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

Maciej Ogrodniczek
Institute of Computer Science
Polish Academy of Sciences
maciej.ogrodniczek@ipipan.waw.pl

Zbigniew Bronk
Institute of Computer Science
Polish Academy of Sciences
zbigniew.bronk@ai.pl

Joanna Bilińska
University of Warsaw
j.bilinska@uw.edu.pl

Witold Kieraś
Institute of Computer Science
Polish Academy of Sciences
wkieras@ipipan.waw.pl

Abstract
Multisłownik is an automated integrator of Polish lexical data retrieved from multiple available online sources intended to be used in various scenarios requiring access to such data, most prominently dictionary creation, linguistic studies and education. In contrast to many available internet dictionaries Multisłownik is WordNet-centric, capturing the core definitions from Słownictwo, the Polish WordNet, and linking external resources to particular synsets. The paper provides details of construction of the resource, discussed the difficulties related to linking different logical structures of underlying data and investigates two sample scenarios for using the resulting platform.

1 Introduction
Multisłownik (Pol. multidictionary) is a linguistic integration platform for Polish lexical data retrieved from multiple available online sources intended to be used in various research and educational scenarios. The difficulty of such setting is clear: lexical data is created for different purposes resulting in various underlying structures and representation formats, tailored to specific requirements of each subfield of linguistics. For instance, morphological dictionaries may not differentiate word senses when inflectional patterns of each sense is the same; in turn, when they are different, senses can be assigned properly but at the same time usage examples from corpora restricted to a given sense may be difficult to retrieve.

The paper presents an attempt of creating such linked resource for Polish using computational methods. Section 2 presents similar attempts for other languages, Section 3 describes the data sources used, Section 4 documents the decisions made during the process of data linking, Section 5 provides two sample scenarios based on the integrated data and Section 6 summarizes the paper and presents the work in progress.

2 Related Work
In contemporary lexicography there can be seen a tendency to integrate dictionaries into portals\footnote{See eg: https://en.oxforddictionaries.com/; http://www.termania.net/; http://dictionaryportal.eu/;} mainly provided as a source of information for the ordinary users rather than linguists and researchers. Usually the idea of such portals is to give maximum data big publicity as possible with a minimal effort.

As compared to FRAN, a Slovenian dictionary of a similar type\footnote{See http://fran.si.}, gathering in-house lexical resources available to the Fran Ramovič Institute of the Slovenian Language ZRC SAZU, the initial assumption was that external resources will be used as well. The reason for such a decision was a desire to present Polish vocabulary in an extensive way which seemed to be impossible while using only open resources or those published by a single unit. Unlike in Slovenia, the main Polish linguistic sources were prepared by various publishing houses and research centres. However, because of the authors' rights, not all the dictionaries could be used in the same way. Therefore some of the dictionary data is only presented as references and information whether the searched word can be found in a given dictionary. By default FRAN presents results 'dictionary by dictionary'.
ordering them from the general one (with definitions) through etymological and historical to more specialised ones (e.g. spelling dictionary, medical lexicon or the dictionary of climber’s language). Online dictionary of the PWN publishing house3 offers a similar approach to Polish: entries from dictionaries of several types are presented “as is” on a single Web page together with language use comments, encyclopaedia entries and corpus-based examples. Even less used-friendly Dictionary Portal of such type4 mainly facilitates searches in various dictionaries providing references to source entries.

MultiSłownik combines the concepts of a dictionary portal and a general dictionary trying emulate a traditional dictionary. Therefore the query results are presented in a form of an automatically generated dictionary-like entry.

3 Sources of Lexical Data

MultiSłownik integrates three different kinds of lexical resources:

1. traditional dictionaries created by philologists and meant for human readers only, either web-based or digitalized

2. electronic datasets created by computational linguists for both human users and automatic processing in NLP implementations

3. community-based lexical collections developed online.

The main two sources of lexical entries, forming the core of MultiSłownik, are plWordNet (Piasecki et al., 2009)5 and Grammatical Dictionary of Polish (Saloni et al., 2012; Saloni et al., 2015; Woliński and Kieraś, 2016)6. Several others contributing to its content are: Polish language version of Wikipedia and Wikisource, Walenty valency dictionary (Przepiórkowski et al., 2014) and National Corpus of Polish (Przepiórkowski et al., 2012, NKJP)7. Various other lexical datasets are linked to each entry.

We briefly characterize these sources below showing their lexical potential and pointing out

their most important features hindering integration.

