Dariah.lab

PAN IJP Institute of Polish Language Polish Academy of Sciences

TEI Modelling of the Lexicographic Data

Krzysztof Nowak

Dorota Mika

Wojciech Łukasik

Fundusze Europejskie Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Outline

The Dariah.Lab Project

Lexicographic Resources

The workflow

Discussion

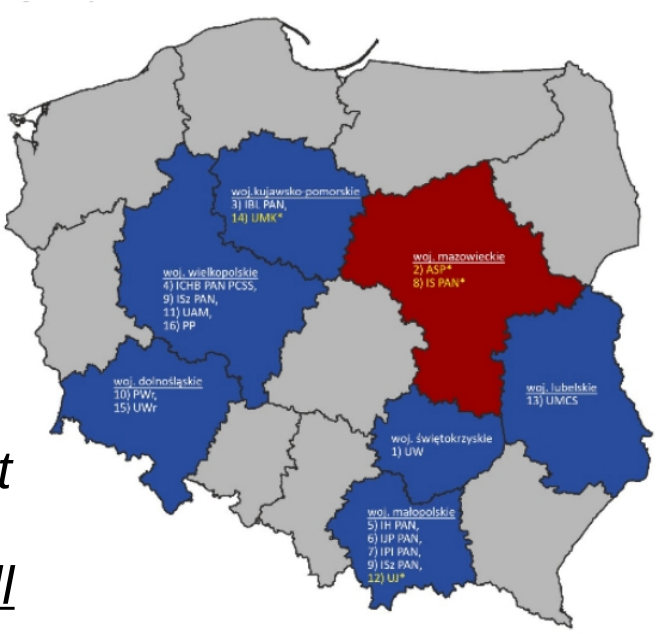
- delimiting lexical units
- defining senses
- representing language use
- placing in time and space

WWW: https://lab.dariah.pl/

Period: January 2021 - December 2023

Funding: European Union

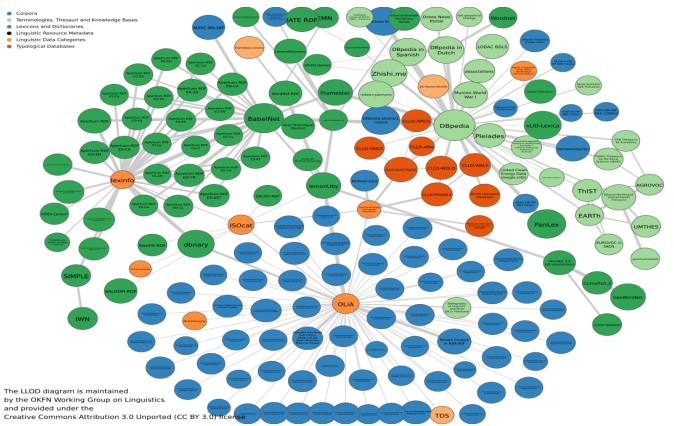
Goal: *Dariah.lab is a research infrastructure for the arts and humanities built in the DARIAH-PL project. It is designed to acquire, store and integrate cultural data from the humanities and social sciences, as well as to process, visualize, and share digital resources.*
(Source: Project's website)

