Walenty: Towards a comprehensive valence dictionary of Polish

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk,
Marcin Woliński, Filip Skwarski, Marek Świdziński

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
{adamp,hajnicz,aep,wolinski}@ipipan.waw.pl

Abstract

This paper presents Walenty, a comprehensive valence dictionary of Polish, with a number of novel features, as compared to other such dictionaries. The notion of argument is based on the coordination test and takes into consideration the possibility of diverse morphosyntactic realisations. Some aspects of the internal structure of phraseological (idiomatic) arguments are handled explicitly. While the current version of the dictionary concentrates on syntax, it already contains some semantic features, including semantically defined arguments, such as locative, temporal or manner, as well as control and raising, and work on extending it with semantic roles and selectional preferences is in progress. Although Walenty is still being intensively developed, it is already by far the largest Polish valence dictionary, with around 8600 verbal lemmata and almost 39 000 valence schemata. The dictionary is publicly available on the Creative Commons BY SA licence and may be downloaded from `http://zil.ipipan.waw.pl/Walenty`.

Keywords: subcategorisation, phraseology, coordination

1. Introduction

The aim of this paper is to present an early version of Walenty, a comprehensive valence dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS).[1] The dictionary is meant to be both human- and machine-readable; in particular, it is being employed by two parsers of Polish, Świgra[2] (Woliński, 2004) and POLFIE[3] (Patejuk and Przepiórkowski, 2012). The former, Świgra, is an implementation of the DCG (Warren and Pereira, 1980) grammar of Polish of Świdziński (1992), consistent with Polish structuralist tradition (cf., e.g., Saloni and Świdziński 1998). The latter, POLFIE, while parasitic on Świgra and – to a lesser extent – an earlier HPSG toy grammar of Polish (Przepiórkowski et al., 2002), 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.

For this reason, Walenty exhibits a number of features which are rare or absent in other valence dictionaries; some of these are described – and illustrated with English examples for the ease of exposition – in §2. Examples of Polish lexical entries, illustrating the formalism assumed in the dictionary, are given in §3. (with a more systematic presentation relegated to the Appendix), and some quantitative characteristics of the current version of Walenty are presented in §4. This version contains almost exclusively morphosyntactic information about valence schemata, but future versions will also include semantic and quantitative information, as mentioned in §5. Finally, §6. concludes the paper.

2. Features

Each lexical entry contains a number of valence schemata,[4] and each schema is a set of specifications of an argument. But are the noun phrase (NP) *a republican* in *Pat became a republican* and the adjective phrase (AdjP) *quite conservative* in *Pat became quite conservative* two different arguments, belonging to separate valence schemata of BECOME, or two different realisations of one argument specified in a single schema? Walenty is explicit about what counts as the same argument, and it employs the **coordination test** to resolve such doubts: if two phrase types can be felicitously coordinated in the same sentence, they are different realisations of the same argument; this is the case here, cf. *Pat became a republican and quite conservative* (Sag et al., 1985, p. 142, ex. (67a)). Hence, arguments are specified disjunctively, by listing phrase types which may occupy a given syntactic position and may in principle be coordinated (see §3.2.).

Specification of nominal arguments includes information about their case. However, this is insufficient in Polish to uniquely determine grammatical **subjects and objects**: the former do not have to be nominal at all, but – depending on the verb – may sometimes be clausal or infinitival, the latter do not have to be accusative, as there are passivisable verbs taking instrumental or genitive objects. For this reason, subject and object arguments are explicitly marked as such (see §3.1.).

Moreover, some nominal arguments cannot be assigned a unique morphological **case**, as it may depend on syntactic context. One such phenomenon is the so-called Genitive of Negation, where – roughly speaking (see Przepiórkowski 2000 for gory detail) – the normally accusative complement of an affirmative verb must bear the genitive case when the verb is negated. The case of such arguments is marked as

[1] An even earlier version of the dictionary is described in Polish in Przepiórkowski et al. 2014.

[2] `http://zil.ipipan.waw.pl/Sk%C5%82adnica`

[3] `http://zil.ipipan.waw.pl/LFG`

[4] As long as the dictionary contains mostly morphosyntactic information, we avoid using the term *valence frame*.

structural, following the generative[5] tradition since Rouveret and Vergnaud 1980 and Chomsky 1981 (see §3.1.).

Another related couple of phenomena widely discussed in the generative literature are **control and raising** (Rosenbaum, 1967; Landau, 2013). For example, valence schemata of verbs such as PROMISE and ORDER are the same at the purely morphosyntactic level – they combine with an NP subject, an NP object and an infinitival phrase, e.g., *John promised/ordered Tom to come* – but they differ in their control properties: with *promised*, the understood subject of *to come* is *John*, with *ordered* – it is *Tom*. 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 (see §3.3.).

