

The electronic corpus of the 17th and 18th century Polish texts (up to 1772) – aims, methods, current state, problems and prospects for development

Włodzimierz Gruszczyński*
wlodekiewa@poczta.onet.pl

* Instytut Języka Polskiego PAN
 Al. Mickiewicza 31, 31-120 Kraków

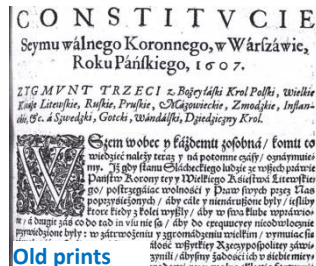
Maciej Ogrodniczuk**

maciej.ogrodniczuk@ipipan.waw.pl
 **Instytut Podstaw Informatyki PAN
 Ul. Jana Kazimierza 5, 01-248 Warszawa

Project factsheet:

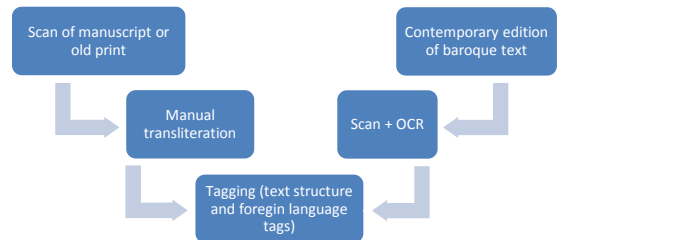
- **Funding:** Polish Ministry of Science and Higher Education, National Programme for the Development of Humanities grant (contract number 0036/NPRH2/H11/81/2012)
- **Duration:** 2013-2018
- **Coordinating body:** Institute of Polish Language, PAS
- **Cooperation:** Institute of Computer Science, PAS
- **Principal investigator:** Włodzimierz Gruszczyński.

Types of source texts and methods of their processing:



Manuscripts

Old prints



Tools:

- Morphological analyser *Morfeusz*¹;
- Tagger *Polita*²,
- *Poliqarp* – new version.

Automatic conversion to TEI XML

Automatic transcription for today's spelling (if necessary)

Morphosyntactic tagging (partly manually partly automatic)

The majority of further processing will be carried out on the latter form, while search results will be retrieved in transliteration.

Applications:

- The data of the resulting corpus is already used in the parallel task of compiling the *Electronic dictionary of 17th and 18th century Polish* (<http://sxvii.pl>).
- Our data will be later used to create the diachronic model of Polish inflection.
- KORBA will make a historical subcorpus of the National Corpus of Polish (<http://nkjp.pl>).

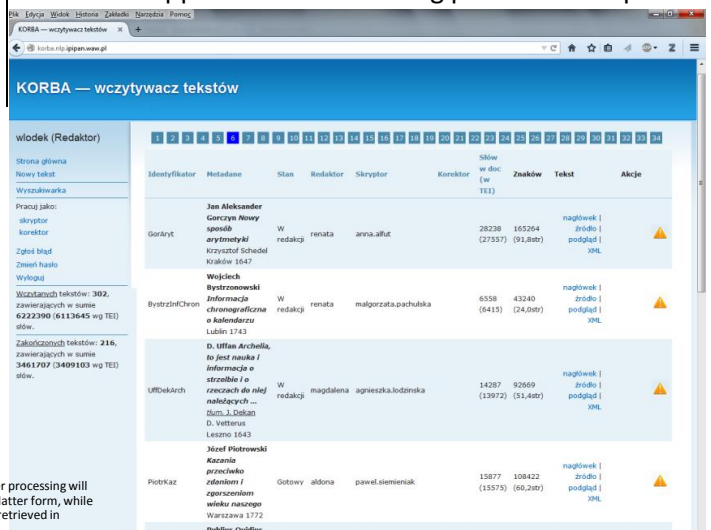


Project aims:

- creation of a fairly balanced corpus of Polish texts dating between 1601 and 1772;

The corpus features:

- **size:** ≈ 12 million tokens;
- **structural annotation** – with rich bibliographic, stylistic, genologic and structural metadata, intended to enable refined search and provide page-aware location of each segment in original text
- **linguistic annotation** – of all foreign elements, with respect to language identification
- **morphosyntactic annotation** – of a 0.5 million token-size subcorpus, planned to be manually annotated and then used to train automated tagger to be applied to the remaining part of the corpus.



The current state of our project:

- over 300 texts,
- over 6,2 M tokens,
- all texts tagged structurally,
- in all the texts are marked all foreign tokens,
- ready morphosyntactic tag set.

The project website will be available soon at:
<http://e-korba.pl>

¹ <http://sgjp.pl/morfeusz/index.html.en>
² <http://zil.ipipan.waw.pl/Polita>