Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus

Maciej Ogrodniczuk
Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

Włodzimierz Gruszczyński
Institute of Polish Language
Polish Academy of Sciences

ICADL 2019

6 November 2019

Interconnecting related language resources

Repositories of related historical language data usually exist independently:

- digital libraries store source objects
- language corpora maintain their textual content
- electronic dictionaries provide linguistic context of individual words/phrases

But how about:

- inter-linking the repositories
- using the content from the corpus and the electronic dictionary to enhance digital library objects

The resources used in the study

The Digital Library:

- CBDU: The Digital Library of Polish and Poland-Related Ephemeral Prints from the 16th, 17th and 18th Centuries
- a thematic digital library of approx. 2,000 Polish and Poland-related pre-press documents
- dated between 1501 and 1729
- managed by EPrints
- planned to contain rich metadata (historical comments, glossaries of foreign interjections, explanations of background details and relations between library objects: translations, adaptations, alterations of the base text, their sources etc.)
- prints available only in image form (PDF files containing scanned originals)

`https://cbdu.ijp.pan.pl`

CBDU: Basic metadata

Relacja potrzeby i wiktorii otrzymanej pod Wiedniem przez wojska króla Jana III

Hide the scan

Full title

[Tytuł nagł., ant.:] RELACYA Potrzeby która trwáłá godźin 14. y Wiktoryey otrzymáney *[kurs.:]* Die 12. Septẽbris. *[ant.:]* nád Nieprzyjaćielem pod Wi-
dniem, przez Woyská Naiáśnieysze-go y Niezwyćiężonego KROLA Iego Mośći Polskiego Wielkiego Monár-chy JANA Trzećiego, z pod Na-miotow Wezyrskich z Obozu wysłá-na, tudźiesz Excerpta z Listu tegoż Naiáśnieyszego Krolá Iego Mośći do Krolowey Iey Mośći pisánego, *[kurs.:]* sub Die 13. Septembris. Anno Domini, 1683.

Metadata

Identifier:	1036
Language:	Polish
Connections:	War. B. Jak war. A z wyjątkiem różnicy w tytule; w. 8 od góry zamiast IANA-IONA. Por. poz 1005, 1006, 1007, 1659.
Place of Publication:	*[B. m. dr.]*
Date of Publication:	*[po 13 IX 1683]*
Year:	1683
Variant:	A
Format:	4°
Sheets:	Knlb. 4
Signing:	Sygn.)(4
Sources:	E. XXVI 204, XXVIII 354. GK. 582
Copies:	BJ 24730 I, BN XVII. 3.

Szer. okna 1 / 12

RELACYA

Potrzeby ktora trwáłá godźin 14. y Wiktoryey otrzymáney *Die 12. Septẽbris.* nád Nieprzyiaćielem pod Wi-dniem, przez Woyská Naiáśnieyszego y Niezwyćiężonego KROLA Iego Mośći Polskiego Wielkiego Monárchy IANA Trzećiego, z pod Namiotow Wezyrskich z Obozu wysłána, tudźiesz Excerpta z Listu tegoż Naiáśnieyszego Krolá Iego Mośći do Krolowey Iey Mośći pisánego, *sub Die 13. Septembris. Anno Domini, 1683.*

BOg y Pan náß ná wieki błogosłá-wiony/ dał Zwyćięstwo y Sławę nieśmier-telną Krolowi Jego Mośći y Narodowi ná-ßemu/ o iákiey