Some obligatory arguments look like typical adjuncts, for example those expressing **location, time or manner**, as in LIVE (somewhere), LAST (some time) or BEHAVE (in some way). Such arguments may be realised by a relatively large number of phrase types, including adverbial and prepositional, with a limited number of prepositions (cf. *John lives there / at Tom's / in the suburb / under the bridge*, etc.). It would not be economical to list all possible realisations of, say, location in valence schemata of all verbs combining with such locative arguments, so – even in this morphosyntactic version – Walenty assumes a number of semantically-defined argument types, including *locative*, *ablative*, etc., *temporal*, *durative* and *manner*. A list of possible morphosyntactic realisations of such arguments accompanies the dictionary proper (see §3.4.).

The final feature to be mentioned here is **phraseology**. Walenty includes special notation for various types of idiomatic arguments, from completely fixed (given as a string) to almost freely modifiable (see §3.5.).

3. Formalism

3.1. Basics

The valence dictionary is a database, with a number of output formats, including the purely textual format illustrated here with a valence schema for the verb KIEROWAĆ 'manage, run':[6]

```
kierować: _: imperf:
subj{np(str)} + obj{np(inst)}
```

There are two arguments mentioned in this schema (separated by +): a subject and an object, both realised as nominal phrases (NPs, marked as `np` in the schema). The subject is defined here as the argument prototypically agreeing with finite forms of the verb, and the object is the argument that can be passivised. As this example illustrates, an object does not have to be accustative in Polish, it bears the instrumental case here. On the other hand, the case of the subject is marked as structural: normally it is nominative, but gerundial forms combine with the genitive case (while the instrumental of the object remains instrumental). It is the task of the grammar that employs this valence dictionary to properly define morphological realisations of structural case depending on the syntactic context.

Each lemma is followed by two parameters: negativity and aspect. In Slavic, aspect is a morphological category of verbs; KIEROWAĆ is marked as imperfective. Negativity is a more ephemeral category signalling whether the schema is valid only for negated, only for non-negated or for any verb forms. Here it is valid for any verb forms (see the underscore, _), but there are verbs which have certain schemata only when negated, e.g., ZNOSIĆ when used as in *nie znosić* 'detest'; for such a frame the value of negativity would be `neg`.

Only the two grammatical functions mentioned above, subject and object, are marked explicitly. For example, one of the schemata for the perfective verb ZWIERZYĆ SIĘ 'confide (sth to sb)' is shown below, where only the subject is explicitly marked. The other two arguments are a (non-passivisable) dative nominal phrase and a prepositional phrase (PP, marked as `prepnp` in the schema) headed by the preposition Z combining with an instrumental NP.

```
zwierzyć się: _: perf:
subj{np(str)} + {np(dat)} +
{prepnp(z,inst)}
```

Note that the lemma contains the so-called reflexive marker `się`, which does not really mark semantic reflexivity here. When it does, as in *myć się* 'wash oneself', where *się* may be replaced by the emphatic reflexive pronoun *siebie samego*, the lemma is MYĆ and the reflexive marker is given as an additional argument, `refl` (the third argument expresses the instrument of washing):

```
myć: _: imperf:
subj{np(str)} + {np(inst)} + {refl}
```

We will omit the lemma and the categories of negativity and aspect in subsequent examples.

3.2. Coordination

All examples given so far involved single morphosyntactic realisations of arguments. As mentioned in §2., one of the unique features of the valence dictionary described here is that it is explict about what counts as a single syntactic position: if two morphosyntactically diverse phrases may be coordinated then they occupy the same syntactic position. If this is an argument position (rather than an adjunct) then they should both be mentioned in the same schema, as different realisations of the same argument.

This is exemplified in the following schema for TŁUMACZYĆ 'explain', as in *Musiałem im tłumaczyć najprostsze zasady i dlaczego trzeba je stosować* 'I had to explain them the most basic principles and why they should be adhered to' involving a coordinated phrase in the object position consisting of an NP (*najprostsze zasady* 'the most basic principles') and an interrogative clause (*dlaczego trzeba je stosować* 'why they should be adhered to'; marked here as `cp(int)`).

```
subj{np(str)} + obj{np(str); cp(int)} +
{np(dat)}
```

[5] We use the term *generative* broadly, as referring to the transformational tradition of Noam Chomsky but also to more formal linguistic theories such as HPSG and LFG.

[6] Such schemata are split into multiple lines here for typographic reasons.

This explains the use of braces {} delimiting morphosyntactic realisations of an argument, alluding to the common notation for sets (although such set elements are separated here by the semicolon ; rather than the usual comma).

Obviously, the number of different realisations is not limited to two. For example, one of the schemata for INFORMOWAĆ 'inform' mentions 7 realisations of one of the arguments, 5 of which are shown below:

```
subj{np(str)} + obj{np(str)} +
{prepnp(o,loc); cp(int); cp(że);
 prepncp(o,loc,int); prepncp(o,loc,że)}
```

Apart from the structural subject and object, analogous to the English *Somebody informed somebody*, there is an argument representing the transferred information: a PP headed by O and involving a locative NP (analogous to *inform about something*), an interrogative subordinate clause (*inform who came*), an indicative subordinate clause headed by the complementiser ŻE (*inform that she came*), and prepositional correlative clauses of the form illustrated by *informować o tym, kto przyszedł*, literally 'inform about it who came' and by *informować o tym, że przyszła*, lit. 'inform about it that she came'.

