

# Accessing and Elaborating *Walenty* —a Valence Dictionary of Polish— via Internet Browser

Bartłomiej Nitoń, Tomasz Bartosiak, Elżbieta Hajnicz

Institute of Computer Science, Polish Academy of Sciences

ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

bartek.niton@gmail.com, tomasz.bartosiak@gmail.com, hajnicz@ipipan.waw.pl

## Abstract

This article presents *Walenty*—a new valence dictionary of Polish predicates, concentrating on its creation process and access via Internet browser. The dictionary contains two layers, syntactic and semantic. The syntactic layer describes syntactic and morphosyntactic constraints predicates put on their dependants. The semantic layer shows how predicates and their arguments are involved in a situation described in an utterance. These two layers are connected, representing how semantic arguments can be realised on the surface. *Walenty* also contains a powerful phraseological (idiomatic) component.

*Walenty* has been created and can be accessed remotely with a dedicated tool called *Slowal*. In this article, we focus on most important functionalities of this system. First, we will depict how to access the dictionary and how built-in filtering system (covering both syntactic and semantic phenomena) works. Later, we will describe the process of creating dictionary by *Slowal* tool that both supports and controls the work of lexicographers.

**Keywords:** valence, valence dictionary, remote access, interface, Polish

## 1. Introduction

*Walenty* (Walenty, 2016) is a comprehensive valence dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS) (Przepiórkowski et al., 2014a; Przepiórkowski et al., 2014b).

The dictionary is meant to be both human- and machine-readable; in particular, it is being employed by two parsers of Polish—*Świga*<sup>1</sup> (Woliński, 2004) and *POLFIE*<sup>2</sup> (Patejuk and Przepiórkowski, 2012). The former, *Świga*, is an implementation of the DCG (Pereira and Warren, 1980) grammar of Polish of Świdziński (1992). The latter, *POLFIE*, is an implementation of an LFG (Bresnan, 1982; Dalrymple, 2001) grammar of Polish. As these parsers are based on two rather different linguistic approaches, the valence dictionary must be sufficiently expressive to accommodate for the needs of both – and perhaps other to come. The dictionary is based on rigorous rules. First, lexicon entries have a strictly defined formal structure. Second, represented syntactic and semantic phenomena should be attested in linguistic reality. Last but not least, the structure of the dictionary should enable flexible access. This concerns several formats (text, XML, PDF) the dictionary can be exported to as well as the possibility of constructing its sub-dictionaries (e.g., phraseological, concerning particular linguistic phenomena).

The assumptions listed above could not be met if *Walenty* had been created in a traditional way. Therefore, it is stored as a database (with a fairly complicated internal structure) and accessed by a dedicated tool called *Slowal*, aimed to insert, correct, process and search lexical data.

The dictionary is developed remotely, by means of an Internet browser. Its entries are elaborated by a highly qualified lexicographers team with precise permissions concerning

the scope of lexical entries edition. On the other hand, browsing the dictionary (without possibility of changing anything) is available for anyone by accessing the page <http://walenty.ipipan.waw.pl/>.

*Walenty* contains valence information for verbs and, to a lesser extent, for nouns, adjectives, and adverbs.<sup>3</sup> It consists of two layers, syntactic and semantic, which are directly connected. Syntactic layer covers a rich phraseological component (Przepiórkowski et al., 2014b), but there is no space in this article to discuss it.

Its advantage over other Polish valence dictionaries is a number of novel features, including the structural case, clausal subjects, distributive PO, complex prepositions, comparative constructions, control and raising and semantically defined phrase types (Przepiórkowski et al., 2014a), and non-standard coordination (Patejuk, 2015). Other important Polish valence dictionaries were created by Polański (1980–1992) and Świdziński (1994). A corpus-based dictionary including some valence information is (Bańko, 2000). Thorough comparison with VALLEX (Žabokrtský and Lopatková, 2007) was conducted by Przepiórkowski et al. (2016). Comparison of the semantic layer of *Walenty* with other formalisms can be found in (Hajnicz et al., 2016).

Furthermore, the formalism used in *Walenty* and *Slowal* tool can be easily adapted for other morphosyntactically rich languages with free word order (e.g. most Slavic languages).

## 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 (e.g., BAĆ SIĘ). If a verb has forms both with reflexive

<sup>1</sup><http://zil.ipipan.waw.pl/Świga>

<sup>2</sup><http://zil.ipipan.waw.pl/LFG>

<sup>3</sup>ca. 12 000 verbs, 2 000 nouns, 1 000 adjectives, and 200 adverbs; ca. 84 000 schemata

mark SIĘ and without it, they will be noted under a single entry (e.g., MARTWIĆ, MARTWIĆ SIĘ)


## 2.1. Syntactic layer

Information about reflexive mark, aspect, predicativity<sup>4</sup> and negation divide entries into subentries. For instance, the entry for BAĆ has one subentry **bać się** (┌, , imperf).

