

Semantic Layer of the Valence Dictionary of Polish *Walenty*

Elżbieta Hajnicz, Anna Andrzejczuk, Tomasz Bartosiak

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

hajnicz@ipipan.waw.pl, anna.andrzejczuk@gmail.com, tomasz.bartosiak@gmail.com

Abstract

This article presents the semantic layer of *Walenty*—a new valence dictionary of Polish predicates, with a number of novel features, as compared to other such dictionaries. The dictionary contains two layers, syntactic and semantic. The syntactic layer describes syntactic and morphosyntactic constraints predicates put on their dependants. In particular, it includes a comprehensive and powerful phraseological component. The semantic layer shows how predicates and their arguments are involved in a described situation in an utterance. These two layers are connected, representing how semantic arguments can be realised on the surface. Each syntactic schema and each semantic frame are illustrated by at least one exemplary sentence attested in linguistic reality.

The semantic layer consists of semantic frames represented as lists of pairs ⟨semantic role, selectional preference⟩ and connected with PLWORDNET lexical units. Semantic roles have a two-level representation (basic roles are provided with an attribute) enabling representation of arguments in a flexible way. Selectional preferences are based on PLWORDNET structure as well.

Keywords: valence, semantic frames, semantic roles, selectional preferences, Polish

1. Introduction

Walenty (Walenty, 2016), a comprehensive valence dictionary of Polish, is developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS) within several projects, but mainly within CLARIN-PL.¹ The dictionary is ment to be both human- and machine-readable; in particular, it is employed by two parsers of Polish, Świgr² (Woliński, 2004) and POLFIE³ (Patejuk and Przepiórkowski, 2012). For this reason, lexicon entries have a strictly defined formal structure. Additionally, syntactic and semantic phenomena represented therein should be attested in linguistic reality.

Walenty is composed of two main layers: syntactic and semantic. The syntactic layer was described in (Przepiórkowski et al., 2014c; Przepiórkowski et al., 2014a), whereas Przepiórkowski et al. (2014b) focuses on its phraseological component. In this paper we want to focus on the semantic layer.

While extending *Walenty* with the semantic layer, we had semantic parsing and textual entailment in mind. Therefore, unified representation of utterances bearing similar information is crucial.

The dictionary comprises over 15 000 entries (composed of about 84 000 schemata), with about 12 000 verbs,⁴ 2 000 nouns, 950 adjectives and 200 adverbs. Each entry is identified by its lemma.

The semantic layer of *Walenty* is strictly connected with PLWORDNET (Piasecki et al., 2009), one of two Polish wordnets. PLWORDNET describes the meaning of a lexical unit by placing this unit in a network of relations (such

as synonymy, hypernymy, meronymy, etc.).

2. Related works

There exist valence dictionaries connecting syntactic and semantic information about predicates and their arguments. The most famous is FrameNet⁵ (Fillmore et al., 2003; Ruppenhofer et al., 2006) based on a theory of meaning called Frame Semantics (Fillmore, 1976; Fillmore and Baker, 2001). It is organised around the notion of a *semantic frame* that is evoked by lexical units representing particular meanings of words (not only verbs). Frames are lists of semantic roles called *frame elements* (FEs). An exemplary frame for transfer situation being evoked by several verbs (*give*, *donate*, *present*, *receive* or even *buy* and *sell*) is presented in (1). Valence of particular verbs is represented by their *valence patterns* ⟨FEs, PTs, GFs⟩ containing corresponding *phrase types* (PTs) and *grammatical functions* (GFs) apart from frame elements. Exemplary valence patterns for verbs GIVE and RECEIVE are presented in (2) and (3), respectively.

(1)	Frame	Transfer			
	FEs	DONOR, THEME, RECIPIENT			
(2)	give	FEs	DONOR	THEME	RECIPIENT
		PTs	NP	NP	NP
		GFs	Ext	Comp	Obj
	give	FEs	DONOR	THEME	RECIPIENT
		PTs	NP	NP	PP-to
		GFs	Ext	Obj	Comp
(3)	receive	FEs	DONOR	THEME	RECIPIENT
		PTs	(PP-from)	NP	NP
		GFs	(Comp)	Obj	Ext

FrameNet contains about 800 hierarchically organised frames evoked by 10 000 lexical units.

¹<http://www.clarin-pl.eu/en/>

²<http://zil.ipipan.waw.pl/Świgr>

³<http://zil.ipipan.waw.pl/LFG>

⁴*Walenty* verbal entries cover 99.8% verb tokens in balanced NKJP subcorpus of 300M tokens, based on frequency list of automatic tagging.

