Possibilities of searching the corpus of baroque texts

Goals and objectives

Renata Bronikowska
Institute of Polish Language
Polish Academy of Sciences

Project factsheet

- Title: Electronic corpus of 17th and 18th century Polish texts (up to 1772)
- Cryptonym: KORBA (**kor**pus **ba**rokowy ‘baroque corpus’)
- 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, Polish Academy of Sciences
- Cooperation: Institute of Computer Science, Polish Academy of Sciences
- Principal investigator: Włodzimierz Gruszczyński

Corpus summary

- A considerably large (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
- Transcription and annotation of the old texts is held in a MS Word document
- MS Word texts are converted to XML TEI P5 representation format
- As a part of NKJP the baroque corpus will be supported by the NKJP search engine named Poliqarp

Sample text – MS Word format

[DOCUMENT BEGIN]

[TITLE PAGE BEGIN]

PIECHOTNE CWICZENIE

Albo WOIENNOSC PIESZA Ktorą Láćinnicy Pedestrem Militiam názywáią. Wodzom, Pułkownikom, y wszelkiey Woienney Stárszynie. Lubo ktoszkolwiek iest Woienney správy Miłośnikiem.

Do wiadomośći podána.

Przez BLAZEIA LIPOWSKIEGO

Nakładem Ierzego Forstera, I.K.M. Bibli opoli.<?>

W KRAKOWIE

W Drukárni v Wdowy Lukaszá Kupiszá, I.K.M.

Typogr. *Roku Panskiego 1660.*

[TITLE PAGE END]

Sample text – XML format

```
- <teiCorpus>
    <xi:include href="KORBA_header.xml"/>
  - <TEI>
      <xi:include href="header.xml"/>
    - <text>
      - <front>
          <pb/>
        - <titlePage xml:id="txt_1-titlePage">
          - <docTitle xml:id="txt_1.1-docTitle">
              <titlePart type="main" xml:id="txt_1.1.1-titlePart">PIECHOTNE CWICZENIE</titlePart>
            - <titlePart type="sub" xml:id="txt_1.1.2-titlePart">
                Albo WOIENNOSC PIESZA Ktorą Láćinnicy
                <foreign xml:lang="la" xml:id="txt_1.1.2.1-foreign">Pedestrem Militiam</foreign>
                názywáią. Wodzom, Pułkownikom, y wszelkiey Woienney Stárszynie. Lubo ktoszkolwiek iest Woienney správy Miłośnikiem.
              </titlePart>
            </docTitle>
            <note xml:id="txt_1.2-note">Do wiadomośći podána.</note>
            <docAuthor xml:id="txt_1.3-docAuthor">Przez BLAZEIA LIPOWSKIEGO</docAuthor>
          - <note xml:id="txt_1.4-note">
              Nakładem Ierzego Forstera, I.K.M. Bibli
              <uncertain reason="unreadable" xml:id="txt_1.4.1-uncertain">opoli.</uncertain>
            </note>
          - <docImprint xml:id="txt_1.5-docImprint">
              <pubPlace xml:id="txt_1.5.1-pubPlace">W KRAKOWIE</pubPlace>
              <publisher xml:id="txt_1.5.2-publisher">W Drukárni v Wdowy Lukaszá Kupiszá, I.K.M. Typogr.</publisher>
              <docDate xml:id="txt_1.5.3-docDate"> Roku Panskiego 1660.</docDate>
            </docImprint>
          </titlePage>
```

Two basic objectives:

- faithful mapping of the notation and structure of old texts
- providing the user with convenient methods of finding information in the corpus

What can be found in the corpus?

- phrases with a graphical form other than contemporary
- information about illegible fragments
- information about the context of the searched expression
- information about the exact location of the searched expression in the text

Differences in spelling

	Old notation	Contemporary notation
Dashed letters	czás śiwy	czas ‘time’ siwy ‘grey’
Lack of diacritics	ktory byl	który ‘which’ był ‘was’
Some letters replaced with another ones	vwaga iest	uwaga ‘comment’ jest ‘is’
Regular abbreviations	meᵒ wiatrẽ	mego ‘my: Gen.sg’ wiatrem ‘wind: Instr.sg’

- Replacement rules implemented to the search engine will enable to find a searched expression regardless of spelling

Irregular differences in spelling

- Deciphered irregular abbreviations
  - talẽt <talent>
  - <choice xml:id="txt_8.4.41.1-choice">
    <orig xml:id="txt_8.4.41.1.1-orig">talẽt</orig>
    <reg xml:id="txt_8.4.41.1.2-reg">talent</reg>
    </choice>
- Corrected spelling mistakes
  - wszięd̨ze <wszędzie>
  - <choice xml:id="txt_7.2.103-choice">
    <orig xml:id="txt_7.2.103.1-orig">wszięd̨ze</orig>
    <reg xml:id="txt_7.2.103.2-reg">wszędzie</reg>
    </choice>

Illegible and questionable fragments

- Information about the number of illegible characters
  - COMED<..>
  - COMED<gap reason="unreadable" unit="chars" quantity="2"/>
- Information about possibility of incorrect interpretation of a word
  - gospodarz <?>
  - <uncertain reason="unreadable" xml:id="txt_7.6.5-uncertain">gospodarz</uncertain>
- Information about surprising yet correct spelling
  - gorzski <!>
  - <sic xml:id="txt_5.2.3.86-sic">gorzski</sic>

Foreign language fragments

- Foreign language fragments within the sentence are transcribed and tagged:
  - WOIENNOSC PIESZA Ktorą Láćinnicy Pedestrem Militiam názywáią.
  - <foreign xml:lang="la" xml:id="txt_1.1.2.1-foreign">Pedestrem Militiam</foreign>
- Such fragments will be excluded from searching so that the search engine will not show foreign words having form identical to Polish words, e.g.
  - Lat. *mali* ‘bad (plural)’ vs. Pol. *mali* ‘small (plural)’

Non-transcribed parts of the text

Structural tag	TEI tag
[ILLUSTRATION]	<figure/>
[EQUATION]	<formula/>
[MUSICAL NOTATION]	<notatedMusic/>
[FOREIGN LANGUAGE FRAGMENT]	<foreign xml:lang="kod"/>
[NON-LATIN ALPHABET]	<foreign xml:lang="x-nlws"/>

Context information – non-searchable fragments

Search query	Search result
przed [] [] laty ‘years ago’ (possible two tokens between the searched words)	**przed** dwiema circiter **láty**
tak [] jako ‘such as’ (possible one token between the searched words)	zostáie tedy **tak**: [EQUATION] **iako** y sam rozum pokázuie

Context information – marginal notes

- If the searched expression is found in the marginal note, it will be displayed together with the paragraph context

Marginal note	Beginning of the paragraph context
Powietrza rozmáitość.	Powietrze morowe/ ábo iest domorodne/ co się domá sámo vrodźi: ábo zánieśione z inszey kráiny.

Pagination

- During the transcribing an identifier is assigned to each page, regardless of whether it occurs in the transcribed text
- As page identifier is regarded:
  - page number (1, 2, 3, 4…)
  - label of sheet parts (A, Av, A2, A2v…)
  - number assigned to a page by a copier (1 nlb., 2 nlb. …)
- Each token will be linked to the appropriate page identifier - this will allow to accurately indicate the location of the searched expression in the text

To conclude:

- Searching of old texts' corpus creates additional difficulties in comparison with corpora of contemporary texts (spelling differences, illegible characters, Latin interjections, marginal notes, various pagination systems)
- Thanks to more accurate marking the search engine will be able to overcome these difficulties and provide a wealth of additional information about the searched texts

Thank you!