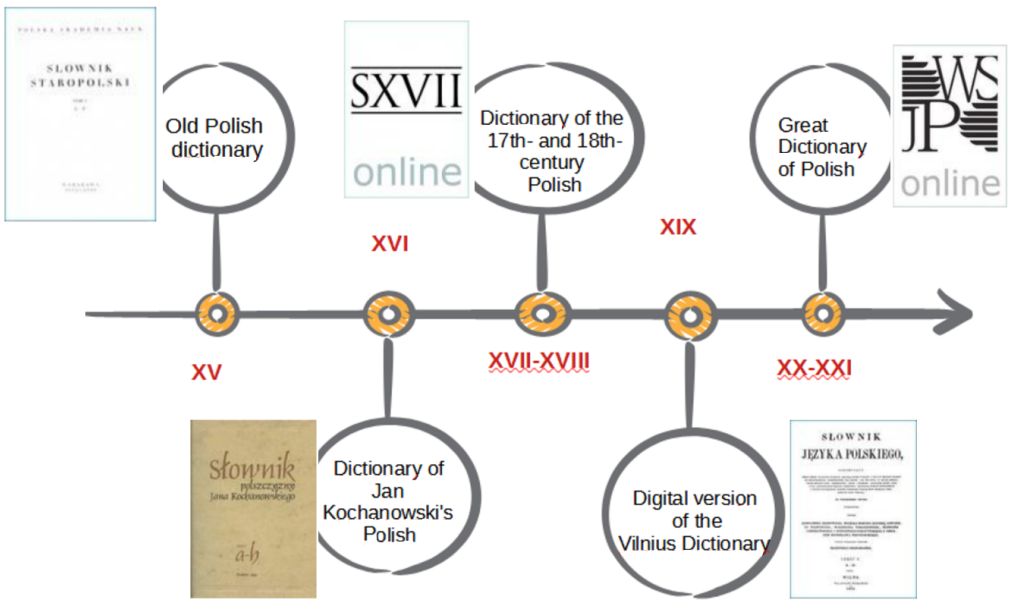


1. The Spoken Component of the Polish Corpus of 2011-2020 (in collaboration with the Institute of Computer Science PAS)

2. Integrated Access to the Institute's Dictionaries and Lexicographic Data

- web platform for data discovery
- API for automated access




SŁOWNIK STAROPOLSKI
Old Polish dictionary
SXVII online
Dictionary of the 17th- and 18th-century Polish
Great Dictionary of Polish
PWN online
XV
XVI
XVII-XVIII
XIX
XX-XXI
Słownik pojęciowy Jana Kochanowskiego
Dictionary of Jan Kochanowski's Polish
Digital version of the Vilnius Dictionary
SŁOWNIK JĘZYKA POLSKIEGO,


diachronic
diatopic
diastratic
SŁOWNIK GWAR OSTRÓDZKIEGO WARMII I MAZUR
SŁOWNICTWO GWAROWE PRZESIEDLEŃCÓW Z UKRAINY
SŁOWNIK GWARY ZE WSI TULIGŁOWY
Gdańsk
Toruń
Białystok
Warszawa
BAĆ SIĘ

Wojciech **Łukasik**
- computational literary studies

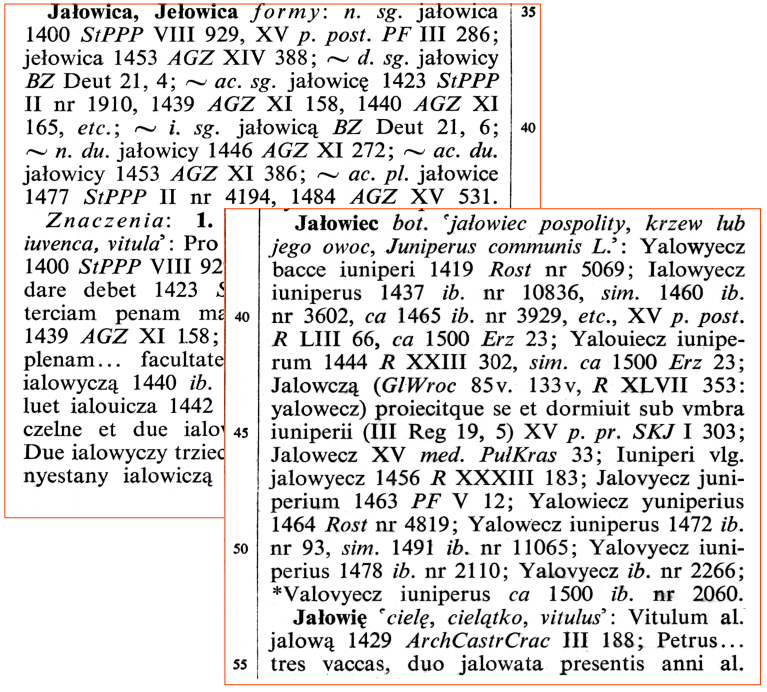
Dorota **Mika**
- linguistics and lexicography of Old Polish

Krzysztof **Nowak**
- linguistics and lexicography of Medieval Latin

I. BŬC *rz* **1.** *'ktoś o wydatnych policzkach, pucołowaty, zwłaszcza dziecko'*: Od ńy̨ fśiske by̨ły̨ take buce. Fśiske ję̨ję̨ *ęeći Bobowo st-gdań.*
2. *'ktoś małomówny'*: *Ostrowce bus.*
3. *przezw 'ktoś niesympatyczny, zarozumiały'*: To taki buc, pśeję̨ły̨ i śy̨ ńy̨ uodezvy̨ *Bóbrka kroś.*
4. *'ktoś niezaradny, niedołęga'*: Të bucu ståri. Ale to bël buc. Ta so vzała ale buca *pn i pd ok Żarnowca wej S I 81.*
5. *'przezw nadawane mieszkańcom wsi'*: Buc, bamber — tak nazywają ludzi ze wsi *Oborniki.*

```
<entryFree n="BUC1" type="hom">
   <form>
      <orth>BUC</orth>
      <gen>rz</gen>
   </form>
   <sense><lbl>1.</lbl><def>ktoś o wydatny
         dziecko</def>: <quote>Od ńy/ fśis
      <usg type="geo">Bobowo st-gdań.</usg
   <sense><lbl>2.</lbl><def>ktoś małomówny
   <sense><lbl>3.</lbl>przezw<usg/><def>kt
         taki buc, pśeį^y i śy ńy uodezv^
      kroś.</usg></sense>
</entryFree>
```