⁵<https://framenet.icsi.berkeley.edu/fndrupal/>

odpovídat^{impf}, **odpovědět**^{pf} v

1 ≈ odvětit; d'ávat odpověď

-frame: ACT₁^{obl} ADDR₃^{obl} PAT_{na+4}^{opt}
EFF_{4,aby,at',zda,že,cont}^{opt} MANN^{typ}

-rfl: cor3, pass.

-rcp: ACT-ADDR

-class: communication

Figure 1: Exemplary entry from VALLEX 2.0 valence dictionary for the verb ODPOVÍDAT ‘answer’

Another important valence dictionary is VerbNet⁶ (Kipper-Schuler, 2005) based on (Levin, 1993) classification of verbs. Each verb class in VerbNet is completely described by semantic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function. VerbNet describes about 5250 senses of 3800 verb lemmas.

There exist several Czech valence dictionaries. First, VerbaLex (Hlaváčková and Horák, 2006), which is connected with the Czech wordnet and contains frames for 6 500 verb synsets (10 500 verb lemmas including orthographic variants). The semantic characteristic of arguments consists of a semantic role (40 elements from EuroWordNet top ontology) and more precise semantic type. Two others, VALLEX (Lopatková et al., 2003; Žabokrtský and Lopatková, 2007) and PDT-VALLEX (Hajič et al., 2003; Urešová, 2009), are based on Functional Generative Description (Sgall et al., 1986). The latter is strictly connected with Prague Dependency Treebank and reflects all linguistic phenomena detected in it (for 7 000 verbal lemmas, 3 700 (predicative) nominal lemmas and 800 (predicative) adjectival lemmas). In particular, this concerns phraseology.⁷ An exemplary frame of VALLEX 2.0 is presented in Fig. 1.

There exist some Polish valence dictionaries as well. The most important are (Polański, 1980 1992; Świdziński, 1994). Only the first one includes semantic information, i.e. abstract selectional restrictions (cf. Fig. 2). A corpus-based dictionary including some valence information is (Bańko, 2000).

LUBIĆ

$$NP_N \rightarrow \left\{ \begin{array}{l} NP_A^1 + \left(\left\{ \begin{array}{l} za \cap NP_A^2 \\ za \cap Ts_A, že \cap S \end{array} \right\} \right) \\ NP_A^3 \\ žeby \cap S \\ IP \end{array} \right\}$$

NP_N → [+Hum]

NP_A¹ → [+Anim] NP_A³ → $\begin{bmatrix} -Abstr \\ -Anim \end{bmatrix}$ [+Abstr]

NP_A² → [+Abstr]

Figure 2: Exemplary entry for the verb LUBIĆ ‘like’ in Polański’s valence dictionary

⁶<https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

⁷The profound comparison of the phraseological component of PDT-VALLEX and *Walenty* was conducted in (Przepiórkowski et al., 2016).

3. Syntactic layer

We start by recalling basic information about the syntactic layer of the dictionary. Each entry is divided into subentries accordingly to their aspect, reflexivity (for verbs), negativity and predicativity (for adjectives and adverbs). Each subentry is a set of syntactic schemata, whereas each schema is a list of syntactic positions. Every syntactic position is a set of phrase types, such as nominal phrases np, prepositional phrases prepnp, clause phrases cp etc. Putting two phrase types into single position means that they can coordinate.

Schemata are illustrated by exemplary sentences. Each sentence is linked to the schema representing its syntactic structure and each phrase type occurring in the sentence is marked in that schema.

Walenty has a rich phraseological component (Przepiórkowski et al., 2014b). It is represented with two special phrase types, *lex* and *fixed*. The latter is used for phrasemes containing forms which cannot be specified with a reference to a lemma and morphosyntactic categories such as case or number, as they violate their usual form. For instance, *wyjsć za mąż* ‘get married (for women)’, contains phraseme *za mąż* composed of the preposition *za* requiring nominal phrase in accusative and the noun *MAŻ* ‘husband’ in nominative.

Standard phraseological constructions are represented using *lex* parameterised with a phrase type of lexicalisation followed by grammatical constraints appropriate for this type (e.g. number, gender), a lemma of the semantic head of the phrase, and information about further admissible modifications. Let us consider the verb *UPIJAĆ SIĘ* ‘get drunk’ and its phraseological usage *upijać się na umór* ‘get drunk into a stupor’. Its lexicalised argument *na umór* forms a prepositional phrase prepnp (na, acc) in singular, its lemma is the noun *UMÓR*,⁸ and no further modifications are allowed (denoted as natr). The construction has a non-lexicalised dependant (subject) as well. The resulting schema is presented in Fig. 3.

upijać się („,imperf):


Schemat:	pewny [25708] 	
Funkcja:	subj	
Typy fraz:	np(str)	lex(xp(mod[prepnp(na,acc)]),sg,'umór',natr)

