Embedding Transcription and Transliteration Layers in the *Digital Library of Polish and Poland-Related News Pamphlets*

Maciej Ogrodniczuk[1(✉)] and Włodzimierz Gruszczyński[2]

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland
maciej.ogrodniczuk@ipipan.waw.pl
[2] Institute of Polish Language, Polish Academy of Sciences, al. Mickiewicza 31, 31-120 Kraków, Poland
wlodzimierz.gruszczynski@ijp.pan.pl

Abstract. The paper presents an experiment intended to overcome the problem of searching for different spelling variants in old Polish prints. In the case of *The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* two concurrent layers of text (transliteration and transcription) underlying selected digital library items are available in the related *Electronic Corpus of the 17th and 18th Century Polish Texts (until 1772)*. Both variants are retrieved and a double-hidden layer representation of a sample item is prepared and made available for textual searching in a PDF containing its scanned image. The experiment can be generalized to other libraries dealing with multiple concurrent textual interpretations of graphical items.

Keywords: Digital library · Transcription · Transliteration · Middle Polish

1 Introduction

The Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries (Pol. *Cyfrowa Biblioteka Druków Ulotnych polskich i Polski dotyczących z XVI, XVII i XVIII wieku*, hereafter abbreviated CBDU[1] [7–9] is a thematic digital library of approx. 2 000 Polish and Poland-related pre-press documents (ephemeral prints—short, disposable and irregular informative publications) dated between 1501 and 1729 (all surviving documents of this kind described in scientific publications, particularly by Zawadzki [11]).

[1] See https://cbdu.ijp.pan.pl/.

The work was financed by a research grant from the Polish Ministry of Science and Higher Education under the National Programme for the Development of Humanities for the years 2019–2023 (grant 11H 18 0413 86, grant funds received: 1,797,741 PLN).

H.-R. Ke et al. (Eds.): ICADL 2021, LNCS 13133, pp. 54–60, 2021.
https://doi.org/10.1007/978-3-030-91669-5_5

The library is managed by EPrints [1], a database management system configured to using extended metadata: apart from the usual ones such as item title. author, publication date etc. CBDU defines extended metadata such as historical comments, glossaries of foreign interjections, explanations of lesser-known background details and relations between library objects (translations, adaptations, alterations of the base text, their alleged sources etc.) For 1404 prints (out of 2011 present in the digital library) scans of actual microfilmed documents are available, stored in multipage PDF files displayed on demand in a side pane.

The PDFs are currently graphical files only but textual transcriptions are planned to be soon added using the process described by [10]. The texts will be acquired from *The Electronic Corpus of the 17th and 18th Century Polish Texts (until 1772)* (Pol. *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do roku 1772)*) [6], also referred to as the *Baroque Corpus* (Pol. *Korpus Barokowy* – hence its acronym KORBA[2] and merged into the PDFs as hidden text using the OCR tools.

The process is straightforward but not complete: the orthography of middle-Polish text may be much different from contemporary Polish, hindering usability of the searchable text even when it is properly embedded in the resulting PDF. In the paper we present the method of overcoming this inconvenience by offering double textual layer with spelling variants of the text.

2 Related Work and Background Information

Although the creators of CBDU intended to add two spelling versions of library items - transliterated and transcribed ones[3], the task proved impossible within the project timeframe and only one sample text was transcribed[4]. The implementation of KORBA corpus allowed for transcribing 40 more prints, amounting to over 200,000 tokens (words and punctuation).

Apart from the rich structural annotation available in the corpus (e.g. marked page numbers for text ranges or foreign language interjections), the texts have been encoded in both transcribed and transliterated forms, cf. *nazajutrz* (identical to a contemporary Polish word) and older *nazaiutrz*):

```
<fs type="morph">
  <f name="orth">
    <string>nazajutrz</string>
  </f>
  <f name="translit">
    <string>nazaiutrz</string>
  </f>
</fs>
```

[2] See also https://korba.edu.pl/overview?lang=en.

[3] The intention of transliteration is accurate representation of the graphemes of a text while transcription is concerned with representing its phonemes.

[4] See *Translation into Contemporary Polish* section of print 1264 at http://cbdu.ijp.pan.pl/12640/.

What must be noted, KORBA creators decided not to follow the extremely faithful transliteration featured e.g. in the IMPACT corpus [3][5] called *strict diplomatic* or *facsimile* transcription, preserving the distinctions irrelevant for many users. For example, minor differences in certain graphical or typographical features were neglected such as different variants of the letter *ę*, all Unicode-encoded as `LATIN SMALL LETTER E WITH OGONEK` (also used in comtemporary Polish) and not as other existing Unicode representations such as `LATIN SMALL LETTER E WITH STROKE`. Obviously, both transliterated and transcribed variants retain their capitalisation as per the original texts.