BUC *rz* **1.** *'ktoś o wydatnych po liczkach, pucołowaty,` zwłaszcza dziecko'*: Od ńy/ fśiske byïi take buce. Fśiske ję̨ję̨/ ęeći Bobowo st-gdań.
2. 'ktoś małomówny': Ostrowce bus.
3. *przezw 'ktoś niesympatyczny, za rozumiały'*: To taki buc, pśeį^y i śy ńy uodezv^ Bobrka *kroś.*



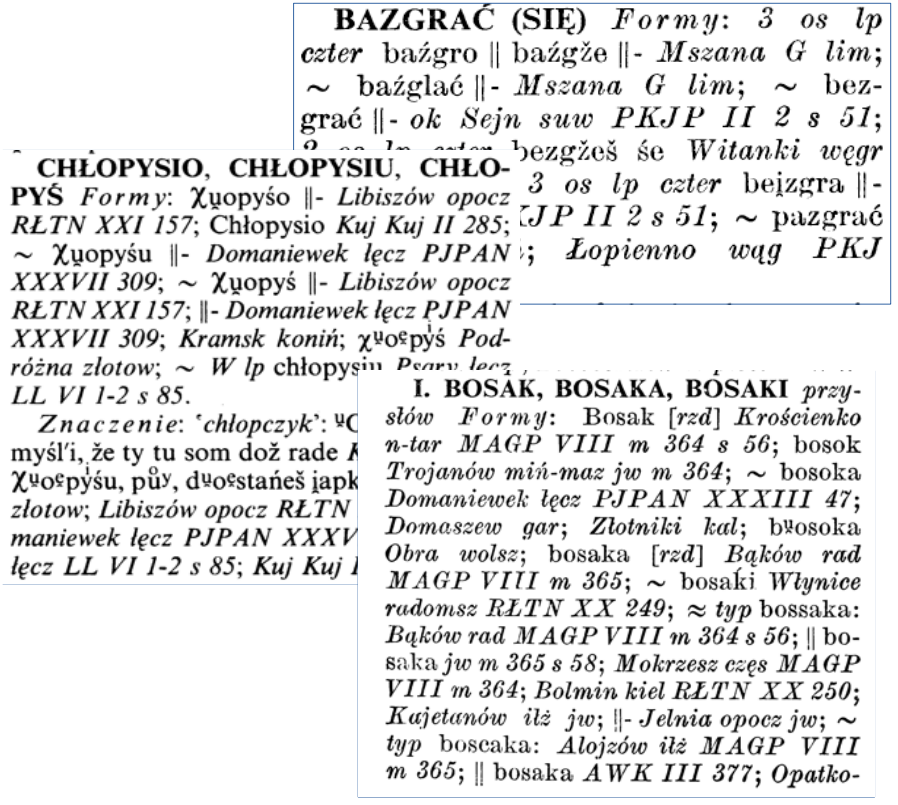
Jałowica, Jełowica *formy*: *n. sg.* jałowica 1400 *StPPP* VIII 929, XV *p. post. PF* III 286; jełowica 1453 *AGZ* XIV 388; ~ *d. sg.* jałowicy *BZ* Deut 21, 4; ~ *ac. sg.* jałowicę 1423 *StPPP* II nr 1910, 1439 *AGZ* XI 158, 1440 *AGZ* XI 165, *etc.*; ~ *i. sg.* jałowicą *BZ* Deut 21, 6; ~ *n. du.* jałowicy 1446 *AGZ* XI 272; ~ *ac. du.* jałowicy 1453 *AGZ* XI 386; ~ *ac. pl.* jałowice 1477 *StPPP* II nr 4194, 1484 *AGZ* XV 531.

