

EVELYN BREITENEDER
DMITRIJ DOBROVOL'SKIJ
ARTEM ŠARANDIN
AAC-Austrian Academy Corpus
Austrian Academy of Sciences
Vienna, Austria / Europe
aac@oeaw.ac.at

AAC - Austrian Academy Corpus and its Major Applications

1. The AAC - Austrian Academy Corpus is a newly founded institution based within the Austrian Academy of Sciences. Its aim is to set up a large text corpus and to conduct research in the field of electronic text corpora. The importance of large electronic text corpora has been commonly agreed upon by linguists and lexicographers. Recently this approach has become a research objective also in the fields of textual and literary studies. The advantages of digital resources for the thoroughly study of literary texts as well as of research based upon various textual sources of historical and cultural significance are currently under debate in various scholarly domains. There is a need for well structured and reliable resources for detailed linguistic and literary research and there is a considerable amount of linguistic corpora available for linguistic and lexicographic purposes, some of which have made valuable contributions to the advance of our knowledge about languages. The AAC is predominantly designed for the study of texts and it will provide substantial research material for linguistic and lexicographic studies as well as for literary studies. With the AAC we will provide an extensive and intricately structured collection of electronic texts with a systematic and scholarly basis. It is planned to build up a complexly structured and carefully selected collection in which textual sources from various cultural domains of the 19th and 20th centuries will be included. Among the variety of different texts and text types that constitute the AAC translations of historical literary significance will also be included (for more detail see Biber, Breiteneder, Dobrovolskij 2002).

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. Among these sources are several influential and notable 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 journal "Die Fackel", published in Vienna, will constitute the core and starting point for future selections of texts. The electronic text corpus will be of a considerable size. 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 AAC working group have been engaged since 1991 in a lexicographic project undertaken by the Austrian Academy of Sciences. The compilation of a phraseological dictionary, "Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«" (Vienna, 1999), and its computer based working procedures and corpus based research techniques focused the lexicographers' attention towards the possibilities and needs of digital resources. These resources have been utilized for the compilation of this particular text-dictionary of the "Fackel", a famous satirical magazine published by Karl Kraus in Vienna between April 1899 and February 1936.

The Austrian Academy Corpus places special emphasis on international cooperation. Being a research unit within the Austrian Academy of Sciences, the AAC is set within a wider framework, in which academic institutions in Austria, Germany, and Switzerland as well as from other countries, such as the Czech Republic, the Russian Federation and the United Kingdom are participating. Efforts are being made to find concerted means to maintain common interests and standards in the subject. In such a wider perspective, the common and individual settings and conditions of the German language in Austria, Germany, and Switzerland have to be taken into consideration, as will their historical and contemporary literatures of various kinds. The Austrian Academy of Sciences has signed a trilateral corpus agreement with its two cousins in the German speaking countries, the Berlin-Brandenburg Academy of Sciences and the Swiss Academy of Humanities and Social Sciences, which will lead to a joint effort in all three German speaking countries to develop and maintain text corpora for various purposes.

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, magazines, journals and newspapers but also on a carefully considered selection of texts from several other cultural and social domains. Text corpora should provide electronically available data for scholarly research in the fields 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 not only generally of historical processes but of also of detailed linguistic features. The knowledge of texts and the accessibility of textual knowledge and of texts on which linguistic, lexicographic and other related research work can be based, can be furthered by means of large text corpora like the AAC.

2. The philosophy of the AAC makes it possible to conduct linguistic research in fields rather untypical for the "traditional" corpus-based approach. First, a certain diachronic depth (not only current texts, but also texts from

the 19th century) enables us to face problems of language usage of the 19th century, and to compare its peculiarities with the norms of modern usage. Second, the principle that not only German original texts can be included into the AAC makes it possible to use translations as an empirical basis. Thus, the relevant specific features of lexical units under consideration can be analysed also from a bilingual perspective.

Within the framework of the AAC a Russian-German parallel corpus is planned, based on Dostoevskij's "Idiot"¹, of which there are some twenty translations into German from the Russian original. With a parallel corpus approach, the base text is aligned at the sentence level with the German translations; for details on alignment principles see (Mörth, 2003). The general aim of the parallel corpus is the comparison of the various translations of the "Idiot" in order to test different theoretical approaches to lexical semantics, such as the hypotheses that the semantic structure of a lexical unit (word, idiom or collocation) is a context-sensitive phenomenon.

