Extended phraseological information in a valence dictionary for NLP applications

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa
{adamp, hajnicz, aep, wolinski}@ipipan.waw.pl

Abstract

The aim of this paper is to propose a far-reaching extension of the phraseological component of a valence dictionary for Polish. The dictionary is the basis of two different parsers of Polish; its format has been designed so as to maximise the readability of the information it contains and its re-applicability. We believe that the extension proposed here follows this approach and, hence, may be an inspiration in the design of valence dictionaries for other languages.

1 Introduction

The starting point of the work reported here is Walenty, a valence dictionary for Polish described in Przepiórkowski et al. 2014 and available from http://zil.ipipan.waw.pl/Walenty (see §1.1). Walenty contains some valence schemata for verbal idioms; e.g., one of the schemata for kuc ‘force’ says that it combines with a nominal subject, a nominal object and a prepositional phrase consisting of the preposition na ‘on’ and the accusative singular form of the noun pamięć ‘memory’ – this represents the idiom ktoś kuje coś na pamięć ‘somebody rote learns something’, lit. ‘somebody forces something onto memory’. The current formalism handles various kinds of verbal phraseological constructions (cf. §1.2), but also has clear limitations. For example, in Polish one may welcome somebody “with arms wide open”, and the current formalism makes it possible to express the “welcome with arms + modifier” part, but not the specifics of the allowed modifier, namely, that it is the adjective meaning “open”, possibly itself modified by an intensifying adverb (cf. §1.3 for details).

The aim of this paper is to propose a new subformalism of Walenty for describing phraseological valence schemata (see §2). To the best of our knowledge, Walenty is already rather unique among valence dictionaries for various languages in paying so much attention to phraseological constructions (among its other rare or unique features), and at the same time it is practically employed in parsing by two different parsers of Polish (cf. §1.1). We believe that these traits make the current proposal to further extend the underlying formalism potentially interesting to the wider audience.

1.1 Walenty

Walenty is a valence dictionary which is meant to be both human- and machine-readable; in particular, it is being employed by two parsers of Polish, Swigra (an implementation of a Definite Clause Grammar description of fragments of Polish syntax; Woliński 2004) and POLFIE (an implementation of a Lexical Functional Grammar description of fragments of Polish; Patejuk and Przepiórkowski 2012). 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.

Each verb is assigned a number of valence schemata and each schema is a set of argument specifications. Walenty is explicit about what counts as an argument: if two morphosyntactically different phrases may occur coordinated in an argument position, they are taken to be 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

---

1 As long as the dictionary contains mostly morphosyntactic information, we avoid using the term valence frame.
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)).

\[ \text{subj} \{ \text{np(str)} \} + \text{obj} \{ \text{np(str); cp(int)} \} + \{ \text{np(dat)} \} \]

There are three argument positions (separated by +) given in this schema: a subject, an object and an
additional argument whose grammatical function is not specified but whose morphosyntactic realisation
is described as a dative nominal phrase (np(dat)). The subject is also described as a nominal phrase
(NP), but its case is specified as structural, i.e., potentially depending on the syntactic context. In Polish,
such subjects are normally nominative, but – according to some approaches – they bear the accusative
case when they are realised as numeral phrases of certain type. Similarly, the nominal realisation of the
object is specified as structural, as it normally occurs in the accusative, unless it is in the scope of verbal
negation, in which case it bears the genitive. Crucially, though, the object is specified here not just as an
NP, but also alternatively as an interrogative (int) clausal argument (cp, for complementiser phrase). A
parser may take this information into account and properly analyse a sentence with unlike coordination
like the one involving tłumaczyć ‘explain’, given in the previous paragraph.

Other features of the formalism of Walenty worth mentioning here, and described in more detail in
Przepiórkowski et al. 2014, are: the representation of control and raising (cf. Landau 2013 and references
therein), specification of semantically defined arguments (e.g., locative, temporal and manner), with their
possible morphosyntactic realisations defined externally (once for the whole dictionary), handling of
various kinds of pronominal arguments, and other types of non-morphological case specifications (apart
from the structural case). While there is no explicit semantic information in the dictionary at the moment
(apart from such semantically defined arguments and control information), i.e., no subdivision of verbal
lemmata into senses and no semantic role information, Walenty is currently being extended to include
such a semantic layer.

