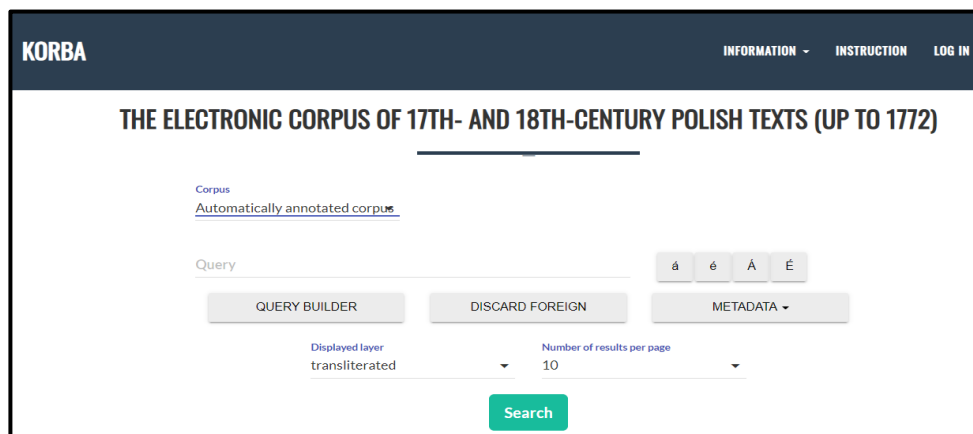


The role of Transkribus in building of the Electronic Corpus of 17th- and 18th-century Polish Texts (KorBa)

Ewa Rodek

The Electronic Corpus of 17th- and 18th-century Polish Texts (KorBa)

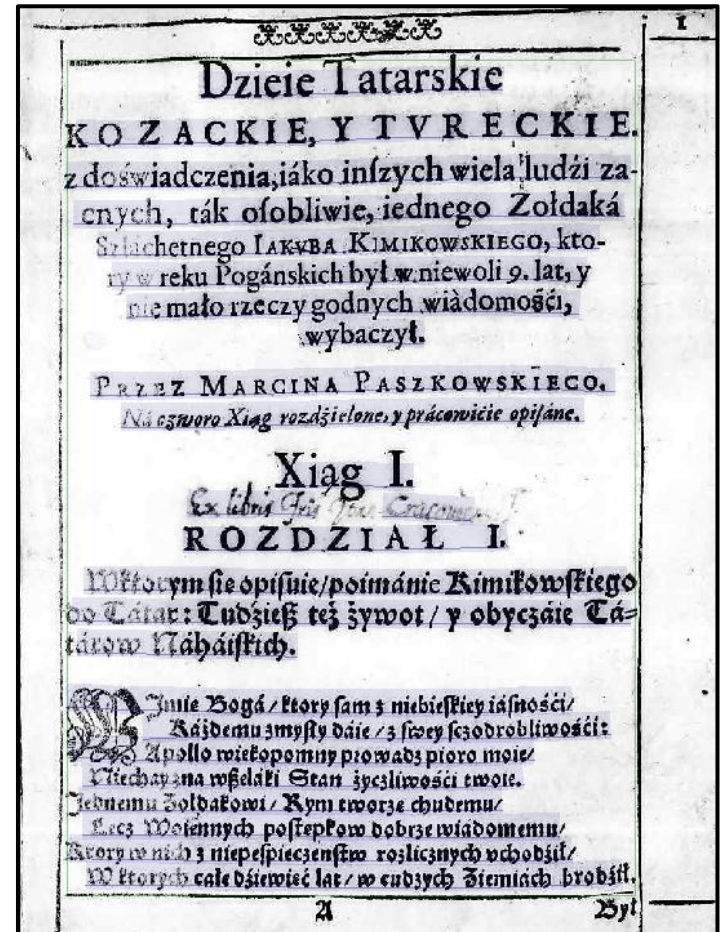
- korba.edu.pl
- 25M in both editions
- the first relatively large corpus of old Polish texts
- the only morphosyntactically annotated (including lemmatization) online corpus of pre-19th-century texts of such size in the Slavic world