Any subentry consists of a number of syntactic valence schemata and each schema is a list 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.

For example, schema (1) for the verb BAĆ SIĘ ‘fear’ has an argument position with realisations of multiple types. In (2) there is a coordinated phrase consisting of an np (*bezrobocia* ‘unemployment’) and a that-clause (*że zabraknie Ci środków na utrzymanie* ‘that you will have not enough money for livelihood’, represented as cp (że)).

(1)

Schemat:	pewny [1178] 	
Funkcja:	subj	
Typy fraz:	np(str)	np(gen)
		cp(że)
		ncp(gen,int)
		ncp(gen,że)
		ncp(gen,żeby)


(2) **Boisz się bezrobocia i że zabraknie Ci środków na utrzymanie?**

fear.SEC.SG self unemployment.ACC.SG and you.DAT miss.TER.SG resources.GEN.PL for livelihood.ACC.SG ‘[You] are afraid of unemployment and that you will have not enough money for livelihood?’

There are two labelled argument positions – subject and object. Usual phrase types are considered, such as nominal phrases (np), prepositional phrases (prepnp), adjectival phrases (adjp), clausal phrases (cp), etc. Phrase types can be further parameterised by corresponding grammatical categories, e.g., np and adjp – 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 (Rosenbaum, 1967; Landau, 2013), implementing the difference between 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. It can be exemplified in schema (3) for the verb BAĆ SIĘ ‘fear’, cf. (4).

(3)

Schemat:	pewny [8] 	
Funkcja:	subj,controller	controllee
Typy fraz:	np(str)	infp(┌)

(4) **Julek boi się wygłupić.**

Named Entity.SG.NOM fear.TER.SG self make a fool.INF ‘Julek is afraid of making a fool of himself.’

The related phenomenon is raised subject E for verbs such as ZACZAĆ ‘start’, which inherits subject structure from its infinitival complement.

Each schema has its assessment attached, indicating its correctness (*pewny* ‘certain’, *wątpliwy* ‘disputable’, *zły* ‘wrong’) and register (*potoczny* ‘colloquial’, *archaiczny* ‘archaic’, *wulgarny* ‘vulgar’).

## 2.2. Semantic layer

The semantic layer is composed of semantic frames. Each frame is a set of semantic arguments represented as pairs ⟨semantic role, selectional preferences⟩.

Each frame is connected to at least one PLWORDNET (pWordNet, 2016) lexical unit (LU) identifying its meaning. If two LUs correspond to the same frame, they are both ascribed to it. On the other hand, exemplary sentences are linked to LUs appropriate for the meaning of the predicate. More information about the semantic layer is presented in (Hajnicz et al., 2016).

Particular phrase types or a whole syntactic position being a syntactic realisation of an argument are linked to the argument, which is marked by the same colouring. All phrase types in a schema adequate for the particular meaning (frame) are linked, indicating that this schema is a syntactic realisation of that meaning.

As in the syntactic layer, each frame has its assessment attached. The only difference is additional opinion value marking a metaphorical meaning of a frame (*metaforyczna* ‘metaphorical’).

## 3. Accessing *Slowal* as guest

A screenshot from *Slowal*<sup>5</sup> interface for unauthenticated users is presented in Fig. 1. On top, main views are listed. **Hasła** (‘Entries’) view allows viewing and editing entries. Editing is possible only for authenticated users with proper permissions.

**Administracja** (‘Administration’) tab allows to view general statistics of the dictionary.

**Rozwinięcia typów fraz** (‘Phrase type realisations’) is used for special phrase types having several realisations. Only super-lexicographers can add, delete and modify those realisations.

**Pobierz słownik** (‘Download dictionary’) is used for downloading a current version of *Walenty* dictionary in textual format.

### 3.1. Entries view

The **Hasła** tab provides access to the actual dictionary. It has three subtabs designed for presenting syntactic layer,

<sup>4</sup>Only for adjectives and adverbs, empty for verbs and nouns.

<sup>5</sup><http://walenty.ipipan.waw.pl/>

The screenshot shows the Slowal dictionary interface. On the left, a list of entries is displayed, with 'blokować' highlighted. The main area shows the syntactic level of the entry 'blokować'. It is divided into two subentries: 'blokować się („imperf):' and 'blokować („imperf):'. Each subentry shows a schema, function, and phrase types. Below this, a 'Przejrzyj przykłady' window displays a table of example sentences.

