coming out of the AAC-Austrian Academy Corpus, is an online edition of the journal "Die Fackel" used by more than 30.000 readers, that offers free access to its 37 volumes, 415 issues, 922 numbers, comprising more than 22.586 pages and six million word forms. It can be regarded as a model, a "Musteredition" (Biber 2015), as it contains a fully searchable database of the entire journal with various indexes, search tools and navigation aids in an innovative and highly functional graphic design interface, in which all pages of the original are available as digital texts and as facsimile images. The satirical journal "Die Fackel" was published by Karl Kraus in Vienna from 1899 until 1936 and was also a model for the literary journal "Der Brenner" published between 1910 and 1954 in Innsbruck by Ludwig von Ficker, which has been made online available as "Brenner online" in cooperation of the AAC with the University of Innsbruck (Biber et al. 2007b). The text of "Der Brenner" consists of 18 volumes and 104 issues, which is just a small segment of the AAC's overall holdings, is about two million running words of carefully corrected text, annotated and provided with additional philological information. Both digital editions are making subsections of the overall corpus holdings available in a way which was determined by combining the advantages of graphic design and corpus-based linguistic exploration for the benefit of scholarly and scientific exploration, with a special emphasis in the study of lexical forms (Biber 2006).

Both exemplary journals, for which exemplary editions have been built, are good examples of culturally and historically significant language use. In particular the satirical texts by the language and social critic Karl Kraus can function as highly interesting focal points into a critical and rather ideological exploration of language change and semantic shifts in language use by analysing the overall corpus and the specific features and contexts of certain lexical items. In the historical period of the AAC "significant changes with remarkable influences on the language and the language use can be observed. The years of the seizure of power of the National Socialists is of specific interest for such language studies, where various documents and significant collocations, lexical items, and figurative linguistic constructions are taken into account." (Biber 2013). "Building a diachronic digital text corpus for historical German language studies of this particular kind is a particularly challenging task for various reasons. First, the technical difficulties of corpus building in dealing with a large historical variety of different text types and genres have to be taken into consideration. Second, the specific historical parameters and the methodological scope of such an investigation has to be taken into account. The German language of the year 1933 is being considered as a historical focal point for which an exemplary corpus-based research methodology for the study of the German language could be developed. The sources of a first exemplary study will cover manifold domains and genres, not only newspapers and political

journals and magazines, which will be at the core, but also several other text types representing the historical communicative strategies will be included. Among them are pamphlets, flyers, advertisements, radio programs, political speeches, but also essays and literary texts as well as administrative, scientific or legal texts, just to name a few examples, which are all difficult to collect. The AAC has started to build up a small collection of ephemera in this field." (Biber and Breiteneder 2013). For this direction of investigation into the language of a specific historical period, but also for a general lexicographic study focussing on general literature, the concept of "container" has been suggested for the structuring process of the corpus (cf. Biber and Breiteneder 2012). Also first suggestions have been made in order to visualize findings within the corpus. (Biber and Barbaresi 2016).

It is possible that corpus research methods based upon a multidisciplinary combination of corpus linguistics, lexicography, historical studies and cultural studies be applied to gain insights into the textual representations of historical collections of such importance. A corpus-based approach is considered promising in this respect, because applying methods of corpus linguistics and testing new strategies of the application of these methods in the context of historical language studies can also be used for studies of the use of metaphorical constructions and idiomatic multi-word units, like idioms that can be regarded as prototypical forms of figurative language, which is particularly interesting for literary studies. In order to name other possible studies done within the framework of the AAC, particular uses of idiomatic expressions have been investigated, as have studies of the use of proverbs been based upon results from the text corpus (Biber 2010) or an analysis of specific thematic texts (Biber 2014), or the specific vocabulary of for example "Austerity in the Thirties" (Biber 2013) or studies of figurative language (Biber 2009), and detailed studies of collocations (Biber, Breiteneder and Dobrovolskij 2002) in a certain historical period, as have academic dictionaries and academic text-dictionaries been compiled with the help of the Austrian Academy Corpus. This text corpus is a highly relevant resource for building lexical resources based upon corpus findings as well as for empirical digital literary studies and beyond.

