"The word expired when that world awoke." New Challenges for Research with Large Text Corpora and Corpus-Based Discourse Studies in Totalitarian Times.

Hanno Biber
Austrian Academy of Sciences
Bäckerstraße 13, 1010 Vienna
Hanno.Biber@oeaw.ac.at

Abstract

The title's opening quotation is a translation of the final line of the famous poem by the satirist Karl Kraus from 1933 that explains in ten lines the limits of language use when violence reigns, something that can be seen as a fundamental research question. The prospects of a possible corpus project based upon the AAC-Austrian Academy Corpus, established in 2001, will be assessed to give answers to questions concerning the research challenges for large digital text corpora in the context of studying totalitarian language. The AAC contains many German language texts from the first half on the 20th century and can be regarded as a valuable diachronic corpus suitable for corpus-based studies focused upon historical sources. Therefore it might be used not only to document the language of the time of the rise of the Nazis in Germany and Austria, but may also lead to giving an example as to how apply such an approach to related issues of addressing the challenges of critically using contemporary text corpora in the context of a new totalitarianism unfolding by the Russian war of extermination in Ukraine and its lexical representations in the discourses of contemporary media full of propaganda and disinformation.

Keywords: propaganda, critical discourse studies, corpus linguistics

1. Research Challenges

In the following paper, first, a report on and a description of the prospects of a corpus project will be given that could possibly be initiated by one of the large historical diachronic digital text corpora hosted by the Austrian Academy of Sciences, the AAC – Austrian Academy Corpus. And second, new potential challenges for corpus linguistic research with large contemporary synchronic digital text corpora will be addressed with a particular reference to lexical representations to be found in text corpora with regard to discourse phenomena to be observed in contemporary media and current news in particular. Furthermore, these challenges for large text corpora will be viewed focusing on phenomena to be observed in digital media that are to a very large extent full of propaganda and disinformation, above all in times that can not only hyperbolically but also have with sound historical reason to be viewed as evolving into new totalitarian times now. On this occasion, however only a rather fine outline for such an extensive framework for corpus research can be made, so that the rough idea presented should rather be regarded as a research proposal and as a suggestion for future projects. First, the digital resources of the AAC – Austrian Academy Corpus, that has been founded in 2001 (cf. Biber and Breiteneder, 2002), which is one of the very valuable examples of considerably large diachronic digital text corpora also suitable for corpus-based discourse studies and for digital corpus-based lexicography grounded upon historical text sources, can be used as a starting point for trying to answer new questions concerning the challenges for doing linguistic research with large digital text corpora in the context of studying totalitarian language use. The questions, as well as the chances and the limits of such an approach, have very obvious actual references to the historic events unfolding today as well as a clearly historical dimension, precisely because the digital text sources that have been created to analyze the German language use of the Nazi-period from 1933 to 1945 can be understood as a model to deal with related questions of contemporary language use, particularly in the context of the new Russian war of extermination in Ukraine of today and particularly how it is represented in contemporary media. It stands to reason that corpus linguistics and historical language studies are not performed in a sphere free from ideological, social, political, nationalistic, and related implications. The dynamics of an increasing influence of information in data-driven societies with industrialized and algorithmically enhanced linguistic activities in the field of political propaganda demands from researchers to be quickly able to address these questions in order to contribute to an enlightened and scientifically tenable perspective on the observed processes, for which large text corpora have already been built and for corpora which are urgently needed to be constructed for such purposes. Linguistic content challenges are particularly prevalent when confronted with various phenomena that are to be observed in historical sources just like in contemporary sources, such as "atrocity propaganda", "industrialized lies", "fear discourse", "alternative facts", "obfuscation techniques", "information warfare" and similar events that are to be detected in large quantities of digital content created, be it in form of diachronic text corpora or in synchronic corpora.

2. Diachronic Text

2.1 AAC – Austrian Academy Corpus

The AAC – Austrian Academy Corpus was founded in 2001 and has been created as a corpus linguistic research project undertaken within the framework of the Austrian Academy of Sciences in Vienna in the first decade of the 21st century. The AAC is a German language text corpus of more than fivehundredmillion tokens and represents a large diachronic digital text corpus with several thousands of German language texts of important historical and cultural significance. The AAC-Austrian Academy Corpus has an emphasis also on literary and political journals as well as