Both layers are available not only in the source corpus files but can also be queried in the corpus search engine[6], featuring e.g. an interface add-on for inputting letters which are absent from modern Polish orthography but which may appear in the transliteration layer, the switch between displayed layer mode (modernized or transliterated) or attributes for querying both layers (in our example, `[orth=nazajutrz"]` and `[translit=nazaiutrz"]` respectively) – see Fig. 1.

Still, the corpus search interface is separated from the digital library and offers access only to generated concordances of text, linked to the *Electronic Dictionary of the 17th–18th Century Polish* (Pol. *Elektroniczny słownik języka polskiego XVII i XVIII wieku*) [4,5] but not to actual scans. What could offer much more flexibility for researchers accessing both layers would be to enable search in them directly in the browser, most conveniently in the PDF file containing both the scanned version of an item and its transcribed/transliterated text.

Table 1 shows the importance of this issue: over 32% of words in the 40 prints from CBDU currently available in the corpus[7] are subject to variation between transcription and transliteration and over 20% of these differences are significant ones (going beyond accent variants corresponding to characters currently unused in Polish alphabet such as *a/á* and *e/é*, as e.g. in *potrzébná*).

Table 1. Counts of differences between transcription and transliteration layers

	Tokens	Percentage
No difference	135 101	67.11%
Punctuation	11 914	5.92%
Difference	66 199	32.89%
Significant difference	40 911	20.32%
All	201 300	100.00%

[5] *Improving Access to Texts* international project, see also http://www.impact-project.eu.

[6] See https://korba.edu.pl/.

[7] KORBA project is being continued until 2023 and several new texts from CBDU will be included in the corpus.

