Embedding transcription and transliteration layers in the Digital Library of Polish and Poland-related News Pamphlets

Maciej Ogrodniczuk, Wlodzimierz Gruszczynski

Institute of Computer Science / Institute of Polish Language
Polish Academy of Sciences

The 23rd International Conference on Asia-Pacific Digital Libraries (ICADL 2021)
December 2, 2021

Our experiment

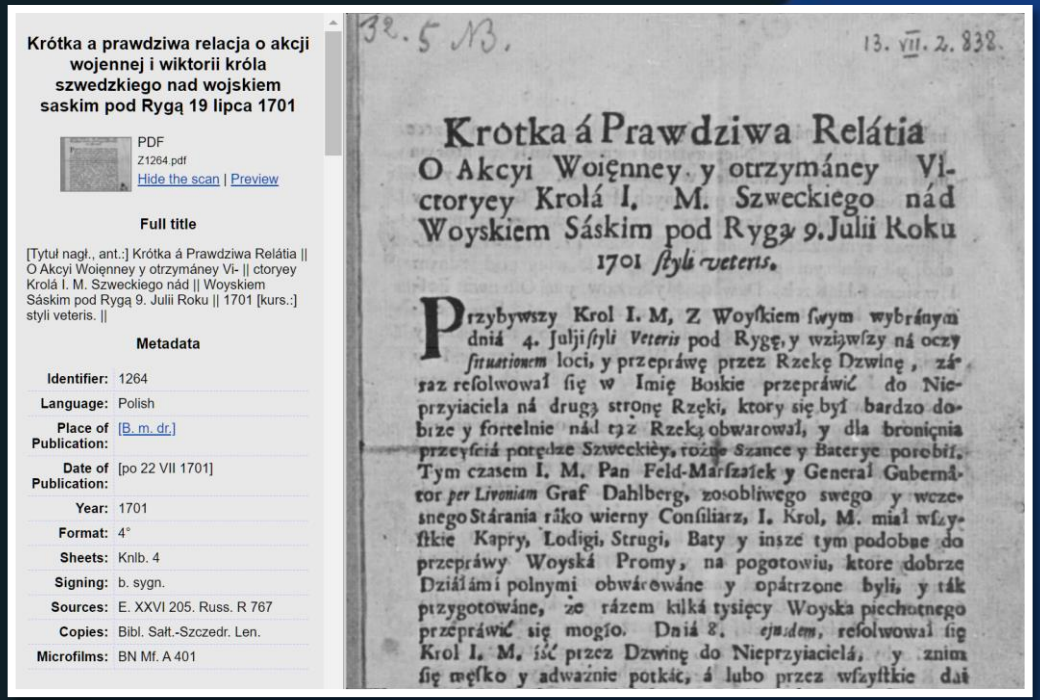
In a nutshell:

- old Polish prints are made available in an existing digital library as graphical PDFs (scanned images of prints)
- an electronic corpus of old Polish texts contains texts from the prints in two variants: transcription and transliteration
- a corpus search engine is available but we need to be able to search also in the digital library
- we can easily search in metadata
- but we would also like to search in both variants inside graphical PDFs

The Digital Library of Polish Prints from the 16th–18th Centuries

The source of images:

- a thematic DL with over 2 000 prepress documents
- managed by Eprints
- soon 200+ items to be extended with texts
- available at https://cbdu.ijp.pan.pl



Krótka a prawdziwa relacja o akcji wojennej i wiktorii króla szwedzkiego nad wojskiem saskim pod Rygą 19 lipca 1701

PDF
Z1264.pdf
Hide the scan | Preview

Full title

[Tytuł nagł., ant.:] Krótka á Prawdziwa Relátia || O Akcyi Woięnney y otrzymáney Vi- || ctoryey Krolá I. M. Szweckiego nád || Woyskiem Sáskim pod Rygą 9. Julii Roku || 1701 [kurs.:] styli veteris. ||

Metadata

Identifier:	1264
Language:	Polish
Place of Publication:	[B. m. dr.]
Date of Publication:	[po 22 VII 1701]
Year:	1701
Format:	4°
Sheets:	Knlb. 4
Signing:	b. sygn.
Sources:	E. XXVI 205. Russ. R 767
Copies:	Bibl. Sałt.-Szczedr. Len.
Microfilms:	BN Mf. A 401

32.5 M3. 13. VII. 2. 838.

Krótka á Prawdziwa Relátia
O Akcyi Woięnney y otrzymáney Victoryey Krolá I. M. Szweckiego nád Woyskiem Sáskim pod Rygą 9. Julii Roku 1701 *styli veteris.*