Here one example to show how different the translations of the same sentence can be.

- (1) *И небось в этом узелке вся ваша суть заключается?*

lit.: „Und dieses Bündel enthält wohl Ihr ganzes Wesen?“

Und in dem Bündel ist vermutlich all Ihr Hab und Gut verborgen? (Scholz 1901)

Und dieses Bündel da enthält wohl alles, was Sie besitzen? (Brauner 1908)

Und dieses Bündel enthält natürlich Ihr ganzes Hab und Gut? (Rahsin 1910)

Und dieses Bündelchen enthält wohl Ihre ganze Habe? (Röhl 1923)

Und nicht wahr, in dem Bündelchen da ist Ihr ganzes Sein beschlossen? (von Walter 1925)

Und in diesem Paket haben Sie wohl Ihre ganze irdische Habe? (Herzog 1925)

Dann enthält wohl auch dies Bündelchen Ihr ganzes Gepäck? (Hoerschelmann 1932)

Und sicherlich enthält dies Bündelchen da wohl Ihre gesamte Habe? (von Guenther 1956)

Und dieses Bündel enthält wohl Euer ganzes Hab und Gut? (Rahsin 1959)

Und in diesem Bündel haben Sie wohl Ihr ganzes Hab und Gut? (Luther 1959)

Und haben Sie in diesem Paket Ihre ganze irdische Habe? (Urban 1968)

Enthält dieses Bündelchen etwa Ihre ganze Habe? (Herboth 1986)

Und in diesem Bündel is' wohl Ihr ganzes Hab und Gut? (Geier 1996)

For this study we used the three translations: two most recent ones, that of Hartmut Herboth [HH 1986] and Svetlana Geier [SG 1996], and the first well-known full translation of the novel by E.K. Rahsin [R 1910].

Abbreviations

[HH 1986] = Translation of "The Idiot" by Hartmut Herboth.

[R 1910] = Translation of "The Idiot" by E.K. Rahsin.

[SG 1996] = Translation of "The Idiot" by Svetlana Geier.

Sources

Dostojewski, F.M. 1910. *Der Idiot*. Aus dem Russischen von E.K.R. In: Dostojewski F.M. Sämtliche Werke. Unter Mitarbeiterschaft von Dmitri Mereschkoski, Dmitri Philossophoff u.a. Hrsg. von Moeller van den Bruck. In 22 Bänden. Erste Abteilung: Dritter und Vierter Band. München/Leipzig: R. Piper & Co. Verlag, s.a. [1906-1919].

Dostojewskij, F. 1996. *Der Idiot*. Aus dem Russischen von S. Geier. 3rd edition. Zürich: Ammann Verlag & Co. (1st edition 1996).

Dostojewskij, F. 1994. *Der Idiot*. Aus dem Russischen von H. Herboth in Dostojewski, F. Sämtliche Romane und Erzählungen, Bde 1-11. Berlin/Weimar: Aufbau-Verlag GmbH (1st edition – in the 20-volume-edition – 1986).

Достоевский Ф. М. 1973. *Идиот* in Достоевский Ф. М. Полное собрание сочинений в тридцати томах, том 8. Ленинград: Наука.

References

Breiteneder, Evelyn/ Dobrovolskij, Dmitrij O. (Hrsg.) (2005): Dostoevskij in Focus: Textlexikographie und Phraseologie. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

¹ Names are transliterated here according to the international transliteration rules, unlike the common popular transcription which is mostly used in German or English translations of Fedor Dostoevskij's work.

- Biber, H./ Breiteneder, E./ Dobrovol'skij, D. (2002): 'Corpus-Based Study of Collocations in the AAC' in A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, Vol. 1*. Copenhagen: Center for Sprogteknologi (CST).
- Mörth, K. (2003): 'Navigating the Markup: Treading New Paths in the Creation of Parallel Corpora' in *COMPLEX-Proceedings*. (7th Conference on Computational Lexicography and Text Research, Budapest 11 April 2003). Budapest.