The screenshot shows the KORBA website interface. At the top, there is a dark blue header with the text "KORBA" on the left and "INFORMATION", "INSTRUCTION", and "LOG IN" on the right. Below the header, the main title reads "THE ELECTRONIC CORPUS OF 17TH- AND 18TH-CENTURY POLISH TEXTS (UP TO 1772)". Underneath, there is a section for "Corpus" with a link to "Automatically annotated corpus". A "Query" input field is present, followed by buttons for "QUERY BUILDER", "DISCARD FOREIGN", and "METADATA". To the right of the input field are four small buttons labeled "á", "é", "Ä", and "É". Below the input field, there are two dropdown menus: "Displayed layer" (set to "transliterated") and "Number of results per page" (set to "10"). A green "Search" button is at the bottom.

Text transcribing

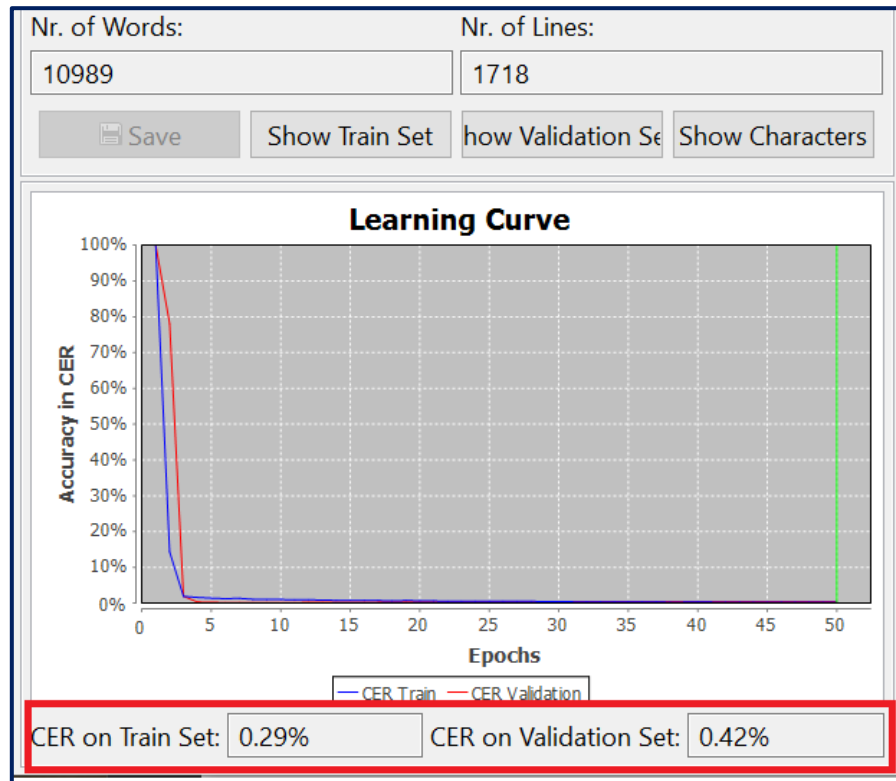
- first edition: materials (718 texts) were transcribed only manually
- second edition: (1316 texts) ca. 75% materials transcribed manually and 25% automatically in Transkribus (but the most difficult – manuscripts and texts with gothic fonts)



