What gooseberries, grapes and (bad) wine have in common?

Linking Dictionaries of Historical Varieties of Polish

Krzysztof NOWAK ❖ Dorota MIKA ❖ Ewa RODEK

eLex 2023

Institute of Polish Language (PAS) in the Dariah.lab Project

Extending National Corpus of Polish 2011-2020

- spoken corpus acquisition
- automatic transcription

Machine-readable dictionary acquisition

- digitisation
- structuring

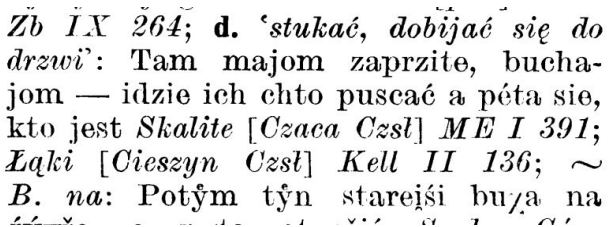
Integrated access

- standardisation
- enrichment
- linking
  - internally
  - externally
- publishing
  - API
  - web interface

Resources

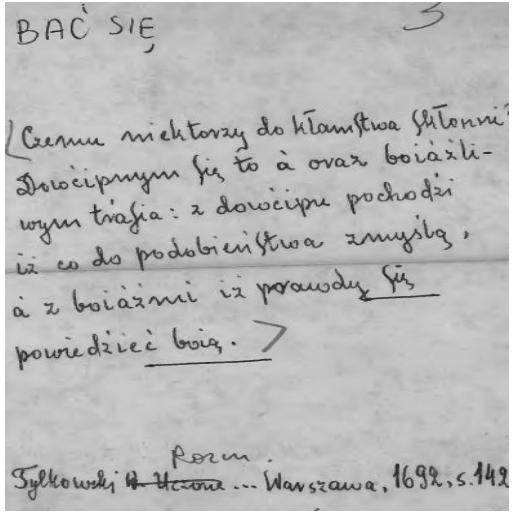
electronic dictionaries

paper-born dictionaries



Zb IX 264; **d.** '*stukać, dobijać się do drzwi*': Tam majom zaprzite, buchajom — idzie ich chto puscać a péta sie, kto jest *Skalite* [*Czaca Czsł*] *ME I 391*; *Łąki* [*Cieszyn Czsł*] *Kell II 136*; ~ *B. na*: Potym tyn starejśi buża na

source data

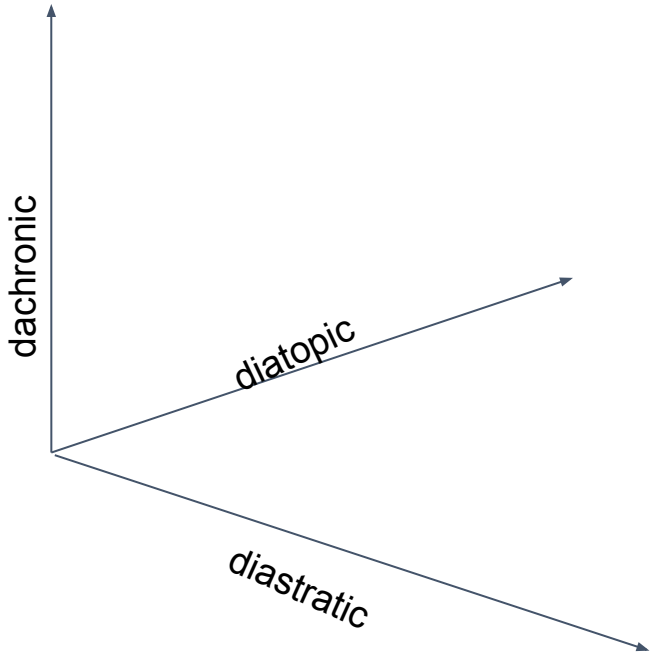


BAĆ SIĘ 3

Czemu niektorzy do kłamstwa skłonni?
Dowcipnym się to a owaz boiaźli-
wym trafia: z dowcipu pochodzi
iż co do podobieństwa zmysłą,
a z boiaźni iż prawdy się
powiedzieć boią.