A similar example with 7 realisations can be given for INTERESOWAĆ 'interest' (as in *something interests somebody*); see the complete schema below:

```
subj{np(str);
 cp(int); cp(że); cp(żeby);
 ncp(str,int); ncp(str,że);
 ncp(str,żeby)} +
{np(str)}
```

There are two arguments in this schema: a subject and a structural NP (not passivisable, hence not an object). The subject has the usual np(str) realisation, but may also be expressed by three types of subordinate clauses (interrogative or headed by one of the complementisers ŻE and ŻEBY) and by three types of nominal correlative phrases, where the correlate TO 'it' bears the structural (here, always nominative) case and is followed by one of the three types of subordinate clause, as in *Interesowało go to, że przyszła*, lit. '(It) interested him it that (she) came'.

3.3. Control and raising

Let us consider the following schemata:

OBIECAĆ 'promise':

```
subj,controller{np(str)} +
{np(dat)} +
controllee{infp(_)}
```

KAZAĆ 'order':

```
subj{np(str)} +
controller{np(dat)} +
controllee{infp(_)}
```

There are three arguments in each of the schemata above: a nominal subject in the structural case, a dative NP, and an infinitival phrase of an unspecified aspect (marked with the underscore _ here; some verbs combine only with imperfective infinitives, i.e., with infp(imperf); see below). In both cases the infinitival argument has its covert subject controlled (as signalled by controllee), but the two schemata differ in what controls it (cf. controller): it is the structural subject in case of OBIECAĆ, but the dative complement in case of KAZAĆ. Hence, *Janek obiecał Tomkowi przyjść* 'Janek promised Tomek to come' means that Janek is supposed to come, while *Janek kazał Tomkowi przyjść* 'Janek ordered Tomek to come' means that Tomek is supposed to come.

Note that in such control constructions the controller has its morphosyntactic makeup defined independently of the controllee; e.g., it is a structurally-cased NP in case of OBIECAĆ 'promise'. This is one of the differences between control (or 'equi') verbs, such as those above, and raising verbs, such as ZACZYNAĆ 'begin' and WYDAWAĆ SIĘ 'seem'. The subject of such a raising verb is simply whatever would count as the subject of its infinitival argument: if the infinitival verb takes a sentential subject, then the higher raising verb also takes a sentential subject, etc. Such raised subjects are marked as E, as in the following example for ZACZYNAĆ 'begin':

```
subj,controller{E} +
controllee{infp(imperf)}
```

This single schema covers uses of ZACZYNAĆ such as *Janek zaczyna biec* 'Janek begins to run', *Zaczyna padać* 'It begins to rain' or *Zaczęło go interesować, że Maria biegnie* 'It began to interest him that Maria is running', where the subordinate clause *że Maria biegnie* 'that Maria is running' is arguably the shared subject of *interesować* 'interest' and *zaczęło* 'begin'.[7]

3.4. Semantically defined arguments

It is well known that some arguments look very much like typical adjuncts in that they express manner, location or time, and have numerous morphosyntactic realisations. For example, the obligatory manner argument of the English TREAT, as in *Somebody treated somebody in some way*, may be realised by a prepositional phrase or by an adverbial phrase, as in *Somebody treated somebody very badly*, or even by a clause, as in *Somebody treated somebody as if he wasn't a human being*. What these realisations have in common is that they express a manner. Similar manner arguments are observed with other verbs, e.g., BEHAVE.

Walenty distinguishes 7 such semantically defined argument types: manner (xp(mod)), 2 broadly temporal types (xp(temp) for points in time and xp(dur) for durations) and 4 broadly locative types (xp(locat) for locations, xp(abl) for ablative arguments, xp(adl) for adlative and xp(perl) for paths). For example, CIĄGNĄĆ 'draw' (as in *Something draws somebody somewhere*) has an adlative argument:

[7] The current version of the dictionary also contains arguments specified as subj{E} (without controller). This notation refers to covert subjects of verbs such as PADAĆ 'rain' and GRZMIEĆ 'thunder', corresponding to expletive subjects in English (cf. *It is raining.*).

```
subj{np(str)} + obj{np(str)} + {xp(adl)}
```

The dictionary comes with a list of possible morphosyntactic realisations of each of the `xp(...)` arguments. For example, adlative arguments can be realised by 13 types of prepositional phrases, including `prepnp(ku,dat)`, where the preposition KU combining with dative NPs usually means 'towards'. The 14th realisation is `advp(adl)`, defined via a number of *lexical* realisations such as TAMŻE '(towards) there' (compare the German HINEIN), DOKĄDKOLWIEK 'towards wherever' (cf. IRGENDWOHIN in German), etc.

