

ASPAC - Amsterdam Slavic Parallel Aligned Corpus

General remarks

The *Amsterdam Slavic Parallel Aligned Corpus* (ASPAC) is compiled to provide material for research in the field of Slavic contrastive linguistics. It contains parallel aligned texts in Russian and at least one other contemporary Slavic language and usually also one or more of the other European languages (Germanic/Romance).

For an up-to-date list of all languages included in ASPAC, cf. part 03 of this Figshare folder, tab 02 “statistics”.

The Slavic languages best represented in the corpus are Russian, Polish, Czech and Serbian/Croatian (all of them may be studied at the University of Amsterdam). It is, however, our aim to represent a given text in as many languages as possible. Obviously, there are some clear limitations to what is possible in this respect, as texts are rarely translated in all Slavic languages. Therefore, only one text - *Le Petit Prince* by De Saint-Exupéry- is represented in all Slavic languages. The collection is incomplete for most other texts as Lower-Sorbian and Molise-Slavic are not represented and in other cases also other Slavic languages. We think, nevertheless, that our collection is as complete as can be, relative to the availability of translated versions. For an up-to-date list of all languages included in ASPAC, cf. part 03 of this Figshare folder, tab 02 “statistics”.

The main part of the present material consists of translations into Slavic from non-Slavic languages (English, German, Dutch, Swedish, French and Italian). For reasons of comparison a number of contemporary translations into Latin, Spanish, Portuguese, Romanian and Modern Greek have also been included. Because it is rather difficult to find translations from a Slavic language into a sufficient number of other Slavic languages, the number of such texts is still rather modest.

All present texts have a narrative character (novels, stories, diaries etc.). A relatively large part consists of classics of children's literature. An advantage of these texts is the fact that they are widely available and have often been translated more than once, by different translators. Whenever this was possible, such “alternative” translations have also been included in the corpus. (For example: Carroll's classic *Adventures of Alice in Wonderland* is represented by six different Russian translations, and Bulgakov's *Master i Margarita* is represented by eight Polish translations.) For an up-to-date list of all literary works and the amount of translations in each of the represented languages included in ASPAC, cf. part 03 of this Figshare folder, tab 03 “distribution”.

Originally a considerable number of texts have been found in various resources on the Internet. By now a growing number is obtained by way of OCR (using the program *FineReader*). I hereby express my great gratitude to all colleagues and friends who have helped me in obtaining texts and sometimes at various stages of processing them (scanning, OCR, spell-checking, aligning).

At present all ASPAC-included texts are in (extended)-ASCII format (“encoded text”). They have all been aligned, usually on the level of paragraphs. Paragraphs longer than about five lines have usually been broken down into smaller parts. The alignment has been done manually (in *Ultra-Edit*). To obtain a certain degree of uniformity some minor adjustments have usually been made, including e.g. the various kinds of quotation marks have been unified into straight upper quotes (‘ and ’). Indentation at the beginning of paragraphs is made by five spaces. Usually italics, used for emphasis in the original text, are replaced by capitals.

The plain text format allows the texts to be easily processed in various programs and apps, used for text-analysis. One such is Michael Barlow's program *ParaConc*, which has proved to be very useful in my own research on multiple languages in parallel.

It must be emphasized that the material is explicitly meant to be used only for purposes of linguistically oriented research. Any distribution for other purposes is strictly forbidden. Researchers that would like to use (parts of) the corpus are kindly invited to contact me.

ASPAC is continually improved and extended. Suggestions for texts to be included are gratefully considered. Also, reports on mistakes and/or irregularities in the texts are very welcome and very much appreciated.

For a full and detailed description of the (composition) of the database on ASPAC, see 02 Description of ASPAC Metadata.

Full metadata on the texts etc. in ASPAC is included in 03 ASPAC Metadata 2023 FS01.

Adrie Barentsen