1.2 Phraseology in Walenty

Two features of the Walenty formalism deal with multi-word expressions. The simpler one is concerned
with complex prepositions such as w kwestii ‘in (some) matter’, na temat ‘on (some) topic’, z powodu
‘because of’ (lit. ‘of reason’), etc. Unlike in case of usual prepositional phrases, parameterised with the
preposition lemma and the grammatical case it governs (e.g., prepnp(z,inst) for a prepositional
phrase headed by z ‘with’ and taking an instrumental NP), such complex prepositions 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:

\[ \text{subj} \{ \text{np(str)} \} + \{ \text{comprepnp(z powodu)} \} \]

Other, more clearly idiomatic arguments are currently specified as fixed, lexnp and preplexnp.
Phrases of type fixed again have just one parameter: the exact orthographic realisation of the phrase;
see the following schema for zabić ‘beat’ (as in He beat them to a pulp), with na kwaśne jabłko meaning
literally ‘into sour apple’:

\[ \text{subj} \{ \text{np(str)} \} + \text{obj} \{ \text{np(str)} \} + \{ \text{fixed(‘na kwaśne jabłko’)} \} \]

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 necessarily
headed by krew ‘blood’ in the singular, and the NP may contain modifiers (cf. atr), illustrates this:

\[ \text{subj} \{ \text{lexnp(str,sg,‘krew’,atr)} \} + \{ \text{preplexnp(w,loc,pl,’żyła’,ratr)} \} \]

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 żyły*
(singular ‘vein’).

The third possible value of the modifiability parameter is natr, for lexicalised arguments that cannot
involve modification. The following schema for *Zmarzać* ‘get cold, freeze’ handles the idiom *Zmarzać
na kość* ‘freeze to the marrow’ (lit. ‘freeze to (the) bone’): note that *kość* ‘bone’ cannot be modified here,
as illustrated by the infelicitous *Zmarzł na gołąb/hwardą kość* *(He) froze to (the) naked/hard bone*:

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

Finally, batr (‘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 we cale taka stara* ‘With my own eyes I saw her smile and that she wasn’t so old at all’:\n
```
subj{np(str)} + {np(str); ncp(str,że)} +
{preplexnp(na,acc,pl,'oko',batr)}
```

We will see below that a more expressive – and more general – scheme for the representation of phrase-
ological valence is needed.

1.3 Limitations

A number of problems were identified with the formalism of Walenty as it was described in
Przepiórkowski et al. 2014 and summarised above. To start with the simplest cases, it is a simplification to
say that complex prepositions (comprepnp above) are internally unanalysable and always combine with
genitive NPs. For example, while *z powodu* ‘because of’ cannot occur without any additional dependents,
it is sufficient for the nominal form *powodu* ‘reason’ to be modified by an appropriate adjectival form for
the whole expression to be complete, e.g., *z tego powodu* ‘because of this’, lit. ‘of this reason’, *z ważnego powodu* lit. ‘of important reason’, etc. This is not a general feature of such complex prepositions, though.
For example, *w trakcie* ‘during’, lit. ‘in (the) course’, must combine with a genitive NP and the nominal
form *trakcie* ‘course’ cannot be modified by an adjective (*w tym trakcie* lit. ‘in this course’).

Second, it is useful for parsers to have more grammatical information about lexically fixed arguments;
for example, *fixed(‘na kwaśne jabło’) clearly has the internal structure of a prepositional phrase.

Third, the current formalism allows for only two types of phraseological phrases to be specified in
more detail: nominal (lexnp) and prepositional (preplexnp). While not so frequent, other kinds
of idiomatic arguments also occur, including adjectival, adverbial and infinitival. For example, one of
the idiomatic uses of *mieć ‘have’ is *mieć przeszkapane* ‘be in the doghouse, be in deep shit’, lit. ‘have
(it) screwed’, with an appropriate form of the adjective *przeszkapany* ‘screwed’. Similarly, the verb
dysząć ‘pant’ may be argued to optionally require the adverb *ledwo* ‘barely’, as in *ledwo dysząć* ‘hardly
breathe’. Also, *kałas Área ‘lie down* typically occurs with the infinitival form of *spać* ‘sleep’. The current
formalism may describe such requirements only using the rather inflexible *fixed* notation.