Rozm.
Sykowski ... Warszawa, 1692, s. 142

Language variation: diachronic dimension



- ➔ Dictionary of Polish Medieval Latin
- ➔ Old Polish Dictionary
- ➔ Thesaurus of Old Polish
- ➔ eDictionary of the 17th- and 18th-century Polish
- ➔ Great Dictionary of Polish

1000

1500

1700

2000

Language variation: diatopic and diastratic dimension

- dictionaries of geographic varieties
  - Dictionary of Polish Dialects
  - Dictionary of Polish Borderlands
- atlases
- dictionaries of proper names
  - Dictionary of Medieval Person Names
  - Dictionary of Polish Place Names

Goals

```
<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>
```

- What words are in use in the Northern Poland?
- What words were employed to insult someone in 15th century?

research

API

```
"@n": "BUC1",
"@type": "hom",
"form": {
   "orth": "BUC",
   "gen": "rz"
},
```

integrated access

BUC
rz

1.ktoś o wydatnych po liczkach, pucołowaty, zwłaszcza dziecko: Od ńy/ fśiske byïi take buce. Fśiske jeji/ ęeći Bobowo st-gdań.

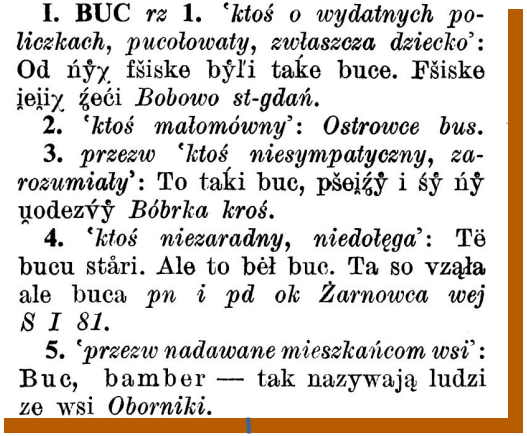
Web presentation

. Miło mi się z Panią śmieje, ale czasem ze mnie	**buc**	wyłazi. Jakbym więc wrzucił zgniły ogryzek
ządzą i co? Nic, afera za aferą, a ty stój jak głupi	**buc**	w kolejkach, bo po to całe życie pracowa
rza. Na sześćdziesiątym kilo policja. Suka? No i	**buc**	, knur kierownik wszystkich, kurwa, zna
tylko w wyszukiwarce Google wpisze się słowo	**buc**	, idiota lub brzydkie słowo na ch lub k, z

buc

pot. pogard.
człowiek zarozumiały, zbyt pewny siebie

Acquiring machine-readable representation



I. BUC *rz* **1.** *'ktoś o wydatnych policzkach, pucołowaty, zwłaszcza dziecko'*: Od ńy̨ fśiske by̨li take buce. Fśiske ję̨i̯y̨ ę̨eći *Bobowo st-gdań.*

2. *'ktoś małomówny'*: *Ostrowce bus.*

3. *przezw 'ktoś niesympatyczny, zarozumiały'*: To taki buc, pśeį^y i śy ńy uodezv^ *Bóbrka kroś.*

4. *'ktoś niezaradny, niedołęga'*: Të bucu ståri. Ale to bël buc. Ta so vząła ale buca *pn i pd ok Żarnowca wej S I 81.*