Identyfikator:	Przykład:	Źródło:	Ocena:
280205	Blokował kolejne uderzenia, uskakiwał przed innymi, zadawał własne.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
280213	Blokuję go stołkiem w kącie pokoju.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
280214	Po uzyskaniu autoryzacji transakcji odpowiednia kwota jest blokowana na rachunku Klienta. [...]	podkorpus zrównoważony NKJP (300M segmentów)	dobry
280206	Przedstawiciele portalu podkreślili, że termin "allah" był blokowany ze względu na nadużycia związane ze spamem, phishingiem i obrazą uczuć religijnych.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
280215	Zaczęły blokować między nimi linie...	podkorpus zrównoważony NKJP (300M segmentów)	dobry
280208	Czasem robiłam też tak, że blokowałam mu tą szczotką możliwość zagryzienia, a sama szybko myłam, smarowałam czy liczyłam zębiska -)	pełny NKJP (1800M segmentów)	dobry

Figure 1: *Slowal* screenshot with syntactic level of entry BLOKOWAĆ (for a guest user)

semantic layer and non-typical examples. The syntactic layer composed of two subentries of entry BLOKOWAĆ ‘block’ is presented in Fig. 1. The list of entries together with their current status<sup>6</sup> appears on the left, the list of schemata grouped into subentries appears on the right. In order to find a particular entry, one can enter its lemma in the search field above the list of entries.

A single schema is presented as a table with columns representing syntactic positions. After clicking on a particular schema, examples connected with it appear at the bottom. Each exemplary sentence is marked with its source information (particular NKJP subcorpus (NKJP, 2012), cf. (Przepiórkowski et al., 2012), linguistic literature, etc.). In turn, clicking a particular sentence highlights phrase types appearing in it.

Semantic layer is visualised together with syntactic layer – semantic frames on the left and syntactic schemata on the right. After clicking a frame, all schemata linked to that frame become coloured accordingly to the semantic arguments represented by particular positions and phrase types (see Fig. 2).

### 3.2. Filtering

Users browsing a dictionary are often interested in particular valence phenomena rather than concrete entry representation. Therefore, *Slowal* allows to use a filtering form (cf. Fig. 3), available by means of button positioned just above the list of entries.

<sup>6</sup>Walenty is still under development. The status shows the stage of work on a particular entry.

The filter form is divided into three tabs grouping filters by scope:

- **Hasło** (‘Entry’) for filtering entries by their general properties (cf. Fig. 3, left)
- **Schematy** (‘Schemata’) for filtering by specified syntactic schema properties (cf. Fig. 3, middle)
- **Ramy** (‘Frames’) for filtering by specified semantic frame properties (cf. Fig. 3, right)

Clicking the button causes restricting the list of entries to those satisfying all the constraints.

#### 3.2.1. Hasło filter tab

Users can filter entries by the lemma (Lemat), part of speech (Część mowy), having phraseological component (Frazologia) and development status (Status).

For defining lemma constraints, one can use regular expressions (without bracketing). This could be helpful for finding, for instance, entries with specified core (e.g., *\*malować* finds all lemmas ending with *malować*). Disjunctions (introduced by |), conjunctions (introduced by &) and negation (introduced by !) are possible. Disjunction has higher priority than conjunction, thus lemma constraints are in disjunctive normal form (DNF).

Other filter fields values in **Hasło** tab are chosen from drop-down lists. For part of speech (Część mowy) one can choose from adjectives (PRZ), nouns (RZ), adverbs (PS) and verbs (CZ).

Frazologia field can be set to one of two values – zawiera (‘contains’) to get only entries with a phraseological schema and nie zawiera (‘does not contain’) to get entries without any.

Schematy [4] Semantyka [2] Przykłady [1] akompaniować (1M=2,300M=376)

**akompaniować-2**

Rama:	brak [10701]	
Rola:	Theme, Foreground	Theme, Background
Preferencje selekcyjne:	dźwięk-1	ALL

**akompaniować-1**

Rama:	brak [9264]			
Rola:	Initiator	Recipient	Instrument	Theme
Preferencje selekcyjne:	ŁUDZIE	ŁUDZIE	instrument muzyczny-1	muzyka-1
				utwór muzyczny-1

**akompaniować (...Imperf):**

Schemat:	wątpliwy [33]		
Funkcja:	subj		
Typy fraz:	np(str)	np(dat)	np(inst)

Schemat:	pewny [9705]			
Funkcja:	subj			
Typy fraz:	np(str)	np(dat)	prepnp(do,gen)	prepnp(na,loc)

Schemat:	pewny [9707]			
Funkcja:	subj			
Typy fraz:	np(str)	np(dat)	prepnp(na,loc)	prepnp(przy,loc)

