Using Digital Tools to Create Modern Multi-Search Engine for Polish Historical Dictionaries

Ewa Rodek

Institute of Polish Language, Polish Academy of Sciences

PAN IJP Institute of Polish Language Polish Academy of Sciences

DH 2023 Collaboration as Opportunity

DATABASE OF HISTORICAL POLISH LEXICONS

Aims:

- Creation of the www service along with the API as a single platform where digitized historical dictionaries can be placed; one place useful for researchers from various fields of science
- Development of a standard for the description of input data enabling the addition of subsequent dictionaries to the database
- Deep digitization of the two most important dictionaries from the 17th and 18th centuries, as well as a dictionary with a novel headword organization
- Combination of a corpus search engine with a presentation of original photos

The main aim of the platform is to expand new digitized dictionaries in the future. Therefore, generally available tools should be used. In historical dictionaries the workshop solutions were not consistently used and standardized. Therefore, the methodology should be standard and as simple as possible, but also covering the dictionary material as broadly as we can.
Tools should be easily accessible, proven, convenient and provide the user with quick results.

The project will be implemented in cooperation with the Institute of Computer Science, Polish Academy of Sciences (IPI PAN). Some of the tools provided by them have already been used by us in other projects.

Actions	Tools
automatic recognition of old prints with multilingual texts, different fonts and layouts	Transkribus (readcoop) with HTR models for Middle Polish and Middle-Aged Latin (proven in previous projects)
structural marking of the material	XML editor with TEI P5 Guidelines
transcription to a version of the modernized orthography	Transcriber (proven in previous projects)
morphosyntactic annotation of the Polish text	Morphosyntactic Analyzer and Tagger e.g. Korbeusz & KFTT (proven in previous projects)
presenting PDF of original scans with a text layer combined with multi-search engine adjusted to linguistic annotations	TEI Publisher

DEEP DIGITIZATION – 3RD LEVEL DIGITIZATION

DIGITAL COPY → HTR → TEI XML ENCODING

Transkribus

This programme gives the opportunity to train models for automatic transcription of old prints on your own. Models adapted to the material are created in a short time and can be transformed and retrained many times.
Transkribus:

- learns the page structure and discards unnecessary elements like page numbers;
- creates PDF files with a text layer, but also xml files for further stages of analysis or editable doc files;
- allows you to work on automatically read scans, such as searching the text on photos or tagging texts.

Korbeusz & KFTT

The morphosyntactic analyzer Korbeusz and tagger KFTT were created by IPI PAN for the Polish language and adapted to grammar of the Middle-Polish. We used these programmes to create the 25-million-token corpus of baroque texts, called KorBa.
Transcriber was also tested in KorBa project.

SCAN ME

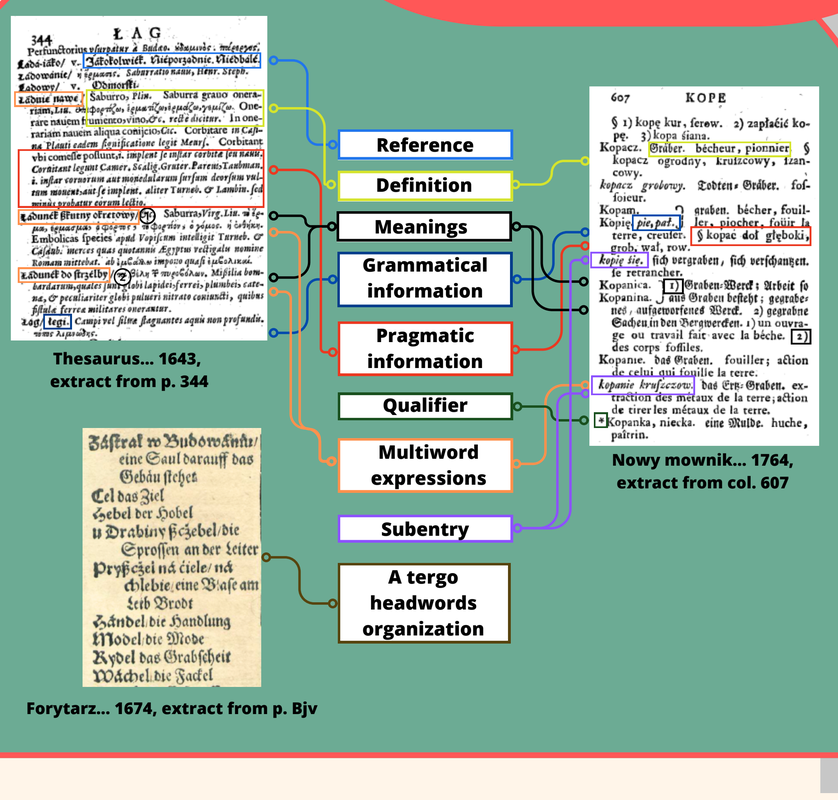
<TEXT ENCODING INITIATIVE>

P5: Guidelines for Electronic Text Encoding and Interchange defines a module for encoding lexical resources of all kinds, not only simlpe glossaries, but also dictionaries with a complicated microstructure.
The use of a standard description language allows the methodology to be adapted to all types of dictionaries and ensures consistency and sustainability.
TEI describes an encoding scheme in the simplest way. Moreover, it accommodates the entire range of structures to allow any element to appear anywhere in a dictionary entry.

