

DEMO: Access to a valence dictionary of Polish *Walenty* via Internet browser

Bartłomiej Nitoń, Tomasz Bartosiak, Elżbieta Hajnicz,
Agnieszka Patejuk, Adam Przepiórkowski, Marcin Woliński

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland

{bartek.niton, tomasz.bartosiak}@gmail.com
{hajnicz, aep, adamp, wolinski}@ipipan.waw.pl

Abstract

The structure of two main layers, syntactic and semantic, of *Walenty*, a comprehensive valence dictionary of Polish, together with their interdependencies, is discussed. The way of accessing lexicon entries by means of a dedicated tool called *Slowal* and how it can be filtered is presented.

1. Introduction

Walenty, a comprehensive valence dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences, is created to a large degree as a part of CLARIN-PL (Przepiórkowski et al., 2014a; Przepiórkowski et al., 2014b).

The dictionary is meant to be both human- and machine-readable — its entries have strictly defined formal structure. All syntactic and semantic phenomena being represented should be attested in linguistic reality. *National Corpus of Polish* (NKJP) (Przepiórkowski et al., 2012) serves as a primary reference, all other sources, including Internet and linguistic literature, are secondary. The structure of the dictionary should enable its various applications in a flexible way. This concerns several formats (text, XML, PDF) the dictionary is exported to as well as the possibility of constructing its sub-dictionaries (e.g., phraseological). In order to meet the assumptions listed above, *Walenty* is stored as a database (with a fairly complicated internal structure) and accessed by a dedicated tool called *Slowal*. It is aimed to insert, correct, process and search lexical data.

Walenty contains valence information for verbs and, to much less extent, for nouns, adjectives and adverbs. It consists of two layers, syntactic and semantic, which are directly connected.

Walenty is the first Polish valence dictionary elaborated according to such rules. Other important Polish valence dictionaries are (Polański, 1980–1992;

Świdziński, 1994). A corpus-based dictionary including some valence information is (Bańko, 2000).

The dictionary is developed remotely via an Internet browser. Browsing the dictionary is available for anyone by accessing the page <http://walenty.ipipan.waw.pl/>. In our demo, we will show the structure of the dictionary and how it is accessible.

2. Basic dictionary structure

Each lexical entry is identified by its lemma (e.g., *AFIRMACJA* ‘affirmation’, *BAĆ* ‘fear’, *POWIEDZIEĆ* ‘say’). It does not contain the reflexive mark *SIĘ* even if it is obligatory (*BAĆ SIĘ*).

2.1. Syntactic layer

Information about reflexivity, aspect, predicativity¹ and negation divide entries into subentries. Any subentry consists of a number of syntactic valence schemata and each schema is a set of syntactic positions. If two morphosyntactically different phrases may occur coordinated in an argument position, they are taken to be different realisations of the same argument. Therefore, a syntactic position is a set of phrase types. There are two labelled positions, subject and object. Usual phrase types are considered. Phrase types can be further parameterised by corresponding grammatical categories, e.g., nominal phrases *np* and adjectival phrases *adjp* are parameterised by information concerning case. Note that the underscore symbol ‘_’ denotes any value of a grammatical category, e.g., *infp*(_) denotes infinitival phrase of any aspect.

A phenomenon connected with whole positions is control and raising, implementing difference between

¹Only for adjectives and adverbs, empty for verbs and nouns.

Work financed by the Polish Ministry of Science and Higher Education, a program in support of scientific units involved in the development of a European research infrastructure for the humanities and social sciences in the scope of the consortia CLARIN ERIC and ESS-ERIC, 2015–2016.