Proceedings of the CMLC-10 Workshop (@LREC 2022), pages 32–35
Marseille, 20 June 2022

on collections of texts that are difficult to obtain in or difficult to integrate into digital text resources. Overall, the texts of the AAC are predominantly German language texts from the period between 1848 and 1989, ranging from the 1848 revolution to the fall of the iron curtain in 1989, but have a focus on the first half of the twentieth century. "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." (Biber, 2020) These parameters are also of empirical and historical as well as of dimensions of linguistic domains. Among the sources of the AAC a very large number of texts of the historical period in question in this report have been collected, digitized, converted into machine-readable text and annotated and been provided with metadata, according to the standards of structural and thematic mark-up applied then, of annotation and mark-up schemes based upon XML almost two decades ago. In a presentation of this project an overview of the necessary methodological considerations and an outline of the research perspectives based upon the principles of corpus linguistics should be given, but as this has been done in several previous presentations, the references to the respective publications should suffice (cf. Biber and Breiteneder, 2002; Biber, 2004; Biber and Breiteneder, 2004; Biber and Breiteneder, 2012; Biber and Breiteneder, 2014).

2.2 Subcorpus of 1933-1945

The topic of the proposed research project to be paid attention here, is focused on the questions of developing a diachronic text corpus of historical significance and establishing a corpus based research environment for language studies of the historical period between 1933 and 1945, with particular emphasis on the year 1933, the year when the Nazis came to power in Germany. As the core of the AAC is from the first half of the twentieth century, the research issue of an analysis of the German language of the time between 1933 and 1945 is feasible and can be done comprehensively. Corpus-based approaches for analyzing the language exploring the historical periods before, during and after Nazi rule together have been rare, despite numerous more detailed scholarly works in the fields of historical studies as well as in German language studies. (cf. Biber, 2010) Building a diachronic digital text corpus for historical German language studies of this particular kind is a challenging task for various reasons. There are certain technical difficulties of corpus building in dealing with a large historical variety of different genres and text types, and the "specific historical parameters and the methodological scope of such an investigation" (Biber and Breiteneder, 2013) have to be taken into consideration. The German language of the years between 1933 and 1945 is being considered as a historical focal point for which an exemplary corpus-based research methodology for the study of the German language can 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" can be integrated in such a research initiative. (Biber and Breiteneder, 2013) Among the text sources to be considered, are not only political speeches, pamphlets, flyers, or advertisements, but also essays and literary texts as well as possibly radio programs, or even administrative, scientific and legal texts, which have been already collected to some extent (cf. Biber and Breiteneder, 2013). In this case not primarily the well-known documents and the evident language of the Nazi period could be included in the analysis, but systematically less easily visible documents and less significant lexical items might also be taken into consideration. This methodological approach is considered as particularly promising by means of applying methods of corpus linguistics and by testing new strategies of the application of these methods in the context of historical language studies. The AAC corpus holdings may provide a large number of interesting resources and lead to corpus-based approaches for investigations into the texts of the historical period in question. (cf. Biber and Breiteneder, 2013) "Quantitative corpus linguistics has proofed to be a valuable technique in many domains of philological, sociological and historical research. The digitized and linguistically annotated corpus is therefore an interesting source for studies in many fields and facilitates the investigation of changing patterns of language use, and how these reflect underlying cultural shifts." (Volk, 2010). And one may ask, if a practical combination of corpus linguistics, lexicography, historical studies, discourse analysis and cultural studies can be used to gain knowledge about the texts of the time in focus.

2.3 1933 and the Following

"The word expired when that world awoke", is the quotation in the title of this paper and a translation by Max Knight (Zohn, 1990) of the last line of the famous poem by the satirist Karl Kraus written in September 1933 and originally published in his satirical journal in the issue of October 1933 (Biber, 2007). Its interpretation is used in order to refer to a crucial challenge for all linguistic research and language studies, to be deliberately formulated here, as indicated, as a particular challenge for corpus research in a particular historical context. Answers to the questions posed with regard to analyzing the language of a certain historical period might be found in analytically making use of the long analytical text written by Karl Kraus between May and September 1993, which has not been published in his journal and for which the poem however functioned as an indicator and index. This long unpublished text from 1933, that was only posthumously published in 1952 for the first time, bears the title "The Third Walpurgis Night" and has only very recently been translated into English. (Kraus, 2021) With the help of this text, that opens up a critical analytical path for language research, exploring the large digital text corpus of German language texts from the first half of the twentieth century, could be achieved in a confident manner. At the Austrian Academy of Sciences a new digital German text edition has been published, for which a register of the many texts and documents quoted and a "register of personal names" (Biber, 2021) has been created of more than 400 perpetrators, victims and witnesses appearing in the text, where the crimes and the language of the time of the rise of the Nazis in Germany and in Austria are documented. The text could function as an example as to how to deal with documents in the context of the mentioned challenges for large text corpora in totalitarian times. "The Third Walpurgis Night" is the most important text and a

