PHRASEOLOGY IN TWO SLAVIC VALENCY DICTIONARIES:
LIMITATIONS AND PERSPECTIVES

Adam Przepiórkowski: Institute of Computer Science of the Polish Academy of Sciences and University of Warsaw (adamp@ipipan.waw.pl)
Jan Hajic: Charles University in Prague, Faculty of Mathematics and Physics (hajic@ufal.mff.cuni.cz)
Elżbieta Hajnicz: Institute of Computer Science of the Polish Academy of Sciences (hajnicz@ipipan.waw.pl)
Zdeňka Urešová: Charles University in Prague, Faculty of Mathematics and Physics (uresova@ufal.mff.cuni.cz)

Abstract

Phraseological components of valency dictionaries for two West Slavic languages are presented, namely, of the PDT-Vallex dictionary for Czech and of the Walenty dictionary for Polish. Both dictionaries are corpus-based, albeit in different ways. Both are machine-readable and employable by syntactic parsers and generators. The paper compares the expressive power of the phraseological subformalisms of these dictionaries, discusses their limitations and makes recommendations for their possible extensions, which can be possibly applied also to other valency dictionaries with rich phraseological components.

1. Introduction

Phraseological dictionaries contain information about phraseological expressions, that is, roughly, combinations of words whose meaning is to some extent unpredictable from the meaning and general properties of words occurring in them and from the productive rules of the grammar. They range from collocations such as strong tea (it is unpredictable that one does not rather say powerful tea with the same meaning) through idioms such as kick the bucket ‘die’ to clichés such as The fat is in the fire ‘trouble is about to start’. Valency dictionaries contain information about arguments of predicates (mostly verbs, but sometimes also lexemes belonging to other parts of speech, e.g. nouns and adjectives). For example, a valency dictionary for English may contain information that the verb sell combines with up to four semantic arguments, let us
call them (after VerbNet; http://verbs.colorado.edu/verb-index/) agent, theme, recipient and asset, as in John sold Mary a car for $200, where John is the agent, Mary is the recipient, a car is the theme and $200 is the asset. Moreover, such a dictionary will specify that, in an active sentence, the agent is a nominal (NP, for noun phrase) subject, the theme is an NP object, the recipient may also be realised as an NP (as in the example above) or as a prepositional phrase headed by to (i.e. by PP[to], e.g. to Mary), and that the optional asset is syntactically realised as a PP[for] (for $200 in the example).

The need for a dictionary combining phraseological and valency information has long been recognised in Slavic linguistics and elsewhere, especially, in the work of Igor Mel’čuk and his colleagues starting in the 1960s and culminating in the development of the concept of an Explanatory Combinatorial Dictionary (ECD) and the publication of ECDs for Russian (Mel’čuk and Zholkovsky 1984) and French (Mel’čuk et al. 1984, 1988, 1992, 1999). An example of the kind of information provided in an ECD entry for the idiom pull the wool over someone’s eyes ‘deceive someone’, given in Mel’čuk 2012: 43, is shown in Figure 1.

**PULL THE WOOL OVER [N₁’s] EYES, verbal idiom**

**Definition**

X pulls the wool over Y’s eyes:
‘X tries to deceive Y in order to hide from Y what X is really doing’.

**The surface-syntactic structure**

![Diagram of the structure](attachment:image.png)

**Government pattern**

<table>
<thead>
<tr>
<th>X ← I</th>
<th>Y ← II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. N</td>
<td>1. of N</td>
</tr>
<tr>
<td></td>
<td>THE EYES OF N</td>
</tr>
<tr>
<td></td>
<td>©possessive</td>
</tr>
<tr>
<td>2. N’s</td>
<td>N’s EYES</td>
</tr>
<tr>
<td></td>
<td>©determin</td>
</tr>
<tr>
<td>3. Aₜ(poss)(N)</td>
<td>Aₜ(poss)(N) EYES</td>
</tr>
</tbody>
</table>

**Examples**

Don’t pull the wool over foreigners’ eyes!

He tried to pull the wool over my (John’s) eyes.

**Figure 1:** Relevant parts of an ECD entry for pull the wool over someone’s eyes.
This entry provides the semantics of the idiom in a constrained subset of English, its syntactic structure (as a dependency tree), the valency pattern and some examples. In particular, in order to provide an exhaustive description of this idiom, valency and phraseological information must be combined: some positions within the idiom (marked as X and Y in the definition) are open and may be filled by any nominal phrases satisfying appropriate selectional restrictions (both typically express humans, etc.), while other parts are strongly lexicalised and are filled by forms of specific lexemes, which do not exhibit their usual meanings in this idiom (pull, wool, over, eye). While this entry illustrates the need to combine phraseological and valency information in a single entry, ECD has a much broader scope and includes both: phraseological expressions which do not have any open valency positions – e.g. clichés such as Money isn’t everything or No parking – and non-phraseological predicates, as in usual valency dictionaries.