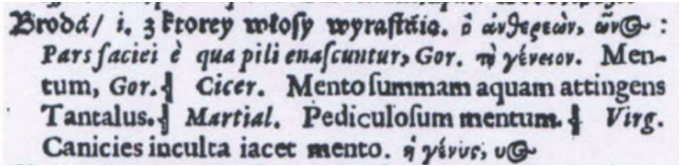
5. *'przezw nadawane mieszkańcom wsi'*: Buc, bamber — tak nazywają ludzi ze wsi *Oborniki.*

```
<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 ję̨i̯i/ ę̨eći 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.*

Issues: data inconsistency and heterogeneity

historic scripts



Brodá/ i. z ktorey włosy wyrastáio. ὁ ἀνθρεών, ῶνος: Pars faciei è qua pili enascuntur, Gor. τὸ γένειον. Mentum, Gor. ‖ Cicer. Mento summam aquam attingens Tantalus. ‖ Martial. Pediculosum mentum. ‖ Virg. Canicies inculta iacet mento. ἡ γένυς, υος

Abecadło, a, lm. a, †Abiecadło, †Obiecadło, [Abeceda] l. *alfabet*. 2. a. ABC przen. *pierwsze początki jakiej nauki, elementarne jej zasady*. Nie zna abecadła filozofji, a wyrokuje o niej. 3. *pewna gra towarzyska*. Gol. <Z nazw głosek abc>

formalism and notational systems

Abramt
1. SSNO: brak; Anti
2. nom. sg. Abramt l
3. XVII: Śl (BoBy).

Comparat. n. sg. m. {
1470 *MamLub* 162, XV
etc.; *f.* gorsza *Rozm* 295;

non-standard characters

3. *przezw* ‘*ktoś niesympatyczny, zarozumiały*’: To taki buc, pšeiźy i šy ńy ụodezvy *Bóbrka kroś.*

on (*sc.* św. Jan) gest bil sze slich ludzy narodzyl, tedyczby mv vøcz *oniy ne bily verzyly *Gn* 11b;

non-standard linguistic varieties

historic

multi-language

destruccio al. ta sskaza zgorzala, ys sezgly ya zly ludze 1435 *Pozn* nr 1577, *sim. ib.*; Aby pothwarzam slych lvdzy (malorum hominum)…

dialectal

aaa, kotki dwa *rzecz. ndm.* ‘spanie, zaśnięcie, utrata przytomności’: Nie pamiętał jednakże, kiedy zamknął oczy i zrobił „aaa, kotki dwa” na zimnym betonie. Hiaas. S, 56

Data modelling: dictionary as text (1)

meta- and micro-structure

Homonimy:

babka I, zdrobn. **babcia, babusia 1.** 'babcia'. *Chadźiajin toj, że jego bapka* [budował dla niej chatę]. *Dźiadunio i babusia* [JS44]. *Byli wze dźieći nasze, było*

```
<entryFree type="hom" xml:id="SGO_entry-n" n="babka1">
    <form type="lemma">
      <orth>babka</orth>
      <lbl>I</lbl>,
    </form>
    <usg type="style">zdrobn.</usg>
    <oRef>babcia</oRef>,
    <oRef>babusia</oRef>
```

Informacja gramatyczna:

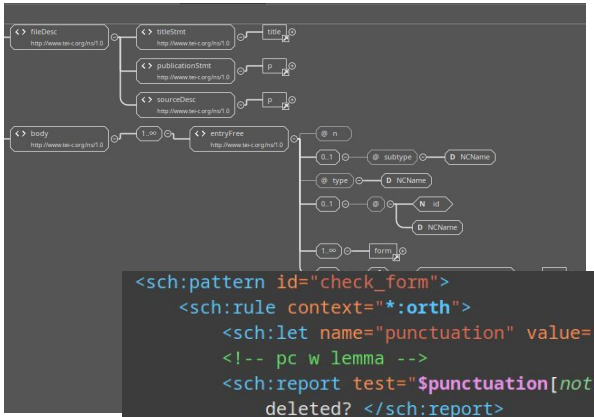
siekać ndk – **posiekać** dk 'siekać, pociąć na drobne kawałki'. [Mama] *kapusty śikali* [K08]. [Co to znaczy *pokryszy*?] *Nożem, posiecze śie* [JS44]. Por. kryszyć

```
<entryFree type="main" xml:id="SGO_entry-n" n="siekać">
    <form type="lemma">
      <orth>siekać</orth>
    </form>
    <gramGrp>
      <subc rend="aspect">ndk</subc>
    </gramGrp>
    <lbl>-</lbl>
    <form>
```

Data modelling: dictionary as text (2)

ODD, RNG, Schematron

XSLT

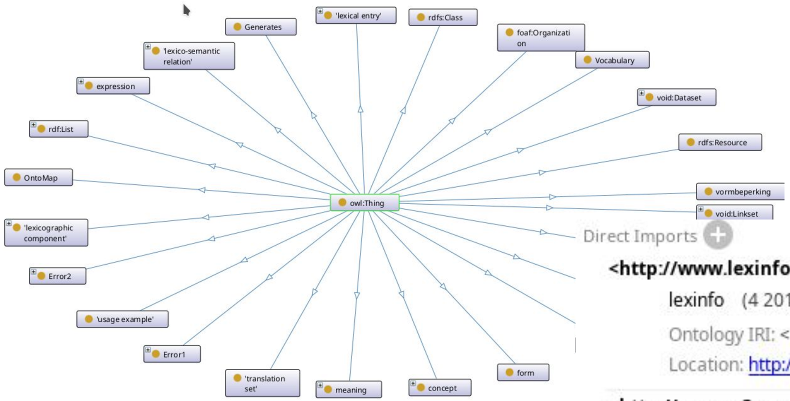