Finally, and perhaps most importantly, modification possibilities within phraseological arguments
are far richer than the four symbols atr, natr, ratr and batr could repre-
sent. One case in point is the idiom already mentioned in §1, namely, *witać kogoś z ot-
wartyymi ramionami* ‘welcome somebody with open arms’. The best representation of the idi-
matic argument realisation *z otwartyymi ramionami* ‘with open arms’ is either *fixed(‘z
otwartyymi ramionami’) or preplexnp(z,inst,pl,’ramię’,atr) or perhaps
preplexnp(z,inst,pl,’ramię’,ratr). The first of these does not allow for any modi-
fication, so it does not cover *z szeroko otwartyymi ramionami* ‘with wide open arms’, etc. The second
mentions the possibility of modifiers of *ramionami* (which is the instrumental plural form of the noun
ramię ‘arm’), but does not constrain this possibility to (a class of) agreeing adjectives, so it would also
cover the non-phraseological *z otwartyymi ramionami Tomka* ‘with Tomek’s open arms’, lit. ‘with open

\[The schema is split into two lines solely for typographic reasons.\]
arms Tomek.gen’. Also, it makes such modification optional, while *witać kogoś z ramionami* ‘welcome somebody with arms’ is at best non-phraseological. Finally, the third possibility makes modification obligatory, but the current meaning of *ratr* is (for good reasons) constrained to possessive modifiers, so it again does not cover the case at hand, where adjectival modification is present.

### 2 Extended phraseological valence

One more objection against the current subformalism for phraseological arguments, apart from those adduced in the preceding subsection, is that it contains some *ad hoc* notation, whose meaning is not transparent. The best examples of this are *ratr* (a possessive modifier, not just any modifier) and *batr* (the modifier is a form of swój ‘self’s’ or *własny* ‘own’). In contrast, we propose a formalism which is not only more expressive, so as to deal with the limitations mentioned in §1.3, but also more transparent. As we will see below, the price to pay for this more expressive and principled formalism is that some argument specifications become more complex.

#### 2.1 Categories of phraseological arguments

The first proposed generalisation is to replace category-specific symbols *lexnp* and *preplexnp* with the single *lex*, whose first parameter indicates the category of the phraseological argument. For example, the *lexnp*(str,sg,’krew’,atr) specification given above could be replaced by *lex*(np(str),sg,’krew’,atr), and *preplexnp*(w,loc,pl,’żyła’,ratr) – by *lex*(prepnp(w,loc),pl,’żyła’,ratr). Note that the first parameter of *lex* expresses more than just the grammatical category of the argument – it is the same specification of the morphosyntactic realisation – here, np(str) and prepnp(w,loc) – as used for non-phraseological arguments. Any (non-lexical, i.e., not *lex*, etc.) morphosyntactic specification used in Walenty could be used here, including also adjectival, adverbial and infinitival.

For example, for *mieć przechlapane* ‘be in the doghouse, be in deep shit’ mentioned above, where the adjective *przechlapany* must occur in the singular neuter accusative and may be modified by an intensifying adverb (*mieć kompletnie przechlapane* lit. ‘have (it) completely screwed’), the appropriate argument realisation could be described as: *lex*(adjp(acc),n,sg,’przechlapany’,atr). Similarly, the adverb *ledwo* ‘barely’ combining with forms of *dysze* ‘pant’ could be described as *lex*(advp(misc),’ledwo’,natr) (recycling the existing morphosyntactic specification advp(misc) for true adverbial phrases), and the infinitival form of *spać* ‘sleep’ co-occurring with *księg* ‘lie down’ – as *lex*(infp(imperf),’spać’,natr) (again, reusing the standard notation for imperfective infinitival phrases, infp(imperf)).