Models

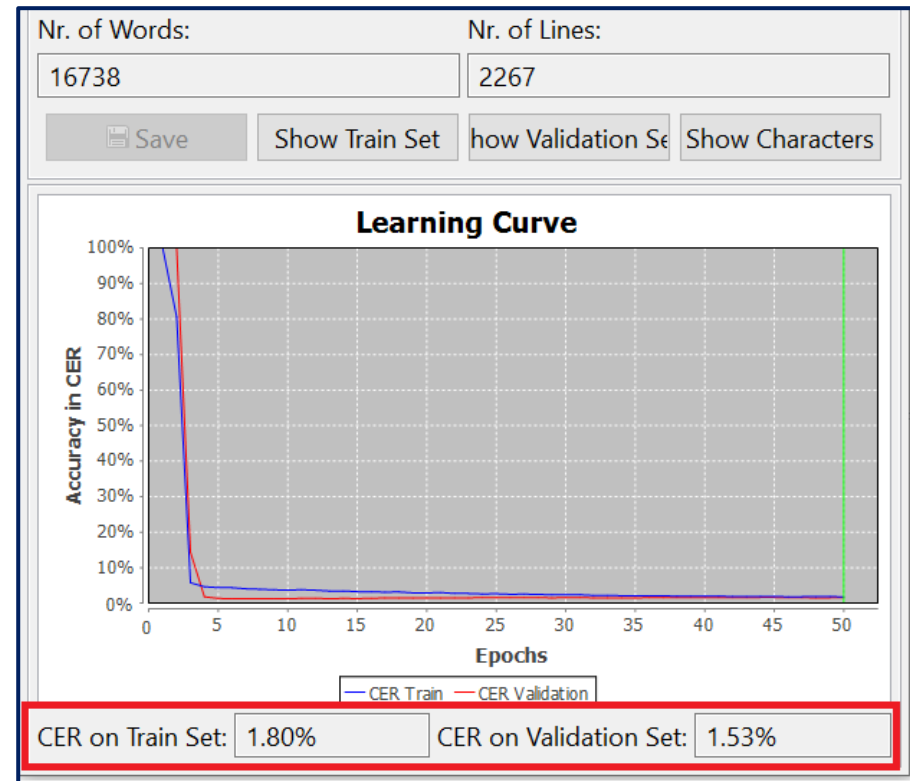
Old prints

CER on Train Set: 0.29%



Manuscripts

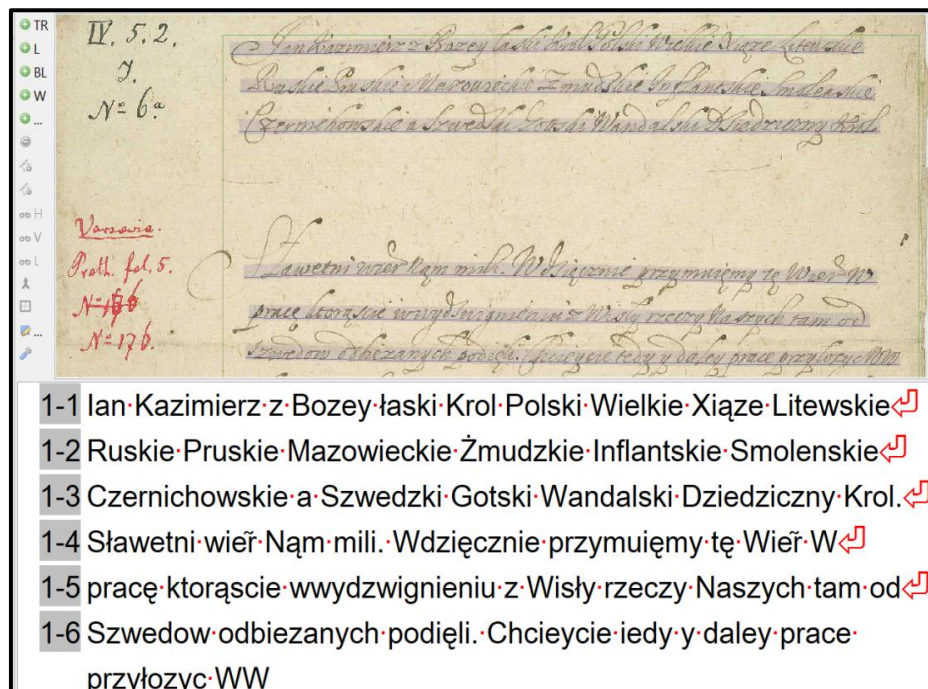
CER on Train Set: 2.25%



Benefits of using Transkribus

- TIME
- MONEY

- DATA QUALITY
- PEOPLE
- NEW OPPORTUNITIES
- EXPERIENCE



Thank you!

korba.edu.pl