Figure 3: A phraseological schema of the verb UPIJAĆ SIĘ

Each schema has its assessment attached, indicating its correctness (*pewny* ‘certain’, *wątpliwy* ‘disputable’, *zły* ‘wrong’) and register (*potoczny* ‘colloquial’, *wulgarny* ‘vulgar’, *archaiczny* ‘archaic’). These assessments enable us to consider (and filter if necessary) even completely erroneous constructions, provided that they are frequently used. For instance, for *HARMONIZOWAĆ* ‘harmonise’, *harmonizować się z czymś* ‘harmonise with sth.’ is considered a

⁸In contemporary Polish functioning only in this phraseme.

wrong schema, while the correct construction is *harmonizować z czymś*, without the reflexive mark *SIĘ*, see fig. 4.

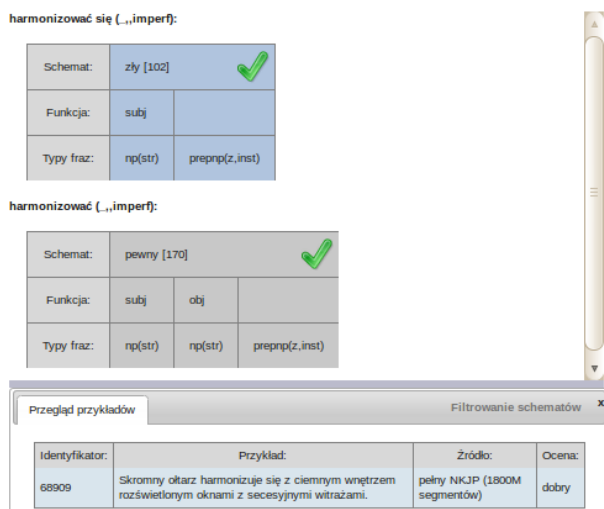


Figure 4: A screenshot with a wrong schema and its correct counterpart

In *Walenty*, we store only the longest schemata, i.e. the longest lists of positions that can appear together in the same utterance and the largest sets of phrase types that can be coordinated in a single position. The reason is that in Polish every position, including subject, can be elided.

4. 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 a meaning of a predicate. Those meanings are identified by PLWORDNET lexical units (LUs). We use version 2.1 (PIWordNet, 2016). It is possible for multiple LUs to correspond to the same frame. In particular, this concerns reflexive and non-reflexive verbs, provided that they represent the same meaning (diathesis alternations, e.g., *ZBIĆ* and *ZBIĆ SIĘ* ‘break’).

On the other hand, some lexical units may be missing in PLWORDNET. In such cases new LUs are added, indicated by capital letters instead of numbers following the lemma of an LU (wordnet standard), in order to differentiate them from the original wordnet LUs. Such new LUs are provided with glosses⁹ as well as synonyms or hypernyms situating them in PLWORDNET structure. This will facilitate including them by PLWORDNET developers.

We assume that there cannot be two identical frames for a single entry, as otherwise there would be no way to distinguish between their meanings.

Exemplary sentences are originally assigned to syntactic schemata, but they are connected to lexical units as well. Thus, they also illustrate semantic frames connected to LUs.

Similarly to schemata, semantic frames are assessed. These assessments concern meaning. Therefore, an archaic or

wrong schema can be connected to a certain frame, if that schema represents an archaic/incorrect use of a syntax for some meaning. On the other hand, an archaic or wrong frame can represent an old or improper meaning of the contemporary, proper syntax. For instance, this concern verbs *ADOPTOWAĆ* ‘adopt’ and *ADAPTOWAĆ* ‘adapt’ often confused with each other.

4.1. Semantic Roles

Semantic layer of *Walenty* is organised around the notion of a semantic role. This notion evolved from the works of (Fillmore, 1968; Fillmore, 1971; Fillmore, 1976). While constructing the set of semantic roles, we took into consideration existing valence dictionaries, including FrameNet, VerbNet and VALLEX. The FrameNet roles are too granulated and their number is too large, whereas the Vallex roles are bound too closely with syntax. Therefore, we mainly modelled on VerbNet. In particular, apart from roles we use selectional preferences, just as in VerbNet.

Basic semantic roles are divided into two groups: main roles (*Initiator*, *Theme*, *Stimulus*, *Experiencer*, *Instrument*, *Factor*, *Recipient*, *Result*) representing situation participants, and auxiliary roles (*Condition*, *Attribute*, *Manner*, *Location*, *Path*, *Time*, *Duration*, *Measure*, *Purpose*) representing its circumstances. This set is supposed to cover both required verb dependants (arguments) and free modifiers (actual adjuncts). In another aspect, roles are divided into three groups: initiating group including roles describing participants that enable a situation to occur, accompanying group composed of roles describing participants that undergo or characterise the situation and ending group including roles describing participants that result from or are influenced by the situation, cf. Fig. 5. Roles have colours assigned to them in a fixed way.

