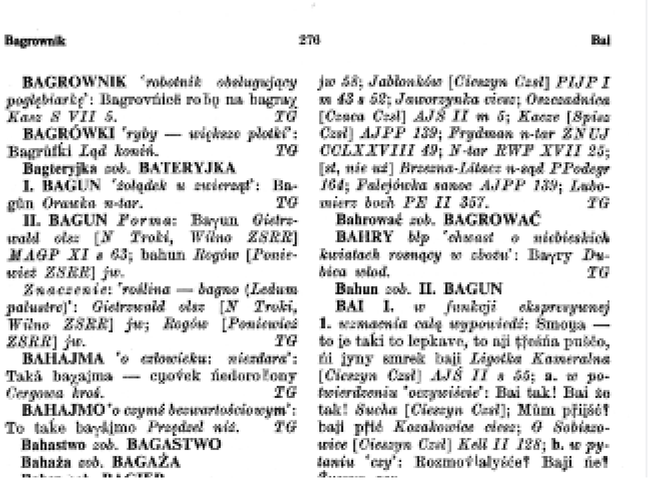
Krzysztof Nowak, Dorota Mika, Wojciech Łukasik

TEI Modelling of Lexicographic Data in the DARIAH-PL Project

The main goal of the "DARIAH-PL Digital Research Infrastructure for the Arts and Humanities" project is building the Dariah.lab infrastructure, which would allow for sharing and integrated access to digital resources and data from various fields of the humanities and arts. Among numerous tasks that the Institute of Polish Language, Polish Academy of Sciences coordinates, we are working towards the integration of our lexicographic data with the LLOD resources (Chiarcos et al. 2012). The essential step of this task is to convert the raw text of a dozen of paper-born dictionaries into TEI-compliant XML format (TEI Consortium).

In this paper we would like to outline the main issues involved in TEI XML modelling of these heterogeneous lexicographic data.

In the first part, we give a brief overview of the formal and content features of the dictionaries. For the most part, they are multi-volume works developed between the 1950s and 2010s with the research community in mind, and as such they are rich in information and structurally complex. They cover diachronic development (from medieval Polish and Latin to present-day Polish) and the functional variation of Polish (general language vs. dialects, proper names).



Bagrownik 270 Bai

BAGROWNIK *'robotnik obsługujący pogłębiarkę'*: Bagrovńicë robją na bagray *Kasz S VII 5.* *TG*

BAGRÓWKI *'ryby — większe płotki'*: Bagrúfki *Ląd konin.* *TG*

Bagteryjka *zob.* **BATERYJKA**

I. BAGUN *'żołądek u zwierząt'*: Bagŭn *Orawka n-tar.* *TG*

II. BAGUN *Forma:* Bayun *Gietrzwałd olsz [N Troki, Wilno ZSRR] MAGP XI s 63*; bahun *Rogów [Poniewież ZSRR] jw.*

Znaczenie: 'roślina — bagno (Ledum palustre)': Gietrzwałd olsz [N Troki, Wilno ZSRR] jw; Rogów [Poniewież ZSRR] jw. *TG*

BAHÁJMA *'o człowieku: niezdara'*: Taká bayajma — cχovek ńedoroŀony *Cergowa krośn.* *TG*

BAHAJMO *'o czymś bezwartościowym'*: To take bayḱimo *Przędzel nis.* *TG*

Bahastwo *zob.* **BAGASTWO**

Bahaża *zob.* **BAGAŻA**

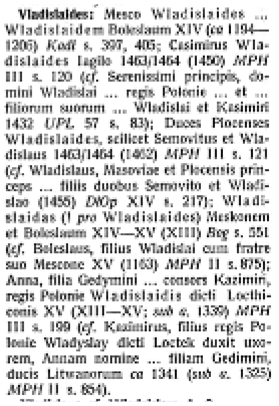
jw 58; Jabłonków [Cieszyn Czsl] PIJP I m 43 s 58; Jaworzynka ciesz; Oszczadnica [Czaca Czsl] AJŚ II m 5; Kacze [Spisz Czsl] AJPP 139; Frydman n-tar ZNUJ CCLXXVIII 49; N-tar RWP XVII 25; [st, nie nż] Brzezna-Lilacz n-sąd PPodegr 164; Palejówka sanoc AJPP 139; Lubomierz boch PE II 357. *TG*

Bahrować *zob.* **BAGROWAĆ**

BAHRY *blp 'chwast o niebieskich kwiatach rosnący w zbożu'*: Baγry *Dubica włod.* *TG*

