Multisłownik: Linking plWordNet-based Lexical Data for Lexicography and Educational Purposes

Maciej Ogrodniczuk, Zbigniew Bronk, Witold Kieraś | Institute of Computer Science, Polish Academy of Sciences
Joanna Bilińska | University of Warsaw

CLARIN-PL
Common Language Resources and Technology Infrastructure

Multisłownik in a nutshell

- an automated integrator of Polish lexical data retrieved from multiple available online sources
- based on Słowosieć, the Polish WordNet
- linking external resources to synsets

Motivation

- a tendency to integrate dictionaries into portals:
  - https://en.oxforddictionaries.com/
  - http://www.termania.net/
  - http://dictionaryportal.eu/
  - http://fran.si/
  - https://sjp.pwn.pl/
- lexicographers still comparing lexical definitions in many online services since:
  - portals are mainly a source of information for the ordinary users rather than linguists and researchers
  - multi-dictionary search (Fran, PWN, Dictionary Portal) only presents entries from component dictionaries 'as is' on a single Web page

Our solution

- combination of a dictionary portal and a general dictionary:
  - automatically generated dictionary-like short entries
  - references to source data
- between the 'dictionary by dictionary' view and complete integration into a single resource

Sources of lexical data

- traditional dictionaries created by philologists and meant for human readers only, either web-based or digitalized
- electronic datasets created by computational linguists for both human users and automated NLP processing
- community-based lexical collections developed online.

Core resources

- plWordNet
- Grammatical Dictionary of Polish (SGJP)
- Wikipedia / Wikisource
- Walenty valency dictionary
- National Corpus of Polish (NKJP)
- SJP.pl

Presence of the entry in other linked sources

- digitized versions or paper dictionaries (e.g. PWN dictionaries: Dictionary of Polish, Dictionary of Foreign Words, Doroszewski's classical dictionary available as scanned pages)
- academia-based electronic dictionaries (e.g. Dictionary of 17th & 18th Century Polish, Great Dictionary of Polish)
- community-based lexical databases (e.g. urban slang dictionary, dictionaries of synonyms, antonyms and crossword definitions)