A slight complication arises in case of some manner arguments, as in the following schema for OBCHODZIĆ SIĘ 'handle, treat' (compare *Somebody handles somebody in a certain way*, but in Polish the non-agentive argument is expressed as a PP):

```
subj{np(str)} +
controller{prepnp(z,inst)} +
controllee{xp(mod)}
```

The use of `controller` and `controllee` is different here than in §3.3.: `controller` marks the argument that certain realisations of `xp(mod)` must be morphosyntactically parallel with, as in the following example, where *jak do dziecka* 'as to (a) child' is the textual realisation of `xp(mod)`: *Ktoś mówi do kogoś jak do dziecka* 'Somebody speaks to somebody like to a child'. Note that the preposition used after *jak* 'like', as must be the same as in the controlling phrase *do kogoś* 'to somebody'. Similarly, for the schema above, the phrase after *jak* would have to be a PP headed by Z combining with the instrumental case: *Ktoś obchodzi się z kimś jak z dzieckiem*, lit. 'Somebody handles with somebody as with (a) child'.[8]

3.5. Phraseology

While plain categories defined in Walenty have no lexical restrictions on how they can be realised, there are three categories which are subject to such constraints: `fixed`, `lexnp` and `preplexnp`.

Phrases of type `fixed` have just one parameter: the exact orthographic realisation of the phrase; see the following schema for ZBIĆ 'beat' (as in *He beat them to a pulp*), with *na kwaśne jabłko* meaning literally 'into sour apple':[9]

```
subj{np(str)} + obj{np(str)} +
{fixed('na kwaśne jabłko')}
```

A more interesting type is `lexnp` with four parameters indicating the case of the NP, its grammatical number, the lemma of the head, and the modifiability pattern. The following schema for PŁYNĄĆ 'flow' (as in *Hot blood flows in his veins*), where the subject is a structurally-cased NP, as usual, but the head of this NP must be KREW 'blood' in the singular and the NP may contain modifiers (cf. `atr`), illustrates this:

```
subj{lexnp(str,sg,'krew',atr)} +
{preplexnp(w,loc,pl,'żyła',ratr)}
```

The final lexical argument type is `preplexnp`, which contains an additional (initial) parameter, namely the preposition. In the above schema, the second argument is a PP headed by the preposition W 'in' combining with a locative NP in the plural. The NP must be headed by ŻYŁA 'vein' and must contain a possessive modifier (`ratr` stands for 'required attribute'). So this schema covers examples such as *Gorąca krew płynie w jego żyłach* 'Hot blood flows in his veins', but – correctly – not the non-phraseological *Gorąca krew płynie w żyłach* (no modifier of 'veins') or *Gorąca krew płynie w jego żyle* (singular 'vein').

Note, incidentally, that phrases of type `preplexnp` also occur in definitions of phrases of type `xp` discussed in §3.4. For example, one of possible realisations of `xp(adl)` is `preplexnp(w,loc,sg,'strona',atr)`, as in *w stronę morza* 'towards the sea' (lit. 'in (the) direction (of the) sea').[10]

The third possible value of the modifiability parameter is `natr`, for lexicalised arguments that cannot involve modification. The following schema for ZMARZNĄĆ 'get cold, freeze' handles the idiom *zmarznąć na kość* 'freeze to the marrow' (lit. 'freeze to (the) bone'):

```
subj{np(str)} +
{preplexnp(na,acc,sg,'kość',natr)}
```

In this idiom, the accusative noun *kość* 'bone' cannot be modified, as illustrated by the infelicitous *Zmarzł na gołą/twardą kość* '(He) froze to (the) naked/hard bone'.

Finally, `batr`, which stands for 'bound attribute', indicates that the NP must involve a possessive modifier meaning 'self' or '(one's) own', i.e., a form of either SWÓJ or WŁASNY. For example, ZOBACZYĆ 'see' is involved in an idiom meaning 'to see with one's own eyes', as in *Na własne oczy zobaczyłem jej uśmiech i to, że nie była wcale taka stara* 'With my own eyes I saw her smile and that she wasn't so old at all':

```
subj{np(str)} + {np(str); ncp(str,że)} +
{preplexnp(na,acc,pl,'oko',batr)}
```

[8] There are plans to abandon such use of `controller` and `controllee`, as it seems that the comparative JAK 'like, as' (and similarly JAKO 'as' and NIŻ 'than') is relatively free in selecting the target of comparison. For example, another possible realisation of `xp(mod)` in the case at hand could be *Ktoś obchodzi się z kimś jak despota* 'Somebody treats somebody like a despot' (example due to Urszula Andrejewicz, p.c.), where the subject of *obchodzi się* is referred to by the *jak* phrase. As various kinds of phrases (or even non-constituents) may follow such comparative elements, the formalism will probably be extended by a new argument type, say, `compar(...)`, with at least three possible values of its single parameter: `jak`, `jako` and `niż`.

[9] The complete Walenty currently contains over 80 schemata using `fixed`, but there are plans to minimise the use of this type and replace it with `preplexnp` with `natr` (see below) whenever it makes sense.

[10] The use of `atr` in this argument specification is imprecise, as it seems that STRONA 'direction' must be modified here; cf. *w tę stronę* 'in this direction' vs. the infelicitous **w stronę*. As this example shows, the modifier does not have to be possessive, so none of the available modifiability patterns can be precisely applied here. Extensions to the phraseological module of Walenty are currently under intensive development.