```
<sch:pattern id="check_form">
    <sch:rule context="*:orth">
        <sch:let name="punctuation" value="*:pc"/>
        <!-- pc w lemma -->
        <sch:report test="$punctuation[not(matches(., '-'))]" sqf:fix="deleteNoise"> Should be
            deleted? </sch:report>
        <sqf:fix id="deleteNoise">
            <sqf:description>
                <sqf:title>Delete noise</sqf:title>
            </sqf:description>
            <sqf:delete match="$punctuation"/>
        </sqf:fix>
    </sch:rule>
```

```
            </xsl:choose>
        </xsl:when>
        <xsl:when test="name(.) = 'vol'">
            <xsl:element name="biblScope">
                <xsl:attribute name="unit" select="'volume'"/>
                <xsl:apply-templates mode="#current" select="node() | text()"/>
            </xsl:element>
        </xsl:when>
        <xsl:when test="name(.) = 'span'">
            <xsl:copy>
                <xsl:apply-templates mode="#current" select="node() | text()"/>
            </xsl:copy>
        </xsl:when>
        <xsl:when test="name(.) = 'number'">
            <xsl:element name="num">
                <xsl:apply-templates mode="#current" select="node() | text()"/>
```

manual correction

Data modelling: dictionaries as linked data (1)



Direct Imports

<http://www.lexinfo.net/ontology/2.0/lexinfo>
lexinfo (4 201 axioms, 1 168 logical axioms)
Ontology IRI: <http://www.lexinfo.net/ontology/2.0/lexinf
Location: http://www.lexinfo.net/ontology/2.0/lexinfo#

<http://www.w3.org/ns/lemon/lexicog>
lexicog (92 axioms, 16 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/lexicog>
Location: http://www.w3.org/ns/lemon/lexicog#

<http://www.w3.org/ns/lemon/lime>
lime (387 axioms, 60 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/lime>
Location: http://www.w3.org/ns/lemon/lime#

<http://www.w3.org/ns/lemon/vartrans>
vartrans (278 axioms, 34 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/vartrans>
Location: http://www.w3.org/ns/lemon/vartrans#

Indirect Imports

<http://lemon-model.net/lemon>
lemon (813 axioms, 239 logical axioms)
Ontology IRI: <http://lemon-model.net/lemon>
Location: http://lemon-model.net/lemon

<http://www.w3.org/ns/lemon/ontolex>
ontolex (521 axioms, 72 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/ontolex>