o iákiey Wieki pierwße nigdy nie ſłychały/ álbowiem/ Działá wßytkie/ Oboz wßytek/ doſtátki nieoßácowá-ne/ doſtały ſie w rece Polſkie/ Nieprzyiaćiel zoſtawßy trupem á prochem/Polá/ Oboz Turecki/ w konfuzyey wielkiey vcieka/ Wielbłądy/ Muły/ Owce/ wßędy po bo-kách Woyſká Chrześćiáńſkie zábieráia/ przy ktorych Tur-kow trzodámi przed ſobą tu pędzą/ drudzy záś Tu-rcy/ oſobliwie Renegátowie/ ná dobrych końiách/ y pie-knie vbrani/ do náßych vćiekáią. Táka ſie to rzecz nie-podobna ſtáłá/ że tegoż dniá/ miedzy Poſpolſtwem iuż tu w Mieśćie/ y w Obozie byłá trwogá/ tak rozu-mieiąc y niemogąc ſobie cále perſwádowáć ináczey ieno że Nieprzyiaćiel názad ſie wraca. Prochow Sámych/ Kul/ y Ammunicyi/ porzućił wiecey niß ná million. Widział Krol Jego Mość tey nocy rzecz cudowna/ ktorą ſobie záwßze widzieć prágnął/ kiedy Chołotá Pol-ſka/ w kilku mieyſcách zápalilá te Prochy/ ktore cále ſadny dźień reprezentowały/ bez namnieyßey ßkody ludzkiey/ z drzeniem ziemi/ pokázały ná Niebie iáko ſie obloki rodzą: ále w tym nießczęśćie wielkie bo bli-ſko millioná vczyniły ßkody. Wezyr ſam tak vćiekł od wßyſtkiego/ że ledwo ná iednym koniu/ y w ie-dney ſukni. Krol Jego Mość zoſtał tego ſucceſſo-rem/ kiedy po wielkiey cześći/ wßytkie mu ſie doſtá-ły po nim ſplendory/ á to tym ſpoſobem/ że będąc w Obozie ná ſámym przodzie/ y tuż zá Wezyrem poſte-puiąc/ przedał ſie do niego ieden pokoiowy Wezyrá/ y pokazał Námioty Wezyrſkie/ táß obßyrne iáko Wár-ßáwá/

CBDU: Extended metadata

Historical Commentary

Opis dotyczy kampanii króla Szwecji Karola XII w Inflantach we wstępnej fazie tzw. Wojny Północnej (1700-1721) między Szwecją a koalicją Rosji, Danii, Saksonii, Prus i Hanoweru oraz od 1704 Rzeczpospolitej. Opisywane zwycięstwo szwedzkie nad połączonymi siłami sasko-rosyjskimi było ukoronowaniem tej kampanii, rozpoczętej nieudaną próbą zdobycia Rygi przez wojska saskie IX 1700. Wspomniana w tekście bitwa pod Narwą miała miejsce 20 XII 1700 i zakończyła się spektakularnym zwycięstwem Szwedów. Dokument ma niewątpliwie charakter propagandowy i skierowany był do zwolenników Karola XII w Rzeczpospolitej.