3.6. Extensions

3.6.1. Pronominal arguments

The formalism distinguishes two cases in which a normally sentential argument may be expressed by a pronoun.

First, verbs such as DECYDOWAĆ 'decide', MAWIAĆ 'speak', SĄDZIĆ 'believe, think', TWIERDZIĆ 'claim' and UWAŻAĆ 'believe, consider' may combine with adverbial pronouns such as TAK 'so' and JAK 'how', e.g., *Jak uważasz?* 'How (do you) think?', *Tak sądzę* '(I) think so'. This possibility is signalled by `advp(pron)`, e.g., for the verb TWIERDZIĆ 'claim':

```
subj{np(str)} + {advp(pron)}
```

Note that this argument is specified in a schema separate from the one that mentions the sentential argument, as sentences such as the following do not seem to be acceptable in Polish: ?**Sądzę tak i że nie przyjdzie* 'I think so and that he won't come'.

Second, a much larger (in the current version of the dictionary) group of verbs allows for replacing the sentential argument with nominal pronouns such as CO 'what', COŚ 'something', NIC 'nothing', TO (SAMO) 'this (same)', etc., e.g.: *Co myślisz?* 'What (do you) think?', *Nic nie odpowiedział* 'He didn't say anything' (lit. 'Nothing (he) not said'), *Pomyślałem to samo* '(I) thought the same', etc. Following Postal (2004), such arguments are called *nonchromatic*, abbreviated to `nonch`, as in the following schema for POMYŚLEĆ (*o czymś*) 'think (about something)':

```
subj{np(str)} + {prepnp(o,loc)} +
{nonch}
```

3.6.2. Complex prepositions

The current formalism distinguishes arguments introduced by complex prepositions such as W KWESTII 'in (some) matter', NA TEMAT 'on (some) topic', Z POWODU 'because of' (lit. 'of reason'), etc., and marks them as `comprepnp`.[11]

Unlike in case of usual prepositional phrases, parameterised with the preposition lemma and the grammatical case it governs (e.g., `prepnp(z,inst)` or `prepnp(z,gen)`), such complex prepositions, when they combine with a nominal phrase,[12] seem to uniformly govern the genitive case, so explicit case information is not needed here. The following schema, for ROZPACZAĆ (*z powodu czegoś*) 'lament (because of something)', illustrates this type of arguments:

```
subj{np(str)} + {comprepnp(z powodu)}
```

3.6.3. Grammatical case

Valence dictionaries usually specify the morphological case of nominal arguments directly. Section 3.1. introduced `str` for the so-called structural case, i.e., case whose morphological realisation depends on the syntactic context and, hence, should be handled in the grammar rather than in the dictionary.

Walenty also explicitly marks potentially partitive arguments, i.e., arguments which are normally accusative but which may take the genitive case to indicate partitivity, as in *Dajcie wina i całą świnię!* 'Give (some) wine and (a) whole pig' (Przepiórkowski, 1999, 175), where *wina* 'wine' occurs in the genitive and *całą świnię* 'whole pig' – in the accusative. Correspondingly, DAĆ 'give' has the following schema (also indicating the beneficiary in the dative case):

```
subj{np(str)} + obj{np(part)} +
{np(dat)}
```

This case specification also occurs in some phraseological schemata, e.g., for the verb NABRAĆ 'take, draw', as used in *nabrać wody w usta* 'keep one's mouth shut', lit. 'draw (some) water into mouth(s)':

```
subj{np(str)} +
{lexnp(part,sg,'woda',natr)} +
{preplexnp(w,acc,pl,'usta',natr)}
```

Another case specification related to `str` is `pred`, for predicative case. This is the case of adjectival arguments of verbs such as CZUĆ SIĘ 'feel', PAMIĘTAĆ 'remember' or WYDAWAĆ SIĘ 'seem'. It is morphologically realised either by the instrumental, or via case agreement, e.g.: *Pamiętam go młodym* '(I) remember him.ACC young.INST' or *Pamiętam go młodego* '(I) remember him.ACC young.ACC'. When the case of such an adjectival phrase agrees with a predicated numeral subject, it may either agree with the genitive noun (e.g., *Pięć osób zostało rannych* 'Five.NOM/ACC people.GEN got hurt.GEN') or with the numeral (e.g., in the following example from the National Corpus of Polish: *…następne kilkadziesiąt metrów było czyste* '…(the) following several-tens.NOM/ACC meters.GEN were clean.NOM/ACC').[13] The source of agreement is marked here as `controller`, so the corresponding schema for WYDAWAĆ SIĘ 'seem' is specified as follows:

```
subj,controller{np(str); ncp(str,int);
 ncp(str,że); ncp(str,żeby)} +
controllee{adjp(pred)} + {np(dat)}
```

This schema takes into consideration sentential subjects introduced by a correlate, as well as a dative argument, as in *To, że umie grać na banjo, wydawało mi się mało istotne* 'It seemed to me of little importance that he can play banjo', lit. 'It that (he) can play on banjo seemed (to) me hardly important'.