	Initial Group	Accompanying Group	Ending Group
Main Roles	<input type="checkbox"/> Initiator <input type="checkbox"/> Stimulus	<input type="checkbox"/> Theme <input type="checkbox"/> Experiencer <input type="checkbox"/> Factor <input type="checkbox"/> Instrument	<input type="checkbox"/> Recipient <input type="checkbox"/> Result
Auxiliary Roles	<input type="checkbox"/> Condition	<input type="checkbox"/> Attribute <input type="checkbox"/> Manner <input type="checkbox"/> Measure <input type="checkbox"/> Location <input type="checkbox"/> Path <input type="checkbox"/> Time <input type="checkbox"/> Duration	<input type="checkbox"/> Purpose
Attributes	<input type="checkbox"/> Source	<input type="checkbox"/> Foreground <input type="checkbox"/> Background	<input type="checkbox"/> Goal

Figure 5: Table of *Walenty*’s roles

Naturally some collections or roles cooccur. This in-

⁹Original PLWORDNET LUs may have glosses in *Walenty* as well.

cludes $\langle \text{Initiator}, \text{Theme} \rangle$,¹⁰ $\langle \text{Initiator}, \text{Theme}, \text{Recipient} \rangle$, $\langle \text{Initiator}, \text{Result} \rangle$ (WRITE), $\langle \text{Initiator}, \text{Theme}, \text{Result} \rangle$ (BUILD) and $\langle \text{Stimulus}, \text{Experiencer} \rangle$.

Let us consider the verb DEGUSTOWAĆ ‘taste’ representing the situation when a person (*Initiator*) is tasting some food (*Theme*), cf. Fig. 6.

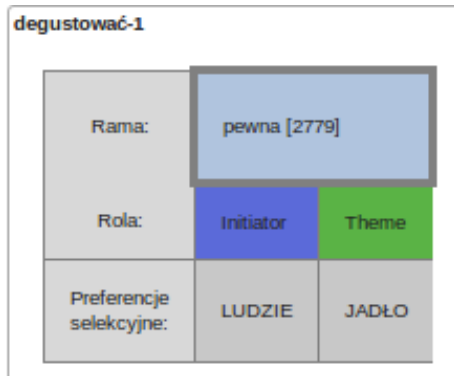


Figure 6: A screenshot with a simple frame

Some roles, like VerbNet *Initial_Time* and *Final_Time* or *Agent* and *Counteragent* are related. In VerbNet (CLE, 2012) they are connected in the hierarchy of roles. Instead, we decided to use two level representation, composed of *basic roles* discussed above and *role attributes*. This solution is more flexible, as we do not assume in advance, which basic role may be equipped with attributes. On the other hand, the relations between roles are apparent without any external knowledge.

Therefore, basic roles can be supplemented with attributes organised into pairs *Foreground, Background* and *Source, Goal*. The first pair is used if various paraphrases change only the focus of an utterance. For instance, the situation that ‘*someone exchanges something for something with someone*’ involves two initiators and two themes. As *Mary exchanged a dress for a jacket with Anna* provides same information as *Anna exchanged a jacket for a dress with Mary*, we will represented them as *Initiator^{Foreground}, Initiator^{Background}* and *Theme^{Foreground}, Theme^{Background}*. Such paraphrase can involve the change of syntax. Consider the verb IRYTOWAĆ ‘irritate’ having 12 syntactic schemata in *Walenty*, three of them used in (4) presenting paraphrases of the same content. It involves a person experiencing irritation and a *Stimulus* causing it. The last can be syntactically expressed by means of two phrases or a single phrase. The corresponding frame is presented in Fig. 7.

- (4) Faceci irytują mnie arogancją.
 guy.NOM.PL irritate me arrogance.INST.SG
 U facetów irytuje mnie arogancja.
 at guy.GEN.PL irritate me arrogance.NOM.SG
 Irytuję się arogancją facetów.
 irritate self arrogance.INST.SG guy.GEN.PL
 ‘Guys irritate me with their arrogance.’

¹⁰Both *Initiator* (WALK, BIVOUAC) and *Theme* (SLEEP, ROT) can appear in frames independently.