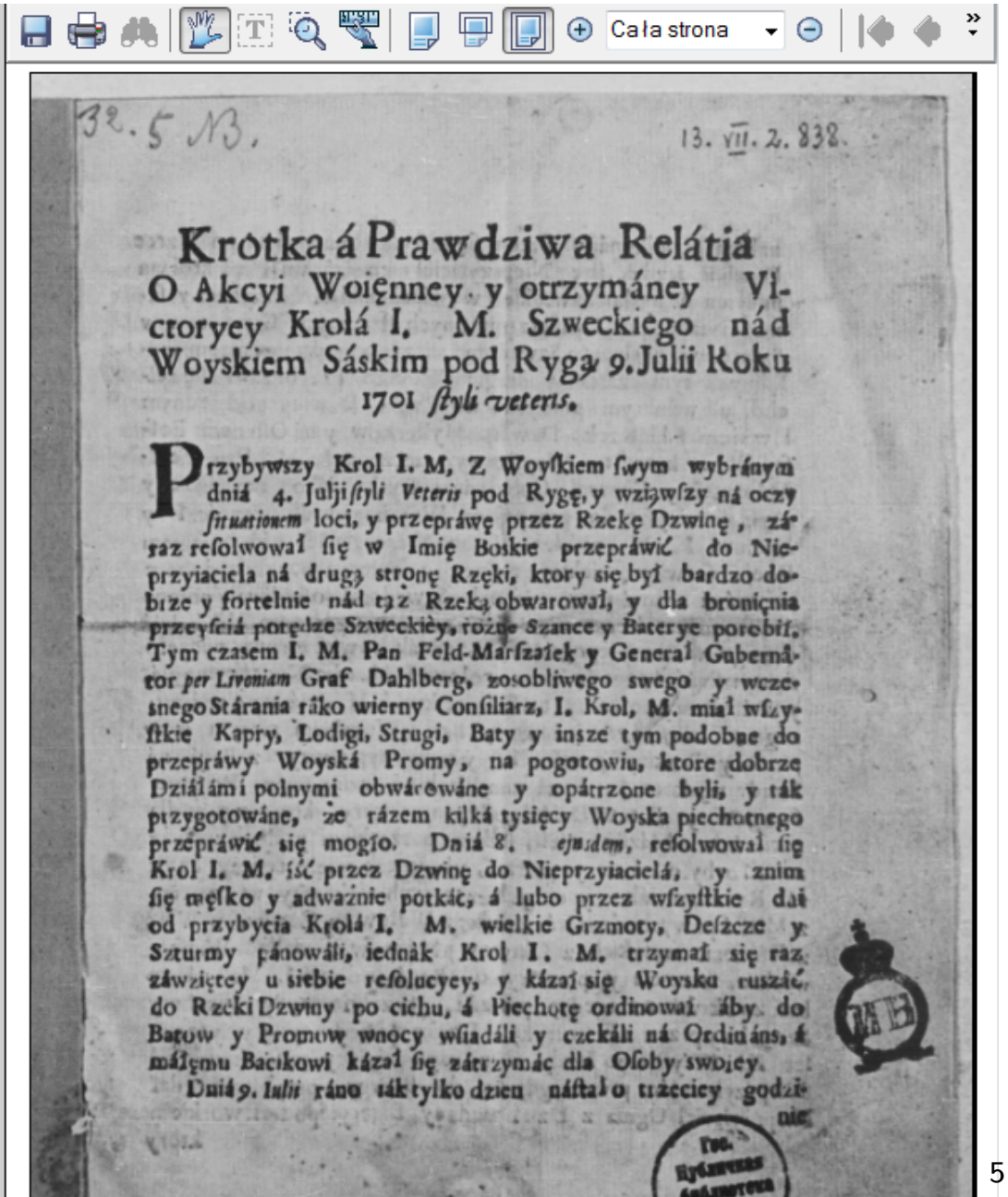
Dictionary

Entry	Explanation
wziąć na oczy	rozpatrzeć się, przyjrzeć się
resolwować się	zdecydować się, powziąć zamiar
fortelnie	rozważnie, sprytnie, pomysłowo
kapra	rodzaj łodzi

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.