Znaczenia: **1.** *iuvenca, vitula*': Pro 1400 *StPPP* VIII 92 dare debet 1423 *S* terciam penam ma 1439 *AGZ* XI 158; plenam… facultate ialowyczą 1440 *ib.* luet ialouicza 1442 czelne et due ialo Due ialowyczy trzied nyestany ialowiczą

Jałowiec *bot.* '*jałowiec pospolity, krzew lub jego owoc, Juniperus communis L.*': Yalowyecz bacce iuniperi 1419 *Rost* nr 5069; Ialowyecz iuniperus 1437 *ib.* nr 10836, *sim.* 1460 *ib.* nr 3602, *ca* 1465 *ib.* nr 3929, *etc.*, XV *p. post. R* LIII 66, *ca* 1500 *Erz* 23; Yalouiecz iuniperum 1444 *R* XXIII 302, *sim. ca* 1500 *Erz* 23; Jalowczą (*GlWroc* 85v. 133v, *R* XLVII 353: yalowecz) proiecitque se et dormiuit sub vmbra iuniperii (III Reg 19, 5) XV *p. pr. SKJ* I 303; Jalowecz XV *med. PulKras* 33; Iuniperi vlg. jalowyecz 1456 *R* XXXIII 183; Jalovyecz juniperium 1463 *PF* V 12; Yalowiecz yuniperius 1464 *Rost* nr 4819; Yalowecz iuniperus 1472 *ib.* nr 93, *sim.* 1491 *ib.* nr 11065; Yalovyecz iuniperius 1478 *ib.* nr 2110; Yalovyecz *ib.* nr 2266; *Valovyecz iuniperus *ca* 1500 *ib.* nr 2060.

Jałowię '*cielę, cielątko, vitulus*': Vitulum al. jalową 1429 *ArchCastrCrac* III 188; Petrus… tres vaccas, duo jalowata presentis anni al.

Dictionary of Old Polish



BAZGRAĆ (SIĘ) *Formy*: *3 os lp czter* baźgro || baźgże ||- *Mszana G lim*; ~ baźglać ||- *Mszana G lim*; ~ bezgrać ||- *ok Sejn suw PKJP II 2 s 51*; bezgżeš śe *Witanki węgr 3 os lp czter* bejzgra ||- *ZJP II 2 s 51*; ~ pazgrać *Łopienno wąg PKJ*

CHŁOPYSIO, CHŁOPYSIU, CHŁOPYŚ *Formy*: χu̯opyśo ||- *Libiszów opocz RŁTN XXI 157*; Chłopysio *Kuj Kuj II 285*; ~ χu̯opyśu ||- *Domaniewek łęcz PJPAN XXXVII 309*; ~ χu̯opyś ||- *Libiszów opocz RŁTN XXI 157*; ||- *Domaniewek łęcz PJPAN XXXVII 309*; *Kramsk koniń*; χu̯oʦpyś *Podróżna złotow*; ~ *W lp* chłopysiu *Psary łęcz LL VI 1-2 s 85*.

Znaczenie: '*chłopczyk*': u̯C myśl'i, że ty tu som dož rade K χu̯oʦpyśu, pů̯y, du̯oʦstańeš japk *złotow*; *Libiszów opocz RŁTN maniewek łęcz PJPAN XXXV łęcz LL VI 1-2 s 85*; *Kuj Kuj I*

I. BOSAK, BOSAKA, BOSAKI *przysłów Formy*: Bosak [*rzd*] *Krościenko n-tar MAGP VIII m 364 s 56*; bosok *Trojanów miń-maz jw m 364*; ~ bosoka *Domaniewek łęcz PJPAN XXXIII 47*; *Domaszew gar*; *Złotniki kal*; bu̯osoka *Obra wolsz*; bosaka [*rzd*] *Bąków rad MAGP VIII m 365*; ~ bosaki *Wlynice radomsz RŁTN XX 249*; ≈ *typ* bossaka: *Bąków rad MAGP VIII m 364 s 56*; || bosaka *jw m 365 s 58*; *Mokrzesz częs MAGP VIII m 364*; *Bolmin kiel RŁTN XX 250*; *Kajetanów iłż jw*; ||- *Jelnia opocz jw*; ~ *typ* boscaka: *Alojzów iłż MAGP VIII m 365*; || bosaka *AWK III 377*; *Opatko-*