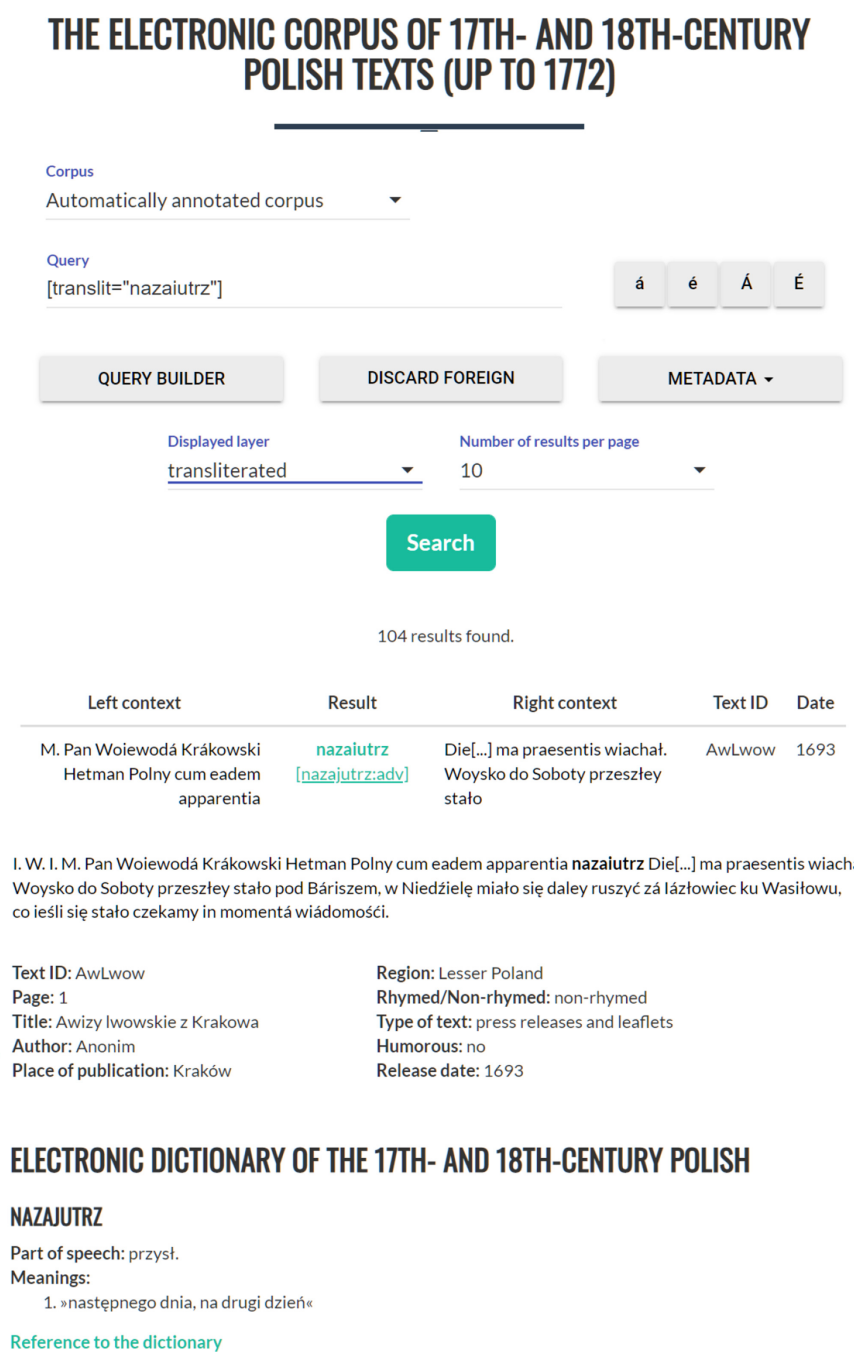

Fig. 1. Corpus search interface: querying transliteration layer and dictionary linking

3 Embedding Transcription and Transliteration Layers in a PDF File

Storing text in PDF along with the scanned version of an item is usually done using the invisible text rendering mode (either by drawing text in the background which is then covered by the scanned image or drawing the invisible text in the foreground of the scanned image). Embedding spelling variants is then possible by manipulating the hidden textual layer in various ways, e.g. by: adding word variants directly in the single hidden layer. This method proves inconvenient for the reader since hidden text is not aligned in the page view (see Fig. 2).

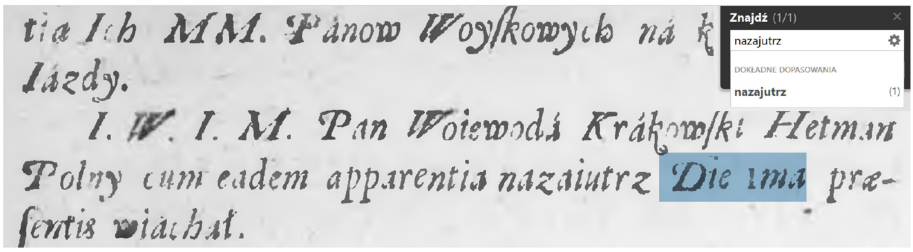

Fig. 2. Word variants added one by one in the hidden layer

A much better solution can be obtained by merging both hidden textual layers into a single file so that they can be searched concurrently. This function is easily obtained using many freely-available PDF manipulation libraries such as *PDFtk*, the PDF toolkit[8] or *CPDF*, Coherent PDF Command Line Toolkit[9] by merging two files each containing a separate hidden layer using *background* or *combine pages* options:

```
pdftk transcription.pdf multibackground transliteration.pdf
      output merged.pdf

cpdf  -combine-pages transcription.pdf transliteration.pdf
      -o merged.pdf
```

The result of this process can be found for print 1179 in CBDU[10] where both transcribed *nazajutrz* and transliterated *nazaiutrz* can be searched for, pointing at exactly the same scan segment as in Fig. 3.

The two tools tested in the process are just examples and several other PDF command-line utilities can be applied to merge multiple textual layers. However, it must be noted that the selected tool should be tested for compatibility with

[8] See https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/.
[9] See https://community.coherentpdf.com/.
[10] See https://cbdu.ijp.pan.pl/id/eprint/11790/.

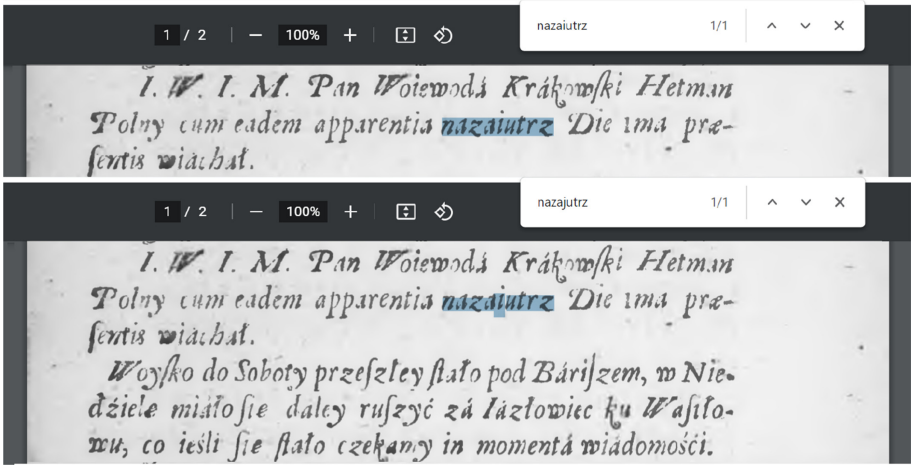

Fig. 3. Two spelling variants available for search in a single item

popular PDF viewers since e.g. files output by *PDFtk* can cause search problems when open offline in Acrobat Reader while they are perfectly processed with the default PDF viewer in the browser[11].