most detailed contemporary account of German literature about the horrifying origins and deadly consequences of National-Socialism, where the murderous reality and the murderous language of the Nazis is documented in many examples as early as May 1933, examples that are quoted and then commented upon, with insights which can be taken as starting points for conducting the proposed research within the frameworks of a large text corpus. "As to Hitler I have nothing to say" is the famous first sentence of this long text that, significantly, concludes after more than 300 pages of analyzing Nazi language and Nazi atrocities with a quotation from Goethe's Faust II, that "may this phantom", the tyrant and his regime, be "hurled among the dead", a sentence said as in the face of total violence the word is inappropriate. (Kraus, 2020) The author decided not to publish his text, but to conserve it for posterity and the text dealing with the language and the violence as well as the consequences of their political developments for the world is just very briefly summarized in the final line of his poem, as quoted above. The AAC – Austrian Academy Corpus contains very many documents from the time and many documents that are also dealt with in the text of 1933, with texts and of "what appeared, day by day, in print, on the radio, and in public forums." (Perloff, 2020). Significant parts are not only dealing with the "rhetoric of respected thinkers" (Perloff, 2020) who were advocating and agitating in favour of Nazism, like the philosopher and university rector Martin Heidegger, or the poet and medical doctor Gottfried Benn, but also treating the former philologist and journalist Goebbels, the Nazi Minister of Propaganda and his language technique, which is more than interesting in this case, where an account and a linguistic discourse analysis of a Reichstag speech is given, where Goebbels "has attitude and empathy, he knows about the stimulus and impetus, application and implication, dramatic presentation, filmic transposition, flexible formulation, and the other aides to radical renewal, he has experience and perspective, indeed for both reality and vision, he has zest for life and world-philosophy, he approves of ethos and pathos but also mythos, he supplies subordination and integration into the living-space and working space of the nation, he embraces the emotional realm of community and the vitalism of personality." (Kraus, 2020) The language of totalitarianism is to be clearly observed in the year in which the Nazis came to power in Germany and because of the analysis by Karl Kraus no one can deny the fact that it had been possible then to predict where this would lead to, to annihilation and finally to extermination.

3. Synchronic Text

3.1 Context of Contemporary Corpora

Digital text corpora of today are to an increasing extent more and more based upon sources from contemporary news cycles, newspapers, media outlets, social media content etc. The communicative frame in which linguistic expression and language is set, becomes more and more important. Contemporary language production, that to a very large extent is of journalistic origin and can possibly function as sources for linguistic research by means of large text corpora, needs to be studied and analysed in detail. This is of particular importance in times when the mechanisms of war-driven communication and the discursive distortions and manipulations of totalitarian propaganda is becoming to be dominant in public discourse. A critical viewpoint is more than necessary in order to successfully analyse also text corpus content yet to be substantially formed into useable text corpora out of the vast amount of contemporary news and discourse cycles. And it is equally important to critically understand and do substantial research about the propaganda propagating in the media of today, let alone those which are openly and aggressively neglecting, ignoring, obfuscating, distorting the atrocities committed in a criminal war. It is obvious that this happens in a principally problematic situation of journalism that has been described even for the times before the First World War or shortly after by the language and media critic. The "rogue profession" that "takes the audience for an idiot" to "outwit their intelligence with its own mindlessness" and "abuses the debility which is called public opinion for any infamy" (Biber 2007), as Karl Kraus describes in still democratic June 1923 the relation of the public and journalism together with politics, demonstrates that journalism in all its forms is the main object of critique, in texts whose conclusions have still validity for old and new media systems alike, as even Jonathan Franzen as translator and commentator has observed in his book "The Kraus Project". (Franzen, 2013) Karl Kraus has blamed and named the "pyramidal dimensions of stupidity" as "the secret of all journalistic seduction", in which the dispositions of the journalistic audience and the journalistic profession meet, and who, in the cited article in his journal, has pointed at "the stolen pathos of the just cause, which no other rogue profession has so easily at hand as this one, that as a means for exploitation always takes the audience for an idiot in order to outwit their intelligence with its own mindlessness and abuses the debility which is called public opinion for any infamy, while the impotence which is called the state authority has become an inactive witness." (Biber, 2007). The satirist views the audience not only as a victim of journalistic practice, but also as an accomplice in the journalistic crime of stupidity and absurdity. The very notion of public opinion is already an absurdity and a logic contradiction for him, who reminds his own audience of the necessary private nature of any opinion and demands not only sincerity from the sender but also a sincere subjectivity and critical awareness from the receiver of a news message, so that an opinion can be formed by the audience and is not at all preformed and engineered by the opinion business beforehand. This complex is to be taken seriously in the context of doing corpus linguistic research with journalistic texts, where the researcher is much more than just another audience of the language production but more of a critical analyst of what can be observed.