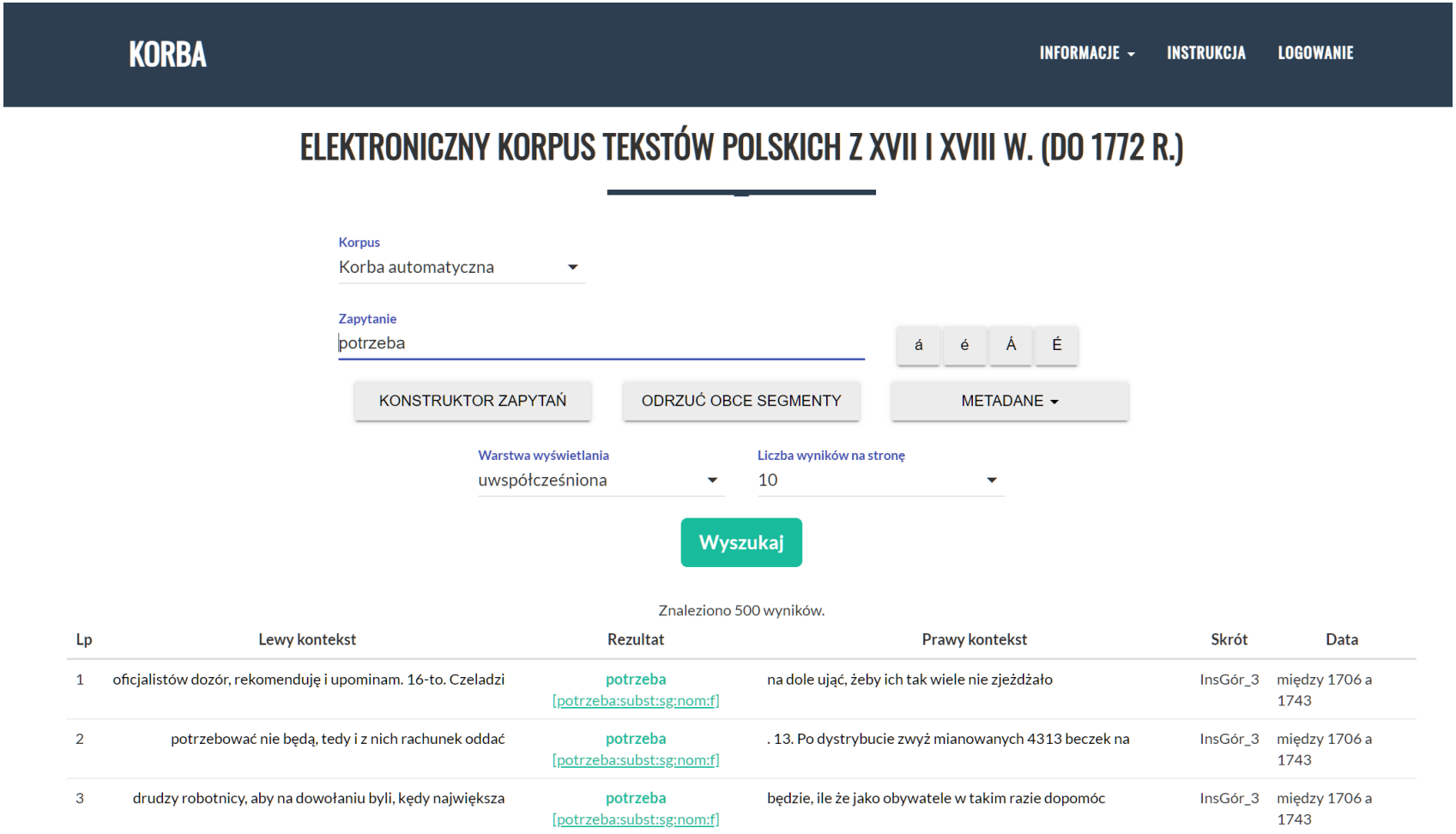
The resources used in the study

KORBA (Baroque) Corpus:

- The Electronic Corpus of the 17th and 18th Century Polish Texts (until 1772)
- over 13M tokens
- diachronic supplement to the National Corpus of Polish
- rich structural annotation (e.g., the location of the search phrase in the appropriate text page)
- text annotation (e.g., tagging words from foreign languages)
- 0.5M subcorpus manually annotated with morphosyntactic tags, then used as training data for a disambiguating tagger, then used for automatic annotation of the whole corpus

`https://korba.edu.pl`

KORBA corpus



Lp	Lewy kontekst	Rezultat	Prawy kontekst	Skrót	Data
1	oficjalistów dozór, rekomenduję i upominam. 16-to. Czeladzi	potrzeba [potrzeba:subst:sg:nom:f]	na dole ująć, żeby ich tak wiele nie zjeżdżało	InsGór_3	między 1706 a 1743
2	potrzebować nie będą, tedy i z nich rachunek oddać	potrzeba [potrzeba:subst:sg:nom:f]	. 13. Po dystrybucie zwyż mianowanych 4313 beczek na	InsGór_3	między 1706 a 1743
3	drudzy robotnicy, aby na dowołaniu byli, kędy największa	potrzeba [potrzeba:subst:sg:nom:f]	będzie, ile że jako obywatele w takim razie dopomóc	InsGór_3	między 1706 a 1743

