

ASPAC - Amsterdam Slavic Parallel Aligned Corpus

Description of ASPAC Metadata database

Published version:

- ASPAC metadata 2023 FSo1.csv

The data in the database are a description of the text material included in the ASPAC text corpus (not published) as well as information about the text corpus that has been compiled and maintained by its creator Adrie Barentsen. The latter data has been selected, translated, updated and presented in the database in 2023 and prepared for publication on UvA-Figshare. The database thus created may be used to note future expansion and developments of the corpus.

The Excel sheet ASPAC metadata 2023 contains the following data, set out here per worksheet.

01 textdata

This worksheet contains the main metadata on the texts included in ASPAC.

Columns included are:

- A. AUTHOR unified (Latin script, transliterated where this applies) [family name, given name, patronymic]

Note that patronymics have usually not been provided, only when customs surrounding a given author's name would require it.

Names whose bearer originally would have used a Cyrillic script have been romanized (International Scholarly System).

- B. TITLE unified (Latin script, Cyrillic transliterated where this applies)

The version of the title here is used throughout the documentation as the basis for reference to a set of original text and its available translations in ASPAC. The title is that of the original publication (sometimes shortened, sometimes including subtitle or description). Titles whose original text uses Cyrillic script have been romanized (International Scholarly System).

- C. Original<>Translation

Two values:

- O = record concerns an original text (and so: original language of compilation);
- T = record concerns a translated text.

- D. ISO 639-3 language code

These standardized language designations are used/added at various instances throughout the database.

There are two language variants which to date lack a separate ISO 639-3 code:

- hrv* is used here for Burgenland Croatian;
- slv* is used here for Resian Slovenian.

- E. ASPAC folder

Name of the folder in which a set of original texts and its translations are stored. The folder names reflect the name of the authors of the respective texts.

- F. ASPAC file name

Note that certain texts have been divided up into part-files. This facilitates operating the concordance software. Examples of part files are Prus, *Lalka* and Kipling *Jungle book*.

G. Size KB [incl. metatext]

Note that the ASPAC text files typically include a very small amount of metatext (version data, retrieval data etc.). The size of the metatext has not been subtracted from the total size data.

H. Word count [incl. metatext]

Note that the ASPAC text files typically include a very small amount of metatext (version data, retrieval data etc.). The size of the metatext has not been subtracted from the word count.

I. ASPAC DATE of inclusion

We have recorded the earliest traceable date. When a date is lacking, the file in question is presently being prepared for inclusion.

Note: when earlier (incomplete) versions have been replaced, the date of the final inclusion is provided here as it marks the date from which a text has been available for research in its entirety.

J. ASPAC NAME text code

The names of the text files usually reflect a title-word, sometimes (part of) the name of the author.

K. ASPAC NAME language code

All ASPAC text file names include an extension for the language of the version. The set of abbreviations used is consistent but not in common use beyond ASPAC.

L. ASPAC NAME extension - additional data

Extensions may be:

- L for Latin script and C for Cyrillic script in the case of Serbian publications.
- [1], 2, 3 for respective translated versions, usually by variant translators. The ordering is per consecutive added text (not by date etc.).
- Translator name abbreviation: e.g. Sh, which in this column is clarified as Bal = Balova, Tat'jana [transator]" to indicate respective translated versions of a text. These extensions are only provided where respective translations occur.

M. Language name in English

Language names in full: this follows Ethnologue as to the exact formulation.

N. Original language

Each record belonging to a set of texts constituting the original and all available translations has been given this indication of the language of the original version: this facilitates the reading of records (especially when filters are used), especially of translated texts as no other records need to be consulted to retrieve the original language.

O. Author(s) as in publication [family name, given name, patronymic]

Original versions of names have been provided, where appropriate in Cyrillic script (no transliterations).

When there are two or more names, they are separated by [space];[space].

P. Title as in ASPAC text file

Original versions of titles have been provided, where appropriate in Cyrillic script (no transliterations).

Q. Translator(s) as in publication [family name, given name, patronymic]

Original versions of names have been provided, where appropriate in Cyrillic script (no transliterations).

R. Dates of author(s)/translator(s)

These have only been provided incidentally. The purpose is to give an impression of the possible date of the recorded text or translation: the dating of texts is important metadata in linguistic research. Surprisingly,

especially translation dates are often omitted in publications and have been hard, sometimes impossible to retrieve. In such instances life-dates of authors/translators may give some insight.

S. Year of publication (year of first publication is different)

These are often lacking due to their absence in actual publications. In instances where ASPAC contains a text other than the version of the first edition, the publication year of the first edition is provided in brackets.

T. Notes

Incidental information that may be of interest.

U. Source as recorded in the text files and original metadata files (for retrieval dates cf. ASPAC DATE of inclusion)

The texts included in ASPAC have been variously collected. Where appropriate, URL's have been provided. For retrieval dates cf. I. ASPAC DATE of inclusion.

02 statistics

This worksheet provides quantificational data of texts per language.

The following columns have been included:

A. STATISTICS / language

List of the languages included in ASPAC.

B. original texts

Note that this is in fact a listing of all text files and so, also the part files rather than complete texts. This may distort actual text numbers somewhat.

C. translated texts

Note that this is in fact a listing of all text files and so, also the part files rather than complete texts. This may distort actual text numbers somewhat.

D. total texts

Note that this is in fact a listing of all text files and so, also the part files rather than complete texts. This may distort actual text numbers somewhat.

E. total size in words

Cf. 01 textdata column G; the metatext contained in each text file is included in this count.

F. total size in KB

Cf. 01 textdata column H; the metatext contained in each text file is included in this count.

G. total size ranking

Gives an indication of the presence of each of the languages included in ASPAC: 1 is the top of the ranking.

03 distribution

This worksheet displays the distribution of all texts over the languages included in ASPAC.

The original languages of ASPAC texts have been listed (column D) but are also marked by shaded cells containing a "1"; unshaded cells represent translations and include the number of translations per language available. Blank cells indicate that a translation is lacking for a particular text.

Columns includes are:

- A. AUTHOR unified (Latin script, transliterated where this applies)
The same as 01 textdata A.
- B. TITLE unified (Latin script, transliterated where this applies)
The same as 01 textdata B.
- C. Original<>translation [hidden]
The same as 01 textdata C. On this sheet this column is hidden and may be manually unhidden.
- D. Original language
The same as 01 textdata D. ISO 639-3 language codes are used here.
- E. Translations, number of
This displays the total number of translations available in ASPAC for each original text listed.

The following columns display the texts available per language.

- F. Belarusian
- G. Bulgarian
- H. Croatian
- I. Croatian, Burgenland
- J. Czech
- K. Dutch
- L. English
- M. French
- N. German
- O. Greek
- P. Italian
- Q. Kashubian
- R. Latin
- S. Macedonian
- T. Polish
- U. Portuguese
- V. Romanian
- W. Russian
- X. Serbian
- Y. Slavomolisano

Z. Slovak
AA. Slovene
BB. Slovene, Resian
CC. Sorbian, Lower
DD. Sorbian, Upper
EE. Spanish
FF. Swedish
GG. Turkish
HH. Ukrainian

05 associated corpora

This sheet provides basic data on the other corpora ASPAC has been associated with.
It contains the following columns:

- A. Corpus organisation
- B. web address
- C. notes