4. Bibliographical References

Biber, H. and Breiteneder, E. (2002): Austrian Academy Corpus: digital resources in textual studies. In: J. Anderson, A. Dunning, M. Fraser (eds.): *Digital Resources for the Humanities 2001–2002. An edited selection of papers.* (Publication 16) London: Office for Humanities Communication, p. 13-18

Biber, H., Breiteneder, E. and Dobrovolskij D. (2002): Corpus-Based Study of Collocations in the AAC. In: A. Braasch, C. Povlsen (Eds.): Euralex Proceedings Vol.1 2002, p. 85-95

Biber, H. and Breiteneder, E. (2004): "The AAC [Austrian Academy Corpus] - An Enterprise to Develop Large Electronic Text Corpora". In: M. L. Lino, M. F. Xavier et al. (Eds.): Proceedings of the 4th International Conference on Language Resources and Evaluation Lissabon 2004. Volume V, Lisbon: ELRA, p. 1803-1806

Biber, H. (2006): Words in "Der Brenner" lexicographic searches in a new scholarly digital edition of the AAC. In: E. Corino, C. Marello, C. Onesti (Eds.): *Atti del XII Congresso Internazionale di Lessicografia*, Torino, 6.-9. 9. 2006, Vol. 1, Torino: Atti, p. 395-398

Biber, H. et al. (Eds.) (2007a): AAC-Austrian Academy Corpus 2007: AAC-Fackel. Online Version: "Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936". AAC Digital Edition No 1, www.aac.ac.at/fackel

Biber, H. et al. (Eds.) (2007b): AAC-Austrian Academy Corpus (2007) and Brenner-Archiv: Brenner online. Online Version: "Der Brenner. Herausgeber: Ludwig Ficker, Innsbruck 1910-1954", AAC Digital Edition No 2, www.aac.ac.at/brenner

Biber, H. and Breiteneder, E. (2008): Words in Contexts: Digital Editions of Literary Journals in the "AAC - Austrian Academy Corpus". LREC 2008, Marrakech: ELRA, p. 339-342

Biber, H. (2009): Hundreds of Examples of Figurative Language from the AAC-Austrian Academy Corpus. In: J. Barnden, R. Moon, G. Philip, A. Wallington (Eds.): Corpus-Based Approaches to Figurative Language. A Corpus Linguistics 2009 Colloquium, Colloquium Companion, School of Computer Science University of Birmingham, *CSRP-09-01 (Cognitive Science Research Papers)*, July 2009, ISSN 1368-9223, p. 13-20

Biber, H. (2010): Corpus-based Studies of Proverbs and Proverbial Expressions. In: Rui J. B. Soares (Ed.): *Interdisciplinary Colloquium on Proverbs*, ACTAS ICP09, Tavira: AIP, p. 106-110

Biber, H. and Breiteneder, E. (2012): Fivehundredmillionandone Tokens. Loading the AAC Container with Text Resources for Text Studies. In: N. Calzolari et al. (Eds.): Proceedings of the International Conference on Language Resources and Evaluation LREC 2012, Istanbul, 23.-25. 5. 2012. Istanbul: ELRA, p. 1067-1070

Biber, H. and Breiteneder, E. (2013): The German Language of the Year 1933. Building a Diachronic Text Corpus for Historical German Language Studies. In: Center for Digital Research in the Humanities (Ed.): Digital Humanities 2013 Proceedings. Lincoln: University of Nebraska, p. 107-109

Biber, H. (2013): Austerity in the Thirties. German examples of historical figurative language of austerity from the AAC Austrian Academy Corpus. In: Gill Philip, et al. (eds.): Corpus-Based Approaches to Figurative Language. Metaphor and Austerity. School of Computer Science University of Birmingham, *CSRP-13-01 (Cognitive Science Research Papers)*, July 2013, p. 22-24

H. Biber, E. Breiteneder (2014): Text Corpora for Text Studies. About the foundations of the AAC- Austrian Academy Corpus. In: H. Biber, et. al (eds.) (2014): Challenges in the management of large corpora (CMLC-2) LREC 2014 Workshop-Proceedings. Reykjavik: LREC, p. 30-34

Biber, H. (2014): Mountains of Text. Analyzing Alpine Literature from the AAC. In: DH2014, Digital Humanities Proceedings 2014. In: DH2014, Digital Humanities Proceedings 2014. Lausanne: EPFL, p. 447-448

Biber, H. (2015): AAC-Fackel. Das Beispiel einer digitalen Musteredition. In: C. Baum und T. Stäcker (Eds.): *Grenzen und Möglichkeiten der Digital Humanities. Sonderband 1 (2015) der Zeitschrift für digitale Geisteswissenschaften*, DOI 10.17175/sb001_019, www.zfdg.de/sb001_019

Biber, H. and Barbaresi, A. (2016): Extraction and Visualization of Toponyms in Diachronic Text Corpora. In: Digital Humanities 2016 Conference Abstracts, Cracow: Jagiellonian University & Pedagogical University, p. 732-734

5. Language Resource References

AAC - Austrian Academy Corpus (2001).