#### 2.2 Modification patterns

The most profound generalisation concerns, however, the specification of modification patterns within idiomatic arguments. We propose to retain three basic indicators, namely, *natr* (no modification possible), *atr* (modification possible) and *ratr* (modification required), but – in the case of the last two – additional information must be given specifying the kind of modification that is allowed or required. Additionally, *atr* and *ratr* are envisaged as variants of *atr* and *ratr* with the additional constraint that at most one such modifier may be present.

For example, instead of *preplexnp*(z,inst,pl,’ramię’,ratr) for *z otwartymi ramionami* ‘with open arms’, the following argument specification could be given, explicitly mentioning that the only possible modifier of *ramionami* ‘arms’ is an adjectival phrase (and not, say, a genitive modifier):\(^4\)

\[
\{lex\{prepnp\{z,inst\},pl,’ramię’,ratr\{adjp\{agr\}\}}\}
\]

Note that morphosyntactic specifications of possible or required modifiers are enclosed in curly brackets, just as in case of direct arguments of verbs, and for the same reason: sometimes multiple morphosyntactic realisations are possible and may be coordinated, which indicates that they occupy the same syntactic position. An example of this is the expression *komuś cierpie skóra na myśl o czymś* ‘something makes

\(^3\)In case of *ratr*, the obligatory nature of modifier together with this constraint mean that exactly one modifier must occur.

\(^4\)The symbol *agr* indicates agreeing case here.
somebody’s flesh creep’, lit. ‘somebody.DAT creeps skin.NOM on (the) thought.ACC about something.LOC’.

The argument na myśl o czymś ‘on (the) thought of something’ may be realised in at least three ways: as a prepositional phrase as here (preppn(o,loc)), as a finite clause introduced by the complementiser źe ‘that’ (cp(że); e.g., komuś cierpienie skóra na myśl, że (to się stało) lit. ‘somebody.DAT creeps skin.NOM on (the) thought.ACC that (this happened.REFL)’), or as a so-called correlative phrase which shares features of the first two realisations, e.g., na myśl o tym, że (to się stało) lit. ‘on (the) thought about this.LOC that (this happened.REFL)’ (preppncp(o,loc,że)). Such a disjunctively specified modification possibility may be expressed as follows (with the line broken for typographic reasons and indented for readability):

\{lex(preppn(na,acc),sg,'myśl',
  rattr({preppn(o,loc);cp(że);preppncp(o,loc,że)}))\}

This specification is still incomplete: the noun myśl may also be modified by an adjectival form, e.g., the adjectival pronoun tę, as in skóra mi cierpienie na tę myśl ‘this thought makes my flesh creep’, lit. ‘skin.NOM me.DAT creeps on this.ACC thought.ACC’. This means that adjp(agr) must be added as a possible modifier type. But the status of this modifier type is different than the three modifier types given above: no two of these three phrases can co-occur unless they are coordinated, but any of them can co-occur (and cannot be coordinated) with adjp(agr), e.g., skóra mi cierpienie na samą myśl o tym ‘the sheer thought makes my flesh creep’, lit. ‘skin.NOM me.DAT creeps on sheer.ACC thought.ACC about this.LOC’. Hence, the two kinds of modification possibilities are analogous to two different arguments (here, actually, dependents) of a predicate occupying different syntactic positions, and the same notation could be used to specify them, with the + symbol:

\{lex(preppn(na,acc),sg,'myśl',
  rattr({preppn(o,loc);cp(że);preppncp(o,loc,że)} + {adjp(agr)}))\}

Such argument specifications involving lex may get even more complex due to the fact that lex may occur inside modification specifications of another lex, as in the following description of arguments such as z otwartymi ramionami ‘with open arms’, more accurate than the one given at the beginning of this subsection:

\{lex(preppn(z,inst),pl,'ramię',
  rattr(lex(adjp(agr),agr,agr,'otwarty',natr))}\}

