



Annotation tools and interfaces for managing transcription of old Polish texts

Włodzimierz Gruszczyński

Dorota Adamiec

Renata Bronikowska

Małgorzata Majewska

Michał Lenart

Maciej Ogrodniczuk

Institute of Polish Studies

Polish Academy of Sciences

Institute of Computer Science

Polish Academy of Sciences

PALC 2014

Łódź, November 22, 2014

We will present:

The **set of tools** currently used for managing transcription of 17th and 18th century Polish texts included into **newly created corpus**.

The tools are:

- the editing environment in a familiar word processor setting
- the Web text management application intended to:
 - collect Rich Text Format files
 - trigger TEI P5 XML converters
 - provide HTML visualization
 - administer correction process.

Project factsheet:

- Title: Electronic corpus of 17th and 18th century Polish texts
- 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
- Principal investigator: Włodzimierz Gruszczyński

Corpus summary:

- a considerably large ($\approx$ 12 million tokens) collection of 17th and 18th century Polish texts (up to 1772)
- featuring annotation of text structure and language
- intended to supplement the National Corpus of Polish (NKJP) with a diachronic aspect
- can come in useful for all forms of historical linguistic studies (e.g. evolution of Polish grammar, regional and social variety of historical Polish), as well as literary studies and editorial work.

Sources:

- **manual transcription!**
- texts offered by IMPACT project
- texts acquired from publishers
- texts acquired from various Internet sources

sometimes already in XIX-, XX- or XXI-century transcription.

Text sample:

A sample of text in a Gothic script, likely from a historical manuscript. The text is written in a dark ink on a light background. The script is highly stylized, with large, ornate letters and a dense, compact layout. The text reads: "Począyna się Historya o porażce Sakramentarzow y zabiciu wodzów ich Książecia de Conde." The text is arranged in three lines, with the first line being the longest and the last line being the shortest.

Począyna się Historya o porażce
Sakramentarzow y zabiciu
wodzów ich Książecia de Conde.

There are many options out there:

- Oxygen-based Ediarum (<http://www.bbaw.de/en/telota/software/ediarum>)
- Web-based Transcription Desk of Transcribe Bentham Initiative (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham)
- many other digital work environments for transcription, editing and publication of old texts.

But we need solution which is:

- much simpler and popular in the humanities
- available for offline usage
- already tested on variety of inexperienced users.

A familiar word-processing environment:

- MS Word-based
- styles used to produce pseudo-XML
- macros implemented to apply basic processing
- keyboard shortcuts
- one template (to rule them all :)

Two types of markup

- single-level structures represented with **character styles**
- markers (e.g. non-textual placeholders: [EQUATION]) and structural markup (e.g. section structure: [DOCUMENT BEGIN] and [DOCUMENT END]) represented with **paragraph styles** (texts in square brackets).

Vertical dimensions

- language, simulated with text colour (making it possible to mark up e.g. a Latin interjection in a text title)
- editor comments (uncertainty of interpretation, clear corrections etc.) marked with angle brackets.

[DOCUMENT BEGIN]

[PAGE BEGIN]

[ILLUSTRATION]

RELATIA WIAZDV SOLENNEGO.

**do Krákowá, y Zamku, ziák naywiększą
Mágnificentią, y Pompą.**

**NAIASNIEYSAEGO, y NIEZWYCIEZONEGO
KROLA IEGOMOSCI
POLSKIEGO; etc. etc.**

**XIAŻĘCIA ELEKTORA SASKIEGO,
etc. etc.**

Odpráwionego, **Die 12. Septembris**, y
**consequenter Praeliminaria Actûs
Coronationis** opisáne.