Dictionary of Polish Dialects

From Image to Text

TIFF

Boruta 395 Borzyszka

stawiają naczynia z potrawami *Przyj-Ludu VI 127.* JO

II. BORUTA *'nazwa krowy'*: Boŕúta *Maguny i ok [Grodno ZSRR] Polgow II 169.* JO

Boruwać *zob.* **I. BOROWAĆ**

Borwinek *zob.* **BARWINEK**

BORYJOWICA *Forma:* Boryjuvica *Przewrotne rzesz.*

Znaczenie: *'gnojówka'*: *jw.* JO

Boryjówica *zob.* **BORYJOWICA**

BORYJÓWKA *'gnojówka'*: Boryjufka *Przewrotne rzesz.* JO

BORYKAĆ (SIĘ) *'plądrować, szukając przewracać'*: Borykać *Krak K I 108.*

Borykać się 1. *'szamotać się, walczyć; iść w zapasy'*: Borykać się *Krak jw.*

2. *'pokonywać trudności, zmagać się z przeciwnościami; męczyć się'*: Tagex śe borykou aż do roku tyśíne ẑevenset štyrycet *ẑpeńć Niebory [Cieszyn Czsł] PIJP I 46; Żył z prácy rąk swojéch,* a i ty prácy nié było dzie pószukać i tak sié w ty biydzié swoji burykáł *Żurawiczki przew LL I 3 s 26;* Caue ẑyće e'ovek śe tylkto b'orykáu, ńiż ńe uzyu *Samocice dąb-tar;* Śe ẑużok caye ẑyće borykou ji tero na stare ḃata ńiż ńi mo *Domaniewek łęcz PJPAN XXXIII 47; Krak K I 108; Złotniki kal:* ~ *B. się z czym:* Boryko śe z ruźnymi ńeszyśćamy *Studzienice rad RLTN XX 249.*

3. *'starac się, zabiegać o coś'*: Juz vlypoveġeli ńi maz roboty, a Io robote śe bloryka/i luże, byli śe *Niedzica n-tar ZNUJ CCLXXVIII 120.* JO

BORYKI *blp 'zapasy'*: Pastuch [...] muż „w boryki z kudłatym" [*'wilkoła-kiem'*] *Bodzentyn kiel Lud I 204.* JO

BORYN
ẑętycy z h
n-tar PIJ

BORYŃ
borynka J
bańka na
woja wad

I. BOR
(Sparassi
PIJPAN

II. Borys *zob.* **PORYS**

BORYSZ *'poczęstunek po dobitym targu'*: Jak zaśli na ten boryś, tu pij a pij *Degucie suw;* Boŕyśa mus'ieli my ẑafśe vłypić', nawet żyt jak kuplovał stawił čv'artka *Łopatowszczyzna [Wilno ZSRR] PJPAN LXI 121; Dobrzyniewo D białoś; Stawęcin choszcz [Michaliszki, Wilno ZSRR] ZPSBJ IV 133.*

Por. BARYSZ. JO

BORYSZEK *'zdr od bór w zn 1'*: Boryšek *Kubeczki rawic PKJP III 3 s 150.* JO

Boryszka *zob.* **I. BORZYSZKA**

BORYSZOWAĆ *'hulać'*: Jak sie dziedzie dowiedział, ze rzońca tak borysuje, dał mu trzy dni folgi *Turów radz-podl PF VI 219.* JO

BORYTASY *zwykle w lm 'rodzaj ornamentu zdobiącego spodnie góralskie'*: Mężczyźni nosili na terenie całego Śpisza spodnie z białego sukna [...] zdobione haftowanymi parzenicami [...] lub [...] uwitymi z dwubarwnego sznura [...] albo z dwubarwnych tasiemek z dodatkiem haftu zw. „borytasami" *N-tar StarPrzew 34; N-tar pd-wsch ASL V 1 s 12.* JO