Note that lex(adjp(agr),agr,agr,'otwarty',natr) replaces adjp(agr) within rattr and it specifies that not just any agreeing adjective phrase may modify the nominal form rękami ‘arms’, but only the simple adjective phrase consisting of an agreeing form of otwarty ‘open’ does.\(^5\)

As discussed above, this is still not a complete description of the range of possibilities here, as the adjective otwarty itself may itself be modified (contrary to the natr specification above), namely, by the adverb szeroko ‘wide’. A closer approximation is given below:\(^6\)

\{lex(preppn(z,inst),pl,'ramię',
  rattr(lex(adjp(agr),agr,agr,'otwarty',natr)
  atr(lex(advp(misc),'szeroko',natr)))\})

Note that this extension, with the possibility of lex recursion (of the centre-embedding type; Chomsky 1959), makes the language of schema specifications properly context-free.

### 2.3 Complex prepositions and fixed arguments

As noted in §1.3 above, the notation comprep(...) (e.g., comprepnp(z powodu)) is not sufficient to model various combinatoric properties of complex prepositions: some of them (e.g., z powodu ‘because of’) may combine with a genitive NP or an agreeing adjective phrase, others (e.g., w trakcie ‘during’) may only co-occur with an NP, etc. We propose to reuse the lex notation to describe complex prepositions in a more satisfactory manner. For example, one of valence schemata of umartwać

---

\(^5\)Recall that if adjp(...) is the first parameter of lex, the next two indicate gender and number, hence the two agrs after adjp(agr).

\(^6\)Sporadic examples of z bardzo szeroko otwartymi ramionami ‘with arms very wide open’ (note the additional bardzo ‘very’) may be found on the internet, but we draw the line here and do not model such occurrences.
się ‘mortify oneself’, which specifies such a `comprepnp(z powodu)` argument, could specify it the following way instead (with ‘_’ indicating any number here):

\[
\{\text{lex}(\text{precomp}(z,\text{gen}),_,\text{‘powód’},\text{ratr}(\{\text{np}(\text{gen});\text{ncp}(\text{gen},\text{że})\} + \{\text{adjp}(\text{agr})\}))\}
\]

In contrast, the requirement of \(w\) \(\text{trakcie}\) (where the noun must be in the singular and adjectival modification is not possible) could be spelled out this way:

\[
\{\text{lex}(\text{precomp}(w,\text{inst}),\text{sg},\text{‘trakτ’},\text{ratr}(\{\text{np}(\text{gen});\text{ncp}(\text{gen},\text{że})\}))\}
\]

As far as we can see, all complex prepositions could be described this way. On the other hand, there are cases of `fixed` arguments which could not be so described simply because they contain forms which cannot be specified with a reference to a lemma and morphosyntactic categories such as case or number. One example is the argument of the form `fixed(‘dęba’)`, with the form \(dęba\) ‘oak\._\text{GEN}\) of the noun \(dąb\) ‘oak’, which \(\text{co-occurs}\) with forms \(\text{stanąć} dęba ‘rear’\) (of a horse), \(\text{lit.} ‘\text{stand oak}\). However, since the form \(dęba\) is not used in contemporary Polish (the contemporary genitive of \(dąb\) is \(dębu\)), this idiomatic argument cannot be expressed as \(\text{lex(np(\text{gen}),\text{sg,‘dąb’},\text{natr})}\), as then parsers would try to analyse the non-phraseological (at best) sequence \(\text{stanąć dębu}\) as idiomatic. Instead, we propose to extend the `fixed` notation and add a parameter describing the general morphosyntax of such an argument, e.g., \(\text{fixed(np(\text{gen}),‘dęba’)}\).