Przybywszy Krol I. M, Z Woyskiem swym wybránym dniá 4. Julji *styli Veteris* pod Rygę, y wziąwszy ná oczy *situationem* loci, y przeprawę przez Rzekę Dzwinę, zá raz resolwował się w Imię Boskie przeprawić do Nieprzyiaciela ná drugą stronę Rzęki, ktory się był bardzo dobrze y fortelnie nád tą Rzeką obwarował, y dla broniçnia przeyściá porędze Szweckiey, rożne Szance y Baterye porobił. Tym czasem I. M. Pan Feld-Marszałek y General Gubernator *per Livoniam* Graf Dahlberg, zosobliwego swego y wczesnego Stárania ráko wierny Consiliarz, I. Krol, M. miał wszystkie Kapry, Lodigi, Strugi, Baty y insze tym podobne do przeprawy Woyská Promy, na pogotowiu, ktore dobrze Działámi polnymi obwárowáne y opátrzone były, y ták przygotowáne, że rázem kilká tysięcy Woyska piechotnego przeprawić się mogło. Dniá 8. *ejusdem*, resolwował się Krol I. M. iść przez Dzwinę do Nieprzyiacielá, y znim się męsko y adważnie potkáć, á lubo przez wszystkie dni

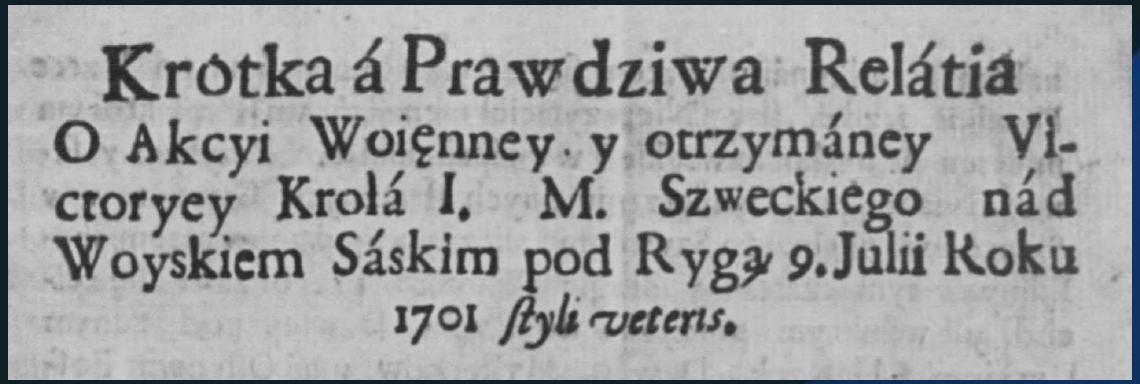
The Electronic Corpus of the 17th and 18th Century Polish Texts

The source of texts:

- nearly 13.5 million words
- TEI-encoded texts following the format of the National Corpus of Polish
- in both transcribed and transliterated forms
- containing 40 prints from the digital library of old prints – ready to be used by the digital library

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

Transcription vs. transliteration: the differences


Krótka á Prawdziwa Relátia
O Akcyi Woiẹnney y otrzymáney Victoryey Krolá I. M. Szweckiego nád Woyskiem Sáskim pod Rygą 9. Julii Roku 1701 *styli veteris.*

Krótka á Prawdziwa Relátia
O Akcyi Woiẹnney y otrzymáney
Victoryey Krolá I. M. Szweckiego
nád Woyskiem Sáskim pod Rygą
9. Julii Roku 1701 styli veteris.

Krótka a prawdziwa relacja
o akcji wojennej i otrzymanej
wiktoriej Króla I. M. Szwedzkiego
nad wojskiem saskim pod Rygą
9. julii roku 1701 styli veteris.

Transcription vs. transliteration: corpus search

Query

[translit="nazaiutrz"]

Displayed layer

transliterated

á é Á É

Left context	Result	Right context	Text ID	Date
M. Pan Woiewodá Krákowski Hetman Polny cum eadem apparentia	nazaiutrz [nazajutrz:adv]	Die[...] ma praesentis wiachał. Woysko do Soboty przeszłey stało	AwLwow	1693

I. W. I. M. Pan Woiewodá Krákowski Hetman Polny cum eadem apparentia **nazaiutrz** Die[...] ma praesentis wiachał. Woysko do Soboty przeszłey stało pod Báriszem, w Niedźielę miało się daley ruszyć zá Iázłowiec ku Wasiłowu, co ieśli się stało czekamy in momentá wiádomośći.

Transcription vs. transliteration: where to store them in the DL?