Boryzna *zob.* **I. BURZYNA**

BORZAN *'mieszkaniec lasu'*: Bořańe a lesáce to je to samo *Kasz S I 59.* JO

Borzeczka *zob.* **PORZECZKA**

Borzuk *zob.* **BORSUK**

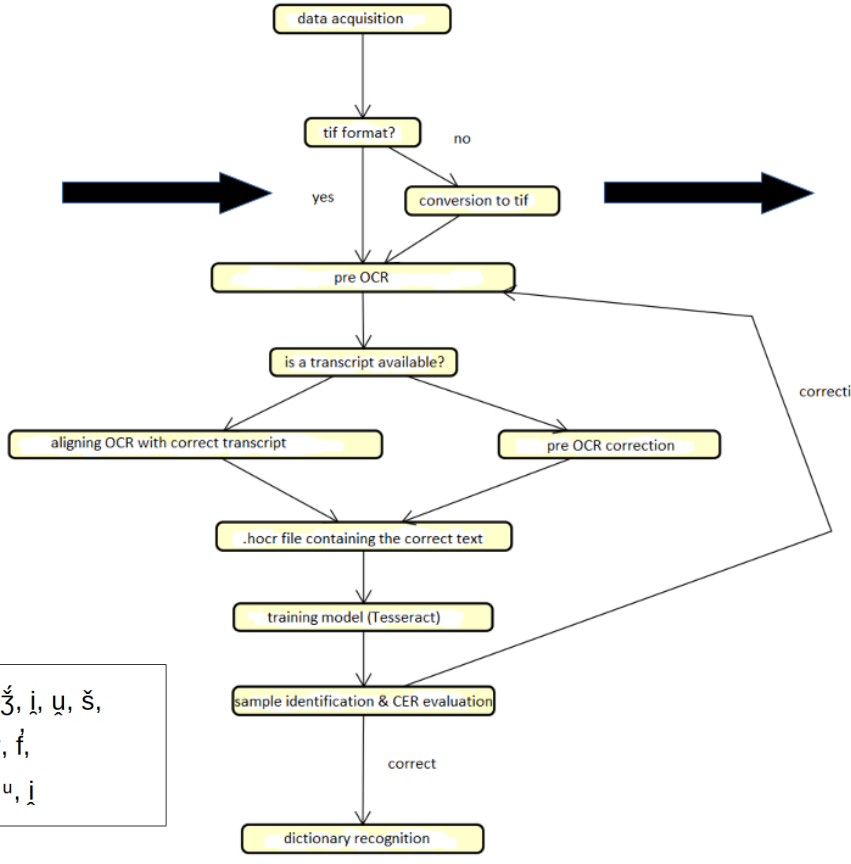
BORZYK *'krzak borówki czernicy (Vaccinium myrtillus)'*: Borzyk *Wólka koln SWM XI 58 m 12.* JO

Borzyna *zob.* **BARZYNA**

BORZYNY *blp 'borówka brusznica (Vaccinium vitis idaea)'*: Bořěně sọ do zapravě dobré *Kasz S I 59.* JO

combining characters: ọ, ʒ, ʒ́, ǯ, ʒ̋, i̯, u̯, š, č, ž, ś, χ́, ǩ, ǵ, ń, ṕ, b̓, t̓, ď, à, å, ý, f́, ẃ, ḿ, ŋ, ń, ţ, ṇ, ḍ, ȯ, å, y̯, e̯, į, o̯, ᵘ, i̯

OCR



Boruta

stawiają naczynia z potrawami Przyj- Ludu VI 127. JO

II. BORUTA "'nazwa krowy": Boróta Maguny i ok [Grodno ZSRR] Polgow II 169. JO

Boruwać 2ob. I. BOROWAĆ

Borwinek zob. BARWINEK

BORYJOWICA Forma: Boryjuvica Przewrotne rzesz.

Znaczenie: "gnojówka": jw.

Boryjówica zob. BORYJOWICA

BORYJÓWKA gnojówka": Boryjufka Przewrotne rzesz. JO

BORYKAĆ (SIĘ) "plądrować, szuka- jąc przewracać: Borykać Krak K I 108. Borykać się 1. 'szamotać się, walczyć; iść w zapasy': Borykać się Krak jw.

2. "'pokonywać trudności, zmagać się 2 przeciwnościami; męczyć się': Tagex śe borykou aż do roku tyślíne ẑevenset štyrycet jpeńć Niebory [Cieszyn Czsł] PIJP I 46; Żył z prócy rąk swojech, a 1 ty pracy nié było dzie pószukać 1 tak sié w ty biydzié swoji burykił Żurawiczki przew LL I 3 s 26; Caue ẑyce e'ovek śe tylkto bYorykiu, ńi ńe uzyu Samocice dąb-tar; Se zużok caye ẑyce borykou ji tero na stare lata ńiĘ ńi mo Domaniewek łęcz PJPAN AXXXIII 47; Krak K I 108; Złotniki kal:; — B. się z czym: Boryko śe z ruźnymi ńescyśćamy Studzienice rad RLTN XX 249.

3. 'starac się, zabiegać o coś': Juz vlypovegeli ńi maz roboty, a Io robote se bloryka/i luże, byli śe Niedzica n-tar ZNUJ COLXXVIII 120. JO

BORYKI blp "zapasy": Pastuch [...] muż „w boryki z kudłatym" [wilkoła- kiem'] Bodzentyn kiel Lud I 204. JO

BORYNA "naczynie do przewożenia żętycy z hali do wsi": Boryna Dzianisz n-tar PIJP IV 41. JO

BORYNKA "dzbanek": Borynka || [st] borynka Rabka-Zaryte n-tar; — 'blaszana bańka na mleko, jagody": Borynka Za- woja wad PIJP XIII 64. JO