IAko zâwsze à **supremo love principium** kâzdey rzeczy, naleŹy, czynić. ták Naiâśnieyszy
KROL iegomość **ab Octava** ieszcze **praesentis**, to iest od NiedŹiele. **Solennis Festivitatis**,
Nativitatis B. V. M. post Confessionem Sacramento peractam. W Koâście SS. Pietró u

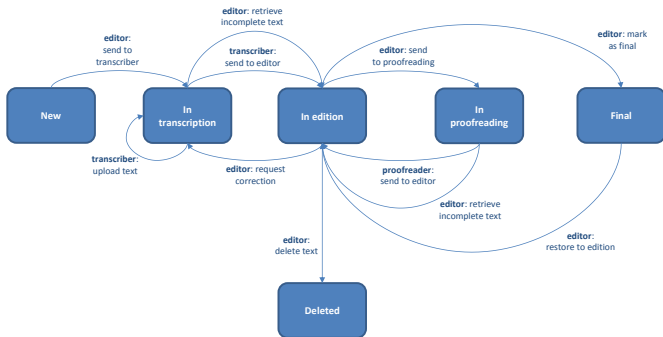
The process involves three user roles:

- editor — initiator and manager of the process,
- transcriber — providing a transcribed, structurally and lexically annotated texts,
- proofreader — verifying the annotation.

Step by step:

- 1 editor: create a minimal metadata description of a new text
- 2 editor: assign it to a transcriber (to maintain text-life-long relationship with the same set of experts)
- 3 transcriber: prepare the transcription using Microsoft Word
- 4 transcriber: save resulting document in RTF
- 5 transcriber: upload the text via Web browser to a dedicated site
- 6 editor: approve the text, forward it to proofreading or send it back to transcription if errors are found

Text processing workflow



Modelled after Drupal CMS with two basic views:

- **list view** showing all texts being processed and providing interface for executing actions (changing text workflow state, displaying its metadata and current content etc.)
- **item view** presenting interface to managing metadata of an individual text.

On upload texts are converted to TEI P5 XML. HTML presentation is also available in the system for easier review.

KORBA - wczytywacz tekstów

renata (Redaktor)

[Strona główna](#)

[Nowy tekst](#)

[Wyszukiwarka](#)

Pracuj jako:

[skrypty](#)

[korektor](#)

[Zgłoś błąd](#)

[Zmień hasło](#)

[Wyloguj](#)

Wczytanych tekstów: 155,
zawierających w sumie 4378233
(4297839 wg TEI) słów.

Zakończonych tekstów: 69,
zawierających w sumie 1014015
(994185 wg TEI) słów.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Identyfikator	Metadane	Stan	Redaktor	Skrypty	Korektor	Słów w doc (w TEI)	Znaków	Tekst	Akcje
PetrSInst	Sebastian Petrycy <i>Instrukcja albo nauka jak się sprawować czasu moru</i> Mikołaj Łob Kraków 1613	Gotowy	renata			11077 (10942)	72526 (40,2str)	nagłówek źródło podgląd XML	przywróć do redakcji 
BirkEgz	Fabian Birkowski O <i>Egzorbanacjach kazania dwoje</i> Andrzej Piotrkowczyk Kraków 1632	Gotowy	renata	renata		8657 (8340)	57010 (31,6str)	nagłówek źródło podgląd XML	przywróć do redakcji 
NowZwyc	Anonim Nowina o <i>zyciestwie cesarza z Wiednia do Lublina przyniesiona 17 września 1634</i> 1634	Gotowy	renata	renata		206 (201)	1512 (0,8str)	nagłówek źródło podgląd XML	przywróć do redakcji
TwarKPoch	Kasper Twardowski <i>Pochodnia miłości bożej</i> Walerian Piątkowski Kraków 1628	Gotowy	renata	renata	renata	10322 (10103)	67241 (37,3str)	nagłówek źródło podgląd XML	przywróć do redakcji 

KORBA - wczytywacz tekstów

renata (Redaktor)

[Strona główna](#)
[Nowy tekst](#)
[Wyszukiwarka](#)
[Pracuj jako:](#)
[skryptyr](#)
[korektor](#)
[Zgłoś błąd](#)
[Zmień hasło](#)
[Wyloguj](#)

[Powrót do strony głównej](#)