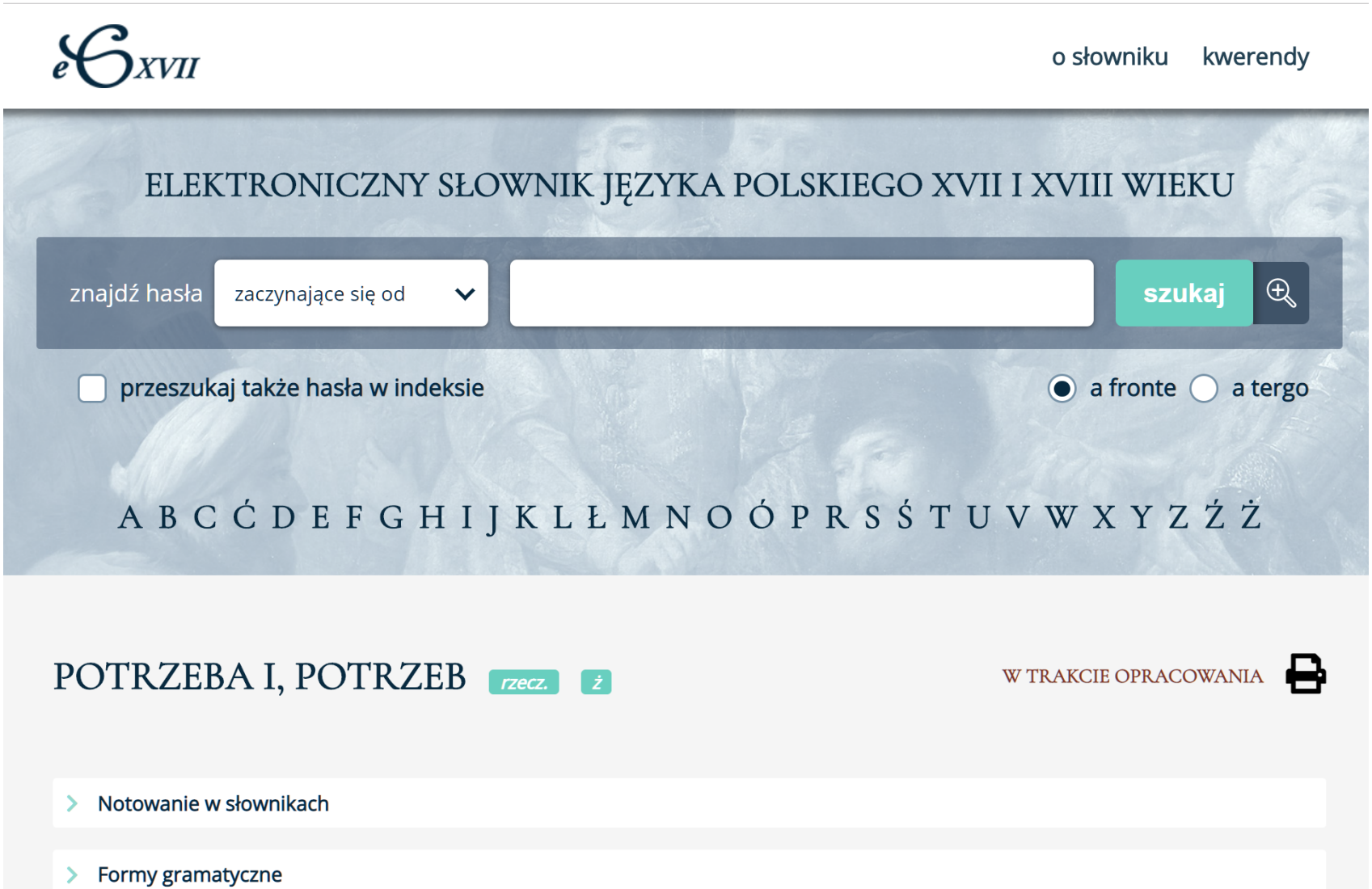
The resources used in the study

The Electronic Dictionary:

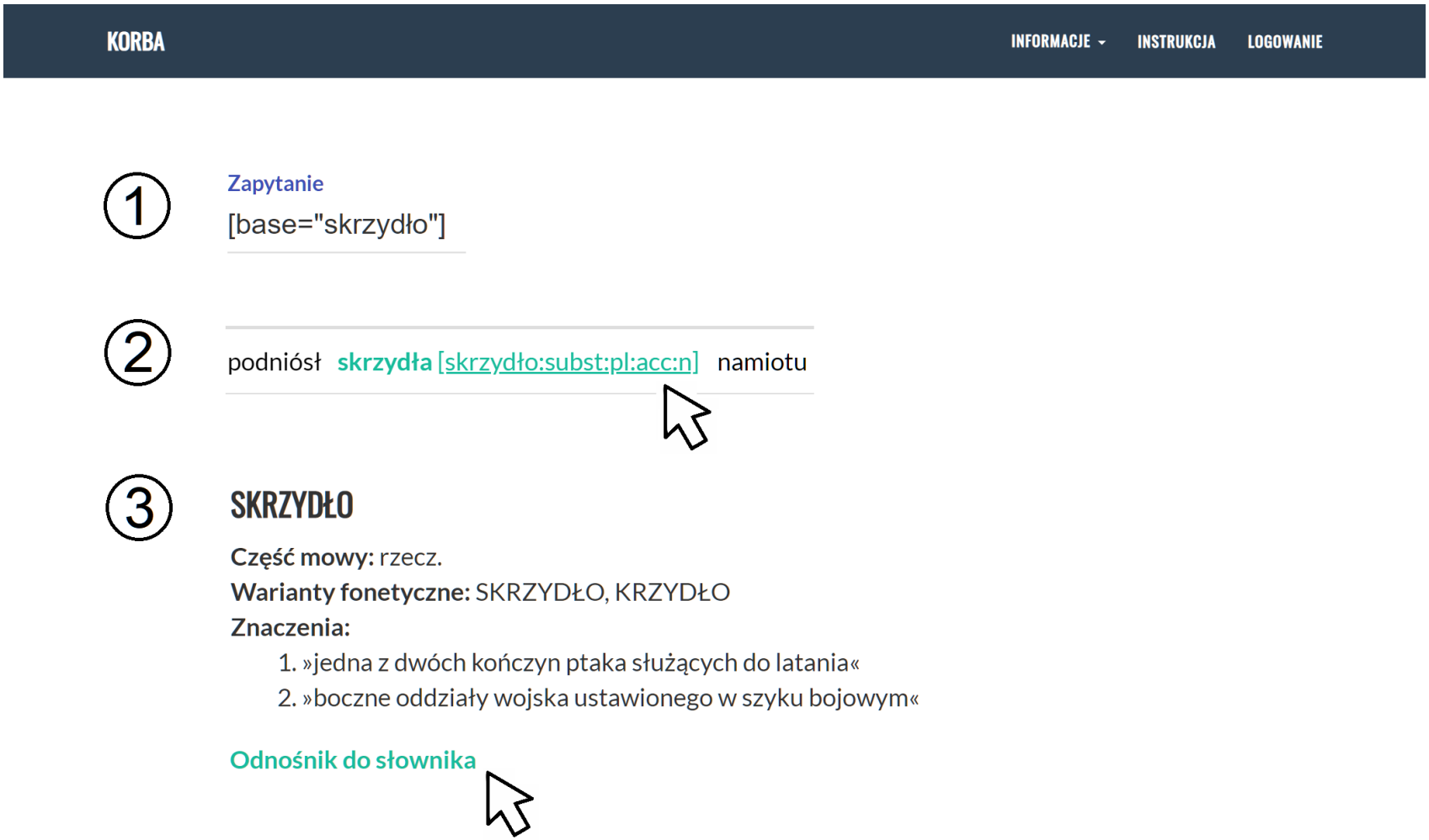
- The Electronic Dictionary of the 17th–18th Century Polish
- started to be published on paper in 1999
- converted to electronic form in 2004 and further developed online
- currently 39k entries in various stages of development
- 77k grammatical forms in the entries (only confirmed inflectional paradigms)
- quotations transliterated from 1100 sources