Identyfikator:	Przykład:	Źródło:	Ocena:
71531	Jeden z uczniów miał gitarę i akompaniował na niej do pieśni nuconej przez apostołów w drodze na Słoneczną Górę.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
71527	To muzyk – Marzena Mikula-Drabek (autorka opracowania muzycznego spektaklu), która nie tylko akompaniuje bohaterem przy śpiewach (jest tego trochę), lecz bierze udział w akcji.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
71526	Wykonawca może też akompaniować sobie do śpiewu.	podkorpus zrównoważony NKJP (300M segmentów)	dobry
71535	Czy akompaniujecie sobie na niej przy kolędach?	pełny NKJP (1800M segmentów)	dobry

Figure 2: *Slowal* screenshot with semantic level of entry AKOMPANIOWAĆ

**Filtrowanie haseł:**

Hasło Schematy Ramy

Lemat: \*

Część mowy: -----

Frazeologia: -----

Status: -----

Filtruj Cofnij Filtrowanie Anuluj

**Filtrowanie haseł:**

Hasło Schematy Ramy

Opinia o schemacie: -----

Typ schematu: -----

Zwrotność: -----

Negatywność: -----

Predykatywność: -----

Aspekt: -----

Zawiera typ frazy: \*

Zawiera pozycje: \*

Odfiltruj niepasujące schematy:

Filtruj Cofnij Filtrowanie Anuluj

**Filtrowanie haseł:**

Hasło Schematy Ramy

Opinia o ramie: -----

Argumenty semantyczne: Dodaj Lub

--- Rola: Theme Atrybut: Source Usun

Preferencje selekcyjne: Relacja Dodaj

+ Relacja: meronimia Do: Theme Goal Usun

**lub** Usun

--- Rola: Instrument Atrybut: ----- Usun

Preferencje selekcyjne: Słowość Dodaj

+ Predefiniowana: CZYNNOŚĆ Usun

+ Słowość: czyn-1 Usun

! Rola: Recipient Atrybut: ----- Usun

Preferencje selekcyjne: Predefiniowana Dodaj

Filtruj Cofnij Filtrowanie Anuluj

Figure 3: *Slowal* filtering form

Since *Walenty* is still under development, one can choose development status while filtering. As the dictionary is developed as a cascade of layers (syntactic first, phraseological second and semantic last), statuses are divided into 3 groups. They are marked by a letter in brackets for any sta-

tus in group other than syntactic— (F) stands for phraseological status, (S) – for semantic.

At each development phase one can distinguish three types of progress stages: ‘w obróbcie’ (eng. *in progress*) when entry is processed, ‘gotowe’ (eng. *ready*) when entry is

ready but has not been checked by the super-lexicographer and ‘sprawdzone’ (eng. *checked*) when entry is finished and ready for next stage of development.

The special ‘załączkowe’ (eng. *embryo*) status is for entries with low frequency which will not be developed beyond syntactic stage.

### 3.2.2. Schematy filter tab

Entries can be further filtered by basic schema properties like: schema assessment (*Opinia o schemacie*) and type of schema (*Typ schematu*), with two possible values – phraseologic (*frazeologiczny*) and normal (*normalny*).

Other schema based filters are related with subentry characteristics such as reflexivity (*Zwrotność*), negation (*Negatywność*), predicativity (*Predykatywność*) and aspect (*Aspekt*).

The most important schema filtering options concern phrase types (*Posiada typ frazy*) and whole positions (*Posiada pozycję*). The simplest way is to type the whole phrase type or position in the corresponding field. For instance, inserting `obj{np(inst)}` into position constraints field finds all verbs having a nominal phrase in instrumental on its object position

Constraints on phrase types and positions work in a similar way as *Lemat* field in the **Hasło** filter tab. They are regular expressions (without bracketing). For instance, inserting `obj{.+}` into position constraints finds all passivisable verbs, whereas inserting `infp(perf)` into phrase type constraints field finds all entries having infinitival arguments in perfect aspect.

Disjunctions (`|`) and conjunctions (`&`) of constraints are possible. For instance, `subj{.*}&obj{.*}` finds verbs having schemata with both subject and object. Particular phrase types or positions can also be forbidden (with `!`). For instance, `subj{.*}&!obj{.*}` finds verbs having schemata with a subject and without an object. Again, disjunction has higher priority than conjunction.

All filter fields from **Schematy** filter tab are used simultaneously, an entry is filtered out if none of its schemata matches all specified constraints. Additionally, if the field *Odfiltruj niepasujące schematy* (eng. *Filter out non-matching schemata*) is chosen, only matching schemata will appear.

