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

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

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
- Core resources
- Motivation
- How to use
- Usage
- Integration
- Presence of the entry in other linked sources
- What’s next?

Usage
- another important application of the Multisłownik is a tool for human users and NLP processing
- a common point of reference serving as the core of the integration — Słowosieć
- how to use:
  - to create a personal collection of lexical data
  - to find interesting definitions of words
  - to create a personal collection of lexical data

Integration
- a common point of reference serving as the core of the integration — Słowosieć
- how to use:
  - to create a personal collection of lexical data
  - to find interesting definitions of words
  - to create a personal collection of lexical data

Presence of the entry in other linked sources
- a common point of reference serving as the core of the integration — Słowosieć
- how to use:
  - to create a personal collection of lexical data
  - to find interesting definitions of words
  - to create a personal collection of lexical data

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 in film or music titles, etc.)