3.2 2022 and the Following

Corpus linguistics and literary studies in the 21st century must be working and doing research in the context of a technologically and ideologically dynamic historical situation and therefore be aware of the difficulties and necessities for the research. The computer technologies and their algorithmic potential, the impacts of social media, digital archives, digital libraries, and many other phenomena of the contemporary world of technological communication and information technology demonstrate how much the ways of thinking, speaking, writing, and communicating are determined by technical acceleration. What it means to read, understand and examine with scientific means the texts in a digitalized world and how the

sources of information must be investigated with new tools in this context of new modes of communication and new forms of text production, also demand new methods for corpus linguistics and language research. For this purpose it is necessary to construct and create an infrastructure of suitable text corpora enabling the researchers to study the described phenomena with the help of corpus linguistic methodologies. The research focus must be determined by well-established methods in combination with a critical apparatus based upon insights of media critique and therefore enable the researchers to find, determine and select as well as carefully analyze the texts in digital corpora. This process of integrating and aggregating texts form various resources should lead to the creation of a new research environment for corpus linguists useful also for synchronic purposes. The German language use of aggression during the Nazi-period and the critical analysis based upon Krausian critique of language use can function as a model to deal with contemporary questions of violent language use, above all in the example of the setting of today's Russian war of extermination in Ukraine. The ways in which and the role how digital and conventional media texts and their linguistic inventories play a role in this war and are used by the aggressor, and how the aggression has been prepared for years by propagandistic interventions in contemporary media, certainly needs to be investigated by means of corpus linguistics as well. Corpus linguistics and historical language studies are not done without ideological, social, political, nationalistic, and other related implications. The dynamics of information warfare in connection with industrialized and algorithmically enhanced linguistic activities in the field of political propaganda demands from researchers to be quickly able to address the issues of coming to terms with language production and linguistic elements that can be followed scientifically by means of corpus linguistics. One has to be aware of the diabolic (obviously Mephistophelian) fact that historical progress is always moving on and that, even when it seems to be over, history will repeat itself and the past will become the future, as the eighth and ninth lines of the quoted poem reminds one: "It cannot last; Later it all was past." (Bridgham and Timms, 2020)" Or as the same lines read in the other translation: "Time marches on; the final difference is none." (cf. Zohn, 1990) New large text corpora need to be built and these text corpora are urgently needed to be constructed for purposes of analyzing texts in the described contexts. The challenges just very briefly explained here are particularly demanding when in contemporary sources discourse phenomena of "atrocity propaganda", "industrialized lies", "fear discourse", "alternative facts", "obfuscation techniques", "information warfare" etc. are to be observed in journalistic texts, all occurrences to be detected in large quantities in large text corpora, diachronic or synchronic.

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. 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 Lisbon 2004. Volume V*, Lisbon: ELRA, p. 1803-1806

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

Biber, H. (2010): "Aufbruch der Phrase zur Tat". Kommunikationsmaßnahmen und sprachliche Formungen der nationalsozialistischen Machtübernahme in Österreich 1938. In: Welzig. W et al. (Eds.) (2010): *Anschluss. März/April 1938 in Österreich*. Vienna: Austrian Academy of Sciences Press, p. 15-37

Biber, H. and Breiteneder, E. (2012): Fivehundredmillionandone Tokens. Loading the AAC Container with Text Resources for Text Studies. In: N. Calzolari et al. (Eds.): P*roceedings 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

H. Biber and 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. (2020): Challenges for Making Use of a Large Text Corpus such as the 'AAC – Austrian Academy Corpus' for Digital Literary Studies. In: Banski, P., et al. (Eds.) (2020): *LREC 2020 Workshop. 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8), Proceedings*. Marseille: ELRA, p. 47-51

Biber, H. (2021): Personenregister. Karl Kraus 1933: Dritte Walpurgisnacht. kraus1933.ace.oeaw.ac.at

Bridgham, F. and Timms, E. (2020): Introduction. In: Kraus, K. (2020): *The Third Walpurgis Night.* New Haven: Yale University Press, p. xix-xxv

Franzen, J. (2013): The Kraus Project: Essays by Karl Kraus. New York: MacMillan

Kraus, K. (2020): The Third Walpurgis Night. [The Complete Text Translated from the German by Fred Bridgham and Edward Timms]. New Haven: Yale University Press

Perloff, M. (2020): Foreword. In: Kraus, K. (2020): *The Third Walpurgis Night.* New Haven: Yale University Press, p. vii-xv

Volk, M. et al. (2010): Challenges in building a multilingual alpine heritage corpus. 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. 1653-1659

Zohn, H. (1990) (Ed.). Karl Kraus - In These Great Times. Chicago: University Press

5. Language Resource References

AAC - Austrian Academy Corpus (2001)