`http://sxvii.pl`

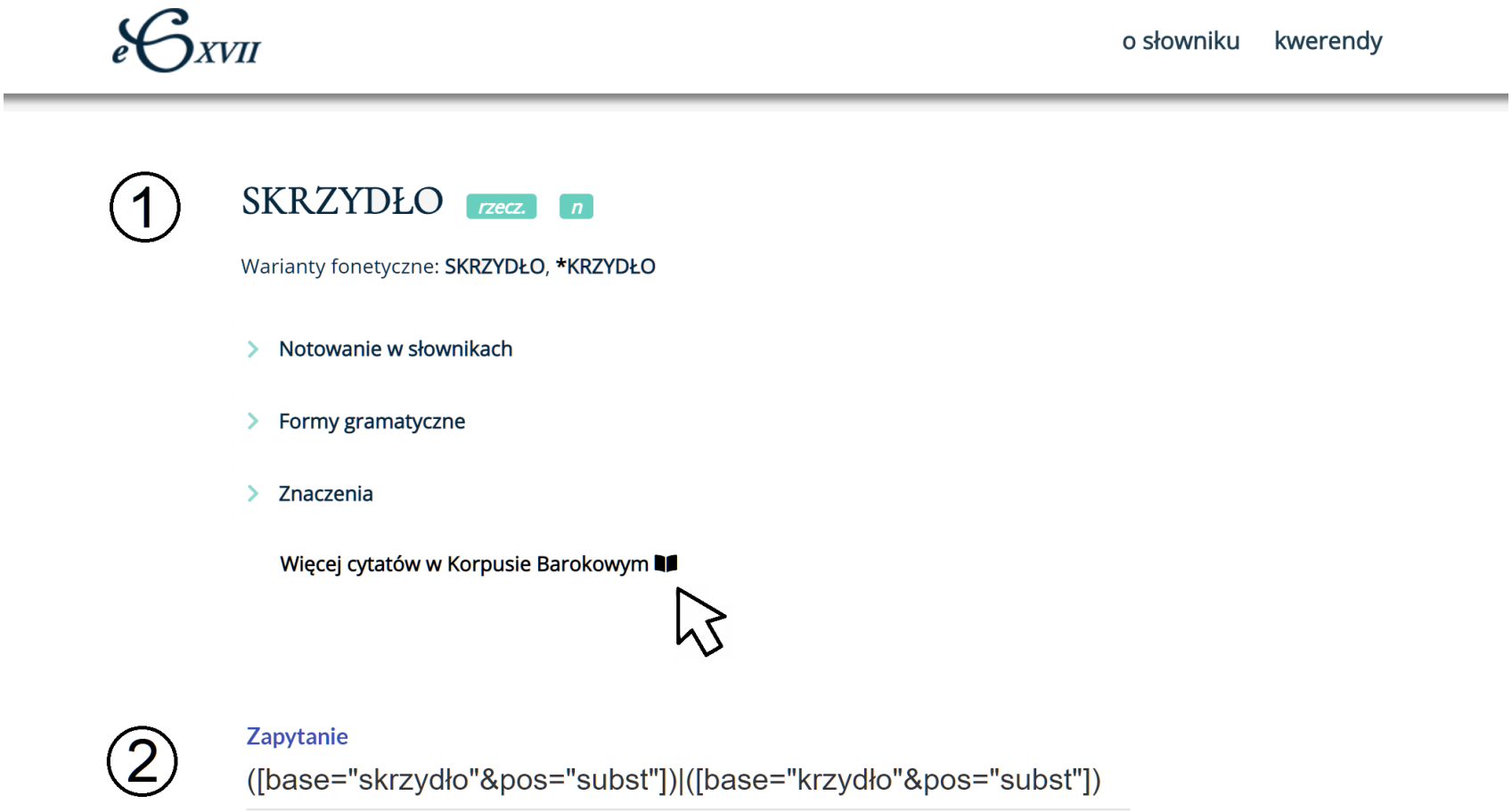
Electronic dictionary



Resource linking: From corpus to dictionary



Resource linking: From dictionary to corpus



Resource linking: From dictionary to corpus

Experiment 1: From corpus to digital library

Initially:

- CBDU prints available only as graphical PDFs
- unsuccessful attempts at dirty OCR

But:

- several prints were transliterated to be included in the corpus
- we could use this content to add the textual layer

Experiment 1: From corpus to digital library

Conversion workflow:

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

Experiment 1: From corpus to digital library

OCR quality for a sample page

	Token count
Actual number of textual tokens on the page	126
Number of tokens in dirty OCR	136
Number of properly OCR-ed tokens	22
Number of tokens in the corresponding corpus text	125
Number of correctly aligned tokens	124
Number of correctly transferred tokens	108

Experiment 1: From corpus to digital library

Transfer errors caused by:

- the quality of the source prints
  - noise in the source image
  - text bleeding through from the reverse page
- overlap of regions and tokens

Experiment 2: Dictionary annotation

The general idea:

- linking from metadata is easy but inconvenient to use
- manual linking requires an enormous amount of work
- dictionary entries could be used
- but some words are still understandable
- so how about filtering them with the contemporary dictionary?