3.1 plWordNet

plWordNet (Piasecki et al., 2009) is a lexicosemantic network reflecting the lexical system of Polish inspired by Princeton WordNet (Miller, 1995)8. It contains sets of synonymous lexical units (synsets) interconnected with lexicosemantic and derivational relations such as synonymy, hypo-hypernymy or mero-holonymy. plWordNet is currently the largest wordnet in the world and contains 178K synsets, 259K word senses and over 600K relations.

Apart from a very rough assignment of part-of-speech category (one of: noun, verb, adjective, adverb) to each lexical unit, plWordNet does not cover any other grammatical information such as grammatical gender for nouns or aspect for verbs. Some of this information may be derived from relations such as verb–noun mpar_vn relation linking verbs and derived gerunds. Currently plWordNet does not cover numerals and uninflected parts of speech.

3.2 SJP.pl

SJP.pl is a Web-based dictionary created by Polish enthusiasts of word games (mainly Scrabble). It aggregates vocabulary from various contemporary printed dictionaries, including spelling and foreign words dictionaries, and classifies them as permitted or non-permitted in word games. Currently it contains ca. 200,000 lexemes. SJP.pl is being developed by the community of its users. As the list of forms noted in SJP.pl is distributed under the terms of open source license it is also used as a data source for spell-checkers. Apart from inflectional forms SJP.pl entries usually also contain short definitions. For MultiSłownik it serves mainly as a supplementary source of lexical and grammatical data, especially when the word searched by the user is not present in SGJP.

3.3 Grammatical Dictionary of Polish

Inflectional information is based on The Grammatical Dictionary of Polish (Pol. Słownik gramatyczny języka polskiego, SGJP) (Saloni et al., 2012; Woliński and Kieraś, 2016). SGJP is the largest existing linguistically elaborated data set of Polish inflectional morphology, from the very

---

3 See http://sjo.pwn.pl.
4 See http://dictionaryportal.eu/.
6 Pol. Słownik gramatyczny języka polskiego, SGJP, see http://sgjp.pl.
7 Pol. Narodowy Korpus Języka Polskiego, see http://nkjp.pl.
beginning developed as an electronic dictionary, now in its third edition turned into Web-based linguistic resource. SGIP serves as a main source of grammatical information for widely used morphological analyzer Morfeusz (Woliński, 2006; Woliński, 2014), as well as for the new general dictionary known as The Great Dictionary of Polish (Pol. Wielki słownik języka polskiego), currently under development (Zmirodzki, 2007).

The integration of morphological data with WordNet senses is hindered by high inflectional variation of Polish lexemes.

3.4 National Corpus of Polish

The National Corpus of Polish (Przępioñowski et al., 2012) is the most prominent corpus of general Polish, providing a balanced representation of contemporary Polish. For Multisłownik it offers real usage examples. To ensure that they represent extensive variety of possible usage of the word it looks for corpus examples for all the possible non-synecratic forms from the inflectional paradigm of the word. For each such form a corpus frequency is also provided.

The corpus data is limited only to NKJP as the largest and most representative corpus of Polish available. Still, closing the dataset in in 2010 makes it less and less up to date each year. As a consequence, NKJP does not reflect the newest Polish vocabulary such as the word prekariat ‘precarit’ which appears in 1-billion-word data set only twice while its actual frequency in daily and weekly newspapers is much higher in the recent years.

3.5 Wiktionary and Wikipedia

Wiktionary9 and Wikipedia10 are open-source, multilingual, community-developed dictionary and encyclopaedia fully available to download in XML format. For Multisłownik they are used as additional sources of lexemes, inflection forms, definitions, examples, collocations, information on pronunciation and etymology.

3.6 Other Linked Sources

Multisłownik also provides information about the presence of a search word in various other lexical resources unable to integrate directly due to licence or format constraints. The list of such resources is extremely heterogeneous. It contains both specialized linguistic dictionaries, both digitalized versions or paper dictionaries and Web-based developments as well as community-based lexical databases. The list of linked sources varies from well known general dictionaries such as PWN dictionaries (Słownik Języka Polskiego PWN, Słownik Wyrazów Obcych PWN, Doroszewski’s classical dictionary, available as scanned pages11, through the electronic Dictionary of 17th & 18th Century Polish (Instytut Języka Polskiego PAN, 2010) to various resources capturing the newest vocabulary, both academia-based (such as the entries from the Language Observatory of the University of Warsaw12) and community-based, e.g. urban slang dictionaries13. Other sources include the Great Dictionary of Polish (Zmirodzki, 2007), dictionaries of Polish personal and place names14 and dictionaries of synonyms, antonyms and crossword definitions15.

Their integration was motivated by practical reasons put forward by lexicographers: it saves user’s time and effort used for searching the word in all these sources separately.