Location: http://www.w3.org/ns/lemon/ontolex

<http://www.w3.org/ns/lemon/ontolex>
ontolex (521 axioms, 72 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/ontolex>
Location: http://www.w3.org/ns/lemon/ontolex

<http://www.w3.org/ns/lemon/ontolex>
ontolex (521 axioms, 72 logical axioms)
Ontology IRI: <http://www.w3.org/ns/lemon/ontolex>
Location: http://www.w3.org/ns/lemon/ontolex

Data modelling: dictionaries as linked data (2)

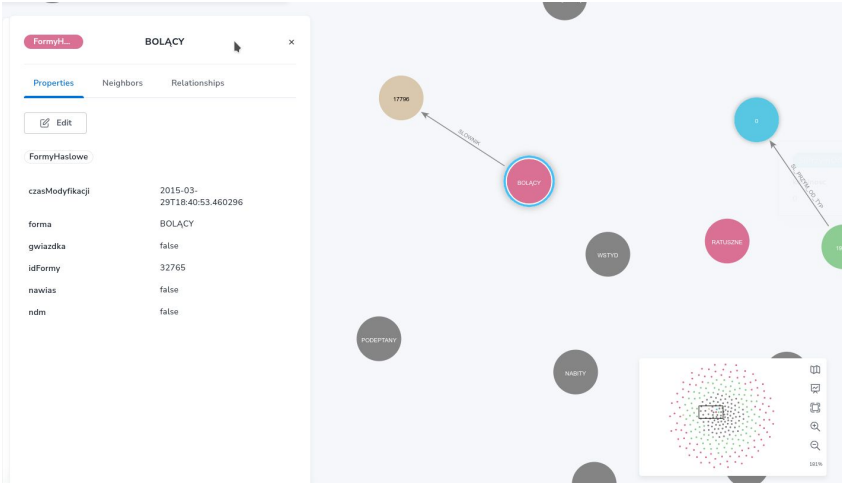
Graph database (*src*)

Data cleaning and reduction (*clean*)

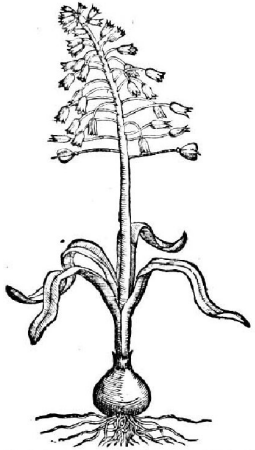
Ontology-based import (*semantic*)

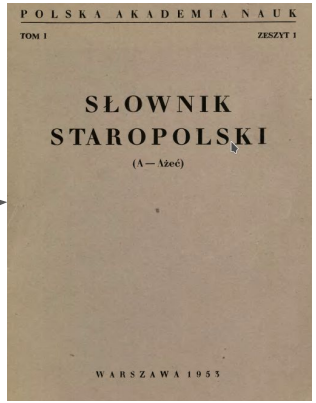
Mapping (*common*)

- internal ressources
- test: Wikipedia



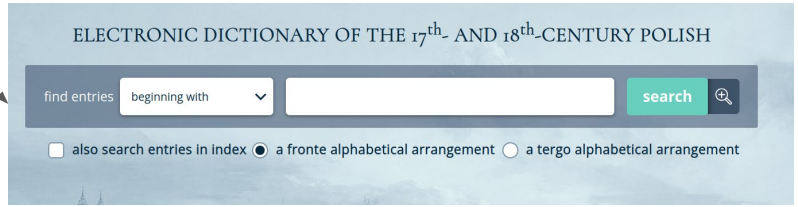
Research scenario: regular polysemy of plant names in diachrony





Dictionary of Old Polish

- regular polysemy
- plant names
- 2 dictionaries
- diachronic perspective



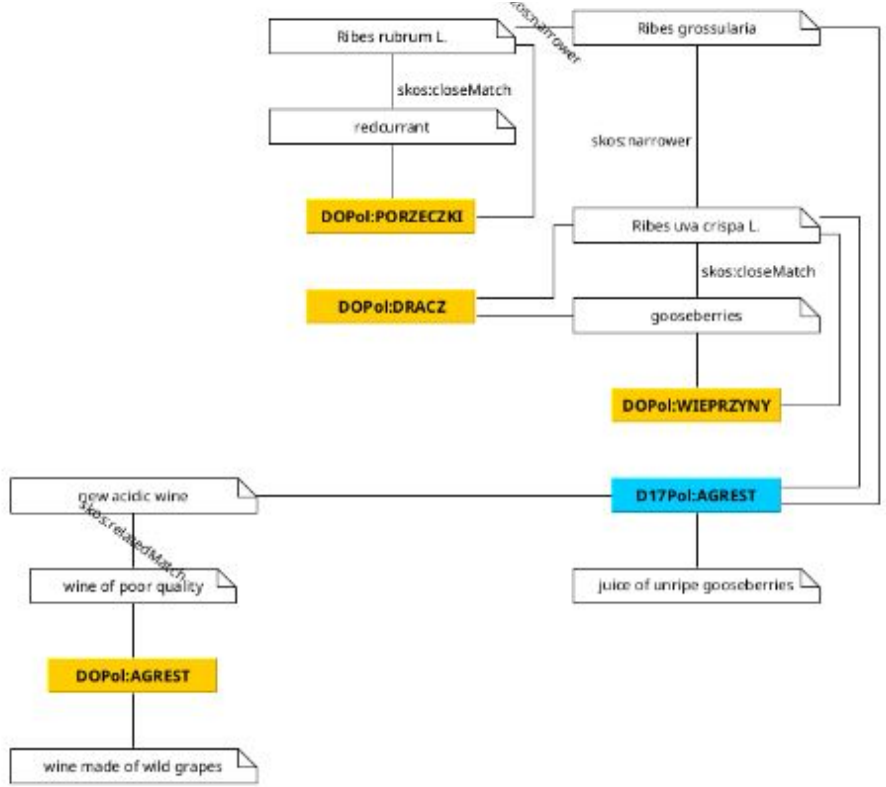
eDictionary of the 17th- and 18th-century Polish

Plant names and historical dictionaries

PORZECZKI

1. *redcurrant, Ribes rubrum L.*
2. *gooseberries, Ribes grossularia*
3. *juice of unripe gooseberries*
4. *new acidic wine*
5. *wine of poor quality*

- ➢ scientific taxonomy
  - identity
  - categorization
- ➢ reference to external knowledge
- ➢ vagueness and ambiguity
- ➢ regular polysemy
- ➢ insight into evolving lexicographic practices



Regular polysemy

Apresjan (1974): Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language,

- there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and
- if a_i and b_i, a_j and b_j are nonsynonymous.

ANIMAL	→	**ITS MEAT**
There's a squirrel.	→	We don't eat squirrel.

(Atkins & Rundell 2008)

Atkins & Rundell (2008): such inter-word relationships are of immediate interest to lexicographers

Regular polysemy markers

- **within a single definition**
  - cumulative and disjunctive conjuctions:

    OAT plant or grain
  - anaphoric expressions:

    PEA round grains; also plant with these grains
- **separate definitions**

  ONION 1. plant 2. edible part of the onion

False friends: vagueness and uncertainty

CHMIEL hop, *Humulus Lupulus L.* – plant or fruit

isPartOf

isSimilarTo

UŚPIWRZÓD *Pulicaria vulgaris Stev*. or one of the species of the *Inula sp. L*. genus

