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
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!