### 3.2.3. Ramy filter tab

The last filter tab serves for filtering entries by semantic phenomena. Its filter fields can be divided into two groups: one representing general features (at this stage limited to the frame assessments (*Opinia o ramie*)), and the other – constrains on semantic arguments.

Semantic arguments filters are defined using disjunctive normal form (DNF). One can add a new argument to the filter by clicking **Dodaj** (‘Add’) button or remove it by pushing **Usuń** (‘Delete’) next to an already defined argument. Arguments are added as conjoined to the last (open) conjunction. Groups of arguments can be separated by ‘or’ (pol. *lub*) relation, which is achieved by clicking **Lub** (‘Lub’) button. Moreover, each argument can be marked as negative for ‘match all, except’ constraint.

Each argument is defined by a semantic role (*Rola*) followed by an attribute (*Atrybut*) and a set of selectional preferences. To add a selectional preference filter, one must select its type on the dropdown list (*Preferencje selekcyjne*) and add it to the argument using **Dodaj** (‘Add’) button. Based on preference type, different fields will be added to the filter form: dropdown list for predefined preferences (*Predefiniowana*), text field with autocorrection for a PLWORDNET synset (*Słowosieć*) and 3 dropdown lists (for selecting relation type and target of this relation) for relation-based selectional preferences (*Relacja*). For more information about selectional preferences and semantic arguments, see (Hajnicz et al., 2016).

### 3.3. Phrase types realisations

Some information that has been intentionally separated from the rest of the dictionary is available under the **Rozwinięcia typów fraz** (eng. *Phrase types realisations*) tab. This concerns composed prepositions *comprenp* such as *na temat* (‘about’, literally ‘on subject’) and semantically-defined argument types *xp*, including *locative*, *ablative*, *temporal*, *manner*, etc.<sup>7</sup> Composed prepositions are represented by mechanisms elaborated for phraseology, similarly as distinguished possessive phrases *possp*. This method allows to simplify the structure of the dictionary and ensures its cohesion.

All realisations of *xp(mod)* (*manner*), being an argument of verbs like *TRAKTOWAĆ* ‘treat’ are presented in Fig. 4.

Rozwinięcia typu frazy xp(mod):	
advp(mod)	pewna
comprenp(na sposób)	pewna
cp(rel(jakby))	pewna
lex(preppnp(w,acc),sg,'sposób',atr(ajp(agr))))	pewna
prepadjp(jako,str)	pewna
prepadjp(jak,str)	pewna
prepadjp(po,postp)	pewna
preppnp(bez,gen)	pewna
preppnp(jako,str)	pewna
preppnp(jak,str)	pewna
preppnp(pod,acc)	pewna
preppnp(z,inst)	pewna
comprenp(na modłę)	archaiczna
cp(rel(jakoby))	archaiczna

Pobierz rozwinięcia typów fraz w postaci pliku tekstowego: **Pobierz**

Figure 4: *Słowa* screenshot with the list of *xp(mod)* realisations

## 4. Dictionary creation procedure

Creating a resource as big as *Walenty* is a serious challenge. It is organised as a chain of performance: syntactic, phraseological and semantic components are elaborated sequentially (by a lexicographer and a supervising lexicographer in each phase). Users have particular access permissions, determining what changes they can make in the dictionary. The super-lexicographers can make particular portions of entries available for particular lexicographers, which en-

<sup>7</sup>Adverbs are grouped in similar way, e.g., *advp(locat)*; *advp(misc)* represents all adverbs.

ables a sort of specialisations (e.g., for verb, noun and adverb subdictionaries). On the other hand, the entries finished in one phase constitute an input for the next one. This way of organisation serves for maximising cohesion of the resource on one hand and improving the parallel work on the other.

#### 4.1. Elaboration of a single entry

Entries are elaborated withing **Hasła** view.

Filtering entries enables (among others) finding entries lexicographer is currently working on or can start working on. *Slowal* interface for the authenticated user (in this case super-lexicographer) can be seen in Fig. 5.

The actual work is performed in the right panel. First, a lexicographer has to assign an (unassigned) entry to himself. **Status** tab allows the lexicographer to assign an entry to himself, to withdraw such assignment and to finish work on an entry. It also shows all users working on that entry during previous stages.

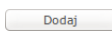
Lexicographers preparing syntactic layer work on the **Schematy** ‘Schemata’ tab, whereas lexicographers preparing semantic layer work on the **Semantyka** ‘Semantics’ tab.

The **Notatki** ‘Notes’ tab displays all notes on an entry, the **Podgląd hasła** ‘Entry preview’ tab serves for viewing other, potentially similar entries, and the **Kontrola zmian** ‘Change control’ tab serves for checking changes made by lexicographers during various phases of work.