The final case specification which does not refer directly to a morphological case is `postp`, i.e., "postprepositional case". It is used in prepositional phrases of type `prepadjp`, where the preposition PO is followed by

[11] Again, there is an ongoing discussion about the possibility of getting rid of this special argument type and treating arguments with complex prepositions as phraseological, using mechanisms described in §3.5.

[12] The other possibility is that the noun constituent of the complex preposition is modified by an adjective, as in *w tej kwestii* 'in this matter'.

[13] The case of the numeral and the agreeing adjective is marked as NOM/ACC here as there is some controversy as to whether this is actually nominative or accusative (see, e.g., Przepiórkowski 1999).

a special adjectival form, as in *po polsku* 'in Polish' or *po męsku* 'manly'. Such arguments are subcategorised for by verbs like CZYTAĆ 'read' (*po angielsku* 'in English') or UMIEĆ 'can (speak)' (*po chińsku* 'in Chinese'):

```
subj{np(str)} + {prepadjp(po,postp)}
```

4. Scope and availability

The dictionary is created manually, with empirical support mainly from the National Corpus of Polish (Przepiórkowski et al. 2010, 2012; `http://nkjp.pl/`) and from other Polish dictionaries; the first versions were based on an unpublished electronic valence dictionary by Świdziński (1998). Each lexical entry is added in two stages: first it is created by a lexicographer, and then verified by one of the (currently) two head lexicographers. In the beginning of March 2014, there were 8587 lexical entries created in the system, including 4009 entries verified by head lexicographers.

It should be noted that lexical entries are understood very broadly here. They are identified by the head lemma, so that valence schemata for different senses, different grammatical aspects or even different values of inherent reflexivity[14] are all included in the same entry. Senses are not distinguished in the current version of Walenty at all (but see the next section). On the other hand, lexical entries contain subentries defined by specific values of aspect and reflexivity; in the version described here, there are 12 132 such subentries, including 5767 already verified. This makes the dictionary already far larger than the previous largest Polish valence dictionary, Polański (1980–1992).

There are 38 645 valence schemata in the dictionary, i.e., 4.5 schemata per lexical entry. But since these are all maximal schemata, with arguments often not strictly obligatory, and one argument may contain a number of morphosyntactic specifications (this is the case for about 12% of schemata), there are usually many possible realisations of each schema, making the dictionary rather comprehensive. The average schema contains almost 3 arguments.

Text versions of the dictionary are available from `http://zil.ipipan.waw.pl/Walenty` on the Creative Commons Attribution ShareAlike licence. In case of the version of 4 March 2014 described here, the downloadable package contains a brief README file (`walenty_03_2014_README.txt`) and two files with the dictionary in the format described here: with all entries (`walenty_03_2014_all.txt`) and with the subset of entries verified by supervising lexicographers (`walenty_03_2014_verified.txt`). Possible realisations of `xp(...)` and corresponding `advp(...)` phrases are defined in `realizations_03_2014.txt`. The verified subset of the dictionary is also available in a perhaps more readable PDF format as `walenty_03_2014_verified.pdf`, together with examples illustrating various subsets of possible realisations of any schema, and a classification of schemata into uncontroversial (*pewna*), colloquial (*potoczna*), vulgar (*wulgarna*), archaic (*archaiczna*), dubious (*wątpliwa*) and downright wrong but attested (*zła*). In the current version of the whole dictionary, out of 38 645 schemata, 34 691 are uncontroversial, 1776 are colloquial, 1689 are dubious, 324 are archaic, 117 are wrong and 48 are vulgar.

5. Outlook

The first version of the dictionary was created within the CESAR project (ICT-PSP PB Pilot Type B, *CEntral and South-east europeAn Resources*, 01.02.2011 – 31.01.2013, `http://cesar.nytud.hu/`, `http://clip.ipipan.waw.pl/CESAR`), a subproject of META-NET, which is also responsible for making publicly available the dictionary of Świdziński (1998), from which Walenty was bootstrapped. Subsequently, the dictionary has been further developed within NEKST (PO IG 1.1.2, 01.04.2009 – 25.02.2014, `http://www.ipipan.waw.pl/nekst/`). Now, one of the largest subprojects of the Polish part of the CLARIN (`http://www.clarin.eu/`) infrastructure, CLARIN-PL (01.04.2013 – 31.12.2015; `http://www.clarin-pl.eu/en/`), is concerned with substantial quantitative and qualitative extensions of Walenty.

First of all, within CLARIN-PL the dictionary will be extended to 15 000 predicates, including 12 000 verbs and 3000 nouns and adjectives. Perhaps more importantly, valence schemata will be paired with semantic argument structures so that corresponding arguments in different schemata for the same meaning of the predicate may be explicitly identified. To give a (classic; e.g., Salkoff 1983) example from English, while two different schemata are at work in the near-synonymous *The garden is swarming with bees* and *Bees are swarming in the garden*, the subject in the former sentence (*The garden*) should be semantically identified with the locative phrase in the latter (*in the garden*), and similarly for the *with*-phrase in the former (*with bees*) and the subject in the latter (*Bees*). Furthermore, the dictionary will be enriched with information about selectional restrictions which, while defeasible (e.g., the subject of BARK is usually – but not necessarily – a canine), are very useful for ranking parses.