TEI Publisher

TEI Publisher is our best candidate to create an application that will present digitized dictionaries and at the same time allow you to search them like a corpus. The programme is very simple to use and significantly reduces the amount of custom code required by other digital edition projects. That Open Source tool ensures data durability and security. Moreover, it is constantly supported and developed.

Different ways of microstructure organization in dictionaries chosen for the deep digitization



Thesaurus... 1643, extract from p. 344

Nowy mownik... 1764, extract from col. 607

Forytarz... 1674, extract from p. Bjv

WE WILL START FROM THE DIGITIZATION OF...

... THE DICTIONARIES WHICH WERE THE MILESTONES OF THEIR TIME:

***Thesaurus Polono-latino-greacus* by Grzegorz Knapiusz** (Cracow 1643) - the first dictionary with Polish as the starting language

- approx. 50,000 entries, alphabetical arrangement, stylistic qualifiers (for colloquial and dialectal words);
- modern, but extensive, multi-element structure of the entry, which today is not fully legible;
- quotations from classical Latin and Greek works;
- fresh borrowings are replaced by neologisms invented by Knapiusz. Some of them are still in use, e.g. *nosorożec* 'rhinoceros';

- **GOALS**: to be a linguistic "treasury" - to store and present the beauty of the Polish language at the time when it was considered inferior, stylistically less efficient than Latin, which was then a pan-European language; didactics and the fight for linguistic purity;
- words used by educated people, without vulgar and obsolete words.

***Forytarz języka polskiego* by Jan Ernesti** (Wroclaw 1674) - the first dictionary of the Polish language organized *a tergo*

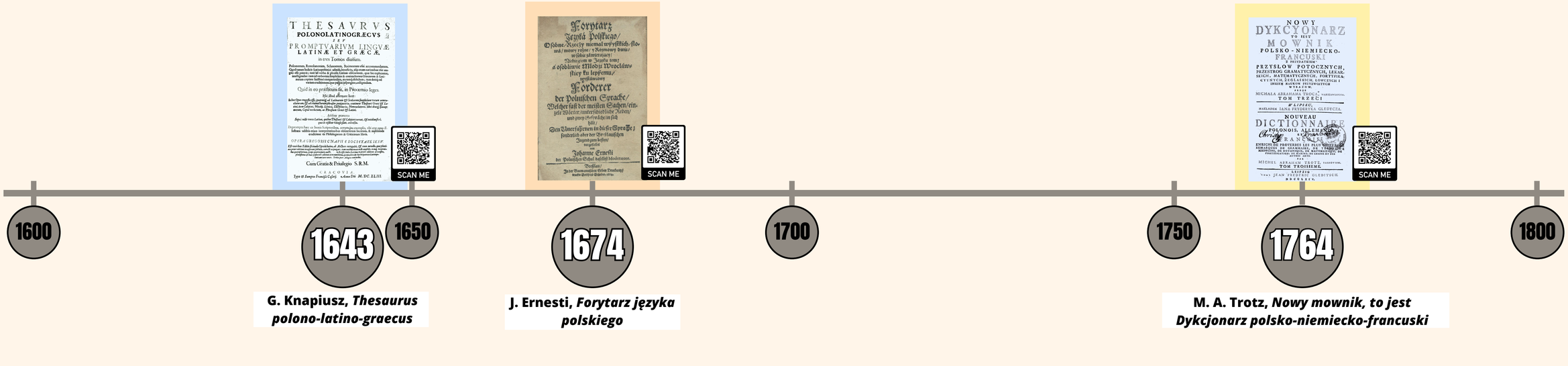
- approx. 5,500 entries;
- grammatical-alphabetical arrangement and *a tergo*: entries were ordered by part of speech, secondly - by other grammatical criteria, and thirdly - alphabetically, but according to word endings;

- **GOAL**: a handy tool for teaching Polish in the 17th c.

***Nowy dykcjonarz, to jest mownik polsko-niemiecko-francuski* by Michał Abraham Trotz** (Warsaw 1764) - the first big dictionary without classical languages

- approx. 45,500 entires with subentires, alphabetical arrangement, 20 types of qualifiers (stylistic, chronological, domain);
- the lexicographer learned from Knapiusz's work, but he also significantly improved the lexicographic method;
- word-formations, foreign words, colloquial words, regionalisms, idioms, proper names, vocabulary in various fields of science and technology;
- clearly separated grammatical information;
- meanings appear within one entry and are arranged in a fixed order;

- **GOALS**: to create modern and useful dictionary, collecting the entire vocabulary of the Polish language without ideological limitations.



G. Knapiusz, *Thesaurus polono-latino-graecus*

J. Ernesti, *Forytarz języka polskiego*

M. A. Trotz, *Nowy mownik, to jest Dykcjonarz polsko-niemiecko-francuski*