##### 4.1.1. Elaboration of the syntax layer

Lexicographers working on syntax layer are supposed to elaborate syntactic schemata appropriate for a particular predicate and illustrate them by corresponding exemplary sentences.

The lexicographer can add, delete, copy and modify elements using buttons available at the bottom of a window or using keyboard shortcuts. They work accordingly to which element is active: the whole schema, a particular position, a particular phrase type or none at all. Elements can be copied from another entry available via **Podgląd hasła** ‘Entry preview’ tab.

Clicking  (‘Add’) button starts the procedure of adding new element. All information concerning any particular element is selected from predefined lists. The only exception are lemmas of the phraseological component.

Double clicking enables edition of a selected element. For instance, for a given position, only `subj` or `obj` marks on one hand and information concerning control on the other hand can be chosen.

Other tabs at the bottom of the window allows editing and viewing exemplary sentences connected to a schema. They are represented as a simple text, with possibility of marking arguments of a corresponding predicate that are present in it.

Some examples can be added from *Składnica* treebank (Woliński et al., 2011) but this concerns only most frequent verbs and their most frequent syntactic constructions. *Składnica* contains information about syntactic dependants which enables semiautomatic assignment of examples to schemata.

##### 4.1.2. Elaboration of the semantic layer

As the syntactic layer is introduced as the first one, the semantic layer is elaborated with regard to it. Elaborating the semantic layer is composed of two tasks. The first one consists of linking sentences illustrating the entry with the PLWORDNET lexical units adequate for the entry (having the same lemma) and adding new units if they are missing. The second task is to create semantic frames.

The semanticist can create a new frame or modify the existing ones. A new frame has to be assigned to at least one lexical unit. The resulting frame is empty. In order to add a new argument to a frame, the semanticist has to choose a whole position or a single phrase type in a schema being its syntactic realisation, and then chose a semantic role from a table (see Fig. 6).

Selectional preferences are added independently afterwards. Both predefined preferences and those defined by means of relations to another argument are chosen from a list. In order to determine selectional preferences by means of a particular PLWORDNET synset, a semanticist has to introduce the lemma of a lexical unit (representing it) manually.

Each argument can be attached to or detached from whole positions or particular phrase types.

#### 4.2. Additional filtering options

For authenticated users, there are some additional filtering options within **Hasło** filter tab. New fields are mainly intended for *Walenty* development organisation. One allows filtering by assigned lexicographers (on various stages of work), another serves for filtering by tranches of entries.<sup>8</sup> Another group of fields serves for filtering by sender of notes added to an entry and by special type of examples, called ‘own’ (pol. *własny*), which is used to mark examples without defined source (usually created by the lexicographer or taken from the Internet).

#### 4.3. Automatic support

Using a dedicated tool for elaborating a valence dictionary enables us to support and control the work of lexicographers.

The simplest part of the control is using predefined lists, which eliminate spelling errors etc. and enforce consistency with the *Walenty* representation language. Next, the consistency of parameters such as aspect of verbs, lemmas of lexical units (with PLWORDNET) and phraseological arguments is checked using the morphological analyser *Morfeusz* (Morfeusz, 2013), cf. (Woliński, 2006; Woliński, 2014). Some other aspects of consistency, such as at most one `subj` and `obj` label in a schema, protecting against assigning the same semantic argument to phrase types from different syntactic positions etc., are controlled.

For exemplary sentences, filling all required fields is checked. Examples with no attested source (from Internet or created by a lexicographer) have to be confirmed by super-lexicographers. Additionally, for non-standard coordination, examples are required.

<sup>8</sup>That was crucial when *Walenty* was expanded by new portions of verbs, adjectives, adverbs or nouns