The dictionary is also employed at another subproject of CLARIN-PL carried out at ICS PAS, concerned with deep parsing. In particular, Walenty has been converted to the XLE (Crouch et al., 2011) format for the purpose of the LFG grammar and parser of Polish already referred to in §1. above.

Finally, let us also mention a new COST Action, PARSEME (*PARSing and Multi-word Expressions*, 08.03.2013 – 07.03.2017, `http://parseme.eu/`), within which the relation between – and possible coupling or even unification of – Walenty and Polish lexica of multi-word expressions will be considered.

6. Conclusion

We described a new open source valence lexicon of Polish, Walenty, already larger than other such dictionaries for Polish. Its design takes into account various phenomena often

[14] As in, e.g., BRAĆ 'take' and BRAĆ SIĘ 'get down to', the latter with the reflexive marker SIĘ.

ignored in valence dictionaries for other languages, and it employs an explicit definition of *argument position* based on the coordination test.

While already a useful resource, the dictionary is still intensively developed both in terms of its empirical scope and the depth of information it contains. We hope that Walenty becomes an inspiration for similar valence lexica for other languages, eventually leading to a linked multilingual resource.

Acknowledgements

The development of Walenty has been partially financed by the projects mentioned in §5. We would like to thank Bartłomiej Nitoń, the author of the application for developing and accessing Walenty, as well as Jakub Szymczak and Sebastian Żurowski, who supervise everyday lexicographic work, for comments and discussions.

Appendix: Formalism specification

This appendix describes the format of the pure text version of Walenty, as of early March 2014.

Each row contains 4 fields, separated by a colon followed by a space, representing 1) a lemma, 2) negativity, 3) aspect and 4) a valence schema, for example:[15]

```
cisnąć się: _: imperf: subj{np(str)} +
 {xp(locat)}
cisnąć się: _: perf: subj{np(str)} +
 {xp(adl)}
sposób: neg: imperf: {cp(żeby)}
żywić: _: imperf: subj{np(str)} +
 obj{np(str)} + {np(inst)}
żywić: _: imperf: subj{np(str)} +
 obj{np(str)} + {prepnp(do,gen)}
żywić się: _: imperf: subj{np(str)} +
 {np(inst)}
```

In case of verbs with more than one valence schema, each schema is given in a separate row, repeating the lemma, etc., as in case of ŻYWIĆ above.

There are three possible values of negativity: `neg` (negation required; cf. SPOSÓB above), `aff` (negation forbidden; this value is currently not used) and `_` (no constraints on negation).[16] Similarly, the possible values of aspect are: `perf` (perfective), `imperf` (imperfective) and `_` (biaspectual, with the same schema in both aspects; cf. DAROWAĆ above).

The final field contains the schema, i.e., a sequence of arguments separated by `+`. Each argument is a set of morphosyntactic realisations (phrase types) delimited by braces, `{` and `}`, and separated by the semicolon `;` (followed by a space). Symbols indicating the subject (`subj`; at most one in any schema), object (`obj`; at most one), `controller` (at most one) and `controllee` (normally at most one, exceptionally two) immediately precede such set specifications. When two (or, exceptionally, three; see below) such symbols co-occur (e.g., for subject controllers), they are separated by commas (e.g., `subj,controller{np(str)}`). In exceptional cases, two controllers may occur in the same schema; the second is marked as `controller2` and the argument it controls – as `controllee2`. It is possible for an argument to be a controllee of another argument and, at the same time, a controller of yet another argument: `obj,controllee,controller2{infp(_)}`.

The following argument types are defined:

- `np(CASE)` – a nominal phrase (noun phrase, numeral phrase, adjectival phrase of a certain type, etc.) with `CASE` equal to one of the following: `nom`, `gen`, `dat`, `acc`, `inst`, `str`, `part` (cf. §3.6.3.);
- `prepnp(PREP,CASE)` – a prepositional phrase (with a nominal argument of the preposition) parameterised by the preposition lemma (`PREP`) and case (`CASE`; as above and `loc`, but not `part`);
- `adjp(CASE)` – an adjectival phrase with `CASE` values as in `np(CASE)`, with the exception of `part` and with the addition of `pred` (cf. §3.6.3.);
- `prepadjp(PREP,CASE)` – a prepositional phrase with an adjectival argument of the preposition; `CASE` values as above, with the additional value `postp` (cf. §3.6.3.);
- `comprepnp(COMPLEX_PREP)` – a prepositional phrase introduced by a specific complex preposition (cf. §3.6.2.);
- `lexnp(CASE,NUMBER,'LEMMA',MOD)` – a nominal phraseologism (cf. §3.5.) headed by the `LEMMA` which must occur in a given `CASE` and `NUMBER` (with values: `sg` for singular, `pl` for plural, and `_` for any number); `MOD` indicates constraints on the modifiability of the lemma and may be one of the following: `atr`, `natr`, `batr`, `ratr`;
- `preplexnp(PREP,CASE,NUMBER,'LEMMA',MOD)` – a prepositional phraseologism (cf. §3.5.); `PREP` is the lemma of the preposition, and the other parameters refer to the nominal argument of this preposition and have the same meaning as for `lexnp`;
- `fixed('STRING OF CHARS')` – a phraseologism given by a string of characters (see §3.5.);
- `cp(TYPE)` – a subordinate clause of a certain type; CP stands for the complementiser phrase, although when the value of `TYPE` is `int` (interrogative) there is no complementiser involved; otherwise the values of the sole parameter are complementisers or representatives of small classes of complementisers: `że`, `żeby`, `żeby2` (for CPs normally headed by ŻE, but alternatively allowing ŻEBY in negative contexts), `gdy`, `jak`, `kiedy`, `jeśli`, `aż`, `jakby`, `jakoby`, `czy`;
- `ncp(CASE,TYPE)` – a subordinate clause of a certain `TYPE` (whose values are as above) with a nominal correlate bearing a certain `CASE`;
- `prepncp(PREP,CASE,TYPE)` – a subordinate clause of certain `TYPE` with a prepositional-nominal