The need for a dictionary combining aspects of valency and the syntax of phraseological expressions was also recognised – albeit a little later, in the late 1970s – in the so-called “lexicon-grammar” approach that is associated with the name of Maurice Gross (e.g. Gross 1984). In this approach, all lexical information is organised in the form of matrices (i.e. tables) which contain data about syntactic and morphosyntactic aspects of predicates and certain types of phraseological expressions. Since all specifications are natively created in a machine-readable format, lexicon-grammar matrices have been used in computational linguistics, e.g. by Gardent et al. 2005 and Tolone and Sagot 2011.

It has now become common for human-readable valency dictionaries to include some phraseological information, but this information is often unstructured and mostly consists of the meaning of the idiom or an example of its use. Thus, the German valency dictionary VALBU (Schumacher et al. 2004) only provides brief definitions of idioms, without any grammatical information, as in case of the lexical entry for geben ‘give’ (pp. 401–404), where the idiom jemandem Recht geben, lit. ‘somebody.DAT right.ACC give.INF’, is described simply as jemandem zustimmen ‘to agree with somebody’. Similarly, the VDE dictionary of English “does not, as a rule, contain idioms” (Herbst et al. 2004: xxxvii), although it provides some explicit information regarding phrasal verbs, as in the entry for give (pp. 348–351), which specifies one of the related phrasal verbs in the following way:

\[ + N_p + \text{up} + N/ADJ \]

At Anfield we used it about twelve times in all and gave it up as unworkable each time. (= stopped using it) At this point the rich families gave Cascia up as a bad job. (= lost confidence in)

The first aim of this paper is to present two electronic valency dictionaries with very rich and fully formalised phraseological components: PDT-Vallex for Czech (see Section 2) and Walenty for Polish (see Section 3). Given that Czech and Polish
are closely related – both being West Slavic languages – it makes sense to compare
the formalisms used to encode phraseological valency information in these two
dictionaries (see Section 4). It turns out that, while these formalisms are already
relatively rich, they are still too weak to accurately describe some more complex
phraseological valencies, so appropriate recommendations for their extension –
and for other such dictionaries – are made (Section 5).

2. Czech: *PDT-Vallex*

http://ufal.mff.cuni.cz/PDT-Vallex/), developed at the Charles University in
Prague, is one of a few electronic valency dictionaries for Czech, the other
ones being *VerbaLex* developed at the University of Brno (http://nlp.fi.muni.
cz/verbalex/htmlDEMO/; Hlaváčková and Horák 2005) and *VALLEX* (https://
ufal.mff.cuni.cz/vallex/; Lopatková 2003, Žabokrtský and Lopatková 2004,
2007, Lopatkova et al. 2008) – also developed at the Charles University in
Prague and sharing with *PDT-Vallex* the common theoretical underpinnings
anchored in the Functional Generative Description (FGD) theory developed in
Prague (Sgall et al. 1969, 1986).

The two Praguian dictionaries also share common origins, but have been
developed independently since December 2001 (Lopatková 2003), following
rather different approaches.³ *VALLEX* aims at providing complete descriptions
of lexemes, so once a lexeme is added to the dictionary, an attempt is made to
describe all its valency frames in some linguistic detail. By contrast, *PDT-
Vallex* has been constructed together with the annotation of the Prague
Dependency Treebank (*PDT*; Böhmová et al. 2003, Hajič et al. 2006; http://
hdl.handle.net/11858/00-097C-0000-0001-B098-5) on the ‘as needed’ basis:
frames were added as they were encountered in the corpus, with no attempt
at modelling some more subtle phenomena such as grammatical control.⁴ On
the other hand, it is the *PDT-Vallex* dictionary that contains rich phraseo-
logical information; while phraseological arguments are also present in some
valency frames of *VALLEX*, they are simply specified as strings which may fill
specific argument positions (Žabokrtský and Lopatková 2007: 50).