Krótka a prawdziwa relacja o akcji wojennej i wiktorii króla szwedzkiego nad wojskiem saskim pod Rygą 19 lipca 1701

Full title

[Tytuł nagł., ant.:] Krótka á Prawdziwa Relátia || O Akcyi Woiẽnney y otrzymáney Vi- || ctoryey Krolá I. M. Szweckiego nád || Woyskiem Sáskim pod Rygą 9. Julii Roku || 1701 [kurs.:] styli veteris. ||

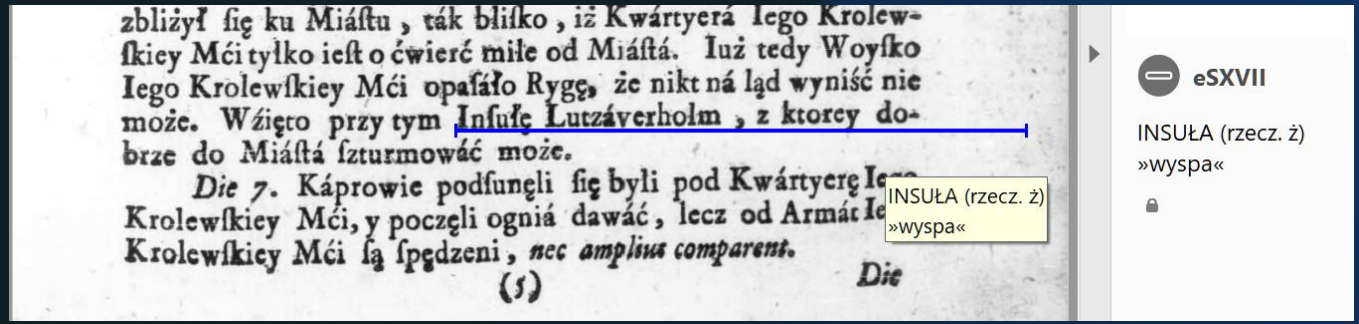
Translation into Contemporary Polish

Transkrypcja: Krótka a prawdziwa relacja o akcji wojennej i otrzymanej wiktoriej Króla I. M. Szwedzkiego nad wojskiem saskim pod Rygą 9. julii roku 1701 styli veteris. Przybywszy Król I. M. z wojskiem swym wybranym dnia 4 julii styli veteris pod Rygę i wziąwszy na oczy situationem loci i przeprawę przez rzekę Dźwinę, zaraz resolwował się w Imię Boskie przeprawić do Nieprzyjaciela na drugą stronę rzeki, który się był bardzo dobrze i fortelnie nad tąż rzeką obwarował y dla bronienia przejścia potędze szwedzkiej różne szańce i baterie porobił. Tymczasem I. M. Pan Feldmarszałek

Back in 2019

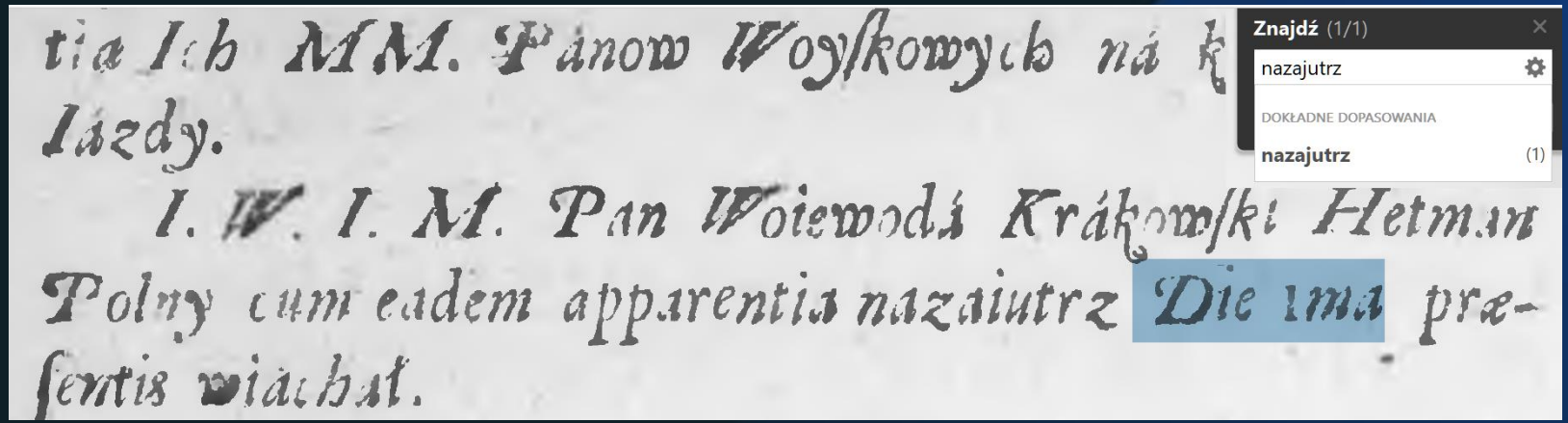
Adding texts from the corpus to PDF:

1. obtaining a dirty OCR of a print
2. retrieval of corpus transliterations
3. token alignment and fine-tuning
4. token replacement and creation of a clean version of the print
5. adding dictionary annotations