4 Integration

Integration of multiple dictionary resources, heterogeneous by nature, poses various problems due to diverse representation and scope of lexical properties, different levels of detail and incompleteness of coverage of lexical entries. For online resources this situation gets additionally hindered by their constant change: new entries are added to lexicons, models are getting restructured and new data sources appear regularly. Based on all these assumptions we believe that the close integration of resources in such setting (such as combining them into a common LMF16 resource) is a myth — the complexity of such resource would need to exceed the complexity of its parts, already very high for most of the resources. Our approach is differ-

---

9 See https://pl.wiktionary.org/wiki.
11 See http://doroszewski.pwn.pl/.
12 See http://nowowyzrys.uw.edu.pl/.
13 See e.g. Słownik miejski, http://www.miejski.pl/.
ent and assumes interfacing related sources rather than absorbing them into a single common 'super-
resource'.

At the same time a common point of reference is
needed to serve as the core of the integration; for
Multisłownik we decided it to be Słownik, the
Polish WordNet (Piasecki et al., 2009), further re-
ferred to as plWordNet, the most extensive freely
available semantic resource offering lexeme–to-
sense mapping. plWordNet contains extensive de-
scription of lexical–semantic relations for Polish
with interlinked synsets and short definitions, cur-
rently featuring over 300K lexical relations, 320K
synsets and 1.2M inter-synset relations. In Mul-
tisłownik it serves as the main source of lexemes
and semantic information.

Since plWordNet and SGIP make the most
prominent resources covering respectively seman-
tic and grammatical layers, comparison of these
resources was of vital importance. As for the
data set, SGIP contains 150K entries which do
not have their counterparts in plWordNet (not tak-
ing into account negated adjectives, representing
in SGIP as separate entries). On the other
hand, plWordNet contains 20K entries absent from
SGIP. plWordNet contains many multiword lex-
ical units (over 30% of the total number) while
SGIP does not cover any multiword entries apart
from hyphenated entries such as vis-a-vis or ping-
pong and a small sample of words functioning to-
day only as parts of fixed phraseological expres-
sions. Homonymy is the main problem of link-
ing plWordNet data to SGIP; the set of homonyms
contains 3450 nouns, 926 adjectives and 586
verbs.

The integration process starts with plWordNet
taking over its semantic domains, lexical relation
and synset relation types. SGIP is the main source
of grammatical data and other resources are used to
populate the entry.

Figure 1 presents a simple Web application in-
terfacing Multisłownik platform. Sections provide
information about pronunciation and etymology of
the entry, its plWordNet senses with SGIP in-
fection variants assigned properly, related words
retrieved from Wikidictionary, concordance from
NKJP and information on presence of the lexeme
in available online sources.

Information on pronunciation is presented in
two formats: IPA and AS. For each sense its do-
main, definition, example and selected semantic
relations as well as English translation are pre-
presented. Grammatical information covers gram-
matical class, selective categories and inflec-
tion pattern symbol. Inflection section presents se-
lected inflectional forms:

- for nouns – singular genitive and locative and
  plural nominative and genitive
- for adjectives – singular nominative feminine
  and neutral and plural nominative masculine
- for verbs – selected personal forms.

Syntax information is presented according to
Walenty model and annotation. Frequency data
and NKJP-based quotations are currently dynam-
ically retrieved using PELCRA search engine.

5 Possible Usage Scenarios

The aggregation platform is intended to reflect a
standard dictionary, therefore the results are pre-
sented in a form similar to a dictionary entry and
reflect its microstructure. Each entry provides a
number of slots for information: headword, pro-
nunciation, etymology, senses/definitions, gram-
mar information (inflectional patterns), transla-
tions into English, derived words and colo-
cates, concordances with quantitative data from
the NKJP. An important part are links to on-
line dictionaries of surnames, geographical names,
antonyms, synonyms, city slang vocabulary and
new vocabulary which makes getting information
about the contents of other sources, popularity or
importance of lemmata very straightforward.

5.1 Lexicographic Scenario

Multisłownik is by its nature a highly heteroge-
neous resource on many levels: it integrates syn-
chronic and diachronic dictionaries, specialist and
general purpose dictionaries, scientific-driven and
crowd sourced lexical databases. Thus it does not
provide a sound lexicographic description but it
can serve as an instant support for a professional
lexicographer working in the field of extending a
specific dictionary or a linguistic text annotation.