Experiment 2: Dictionary annotation

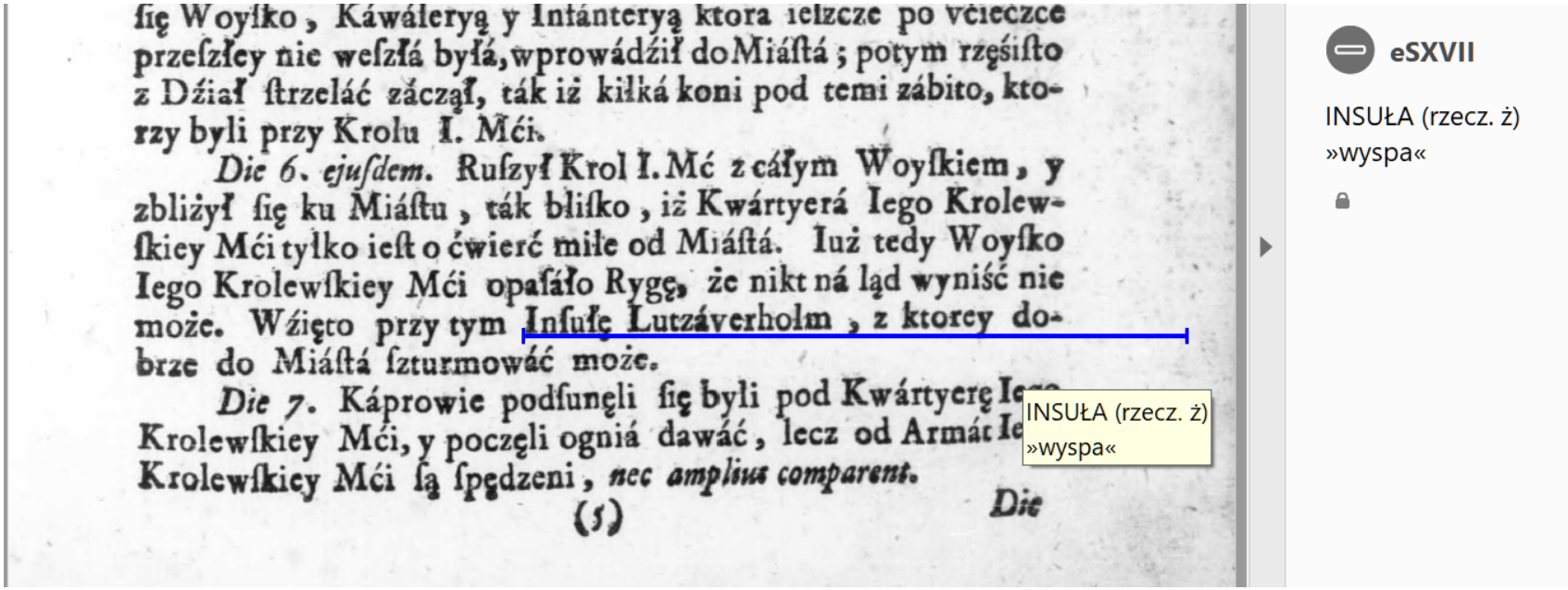
Annotation workflow:

1. Filtering out modern vocabulary from e-SXVII using PoliMorf dictionary

Subset of the e-SXVII vocabulary	Entries
Meaningful dictionary entries	15 702
Lemmata present in the corpus	10 440
Lemmata older than 20th century	2 652
— with gerund correction	2 479

2. Creating dictionary annotations

Experiment 2: Dictionary annotation



Experiment 2: Dictionary annotation

Observations for a sample print (29 entries in the manually created dictionary):

- 21 entries cannot be suggested:
  - 14 entries are absent from e-SXVII
  - 5 entries are multi-word
  - 2 entries are stubs
- 3 entries are unexplained in the manual dictionary
- 4 entries can be found in the modern dictionary
- one entry is valid to be suggested by our mechanism

Conclusions

And future work:

- the idea can be applied to digital libraries using other layered representation formats such as DjVu
- the proposed tools can be replaced
- linking individual content objects with external resources is independent from any technical solution used for digital library management
- was it worth it? yes! digital libraries play the central role in linguistic research!

Thank you!

And the funding institutions:

- 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)
- The conference fee was sponsored by the grant holder, the Institute of Polish Language, Polish Academy of Sciences
- The travel to ICADL 2019 was sponsored by the Institute of Computer Science, Polish Academy of Sciences