Experiments in adding word variants

1. adding word variants directly in the single hidden textual layer
   → wrong positioning of text

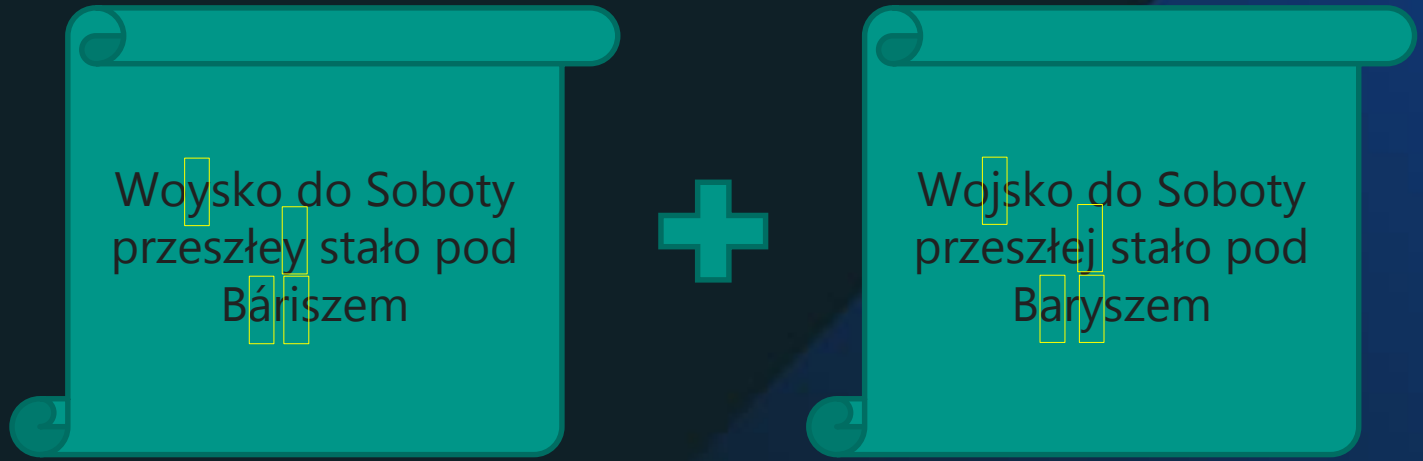


So what about...

1. checking whether it is possible to store more than one layer of the text in the PDF
2. testing which tools are capable of merging both layers
3. checking whether the most common working environment (for our users) can make use of that

Experiments in adding word variants

2. merging two hidden textual layers to make them concurrently searchable



How does it work?

The tools used for merging both layers:

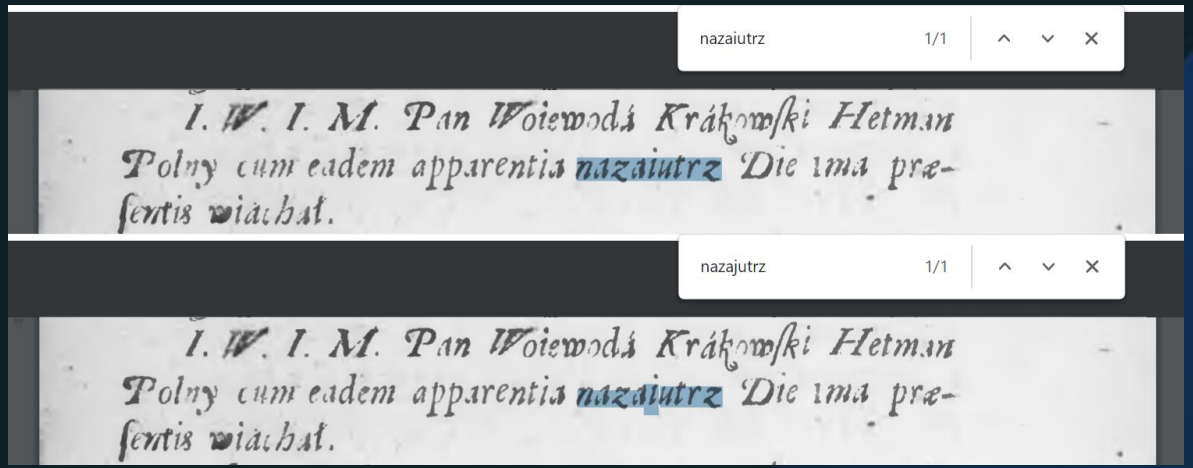
- CPDF
- PDFTK

The tools used to search in both layers:

- standalone Acrobat Reader
- Acrobat Reader plugin for Edge, Firefox and Chrome

The result

1. transliterated and transcribed text is searchable in the browser and offline
2. position of the text is properly shown:



3. only one layer is available for copying text

Thank you!