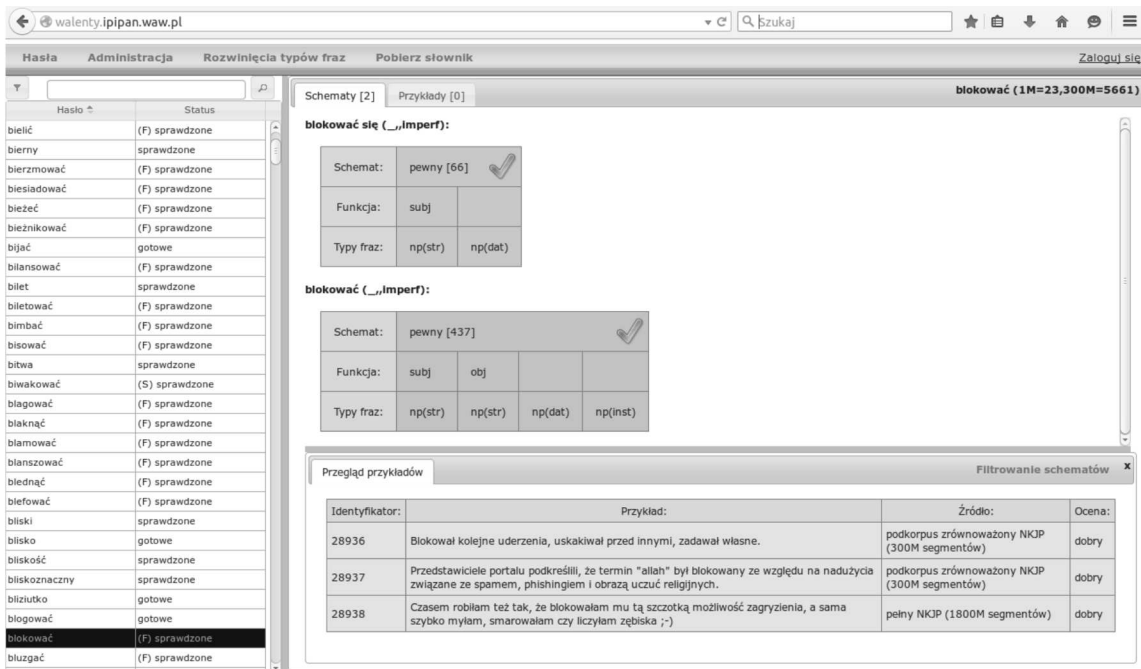


Figure 1: *Słowa* screenshot with entry BLOKOWAĆ

KAZAĆ ‘order’ and OBIECAĆ ‘promise’. The corresponding positions are labelled with controller and controllee. In Polish, this distinction does not only matter for semantic interpretation, but is also correlated with certain agreement facts, i.e., it is useful even for purely syntactic parsers. A related phenomenon is raised subject E for verbs such as ZACZAĆ ‘start’, which inherits subject structure from its infinitival complement.

Each schema has an assessment attached, indicating its correctness (e.g., wątpliwy ‘disputable’) and register (e.g., potoczny ‘colloquial’).

2.2. Semantic layer

Semantic layer consists of semantic frames represented by lists of semantic arguments. Each argument is a pair ⟨semantic role, selectional preferences⟩. Each semantic frame is connected with a list of adequate PIWordNet (Piasecki et al., 2009) lexical units. Selectional preferences are represented by means of PIWordNet synsets and relations.

Semantic roles are two-level entities. Main roles are divided into two groups: basic roles (*Initiator, Stimulus, Factor, Theme, Experiencer, Instrument, Result, Recipient*) and auxiliary roles (*Condition, Manner, Attribute, Measure, Location, Path, Time, Duration, Purpose*). They can be equipped with attributes, forming two pairs *Foreground, Background* and *Source, Goal*, which are used when the same main role appears in a frame twice.

Each main role is marked by a unique colour, which is used in visualisation of the semantic layer.

3. View of entries

The main tabs **Hasła** ‘Entries’, **Administracja** ‘Administration’, **Rozwinięcia typów fraz** ‘Phrase types realisations’ and **Pobierz słownik** ‘Download dictionary’ are placed on the top bar of the page <http://walenty.ipipan.waw.pl/>. The tab **Administracja** contains some statistics about the dictionary.

The tab **Hasła** providing access to the actual dictionary have three subtabs, **Schematy** ‘Schemata’, **Semantyka** ‘Semantics’ and **Przykłady** ‘Examples’ with sentences not assigned to any schema. The view of the first tab with two subentries of entry BLOKOWAĆ ‘block’ is presented in Figure 1. The list of entries together with their current status² appears on the left, the list of schemata grouped into subentries appears on the right. Note the frequency in manually annotated (1M) balanced (300M) NKJP subcorpora information.

A single schema is presented as a table with columns representing syntactic positions. After clicking on a particular schema, the examples connected with it appear at the bottom.


In order to find a particular entry, one enters its lemma in the search field above the list of entries.

If one chooses the tab **Semantyka**, the main window is divided in two, with semantic frames appearing on the left and syntactic schemata appearing on the right. After clicking a particular frame, all its arguments appear in the corresponding roles colours. This concerns the phrase types of the schemata being the syntactic

²Walenty is still under development. Status shows the stage of work on a particular entry.

realisations of the frame, which shows the interdependencies between the syntactic and semantic layer. The sentences connected to the lexical units represented by the frame are visible at the bottom of the window.