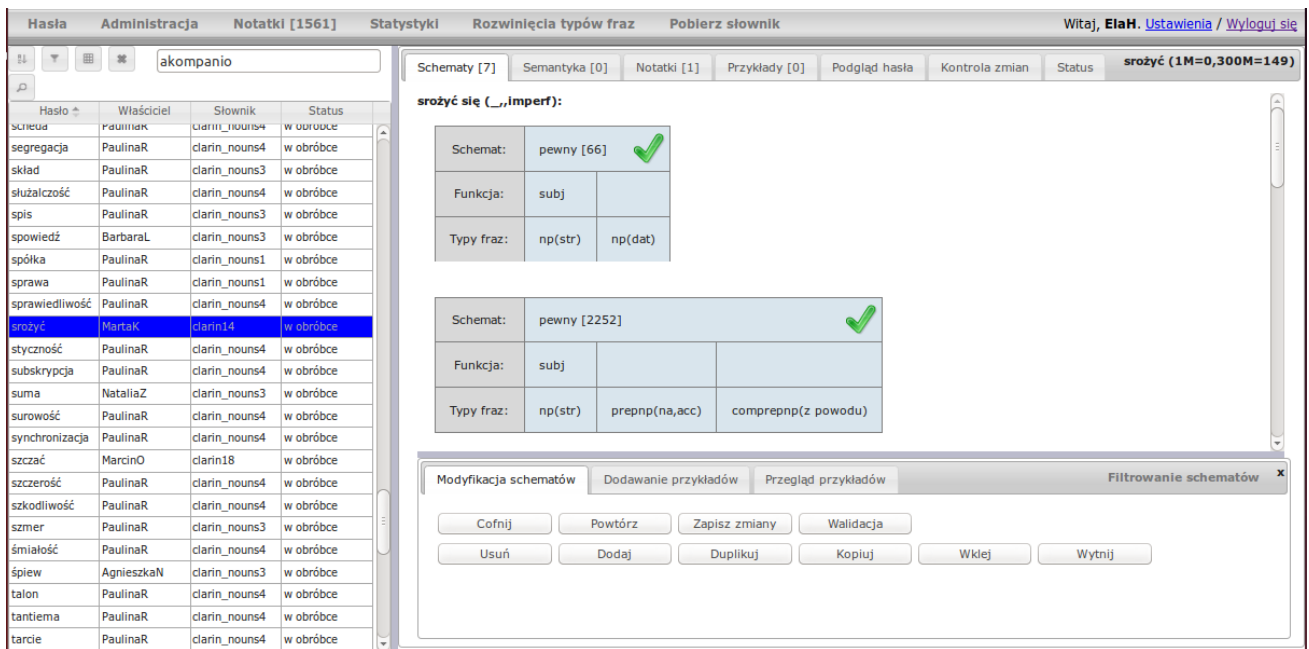


Figure 5: *Slowal* screenshot for the authenticated user (here super-lexicographer)

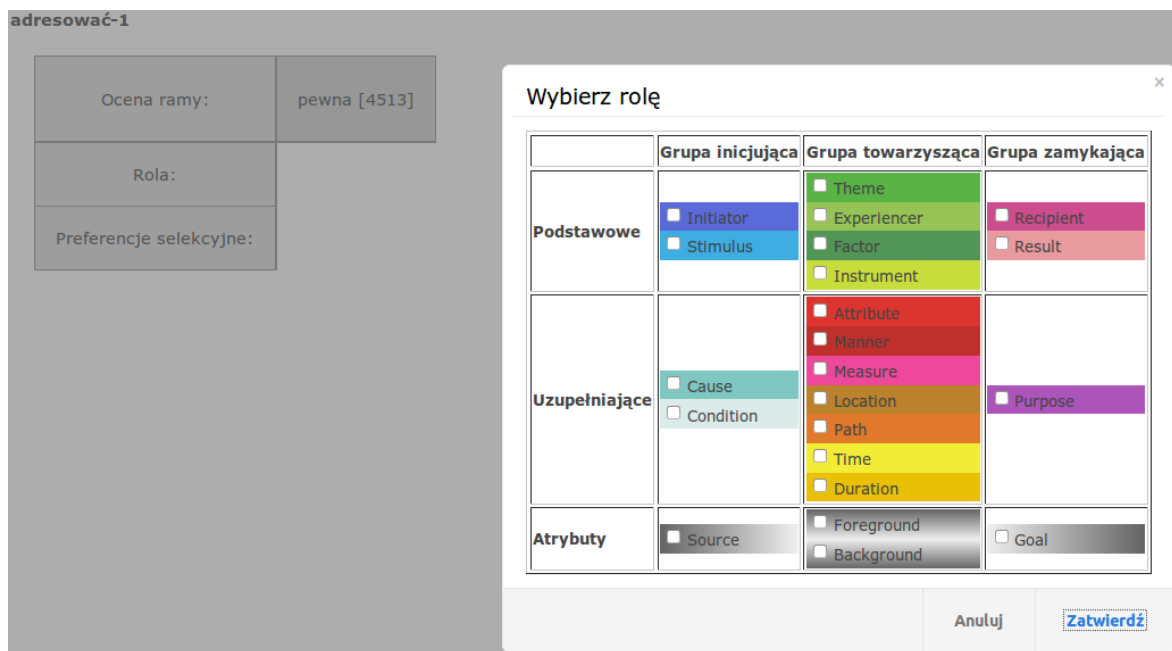


Figure 6: *Slowal* screenshot for adding a new argument

The above rules cannot be violated. There is also some support and control in form of suggestions, based on the dictionary statistics. For instance, this concern frequent coordinations. First, they are suggested while adding a new phrase type to a position. Second, a suggestion of merging two schemata differing only in frequently coordinating phrase type appears during consistency checking.