[15] Only selected valence schemata are shown here for each verb. All rows but one are broken here for typographic reasons.

[16] There are plans to replace `neg` with two values: `mneg` for morphological (immediately preceding, prefixed; cf. Kupść and Przepiórkowski 2002) negation and `sneg` for syntactic (not necessarily immediately preceding) negation.

correlate (for a certain PREPosition and a certain CASE of the nominal pronoun);

- `infp(ASPECT)` – an infinitival phrase headed by a verb with a given ASPECT value: `imperf`, `perf` or – most commonly – any aspect _;
- `xp(SEM)` – an “adverbial” phrase of a given SEMantics (cf. §3.4.): `locat`, `abl`, `adl`, `perl`, `temp`, `dur`, `mod`; possible morphosyntactic realisations of such phrases, including various prepositional realisations, are defined separately;
- `advp(SEM)` – an adverb of a certain semantic type (as above); currently only used in the list of possible realisations of the corresponding `xp(SEM)`;
- `advp(misc)` – a true adverbial phrase (with no prepositional realisations);
- `advp(pron)` – an adverbial pronoun (cf. §3.6.1.);
- `nonch` – a nonchromatic phrase (cf. §3.6.1.);
- `or` – an *oratio recta* (direct speech) argument;
- `refl` – the reflexive marker *się* indicating semantic reflexivity (cf. §3.1.);
- `E` – raised subject (cf. §3.3.) when used as in `subj,controller{E}`; covert expletive subject (cf. fn. 7) when used as in `subj{E}`.

References

Bresnan, J., editor (1982). *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE documentation. `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Academic Press, San Diego, CA.

Kupść, A. and Przepiórkowski, A. (2002). Morphological aspects of verbal negation in Polish. In Kosta, P. and Frasek, J., editors, *Current Approaches to Formal Slavic Linguistics: Proceedings of the Second European Conference on Formal Description of Slavic Languages, Potsdam, 1997*, pages 337–346, Frankfurt am Main. Peter Lang.

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 Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, Istanbul, Turkey. ELRA.

Polański, K., editor (1980–1992). *Słownik syntaktyczno-generatywny czasowników polskich*. Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Cracow.

Postal, P. M. (2004). Chromaticity: An overlooked English grammatical category distinction. In Postal, P. M., editor, *Skepical Linguistic Essays*, pages 138–158. Oxford University Press, Oxford.

Przepiórkowski, A. (1999). *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. dissertation, Universität Tübingen, Tübingen.

Przepiórkowski, A. (2000). Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, 8:151–189.

Przepiórkowski, A., Kupść, A., Marciniak, M., and Mykowiecka, A. (2002). *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Przepiórkowski, A., Skwarski, F., Hajnicz, E., Patejuk, A., Świdziński, M., and Woliński, M. (2014). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, 33. To appear.

Rosenbaum, P. (1967). *The Grammar of English Predicate Complement Constructions*. The MIT Press, Cambridge, MA.

Rouveret, A. and Vergnaud, J.-R. (1980). Specifying reference to the subject: French causatives and conditions on representations. *Linguistic Inquiry*, 11(1):97–202.

Sag, I. A., Gazdar, G., Wasow, T., and Weisler, S. (1985). Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171.

Salkoff, M. (1983). Bees are swarming in the garden: A systematic synchronic study of productivity. *Language*, 59(2):288–346.

Saloni, Z. and Świdziński, M. (1998). *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 4th (changed) edition.

Świdziński, M. (1992). *Gramatyka formalna języka polskiego*, volume 349 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Świdziński, M. (1998). Syntactic dictionary of Polish verbs. Version 3a. Unpublished manuscript, University of Warsaw.

Warren, D. H. D. and Pereira, F. C. N. (1980). Definite clause grammars for language analysis — a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13:231–278.

Woliński, M. (2004). *Komputerowa weryfikacja gramatyki Świdzińskiego*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.