4. Filtering

Slowal allows to use a simple filtering form (cf. Figure 2) for searching particular valence phenomena, available by means of button  positioned just above the list of entries on the left side.

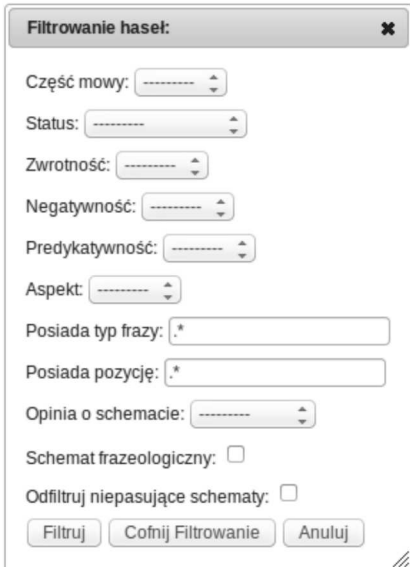


Figure 2: *Slowal* filtering form

Entries can be filtered by their general properties, such as part of speech (pol. *część mowy*), aspect, reflexivity (pol. *zwrotność*), negation (pol. *negatywność*), predicativity (pol. *predykatywność*) (only for adjectives and adverbs) and the schema assessment (pol. *opinia o schemacie*). They are chosen from drop-down lists. However, the main filtering concerns phrase types (pol. *posiada typ frazy*) and whole positions (pol. *posiada pozycję*). The simplest way of filtering is to type the whole phrase type/position in the corresponding field. For instance, inserting `obj{np(inst)}` into position constraints field causes filtering all verbs having nominal phrase in instrumental on its object position.

Constraints on phrase types and position are regular expressions (without bracketing). For instance, inserting `obj{.+}` into position constraints field causes filtering all passivisable verbs. Constraints are inserted in text format.

Disjunctions (introduced by `|`), conjunctions (`&`) and negation (`!`) of constraints are possible. For instance, `subj{.*}&obj{.*}` filters verbs having schemata with both subject and object position.

Clicking the  button causes restricting the

list of entries to the ones satisfying the constraint. Additionally, if the field *Odfiltruj niepasujące schematy* 'Filter out non-matching schemata' is chosen, only schemata matching the constraint appear.

5. Phrase types realisations

Some information that has been intentionally separated from the rest of the dictionary is available under the tab *Rozwinięcia typów fraz* 'Phrase types realisations'. This concerns composed prepositions `comp prepnp` such as *na temat* ('about', literally 'on subject') and semantically-defined phrase types `xp`, including *locative*, *ablative*, *temporal*, *manner*, etc.³ Composed prepositions are represented by mechanisms elaborated for representing phraseology, similarly as possessive phrases `poss p`. This method of representation enables to simplify the structure of the dictionary and ensures its cohesion.

6. References

- Bańko, Mirosław (ed.), 2000. *Inny słownik języka polskiego*. Warsaw, Poland: Wydawnictwo Naukowe PWN.
- Piasecki, Maciej, Stanisław Szpakowicz, and Bartosz Broda, 2009. *A Wordnet from the Ground Up*. Wrocław, Poland: Oficyna Wydawnicza Politechniki Wrocławskiej.
- Polański, Kazimierz (ed.), 1980–1992. *Słownik syntaktyczno-generatywny czasowników polskich*, volume I–V. Wrocław · Warszawa · Kraków · Gdańsk, Poland: Zakład Narodowy imienia Ossolińskich.
- Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.), 2012. *Narodowy Korpus Języka Polskiego*. Warsaw, Poland: Wydawnictwo Naukowe PWN.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, Filip Skwarski, Marcin Woliński, and Marek Świdziński, 2014a. Walenty: Towards a comprehensive valence dictionary of Polish. In *LREC-2010 Proceedings*. Reykjavík, Iceland: ELRA.
- Przepiórkowski, Adam, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński, 2014b. Extended phraseological information in a valence dictionary for NLP applications. In *LG-LP 2014 Proceedings*. Dublin, Ireland.
- Świdziński, Marek, 1994. *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.

³Adverbs are grouped in similar way, e.g., `advp(locat)`; `advp(misc)` represents all adverbs.