Consistency checking is performed when a lexicographer finishes his work on an entry (changing the status) or on demand (using  'validation' button).

## 5. Conclusions and future work

This presentation describes a valence dictionary of Polish *Walenty* from the point of view of accessing and elaborating it via Internet browser. The dictionary is still under development, hence the dedicated tool *Slowal* is focused on editing functionalities rather than presentation. It is intended to maximise cohesion, minimise error rate, and to make the lexicographers' work as easy as possible. Therefore, the future work will focus on increasing the readability of presentation in *Slowal*.

In the nearest future we also plan to add phraseology to semantic layer, i.e., extracting multi-word lemmas from lexicalised arguments and linking them with PLWORDNET

multi-word lemmas. The automatic support of this layer based on frame statistics and PLWORDNET is planned as well.

**Acknowledgements** This research was 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.

## 6. Bibliographical References

- Mirosław Bańko, editor. (2000). *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.
- Joan Bresnan, editor. (1982). *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.
- Nicoletta Calzolari, et al., editors. (2014). *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavík, Iceland. ELRA.
- Dalrymple, M. (2001). *Lexical-Functional Grammar*. Academic Press.
- Hajnicz, E., Andrzejczuk, A., and Bartosiak, T. (2016). Semantic layer of the valence dictionary of Polish *Walenty*. Proceedings of the LREC 2016 conference (same proceedings).
- Landau, I. (2013). *Control in Generative Grammar: A Research Companion*. Cambridge University Press, Cambridge.
- Patejuk, A. and Przepiórkowski, A. (2012). Towards an LFG parser for Polish. an exercise in parasitic grammar development. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3849–3852, Istanbul, Turkey. ELRA.
- Patejuk, A. (2015). *Unlike coordination in Polish: an LFG account*. Ph.D. dissertation, Institute of Polish Language, Polish Academy of Sciences, Cracow.
- Pereira, F. C. N. and Warren, D. H. D. (1980). Definite clause grammars for language analysis. *Artificial Intelligence*, 13(3):231–278.
- Kazimierz Polański, editor. (1980–1992). *Słownik syntaktyczno-generatywny czasowników polskich*, volume I–V. Zakład Narodowy imienia Ossolińskich, Wrocław · Warszawa · Kraków · Gdańsk, Poland.
- Adam Przepiórkowski, et al., editors. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., Skwarski, F., Woliński, M., and Świdziński, M. (2014a). *Walenty: Towards a comprehensive valence dictionary of Polish*. In Calzolari et al. (Calzolari et al., 2014), pages 2785–2792.
- Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014b). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland.
- Przepiórkowski, A., Hajič, J., Hajnicz, E., and Urešová, Z.

- (2016). Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, 29. Forthcoming.
- Rosenbaum, P. (1967). *The Grammar of English Predicate Complement Constructions*. The MIT Press, Cambridge, MA.
- Świdziński, M. (1992). *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw, Poland.
- Świdziński, M. (1994). *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.
- Woliński, M., Głowińska, K., and Świdziński, M. (2011). A preliminary version of *Składnica* — a treebank of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference*, pages 299–303, Poznań, Poland. Fundacja Uniwersytetu im. A. Mickiewicza.
- Woliński, M. (2004). *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.
- Woliński, M. (2006). *Morfeusz* — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, et al., editors, *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'06*, Advances in Soft Computing, pages 503–512, Ustroń, Poland. Springer-Verlag.
- Woliński, M. (2014). *Morfeusz reloaded*. In Calzolari et al. (Calzolari et al., 2014), pages 1106–1111.
- Žabokrtský, Z. and Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, 87:41–60.

## 7. Language Resource References

- plWordNet: G4.19 Group at Department of Artificial Intelligence, Wrocław University of Technology. (2016). *Polish wordnet* plWordNet. Department of Artificial Intelligence, Wrocław University of Technology, <http://plwordnet21.clarin-pl.eu/>, ver. 2.1.
- NKJP: ZIL Group at Institute of Computer Science PAS. (2012). *Narodowy Korpus Języka Polskiego*. Institute of Computer Science PAS, <http://nkjp.pl>.
- Morfeusz: ZIL Group at Institute of Computer Science PAS. (2013). *Morfeusz – a morphological analyser of Polish*. Institute of Computer Science PAS, <http://sgjp.pl/morfeusz/>, ver. 2.0.
- Walenty: ZIL Group at Institute of Computer Science PAS. (2016). *Walenty – valence dictionary of Polish*. Institute of Computer Science PAS, <http://walenty.ipipan.waw.pl/>, ver. 0.8.