However, such `fixed` arguments with forms not attested in contemporary Polish are extremely rare and we envisage that almost all other specifications currently involving `fixed` can be translated into perhaps more precise specifications involving `lex`. For example, \(\text{fixed(‘na kwaśne jabłko’)\), used in zbić na kwaśne jabłko ‘beat into a pulp’, literally meaning ‘into sour apple’, may be specified as follows:

\[
\{\text{lex}(\text{precomp}(\text{na, acc}),\text{sg,‘jabłko’}, \text{ratr}(\{\text{lex}(\text{adjp}(\text{agr}),\text{agr, agr,‘kwaśny’},\text{natr})\}))\}
\]

### 2.4 Syntactic sugar

There are two types of syntactic sugar that we would like to propose together with the above extensions. First of all, while it seems that the `comprepnp` notation for complex prepositions should be replaced by `lex`, we propose to leave such `comprepnp` specifications in the dictionary proper and define them in terms of `lex` specifications separately. The reason for this is that once a complex preposition occurs in the specification of a valence schema, it has the tendency to occur in many schemata; for example, `comprepnp(z powodu)` occurs in 126 schemata in the March 2014 version of Walenty. Replacing all these occurrences with the considerably more complex `lex` specification given above would diminish the readability of the dictionary. Instead, `comprepnp(‘z powodu’)\) (perhaps with inverted commas, to increase notational consistency) should be left in particular schemata and it should be defined in terms of `lex` once for the whole dictionary.\(^7\)

The other kind of abbreviatory notation is best illustrated with idiomatic arguments which have so far required the `batr` modification indicator. Recall that `batr` means that a given noun may be modified by forms of either of the two adjectives meaning ‘self’s, own’, i.e., forms of \(\text{swój}\) and \(\text{własny}\). One example is the verb `spróbować ‘try’, which may combine with the expression `swoich/własnych sił ‘one’s power’ rendering `spróbować swoich/własnych sił w czymś ‘try one’s hand at sth’, lit. ‘try one’s powers in something’. Given the notation introduced so far, this argument would have to be specified as follows:

\[
\{\text{lex}(\text{np}(\text{gen}),\text{pl,‘siła’}, \text{ratr}(\{\text{lex}(\text{adjp}(\text{agr}),\text{agr,agr,’swój’},\text{natr})\} + \{\text{lex}(\text{adjp}(\text{agr}),\text{agr,agr,’swój’},\text{natr})\}))\}
\]

This is not only hardly readable, but also misses the generalisation that the two possible modifiers are just the same kinds of adjectival phrases differing only in the lexical realisation of the adjective meaning ‘self’s, own’. We propose abbreviating such specifications as follows, with the use of `OR`:

\[
\{\text{lex}(\text{np}(\text{gen}),\text{pl,‘siła’}, \text{ratr}(\{\text{lex}(\text{adjp}(\text{agr}),\text{agr,agr,’swój’},\text{natr})\}))\}
\]

\(^7\)This move would be analogous to specifying externally morphosyntactic realisations of semantically defined arguments such as `xp(locat)` or `xp(temp)` (see Przepiórkowski et al. 2014 for details).
While we could have reintroduced the $batr$ notation to make this even more readable, this symbol is not used in Walenty uniformly, so it would make sense to replace it with more explicit notation involving $lex$. In particular, some argument specifications mentioning $batr$ actually allow only for forms of $własny$, not swój. This is, e.g., the case with $doręczyć$ ‘hand (over)’, as in $doręczyć do rąk własnych$ (but not $doręczyć do rąk swoich$) ‘deliver as hand delivery’, lit. ‘hand to own hands’, where the specification of the relevant argument in terms of $lex$ explicitly mentions only one of these two adjectives:

$$\{lex\text{(prepnp\{do,gen\},pl,\text{\'ręka\}'\},}$$

$$ratr\{\{lex\text{(adjp\{agr\},agr,agr,\text{\'własny\'},natr)}\}\}\}$$

Note finally that this shorthand notation is useful not just in cases involving $batr$ in the old formalism. One case in point is the expression $coś strzelić komuś do głowy$ ‘something came over somebody’, lit. ‘something. nom shot sambody.dat to head.gen’, where the form of $głowa$ ‘head’ may be replaced by analogous forms of other nouns with the same meaning, including $leb$ and $lepetyna$, as specified below: 8

$$\{lex\text{(prepnp\{do,gen\},sg,XOR\{\text{\'głowa\'},\text{\'łeb\'},\text{\'lepetyna\}'\},atr\{\text{adjp\{agr\}}\})}\}$$