Edytuj metadane

Id: *	PetrSInst	
Tytuł: *	Instrukcja albo nauka jak się sprawować czasu moru	
Autor: *	Sebastian Petrycy	☐ Anonimowy
Tłumacz:		☐ Anonimowy
Drukarnia:	Mikołaj Lob	
Rok:	1613	
Rok wydania niepewny (wydano nie wcześniej, niż podany rok):	☐	
Miejsce:	Kraków	
Wydanie uwspółcześnione:	☐	
Pochodzi z korpusu IMPACT:	☐	

Aktualizuj

National Corpus of Polish-based:

- stand-off annotation
- XML TEI P5

text_structure.xml:

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus>
  <xi:include href="KORBA_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <front>
        <!-- title and post-title page -->
      </front>
      <body>
        <!-- subsequent pages -->
```


Metadata

- structural elements modelled explicitly (all content preserved as it appeared in the document)
- document metadata may repeat these values.

Title page fragment:

```
<front>
  <titlePage>
    <docTitle>
      <titlePart type="main">
        The DUNCIAD, VARIOURVM.</titlePart>
      <titlePart type="sub">
        WITH THE PROLEGOMENA of SCRIBLERUS.</titlePart>
      <docAuthor>Gal Anonim</docAuthor>
    </docTitle>
    <docImprint>
```

Represented with @typed <div> elements:

```
<body>
  <div type="volume">
    <p>...</p>
    <div type="part_level1">
      <!-- Section 1 -->
      <p>...</p>
    </div>
    ...
  </div>
</body>
```

Several division types are distinguished: volume, four section levels, motto, preface, epilogue, index and table of contents.

Represented with typed `<head>` of the title:

```
<head type="title">
```

```
  0 nieporządneý chćiwośći iedzenia,</head>
```

```
<head type="subtitle">
```

```
  to iest, o apetićie przećiwnym ludzkiemu  
  przyrodzeniu.</head>
```

```
<head type="summary">
```

```
  POETA idąc zą zdánien dawnych Pogan, trzy początki  
  świátá, y rzeczy będących ná nim, bydz rozumiał:  
  Bogá, Lepsze przyrodzenie, y Chaos.</head>
```

Absent in NKJP:

- preserved with `<pb n="number of the next page"/>` tag
- explicit indication of the first page: `<pb n="1"/>`
placed also before the `<titlePage>`
- page breaks can occur inside the paragraph, logically separating pagination and structural divisions.

Marginal notes preserved inline:

```
<p>Wielkość zamyka w sobie/  
    wielkie ábo nie wielkie iedzenie/  
    y częste iedzenie. Co się tknie pierwszego.  
    <note type="margin">Iák wiele ieść</note>  
    Iák wiele mamy ieść/  
    naznacza nam všmierzenie głodu/  
    to iest/ gdy się zniešie przez wzięty pokarm  
    przyrodzona chćiwość iadłá.</p>
```

Other:

- `<figure/>` or `<formula/>` placeholders preserved in place
- foreign interjections embedded in `<foreign>` tag with language codes preserved in `xml:lang`.

RTF-to-TEI converter:

- preserves all texts present in the source
- detects potential errors in markup, mostly signalling structural inconsistencies or missing vital information.

Error/warning types detected:

- broken tags (inability to parse text to a valid XML tag)
- misplaced tags (ending of an unopened tag or tags out of context such as new page before end of a previous one)
- fragments of text outside the start- and end-of-page tags (probably resulting from missing tags)
- inconsistent foreign language marking (very common warning; when only part of the foreign word or punctuation is marked).

To conclude:

- the setting proves very effective in the transcription process
- it enables a precise structural annotation
- thanks to automatic XML conversion the ultimate text format is consistent with the widely accepted TEI standard
- the application organizes a fluent workflow between all users in the transcription process and facilitates error correction
- the solution can be easily customized to handle other types of texts which makes it suitable for other corpus building projects.

Thank you!



I
MERKVRIVSZ
P O L S K I,

Dzieie wſzytkiego ſwiátá w ſobie zamykáiaący
dla Informácyey poſpolitey.

W Krákwie 3. Ianuarij 1661.