Since Polish is a highly inflectional language,
morphological resources are crucial to almost any
natural language processing task. For this reason
grammatical data sets need constant development
especially in reference to new vocabulary. A lex-
icographer working on this task needs to deter-
mine both grammatical features of the lexical en-
try (such as gender for nouns and aspect for verbs)
PLYWAK

Wymowa
Sł. pol. : ąwak, AC: pływak

Pochodzenie
pol. pływać

Znaczenia, charakterystyka gramatyczna

- Ruszczek niespójny,łów odmiany: ELK
  Charakterystyczne formy fikcyjne:
  ip D. : pływak
  Im M. : pływać
  dep : pływać
  zrodz. (z tablicy odmiany): SGIR

pływać, - (osoba) "mężczyzna uprawiający pływanie"  

Nikolaj był najszyszym pływakiem w polskiej reprezentacji olimpijskiej.
Synonimy:

- wędkować

Hiponimy:
  nurka, krawia, żabka, grzebić, zimionka, zimionka

Tłum. ang.:
  swimmer

Agencjusub.:
  płynąć

Znaczenie:
  płynaczkę

Derekwencja:
  płynowatka

Synonimia międzyparadynmatyczna dla relosyjnych: złotwiak

- Ruszczek spójny,łów odmiany: ELK
  Charakterystyczne formy fikcyjne:
  ip D. : pływak
  Im M. : pływać
  dep : pływać
  dep : pływać
  zrodz. (z tablicy odmiany): SGIR

pływać, - (osobn.) "bójka"

pływać, - (osobn.) "sępik wędrowca"

pływać, - (osobn.) "załadowany zbiornik, wypłynięty powłokiem lub gazem, służący do otrzymywania samolotu na powierzchni wody"

pływać, - (osobn.) "wznosząca się wody rzecek i rzek, utrzymująca się w pozycji poziomej i na zamieszczonej głębokości"

pływać, - (osobn.) "rozうことę przerzucono służącego do oznaczania poziomu wody w lekkim zbiorniku"

- Ruszczek spójny,łów odmiany: ELK
  Charakterystyczne formy fikcyjne:
  ip D. : pływak
  Im M. : pływać
  dep : pływać
  dep : pływać
  zrodz. (z tablicy odmiany): SGIR

pływać, - (wzięte) "woda płynąca z rzeźby wodny"  
  czas. pływać, płynąć

przyn.: płynak, płynowaty

Wynpy pokrewnie

- rzecz: płynina (n), płynówka (f), płynawka (f)
- czas. pływać, płynąć
- przyn. płynak, płynowaty

Zawarte wyrazowe:
  płynak, płynowata, płynak i społeczność, zwartac pływać, zwolnienie

Freqwencja form i cytaty z NIKP (źródłowego)

pływać: 207
  ... rok, czas po Poddębickiem, wypływał się świtem. Zauważyłem się jak płynąć w morze nadziei. Nie odczuwałem, że na dźwięk słowa "odwóć" mogę...
  ... Geralt wyrwał się, zespód w przód negocjując śmieci pięści jak płynąć wodę, ruchom w całej sile, z góru, z mocą napadł na ostro...
  ... niezmiernego, „z przednimi” powabliwymi rynią - płynąć od wody rzeki. Gdy już podetrzymał się, a dnia wielu czasie...
  ... teraź. Będu, dokładnie zaczęły się czternaście jeszcze przez czas płynąć podetrzymać w gotyce zastępowane. Roziała Michalakowska...
  ... jej kinu. Kobieta pochłonęła na dół, walczy z wronami, płynie płynąć z ruchomych balonów. Bogaty popyt pominał kolej i starsza...
  ... wywołał równoczesnie gwałtowne naczyń naczyń i najemniki, jak płynąć, ale w tym nie poprawiło to jego rozpoznawane sytuacje...

Występowanie w słownikach i encyklopediach

<table>
<thead>
<tr>
<th>źródło</th>
<th>Zawartość</th>
<th>Linki</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGIR</td>
<td>gramatyka, leksyka</td>
<td>płynak</td>
</tr>
<tr>
<td>Wikipédia</td>
<td>encyklopedia</td>
<td>płynak</td>
</tr>
<tr>
<td>WBWP</td>
<td>słownik, odnose, wymogów</td>
<td>płynak</td>
</tr>
<tr>
<td>SL</td>
<td>słownik, odnose i ich wstępowanie w Polski</td>
<td>płynak</td>
</tr>
<tr>
<td>SL</td>
<td>słownik, odnose, etymologii</td>
<td>płynak</td>
</tr>
<tr>
<td>Obserwatorium UW</td>
<td>słowo w słownikach polskich</td>
<td>płynak</td>
</tr>
<tr>
<td>Słowniki.NET</td>
<td>słownik, synonimy</td>
<td>płynak</td>
</tr>
<tr>
<td>Krzyżkowska.NET</td>
<td>słownik, definicja</td>
<td>płynak</td>
</tr>
</tbody>
</table>