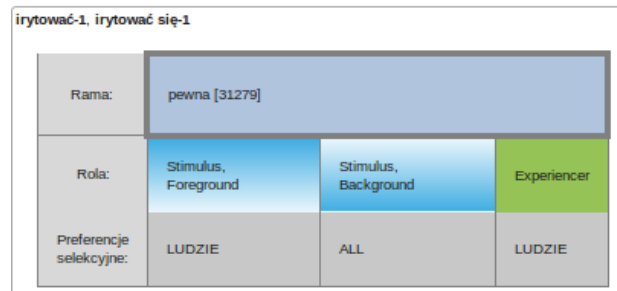


Figure 7: *Foreground, Background* attributes of role *Stimulus* for the verb IRYTOWAĆ

The other pair of attributes (*Source, Goal*) is used when a direction between actants can be found (in the wide sense). The most natural example here is *Location* for verbs of movement, e.g. we have *Location^{Source}, Location^{Goal}* in ‘*to go from somewhere to somewhere*’. However, we use them also for representing a BUY/SELL situation, i.e. ‘*someone buys something from someone*’, ‘*someone sells something to someone*’. We think that this situation differs from GIVE/RECEIVE situation, as a buyer pays a seller for goods he is receiving. We assign *Initiator^{Source}, Initiator^{Goal}* in assumption that the direction of transferring goods is more important than the direction of transferring money.

We assume that attribute roles form genuine pairs aimed at distinguishing two occurrences of the same role in a frame, hence we forbid including only one of them in it. The only exception is *Location*, as the initial location is something special in spite of whether there exists the final location (and vice versa). The other reason is that we have distinguished ablative and adlative constructions on the syntax level as well, cf. (Przepiórkowski et al., 2014a, §3.4, p.2787). For instance, POCHODZIĆ ‘come from’¹¹ has an obligatory source location and an impermissible goal location, cf. (5).

- (5) pochodzić-1: *Theme* *Location^{Source}*
 subj{np(str)} {xp(abl)}

Obie panie pochodziły z Europy Środkowej.
 Both woman.PL come.PL.PAST from Europe Central.
 ‘Both women came from Central Europe.’

4.2. Selectional preferences

Arguments, identified by semantic roles, are provided with selectional preferences (Katz and Fodor, 1964; Resnik, 1993). Unlike many other dictionaries, we do not use a fixed set of qualifiers, like *abstract/concrete, solid/liquid/gaseous* etc. We want to be much more precise, hence we use PLWORDNET synsets (represented by LUs) and relations to represent selectional preferences. Therefore, only *dogs* can BARK, we DRINK only *beverages* (not all *liquids*), and we only use *bandages* to BANDAGE (not every *cloth*).

The selectional preferences are represented as a list of elements of the following four types (elements of different

¹¹Meaning *to be born/raised somewhere*, not *to arrive from somewhere*.

types can cooccur in the same list):

1. a predefined set of synsets,
2. a PLWORDNET synset,
3. a PLWORDNET relation to another argument,
4. a PLWORDNET relation to another synset.

The most basic way to represent selectional preferences is a direct use of PLWORDNET synsets. For instance, the frame of the verb AKCENTOWAĆ ‘accentuate’ (a word on the syllable) with a strictly constrained meaning is presented in Fig. 8. The complete frame including selectional preferences for (5, previous page), is presented in (6). Since *country* is connected neither to *administrative district*, nor to *geographical region*, it has to be considered separately.

akcentować-3			
Rama:	pewna [23388]		
Rola:	Theme, Background	Theme, Foreground	Initiator
Preferencje selekcyjne:	sylaba-1	leksem-1	istota ludzka-1

Figure 8: A frame for the verb AKCENTOWAĆ with PLWORDNET selectional preferences only

- (6) *Theme* *Location*^{Source}
osoba-1 {jednostka administracyjna-1 ‘administrative district’,
obszar-1 ‘region’,
państwo-1 ‘country’}

However, in many situations, groups of PLWORDNET synsets commonly occur together in a single selectional preference. For example, both foods and drinks can be tasted or pasteurised. Similarly, both people and organisations/companies can buy, sell or store goods. As such semantically connected concepts are composed of many unrelated PLWORDNET synsets, we decided to add symbols representing such common combinations. For example, *JADŁO* ‘victuals’ represents the set {*jedzenie-1* ‘food’, *napój-1* ‘drink’}, *LUDZIE* ‘PEOPLE’ – {*człowiek-1* ‘human’, *grupa ludzi-1* ‘group of people’} and *PODMIOTY* ‘legal subjects’ – {*człowiek-1* ‘human’, *grupa ludzi-1* ‘group of people’, *podmiot-2* ‘legal subject’ (grouping organisations, companies and firms)}. Their main advantages are that they can be rearranged in the future (if appropriate) and that they simplify processes of creating and reading the dictionary. Additionally, *ALL* means that no selectional preferences can be defined. Such predefined sets are used in frames presented in Figs. 6 and 7.