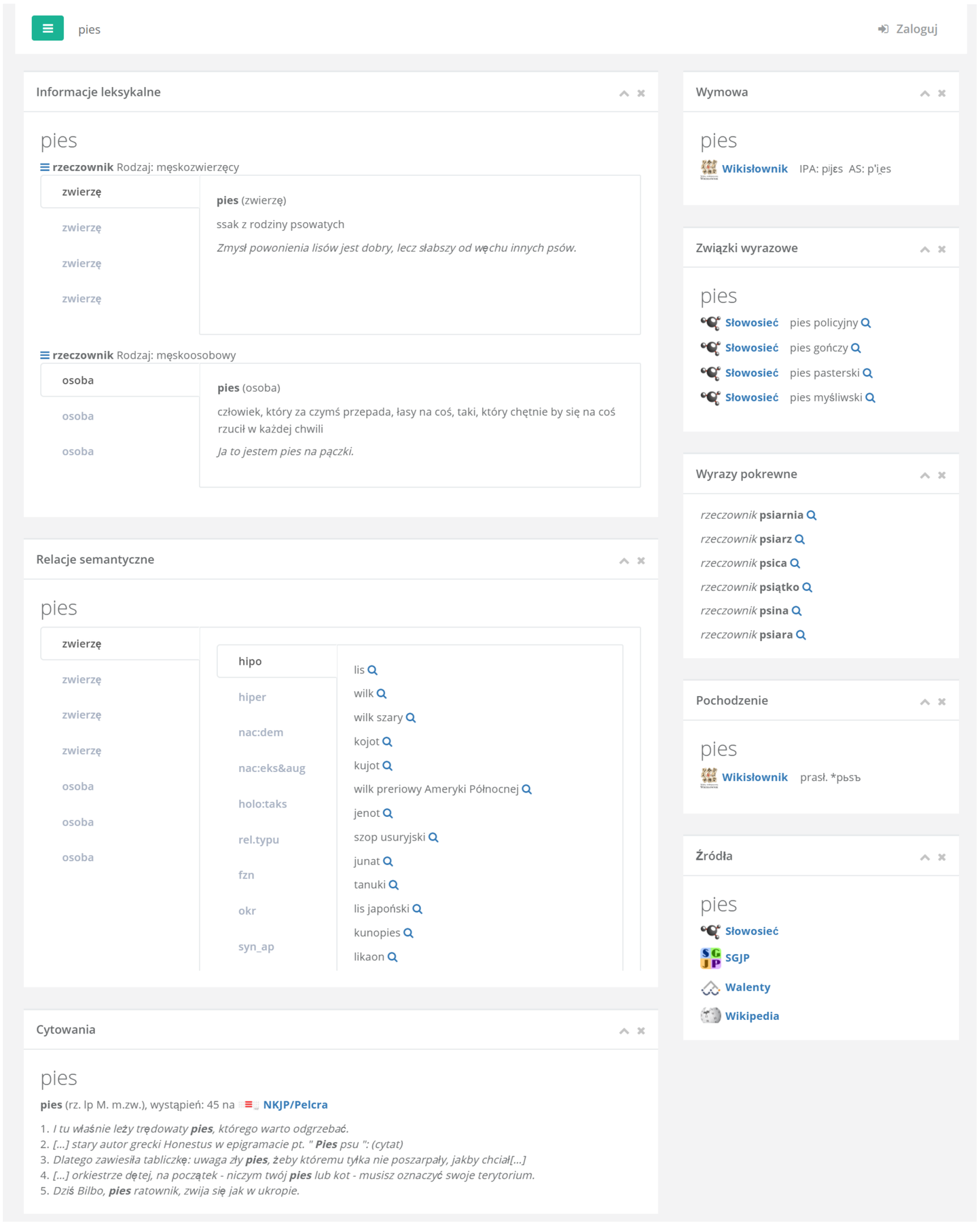
Integration

- a common point of reference serving as the core of the integration — Słowosieć
- difficulties:
  - heterogeneity of resources
  - different levels of detail and incompleteness of coverage of lexical entries
  - constant change of online resources
- complete integration unfeasible — and unnecessary

Usage

- lexicographic scenario: an instant support for a professional lexicographer working in the field of extending a specific dictionary or performing linguistic annotation.
- educational scenario: teaching students the differences between dictionaries by looking up words

New interface



A sample educational scenario

1. Check the word KAFAR and PROMULGOWAĆ in Google and in Multisłownik — what are the differences, information given, which source gives you more information on the lemma in the first hit (without further clicking)?
2. What is GEN.PL of MECZ or DAT.SG of MUCHA? (results from the grammatical dictionary)
3. What are the possible lemmata for the word form "danie" (the grammatical dictionary)
4. Which animals groups are called STADO? (the National Corpus of Polish)
5. Who is KALETNIK (plWordNet)
6. What are the other words derived from SEKRET (plWordNet)
7. What are the antonyms of the word SEKRET? (the dictionary of antonyms)
8. Is the form "Dania" in "Dania jest piękna" and "Dania hiszpańskie są smaczne" pronounced in the same way? (Wikisłownik)
9. What is the difference in meaning of NYGUS in general Polish and in the city slang? (plWordNet, slang dictionary)
10. Is the word form ŁABĄDŹ always incorrect? (dictionary of surnames and 16–17 century dictionary)
11. What is the origin of the words KSIĘŻYC and ŁABĘDŹ? (Wikisłownik)
12. Is there a place (city, town, village) called "Łabędź" in Poland? (dictionary of surnames)
13. What does the word TRZECIOTEŚCIK mean? (language observatory)
14. What are the synonyms for the DOM? (plWordNet)
15. Which case is "tysiącpięćsetletniemu"? (grammatical dictionary)

What's next?

- more data to improve coverage (historical corpora, modern slang, ...)
- more sophisticated methods of lexical search (non-lemmatized entries, phraseology etc.)
- lexically motivated applications such as "cultural traces" of a given word or phrase — tracking its references to important artwork (e.g. its presence novel and movie titles, lyrics of popular song or famous quotes)

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.
http://clarin-pl.eu/en/