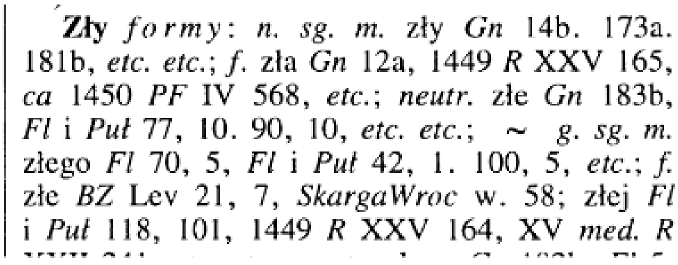
Bahun *zob.* **II. BAGUN**

BAI I. *w funkcji ekspresywnej* **1.** *wzmacnia całą wypowiedź*: Šmoya — to je taki to lepkave, to nji třeśńa puśćo, ći jyny smrek baji *Ligotka Kameralna [Cieszyn Czsl] AJŚ II s 55*; **a.** *w potwierdzeniu 'oczywiście'*: Bai tak! Bai to tak! *Sucha [Cieszyn Czsl]*; Mům přijść baji přić *Kozakowice ciesz; G Sobieszowice [Cieszyn Czsl] Kell II 128*; **b.** *w pytaniu 'czy'*: Rozmoŀalyśće? Baji ńe?



Vladislaides: Mesco Wladislaides ... Wladislaidem Boleslaum XIV (*ca* 1194—1206) *KodI* s. 397, 405; Casimirus Wladislaides Iagilo 1463/1464 (1450) *MPH* III s. 120 (*cf.* Serenissimi principis, domini Wladislai ... regis Polonie ... et ... filiorum suorum ... Wladislai et Kasimiri 1432 *UPL* 57 s. 83); Duces Plocenses Wladislaides, scilicet Semovitus et Wladislaus 1463/1464 (1462) *MPH* III s. 121 (*cf.* Wladislaus, Masoviae et Plocensis princeps ... filiis duobus Semovito et Wladislao (1455) *DlOp* XIV s. 217); Wladislaidas (! *pro* Wladislaides) Meskonem et Boleslaum XIV—XV (XIII) *Bog* s. 551 (*cf.* Boleslaus, filius Wladislai cum fratre suo Mescone XV (1163) *MPH* II s. 875); Anna, filia Gedymini ... consors Kazimiri, regis Polonie Wladislaidis dicti Locthiconis XV (XIII—XV; *sub a.* 1339) *MPH* III s. 199 (*cf.* Kazimirus, filius regis Polonie Wladyslay dicti Loctek duxit uxorem, Annam nomine ... filiam Gedimini, ducis Lituanorum *ca* 1341 (*sub a.* 1325) *MPH* II s. 654).

Dictionary of Old Polish Personal Names



Zły *formy*: *n. sg. m.* zły *Gn* 14b. 173a. 181b, *etc. etc.*; *f.* zła *Gn* 12a, 1449 *R* XXV 165, *ca* 1450 *PF* IV 568, *etc.*; *neutr.* złe *Gn* 183b, *Fl* i *Puł* 77, 10. 90, 10, *etc. etc.*; ~ *g. sg. m.* złego *Fl* 70, 5, *Fl* i *Puł* 42, 1. 100, 5, *etc.*; *f.* złe *BZ* Lev 21, 7, *SkargaWroc* w. 58; złej *Fl* i *Puł* 118, 101, 1449 *R* XXV 164, XV *med.* *R*

Dictionary of Old Polish

On a practical level, this means that, first, substantial effort had to be put into optimizing the quality of the OCR output. Since, except for `grobid-dictionaries` (Khemakhem et al. 2018), there are no tools at the moment that would enable easy conversion of lexicographic data, the subsequent phase of structuring of dictionary text had to be applied on a *per resource* basis.

TEI XML annotation has three main goals. First, it is a means of preserving the textuality of dictionaries which make heavy use of formatting conventions to convey information and employ a complex system of text-based internal cross-references. Second, TEI modelling aims at a better understanding of each resource and its explicit description. The analysis is performed by lexicographers who may, however, come from a lexicographic tradition different from the one embodied in a particular dictionary, and thus need to make their interpretation of the dictionary text explicit. Regardless, this way we may also detect and correct editorial inconsistencies, which are natural for collective works developed over many years. Third, the annotated text is meant to be the input used in the alignment and linking tasks, it is therefore crucial that functionally equivalent structures are annotated in a systematic and coherent way. As we plan to provide an integrated access to the dictionaries, the TEI XML representation is also where the first phase of data reconciliation takes place. It does not only concern the structural units of a typical dictionary entry, such as `<sense/>` or `<form/>`, but also mapping between units of analytical language the dictionaries employ, such as labels, bibliographic reference system etc.

References

Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), Linked Data in Linguistics. Representing Language Data and Metadata, Springer, Heidelberg, p. 201-216.

Mohamed Khemakhem, Axel Herold, Laurent Romary. 2018. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In: GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan. 2018.

TEI Consortium, eds. “9 Dictionaries.” *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* [Version number]. [Last modified date]. TEI Consortium. https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html (10 June 2022).