Complicated structure of PLWORDNET made us also introduce PLWORDNET relations to another synset as a way of representing selectional preferences. For instance, an *Instrument* for *PISAĆ* ‘write’ could be a *pen*, a *ballpen*, a *pencil* etc. However, in PLWORDNET their direct hypernym is *artykuł papierniczy-1* ‘writing materials’ which is evidently too wide (as it includes, e.g., ‘notebook’). They

are correctly joined by the *holonymy (collection)* relation to *przybory do pisania-1* ‘writing implements’, as this term is used in Polish only in plural.

For some predicates, arguments considered separately represent a wide class of entities, but actually they are closely related to each other. For instance, one meaning of *MLEĆ* ‘mill’ concerns objects moving their parts through some substance. For example, windmill can mill air with its sails, while water wheel can mill water with its blades (but not with sails as it has none). Classic selectional preferences tell us nothing about what can be used by those objects for milling, but we can clearly see that they have to have to be internal parts of original object. Therefore, we introduced selectional preferences determined by means of relations to another argument. Meronymy seems to be an appropriate relation here, cf. Fig. 9.

mleć-A			
Rama:	brak [42655]		
Rola:	Instrument, Background	Theme	Instrument, Foreground
Preferencje selekcyjne:	urządzenie-5	substancja-1	meronimia (typu część) -> [Instrument, Background]

Figure 9: Selectional preferences based on relations between arguments for the verb MLEĆ

Sometimes, such a relation between arguments is not strictly defined. Let us consider the verb *SKŁADAĆ SIĘ* ‘consist of’. It may concern concrete objects, e.g., various devices, groups of concrete people, events having their phases, etc. Therefore, we consider a symbol *RELAT* meaning any close PLWORDNET relation between LUs occurring in text. The frame for *SKŁADAĆ SIĘ* is presented in (7).

- (7) *składać się-1*
Theme^{Foreground} *Theme*^{Background}
ALL {część-3 ‘part’,
RELAT → *Theme*^{Foreground} }
 {subj{np(str)} {prepn(z, gen)}
Obiad składał się z drugiego dania i kompotu.
 ‘Dinner consisted of a main dish and compote.’
Jury konkursowe składało się ze znanych dziennikarzy.
 ‘Jury consisted of famous journalists.’
Turniej składał się z wielu konkurencji.
 ‘Tourney consisted of many competitions.’

5. Connecting both layers

In *Walenty*, syntactic and semantic valence information are represented separately. Nevertheless, they are closely connected, but this relation is a many-to-many one. On one hand, one semantic frame can be syntactically implemented by several schemata (diathesis alternation). On the other, one schema can be used in several frames.

As mentioned before, exemplary sentences are linked to corresponding schemata (with particular phrase types used in them marked). On the semantic level, the same examples

datować-1, datować się-1			
Rama:	pewna [18334]		
Rola:	Time	Theme	Initiator
Preferencje selekcyjne:	CZAS	KOMUNIKAT	PODMIOTY
		KONCEPCJA	
		SYTUACJA	

datować-2			
Rama:	pewna [18335]		
Rola:	Time	Theme	Initiator
Preferencje selekcyjne:	CZAS	WYTWÓR	PODMIOTY

datować się („imperf):	
Schemat:	pewny [56]
Funkcja:	subj
Typy fraz:	np(str) prepnp(na,acc)

datować („imperf):	
Schemat:	pewny [2077]
Funkcja:	subj
Typy fraz:	np(str) xp(temp)

datować („imperf):		
Schemat:	pewny [70]	
Funkcja:	subj	obj
Typy fraz:	np(str)	np(str) prepnp(na,acc)

datować („imperf):		
Schemat:	pewny [4878]	
Funkcja:	subj	obj
Typy fraz:	np(str)	np(str) xp(temp)

Figure 10: A screenshot with two semantic frames and schemata being syntactic realisations of the first

are connected to the corresponding lexical units and, as result, they illustrate semantic frames. This dependency links schemata to frames, but such link is by all means insufficient.

We directly link semantic arguments with corresponding syntactic positions. We assume that the fact that all phrase types composing a single position can coordinate means that they cannot represent different semantic arguments, but, in various situation, not all phrase types composing a position must be connected with an argument. Similarly, not all positions must be connected with all frames adequate for a particular lemma (due to storing longest schemata).