Figure 1: Test front-end of Multisłownik
and some specific word endings. Consider for example a noun PARKOUR 'a training discipline', which does not appear in the Grammatical Dictionary of Polish. Since she is dealing with an obvious loanword the lexicographer needs to determine, whether the noun declines or it has all its forms homonymous. If it declines, some alternative word endings need to be determined, such as -u or -a in genitive singular (both are possible). Also a grammatical gender needs to be assigned (could be either neuter or masculine inanimate). Since the word refers to a rather niche sport activity, a regular lexicographer cannot rely on her own experience and needs to consult some external lexical resources. By simply typing the word parkour in Multisłownik's search bar the lexicographer gains access to

1. basic definition (provided by plWordnet)
2. characteristic inflectional forms and hypothetical gender value (provided by Multisłownik's own heuristic algorithms)
3. usage examples for four different inflectional forms including their frequencies (found in the National Corpus of Polish).

Based on these informations a proper grammatical description of the word can be formulated and included in the dictionary.

On the other hand a human annotator conducting a morphological, syntactic or semantic text annotation needs a constant access to large lexical data sets supporting her work. Text samples often do not provide a sufficiently large context to determine the proper meaning of a text token or the annotator simply does not have enough specialist knowledge to determine i.e. a lemma of a word. Consider a locative phrase w Sycowie ("in Syców/Sycowo") in which a proper name can be lemmatized either as SYCÓW or SYCOWO. Both endings (-ów and -owo) are correct and both are very common in Polish names of settlements, both form a locative case form ending with -owie but only one of the resulting base forms actually exists and refers to a small town Syców in southwestern Poland. The proper lemma can be easily determined in Multisłownik in which a proper names' declension dictionary is integrated.

5.2 Educational Scenario

Although the platform is aimed at the linguistically- and lexicographically-aware user, it can also be an attractive source of information for wider audience, for instance high school pupils. Searching for random words can be a good start point to teach the students what is the dictionary microstructure and how it can differ between dictionaries. After this stage we plan to present the dictionary by looking up the words. We would suggest following queries for teaching purposes, aiming to present the platform to the young people:

1. Check the word KAFAR and PROMULGORWAĆ 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 NGUS in general Polish and in the city slang? (plWordNet, slang dictionary)
10. Is the word form ŁABĘDZ always incorrect? (dictionary of surnames and 16–17 century dictionary)
11. What is the origin of the words KSIĘŻYC and ŁABĘDZ? (Wikisłownik)
12. Is there a place (city, town, village) called "Łabędź" in Poland? (dictionary of surnames)
13. What does the word TRZECIOTRZÉŚCIK mean? (language observatory)
14. What are the synonyms for the DOM? (plWordNet)
15. Which case is "tysiącpięćsetletniemu"? (grammatical dictionary)
The classes on using the dictionary portal would be even more attractive to students when cross-words or other word games (e.g. Scrabble) are used as search targets. One of such activities could be deciphering a coded information with the usage of Multisłownik conducted in a following way:

- Formulating a question that needs to be answered.
- Providing the coded answer with some or all characters replaced with numbers connected to the questions that lead to decoding the secret characters.
- Possible types of questions:
  - “The last letter of the synonym of the word SEKRET that ends with letter T”.
  - “What is the origin of the word KUŚNIERZ? The first letter of the original language name is the secret character number X”.
  - “Is there a surname Łabądź in Polish? If yes, the secret letter is N, if no, the secret letter is C”.

6 Conclusions and Further Steps

Multisłownik already proved useful in many scenarios related to combining lexical information by offering a simple yet practical method of referring to multiple sources at the same time.

The most obvious further direction for extension of Multisłownik is adding more data; it occurs that even resources less relevant to the current task, e.g. numerous historical corpora can help lexicographers retrieve usage examples from historical texts to trace back the change of word meanings.

Another type of interesting functionality of Multisłownik would be searching for so called “cultural traces” of a given word. Apart from offering the user extensive dictionary-based grammatical and semantic information also references of a given word or phrase to important artwork (e.g. its presence novel and movie titles, lyrics of popular song or famous quotes) could be tracked. This would require building much larger datasets based on library catalogues, movie databases and Wikiquoute, integrated and sorted according to its impact on both high and popular culture.

Acknowledgments

The work reported here was carried out within the research project financed by the Polish National Science Centre (contract number 2014/15/B/HS2/00182) and was partially financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References


Marcin Woliński and Witold Kieraś. 2016. The online version of Grammatical Dictionary of Polish.