Experiments

- grobid-dictionaries
- other ML solutions

Solution: custom Python (*CLARIN-PL) and XSLT scripts

- punctuation marks
- anchoring strings: labels, abbreviations, initials etc.
- location
- no font style info

Motivation

- (slightly more) formal resource's description
- documenting
  - annotation choices and semantics
  - original rendering
- generating scheme and documentation

Principles

- preserving textuality for quoting and display
- exposing data to enrichment
  - aligning dictionary parts
  - dates, places, sources etc.
  - reconciling meta-languages

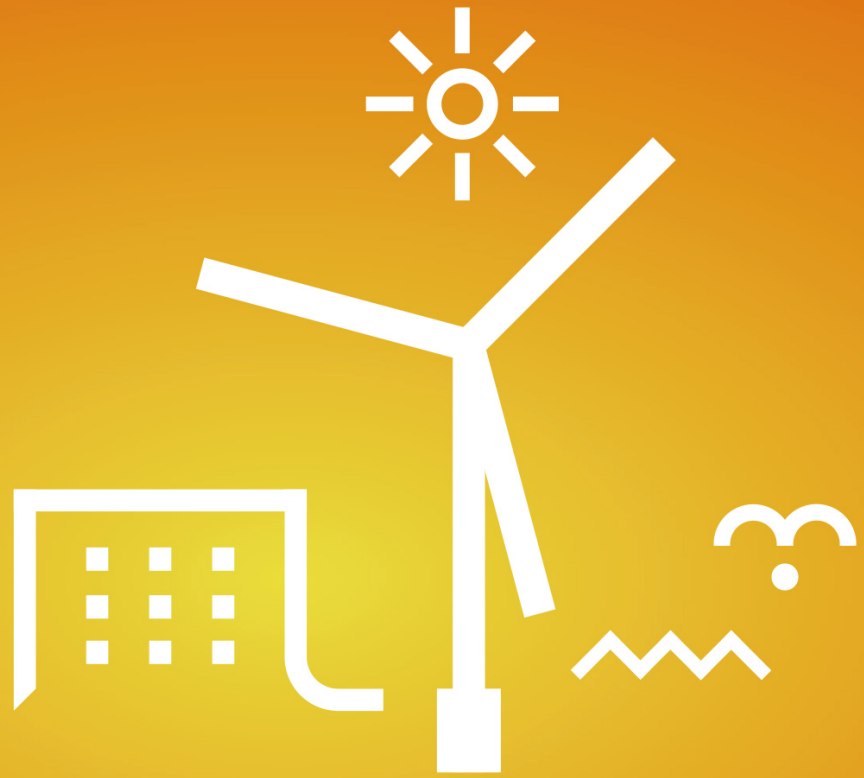
```
<schemaSpec ident="ijpsgp"
                source="ijp_all.compiled.odd">
  <elementSpec ident="entryFree" mode="change">
    <gloss xml:lang="pl">hasło</gloss>
    <desc xml:lang="pl">hasło słownikowe</desc>
    <attList org="group">
      <attDef ident="type" mode="change">
        <valList type="closed" mode="change">
          <valItem ident="main" mode="change">
          <gloss xml:lang="pl">zwykłe</gloss>
          <desc xml:lang="pl">zwykłe hasło</desc>
          </valItem>
        </attDef>
    </attList>
  </elementSpec>
</schemaSpec>
```

Delimiting lexical units

Linguistic evidence: time and space

Taking stance: critical apparatus

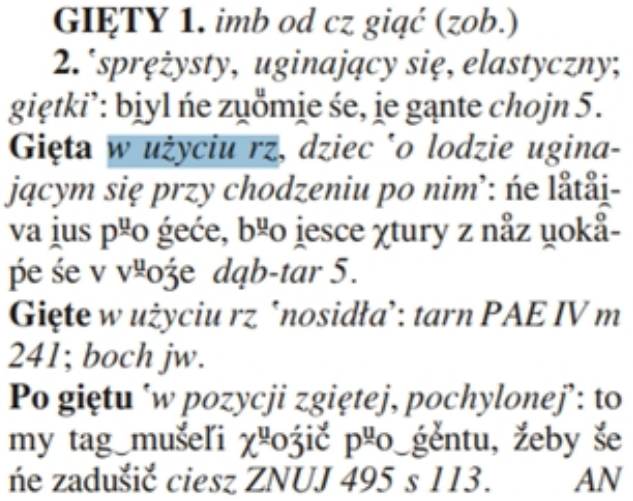
Citation strategies



*BAŁAŁAJKA

```
<entryFree type="empty">
   <form type="lemma">
      <lbl>*</lbl>
      <orth norm="bałałajka">BAŁAŁAJKA</orth>
   </form>
</entryFree>
```