Let us consider the verb DATOWAĆ ‘date’ with two meanings: *datować-1* ‘determine time of appearance’ and *datować-2* ‘put a date’. The first one has its reflexive counterpart as well: *datować się-1* ‘have a determined time of appearance’. The frames corresponding to these two meanings together with schemata being realisations of one of these frames (for *datować-1* and *datować się-1*) are presented in Fig. 10. On the left one can see semantic frames, with PLWORDNET lexical units connected to it just above. The roles of “active” frames are distinguished by corresponding colours. Syntactic schemata are positioned on the right. Phrase types and syntactic positions they belong to are coloured accordingly to the role they are connected to. The paraphrase could appear within the same syntactic schema, the more so as we represent the longest schemata. For example, let us consider the verb PORASTAĆ ‘overgrow’. The subject of the verb can be a *plant* (*Theme^{Source}*) overgrowing an area (*Theme^{Goal}*) as some layer (*Attribute*), cf. (8). In (8a) TRAWA ‘grass’ appears as a subject in nominative, whereas ZBOCZE ‘slope’ appears as an object in

accusative (together with DYWAN ‘carpet’ in instrumental). Contrary, in (8b), TRAWA appears in instrumental, whereas ZBOCZE appears in nominative (an accusative object is impossible here; furthermore, there is no way to include information about *Attribute*). Therefore, the same phrase type may be used in text in two different ways.

- (8) a. Trawa porasta zbczce
grass.NOM.SG overgrows slope.ACC.SG
zielonym dywanem.
green.INST.SG carpet.INST.SG
- b. Zbczce porasta trawą.
slope.NOM.SG overgrows grass.INST.SG
‘Grass overgrows a slope (with a green carpet.)’

The snapshot of the program visualisation of this entry is presented in Figure 11. Namely, NP in nominative (subject) of the upper schema [437] is used as *Theme^{Source}* or *Theme^{Goal}*, whereas NP in instrumental is used as *Theme^{Goal}* or *Attribute*. We call such phenomenon *autoalternation*. It is shown on screen as duplication of syntactic positions (function line and phrase type lines). Each duplicate is connected with the frame in its own, independent way. Currently, only one autoalternation per schema is allowed.

Note that all exemplary sentences connected to the corresponding lexical units appear at the bottom.

Schemata representing phraseological usage of predicates should be interpreted semantically as well. Two different cases should be considered. The simpler is when a lexicalised dependant does not change its meaning and represents a fixed form of an argument (or a modifier). This is the case for the lexicalisation *upijać się na umór* (cf. section 3.), as this shows the manner of getting drunk. There-

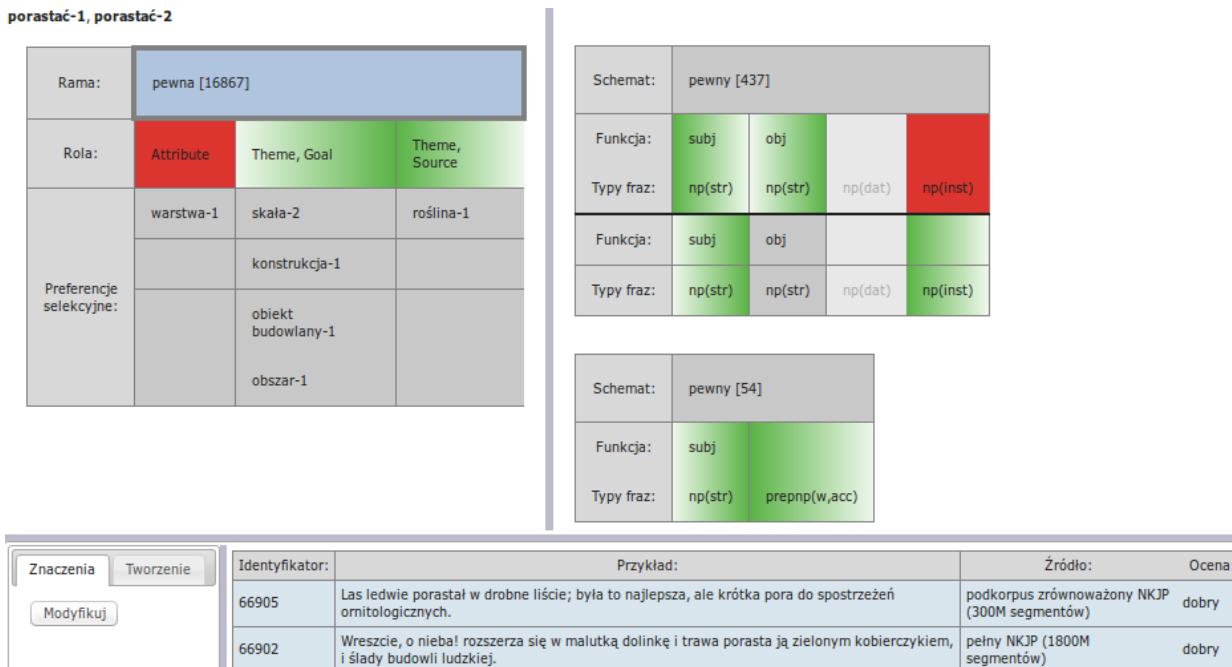


Figure 11: A screenshot with a semantic frame and schemata being its syntactic realisation

fore, *na umór* is treated as syntactic realisation of *Manner*, cf. Fig. 12. Observe that a non-lexicalised phrase type $x_p(\text{mod})$ (from another schema) is connected with it as well. For arguments having solely lexicalised realisations selectional preferences are not provided. The other concerns a lexicalised dependant modifying the meaning of a verb.