3 Case study

In the current (as of the end of April 2014) version of Walenty, there are 7 complex prepositions used 691 times (in the schemata of 367 different verbs), 17 fixed phrases used 82 times (36 verbs), 177 $lexnp$ phrases used 686 times (393 verbs) and 238 $preplexnp$ phrases used 1133 times (496 verbs). The last two contain 1182 $natr$ parameters, 217 $ratr$ parameters, 40 $batr$ parameters and 406 $atr$ parameters. Summing up, there are 439 different lexicalisations used in 2567 schemata of 659 verbs. This means that the representation of idiomatic schemata is already non-negligible, and it is bound to increase, as more emphasis is put on such schemata in the current development of Walenty. Hence, the proposed changes – if adopted – will involve substantial interference into an existing resource, and may have a potentially adverse impact on the development of the lexicon, if the formalism proves to be too difficult for lexicographers. This section describes an experiment investigating this issue.

We selected 84 schemata of 36 verbs and asked two main lexicographers involved in the development of Walenty to rewrite these schemata using the proposed formalism. The schemata include 38 fixed arguments, 10 $comp\text{prepnp}$ arguments, 17 $lexnp$ arguments and 28 $preplexnp$ arguments. The last two contain 22 $natr$ parameters, 6 $ratr$ parameters, 4 $batr$ parameters and 8 $atr$ parameters. The schemata were selected manually, taking into account their frequency, diversity of types of lexicalisations and their parameters, as well as the expected difficulty of rewriting them. This is the reason for the over-representation of fixed phrases, which need to be completely reanalysed. We chose multiple schemata for the same verb, to give lexicographers the possibility to join them into a single schema, given the more expressive new formalism. In particular, all 12 lexicalised schemata for the verb $stać$ ‘stand’ were selected for the experiment.

The two lexicographers worked on the textual format of the dictionary (cf. http://zil.ipipan.waw.pl/Walenty) without any support from a tool verifying the syntax of the schemata, etc. Correspondingly, when comparing their results, we ignored purely syntactic errors, including differences in bracketing etc., as such errors can be prevented by such a dedicated tool.

After ignoring such trivial differences, 34 of 84 schemata were found to be encoded differently by the two lexicographers. The differences included 3 cases of a wrong lemma, 5 cases of different values of grammatical categories of case, number, etc., and 7 differences concerning the introduction of new non-lexicalised arguments or merging schemata on the basis of the coordination criterion mentioned in §1.1. These differences are not directly connected with the proposed changes of the formalism for lexicalisations. Moreover, 9 differences concerned using $(r)\text{atr}$ instead of $(r)\text{atr1}$ (cf. §2.2) where a single realisation of a modifier is possible, as in the following (correct) argument specification for $przemarząć$ ‘freeze’ surfacing in $przemarząć do szpiku kości$ ‘freeze to the bone’, lit. ‘freeze to (the) marrow (of) bone(s)’:

8This argument specification uses XOR instead of OR as only one of the lexemes meaning ‘head’ may be used at a time, unlike in the previous case, where forms of both swój and $własny$ could be used simultaneously: $spróbować swoich własnych$ $sił w czymś$ lit. ‘try one’s own forces in something’.
Since koścımust appear exactly once, ratsl instead of rats should be used here. Obviously, guidelines for lexicographers should emphasise this point.

Finally, 17 differences concerned core aspects of the new formalism, such as different modification patterns (5), the lack of the morphosyntactic type for fixed (3) and an incorrect specification of the morphosyntactic type of lex, e.g., lack of aspect of infp (2). We judged 6 of them as considerably difficult cases of fixed lexicalisations, rewriting of which was not at all obvious. One such difficulty concerned an idiomatic use of wyrób ‘exit’ as in ktoś wyszedł za kogoś za mąż ‘somebody married somebody else’ (of a woman marrying a man, not the other way round), lit. ‘somebody.NOM exited prep9 somebody.ACC prep husband’. The problem lies in the za mąż ‘PREP husband’ part, where mąż could be analysed as the regular nominative, but then it would be unexpected that the preposition za occurs with a nominative noun (it normally combines with the accusative and the instrumental), or as an idiosyncratic accusative form only occurring in this idiom, similarly to dęba mentioned above occurring in stanąć dęba ‘rear’ – in this case the exceptional use of fixed would be justified, even though the use of fixed was explicitly discouraged in this experiment. The two specifications of the argument za mąż given by the two lexicographers are cited below10.