Regular polysemy: diachronic patterns

plant X → part of X (1)

1. (ornamental) plant X → flower of X

RÓŻA (‘rose’) 1. **flower**, symbol of beauty 2. bot. *Rosa alba L.* /Sstp/

RÓŻA (‘rose’) 1. thorny flowering bush. a. its **flower**. /eSXVII/

2. (fruit) bush / tree → fruit of X

MIGDAŁ almond, plant or its **fruit**, *Amygdalus communis L.*’ /Sstp/

MORELA (‘apricot’) 1. tree with bright orange, tasty fruits. a. its **fruit**. /e-SXVII/

plant X → part of X (1)

3. cereal grass X → grain(s) of X

JĘCZMIEŃ barley (plant, **grains**), *Hordeum vulgare L.* /Sstp/

JĘCZMIEŃ ~~cereal~~ [entry in preparation] /e-SXVII/

4. plant X → part of X (different than fruit, grain, or flower)

PALMA (‘palm’) plant; ~ about its **leaf** or branch /Sstp/

PALMA tropical tree; also refers to its **leaves** /e-SXVII/

plant X → product made of X (1)

1. (edible) plant X → food made of X

GROCH peas, *Pisum sativum L.* – plant or its fruit; ~ peas as **a dish** /Sstp/

GROCH 1. pulse. 2. its grains and **a dish** made of them /e-SXVII/

2. tree X → wood of X

LESZCZYNA (‘hazel’) *Corylus avellana L.*, also its **wood** [...]’ /Sstp/

LESZCZYNA ~~bushes of nut-trees~~ /e-SXVII/

plant X → product made of X (2)

3. plant X→ spice made of X

PIEPRZ 1. black pepper. 2. **seeds** of black pepper used as a **spice** /Sstp/

4. plant X → other product of X

NARD (‘nard’) plant of valerian family, also **essential oil** made of it /Sstp/

ALOES (‘aloe’) 1. plant of liliacae family; **juice** of its leaves /e-SXVII/

plant X → group of X

1. plant X → plant Y [← similar to X]

GŁOGOBICZ **various** thorny bushes: blackberries, hawthorn etc. /Sstp/

OPICH celery; *Apium*; also other **similar plants** /e-SXVII/

2. plant, tree X → group of X [← consisting of X or of plants similar to X]

BLUSZCZ (‘ivy’) climbing plant; also waterside **bushes** /e-SXVII/

DĘBINA oak wood; oak **forest** /e-SXVII/

Conclusions and perspectives

- the automatic analysis is not entirely reliable
  - granularity and quality of dictionary description
  - quantitative analysis: no evidence ≠ no linguistic event
- identifying major patterns of regular polysemy
  - productivity
  - changing categorization
- more data
  - language change preserved in dialects
  - contemporary Polish
  - cross-linguistic comparison