Figure 12: A connection between a phraseological schema and a frame for the verb UPIJAĆ SIĘ

6. Conclusions and future work

In this paper we presented the semantic layer of Polish valence dictionary *Walenty*, and how it is linked with the existing syntactic layer. We listed the set of semantic roles and discussed how they (together with selectional preferences) form semantic arguments and subsequently frames.

Walenty is still under development. While its syntactic layer is nearly finished, the work on the semantic layer is

in progress, only about 6 300 (52%) entries are semantically elaborated. In particular, integrating the phraseology in *Walenty* and PLWORDNET should be developed. This concerns only the idiomatic arguments changing the meaning of the verb, like *drzeć koty*, ‘to squabble’ (literally ‘to tear cats’). They are represented in PLWORDNET as separate LUs with multi-word lemmas. The word *koty* is not an argument here, hence the corresponding syntactic position has to be marked as a part of a lemma. Therefore, we have to generate a multi-word lemma from the morphosyntactic characteristics of its elements. The problem is that such expressions are conventionalised in a certain form (including an order) which is not represented in *Walenty* to the full extent, cf. (Przepiórkowski et al., 2016).

Till now, we have focused on semantic representation of verbs. The representation for other predicates is the same. The only difference is that we plan to connect derivatives, such as the noun ROZKAZ and the verb ROZKAZAĆ ‘order’ or the adjective DOJRZAŁY, the verb DOJRZEĆ ‘mature’ and the noun DOJRZAŁOŚĆ ‘maturity’, with the same frame.

In the future, we plan to use semantic information included in *Walenty* for multiple tasks, including semantic parsing and anaphora resolution. We also plan to further extend the resource, mainly its nominal part.

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.

7. Bibliographical References

Mirosław Bańko, editor. (2000). *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.

- Center for Computational Language and Education Research, University of Colorado, Boulder, CO, (2012). *VerbNet Annotation Guidelines*.
- Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of the WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL.
- Fillmore, C. J., Johnson, C. R., and Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.
- Fillmore, C. J. (1968). The case for the case. In Emmon Bach et al., editors, *Universals in linguistic theory*, pages 1–88. Holt, Rinehart and Winston, New York, NY.
- Fillmore, C. J. (1971). Some problems for Case Grammar. In *Report of the 22nd Annual Roundtable Meeting on Linguistics and Language Studies*, pages 35–56. Georgetown University Press, Washington, DC.
- Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.
- Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. *Mathematical Modelling in Physics, Engineering and Cognitive Science*, 9:57–68.
- Hlaváčková, D. and Horák, A. (2006). VerbaLex — new comprehensive lexicon of verb valences for Czech. In *Proceedings of the Third International Seminar on Computer Treatment of Slavic and East European Languages*, pages 107–115, Bratislava, Slovakia.
- Katz, J. J. and Fodor, J. A. (1964). The structure of a semantic theory. In J. A. Fodor et al., editors, *The Structure of Language*, pages 479–518. Prentice Hall.
- Kipper-Schuler, K. (2005). *VerbNet: A broad coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, University of Pennsylvania.
- Levin, B. (1993). *English verb classes and alternation: a preliminary investigation*. University of Chicago Press, Chicago, IL.
- Lopatková, M., Žabokrtský, Z., Skwarska, K., and Benešová, V. (2003). VALLEX 1.0 valency lexicon of Czech verbs. Technical Report TR-2003-18, ÚFAL/CKL MFF UK, Prague, Czech Republic.
- 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.
- Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland.
- 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., Hajič, E., Patejuk, A., Skwarski, F., Woliński, M., and Świdziński, M. (2014a). Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2785–2792, Reykjavík, Iceland. ELRA.
- Przepiórkowski, A., Hajič, 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., Skwarski, F., Hajič, E., Patejuk, A., Świdziński, M., and Woliński, M. (2014c). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII:159–178.
- Przepiórkowski, A., Hajič, J., Hajič, E., and Urešová, Z. (2016). Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, 29. to appear.
- Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, December.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., and Johnson, C. R. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.
- Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, Holland.
- Świdziński, M. (1994). *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.
- Urešová, Z. (2009). Building the PDT-Vallex valency lexicon. In *Proceedings of the 5th Corpus Linguistics Conference*. University of Liverpool.
- Woliński, M. (2004). *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.
- 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.

8. Language Resource References

- PIWordNet: 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.
- 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.