The hidden layers stored this way do not have any negative impact on copying of text from the PDF file (i.e. the text is not copied twice which happens with using the word variants method) and the layer available for copying can be selected by placing it in the foreground.

4 Conclusions and Future Plans

The presented experiment intended to show the method of supplementing CBDU but also similar libraries of old texts with several interpretation layers. Whenever suitable textual data is available, it can be encoded directly in the file to provide the best user experience and make the PDF file function independently on any digital library management system.

An additional experiment showed that embedding more than two layers in a single PDF file is also possible which gives many opportunities for encoding various interpretation over a single graphical layer of an item.

Another path of implementation could lead to integration of search in the digital library with both the corpus and the dictionary. From the point of view of the corpus search user it might be useful to view the retrieved concordance directly on the scan of the source document. Similarly, search in the dictionary could be illustrated with examples shown in actual context, on the respective page of the scanned item.

[11] Tested with Chrome 91.0.4472.124, Firefox 90.0 and Edge 91.0.864.67.

Last but not least, search in multiple PDF files in the form of so called graphical concordance (a Key Word in Context index with the scan snippets created on the fly) could be implemented following the method used by *Poliqarp for DjVu* [2], linking search results to a series of scans with highlighted hits.

References

1. EPrints Manual (2010). http://wiki.eprints.org/w/EPrints_Manual
2. Bień, J.S.: Efficient search in hidden text of large DjVu documents. In: Bernardi, R., Chambers, S., Gottfried, B., Segond, F., Zaihrayeu, I. (eds.) AT4DL/NLP4DL -2009. LNCS, vol. 6699, pp. 1–14. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23160-5_1
3. Bień, J.S.: The IMPACT project Polish Ground-Truth texts as a Djvu corpus. Cogn. Stud. 75–84 (2014)
4. Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M., Woliński, M.: The use of electronic historical dictionary data in corpus design. Stud. Pol. Linguist. **11**(2), 47–56 (2016). https://doi.org/10.4467/23005920SPL.16.003.4818
5. Gruszczyński, W. (ed.): Elektroniczny słownik języka polskiego XVII i XVIII w. (Electronic Dictionary of the 17th and the 18th century Polish, in Polish). Institute of Polish Language, Polish Academy of Sciences (2004). https://sxvii.pl/
6. Gruszczyński, W., Adamiec, D., Bronikowska, R., Wieczorek, A.: Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. - problemy teoretyczne i warsztatowe. Poradnik Językowy (8/2020 (777)), 32–51 (2020). https://doi.org/10.33896/porj.2020.8.3
7. Gruszczyński, W., Ogrodniczuk, M.: Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski dotyczących z XVI, XVII i XVIII w. w nauce i dydaktyce (Digital Library of Poland-related Old Ephemeral Prints in research and teaching. In: Polish). In: Materiały konferencji Polskie Biblioteki Cyfrowe 2010 (Proceedings of the Polish Digital Libraries 2010 Conference), Poznań, Poland, pp. 23–27 (2010)
8. Ogrodniczuk, M., Gruszczyński, W.: Digital library of Poland-related old ephemeral prints: preserving multilingual cultural heritage. In: Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, pp. 27–33 (2011). http://www.aclweb.org/anthology/W11-4105
9. Ogrodniczuk, M., Gruszczyński, W.: Digital library 2.0 – source of knowledge and research collaboration platform. In: Calzolari, N., et al. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavík, Iceland, pp. 1649–1653. European Language Resources Association (2014). http://www.lrec-conf.org/proceedings/lrec2014/pdf/14_Paper.pdf
10. Ogrodniczuk, M., Gruszczyński, W.: Connecting data for digital libraries: the library, the dictionary and the corpus. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) ICADL 2019. LNCS, vol. 11853, pp. 125–138. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34058-2_13
11. Zawadzki, K.: Gazety ulotne polskie i Polski dotyczące z XVI, XVII i XVIII wieku (Polish and Poland-related Ephemeral Prints from the 16th-18th Centuries, in Polish). National Ossoliński Institute, Polish Academy of Sciences, Wrocław (1990)