‘a balalaika’



GIĘTY 1. *imb od cz giąć (zob.)*
2. *'sprężysty, uginający się, elastyczny; giętki'*: biyl ńe zu̯o̯mie śe, je gąnte *chojn 5*.
Gięta *w użyciu rz, dziec 'o lodzie uginającym się przy chodzeniu po nim'*: ńe låtå̯iva jus pu̯o ǵeće, bu̯o jesce χtury z nåz u̯okå̯pe śe v vu̯oźe *dąb-tar 5*.
Gięte *w użyciu rz 'nosidła'*: *tarn PAE IV m 241; boch jw.*
Po giętu *'w pozycji zgiętej, pochylonej'*: to my tag_mušeľi χu̯oźič pu̯o_ǵẽntu, żeby še ńe zadušič *ciesz ZNUJ 495 s 113.* AN

```
<!-- main entry: a participle -->
<entryFree>
    <!-- related entry: a derived noun-->
    <re/>
    <!-- related entry: a derived noun -->
    <re/>
    <!-- related entry: a derived prepositional phrase -->
    <re/>
</entryFree>
```

```
<sense>
    <lbl>~</lbl>
    <usg type="colloc">jałowcowe drzewo</usg>
    <usg type="dom">bot.</usg>
    <def xml:lang="pl">jałowiec pospolity</def>
    <def xml:lang="la">Juniperus communis L.</def>
</sense>
```

‘juniper tree’

Doubting about...

```
<sense xml:id="Sstp_Jałmużnik_sense_1">
    <def xml:lang="pl">dający jalmużnę lub z urzędu
                      rozdzielający jałmużnę</def>
    <def xml:lang="la">qui eleemosynam dat vel ex officio
                      distribuit</def>
    <lbl>(?)</lbl>
    <certainty locus="value" degree="0.5" match="orth"/>
</sense>
```

… word senses

```
<form type="lemma">
    <orth norm="grzycka">GRZYCKA</orth>
    <lbl>[?]</lbl>
    <certainty degree="0.5" locus="value" match="orth"/>
</form>
```

… word forms

Space (in time)

```
<cit xml:lang="pl">
    <quote>Boryna</quote>
</cit>
<usg type="geo">
    <placeName>
        <settlement>Dzianisz</settlement>
        <lbl type="geo>
            <region type="county" when="1986">
                n-tar
            </region>
        </lbl>
    </placeName>
</usg>
<cit>
    <bibl>
        <title type="abbr">PIJP</title>
        <biblScope unit="volume">IV</biblScope>
        <biblScope unit="page">41</biblScope>
    </bibl>
</cit>
```

```
<cit>
    <quote  xml:id="quote-ialmuzna-1"> ...</quote>
        <bibl>
            <title type="abbr" xml:id="BZ-1">BZ</title>
            <biblScope unit="chapter">2</biblScope>,
            <biblScope unit="line">16</biblScope>
        </bibl>
</cit>
```

```
<cit>
    <ref corresp="#quote-ialmuzna-1" type="sim">sim.</ref>
    <bibl>
        <ref target="#BZ-1">ib.</ref>
        <biblScope unit="chapter">12</biblScope>,
        <biblScope unit="line">9</biblScope>
    </bibl>
</cit>
```

Interpretation hints

```
<cit>
    <quote>Tegodlya,
        <seg xml:id="seg-4">kyedy czynysch yamvzna</seg>
        <note>
            (<gloss target="#seg-4" resp="#lexicographer">
                cum... facis eleemosynam
            </gloss>
            <bibl type="bible">
                <title>Mat</title>
                <biblScope unit="chapter">6</biblScope>
                <biblScope unit="line">2</biblScope>
            </bibl>)
         </note>, nye day przed sobą trąbycz ...
    </quote>
    <bibl>
        <title type="abbr">Rozm</title>
        <biblScope unit="page" from="271" to="272">271-2</biblScope>
    </bibl>
</cit>
```

```
<cit>
    <quote>
        Ialmvsznym
        <corr resp="#lexicograph" xml:id="corr-1">nacznem</corr>
        <note>(
            <lbl>leg.</lbl>
            <sic corresp="#corr-1">nędznem</sic>
        )</note> nye daval
    </quote>
</cit>
```

Conclusions

Keep track of advances in token-level text structuring

- sharing tools and expertise

Discuss with content creators

Take annotators' feedback into account

Work with ODD

- describe the resource
- document annotation choices
- explain assumptions about document semantics
- generate multi-lingual documentation

Know when to stop