{fixed(prepnp(za,acc),’za mąż’)}
{lex(prepnp(za,nom),sg,’mąż’,nattr)}

In summary, we feel that the experiment showed that the new formalism is relatively clear and can be learnt by lexicographers, given some training. Support provided by a dedicated lexicographic tool should reduce the number of errors, especially syntactic inconsistencies. On the other hand, the experiment confirmed that some of the most difficult lexicalisations are those currently marked as fixed, and they clearly require special attention.

4 Discussion and conclusion

Many valence dictionaries mark some valence schemata as idiomatic – this is true, e.g., of the VALBU valence dictionary for German developed at the Institut für Deutsche Sprache (Schumacher et al. 2004; http://hypermedia.ids-mannheim.de/evalbu/), of the VALLEX dictionary of Czech developed at the Charles University in Prague (Lopatková et al. 2006; http://ufal.mff.cuni.cz/vallex/), as well as some previous valence dictionaries of Polish, including Polański 1980–1992 and the dictionary that was used to bootstrap the first version of Walenty, i.e., Świdziński 1998. However, we are not aware of another valence dictionary that would explicitly describe lexicalised arguments at the same level of detail as the version of Walenty presented in Przepiórkowski et al. 2014. Regardless of this level of detail, though, the formalism employed in that version suffers from a number of problems discussed in §1.3, limiting its ability to describe phraseological constructions precisely.

In this paper, we propose a far-reaching extension of the formalism of Walenty, making it possible to describe the syntactic structure of a lexicalised argument to any necessary depth. As noted in §2.2, the proposed extension makes the description language properly context-free, but this does not seem to be a problem for parsers employing the valence dictionary. On the contrary, as the parsers of Polish become more sophisticated and are developed with full semantic parsing in view, they need precise description of valence schemata that makes it possible to reliably distinguish idiomatic arguments from non-idiomatic compositional constructions.

The need for such deeper description of phraseological arguments in valence dictionaries has occasionally been expressed in the literature, e.g., by Žabokrtský (2005, 65–66, fn.20), who notes that “[i]n case of multiword parts of phrasemes, the tree (and not only the sequence of forms) representing this part should be ideally captured in the lexicon…” . This makes us hope that the current proposal will prove interesting also for the developers of valence lexica for languages other than Polish.

9The preposition za has a number of different uses and may be translated as ‘behind’, ‘for’, ‘per’, ‘by’, ‘as’, etc.
10Intuitively, the first representation seems appropriate to us, but we see no strong arguments supporting this intuition.
References

Noam Chomsky. 1959. On certain formal properties of grammars. *Information and Control* 2, 137–167.

Idan Landau. 2013. *Control in Generative Grammar: A Research Companion*. Cambridge: Cambridge University Press.

Markéta Lopatková, Zdeněk Žabokrtský, and Karolína Skwarska. 2006. Valency Lexicon of Czech Verbs: Alternation-Based Model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC* 2006, pages 1728–1733, ELRA, Genoa.

Agnieszka Patejuk and Adam Przepiórkowski. 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, ELRA, Istanbul, Turkey.

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

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. 2014. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC* 2014, pages 2785–2792, ELRA, Reykjavik, Iceland.

Helmut Schumacher, Jacqueline Kubczak, Renate Schmidt, and Vera de Ruiter. 2004. *VALBU – Valenzwörterbuch deutscher Verben*, volume 31 of *Studien zur deutschen Sprache. Forschungen des Instituts für Deutsche Sprache*. Tübingen: Narr.

Marek Świdziński. 1998. Syntactic Dictionary of Polish Verbs. Version 3a, unpublished manuscript, University of Warsaw.

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

Zdeněk Žabokrtský. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. dissertation, Charles University, Prague.