Let us compare two corresponding valency frames for the verb *brát si*, lit.
‘take refl’, in *VALLEX* and *PDT-Vallex*, as they are displayed on the web
pages of the two dictionaries:

(1) *VALLEX:*⁵

\[ \text{soustředit se (idiom)} \]

-frame: \( \text{ACT}^{1}_{\text{obl}}, \text{PAT}^{4}_{\text{obl}}, \text{DPHR}^{\text{obl}}_{\text{na paškál, na mušku}} \)
-example: \( \text{impr.: brát si někoho na mušku pt: vzel si něco na mušku} \)
In both dictionaries, the frame corresponds to the third meaning of \textit{bráť si}, as indicated by the boxed ‘3’ in (1) and by the superscript ‘3’ in (2).\textsuperscript{8} In both, three arguments are postulated, labelled as \textit{ACT(or)}, \textit{PAT(ient)} and \textit{DPHR} (Dependent PHRaseme). The single digits that follow \textit{ACT} and \textit{PAT} indicate cases: 1 stands for the nominative and 4 – for the accusative, so the morphosyntactic information for \textit{ACT} and \textit{PAT} is the same in both dictionaries: the former must be realised as a phrase in the nominative, and the latter – in the accusative case. Additionally, \textit{VALLEX} explicitly marks all arguments\textsuperscript{9} as either obligatory (all arguments in this frame), optional or typical (Lopatková 2003), while \textit{PDT-Vallex} only explicitly marks optional arguments with a preceding question mark ‘?’ (not shown here, but see (15) below) – all other arguments are assumed to be obligatory. A difference that is important in the context of this paper concerns the \textit{DPHR} argument. \textit{VALLEX} provides two strings that may occur in this position (corresponding to two different idioms): \textit{na paškál} ‘to task’ (as in ‘to take somebody to task’) and \textit{na mušku}, ‘on aim’, lit. ‘on front sight’, while \textit{PDT-Vallex} notes only one of these, but also provides its internal structure: the head \textit{na} ‘on’ (in its first meaning, i.e. as a preposition, hence ‘na-1’) and its dependent which is a form of \textit{muška} ‘front sight’ in the singular (‘S’) and in the accusative case (‘4’) (i.e. the form \textit{mušku}).

Another difference between the two dictionaries is that \textit{PDT-Vallex}, but not \textit{VALLEX}, makes use of another phraseological type of argument, namely, \textit{CPHR} (Compound Phraseme) (i.e., roughly, the nominal element in a light-verb construction). For example, one of the frames for the verb \textit{učinit} ‘make’, which may occur in many light-verb constructions, is given in (3) below. Within the \textit{CPHR} argument, many nouns carrying the main meaning of the construction are listed, including the noun \textit{rozhodnutí} ‘decision’; the ellipsis (‘…’) signals that this is not a closed list and ‘4’ again indicates the accusative case. This frame corresponds to the sentence in (4).\textsuperscript{10}

(3) \textit{PDT-Vallex}:\textsuperscript{11,12}

\begin{verbatim}
uczinit\textsuperscript{3}20x, 106x ACT(1) CPHR{{expertiza, chyba, kapitulace, kontrola, krok, návrh, objev, odhad, omezeni, opatreni, oznámeni, pokrok, pokus, poznamka, prohláseni, prověrka, připomínka, příznání, rozhodnutí, sázka, slob, volba, vyjádření, zátah, závazek, změna,...},4)
\end{verbatim}

(činit, dělat) • \textit{uczinit pokus; u. rozhodný krok; u. opatření; při řízení učinil chyby, lékař učinil objev genu řídícího růst buněk sítnice; připomínky, které učinil loni v prosinci; mluviči společnosti neučinil vyjádření}

(4) Učinil rozhodnutí.

\textit{made.MASC.SG.PAST decision.ACC}

‘He made a decision.’
As a separate module, the phraseological component of PDT-Vallex has been documented rather scantily so far. The main reason is that, apart from the use of the special argument names DPHR and CPHR, there is no separate formalism for the description of the surface realisation of phraseological arguments. Rather, the same formalism is used as in case of all other arguments, although its real strength is most conspicuous in case of DPHR arguments, whose specification may be very complex. Below, we will describe this formalism on the basis of Urešová 2009 and phraseological examples from PDT-Vallex.

There are three arguments in the following frame for (the second meaning of; we omit superscripts here) the verb ZVLÁDNOUT ‘manage, master’:

(5) zvládnout ACT(1) DPHR(na-1[výborný_FS4@1$11<A>]) PAT(4)

Apart from the nominative actor and the accusative patient, there is a phraseological argument headed by the preposition NA, just as in case of (2). Its dependent must be a form of the adjective VÝBORNÝ ‘excellent’ in the feminine gender (♀), singular number (S), accusative case (4) and positive degree (@1). Other possible values for gender are masculine animate (m), masculine inanimate (I) and neuter (N); another value for number could be plural (P); other Czech cases are nominative (1), genitive (2), dative (3), vocative (5), locative (6) and instrumental (7); and other values for degree are comparative (@2) and superlative (@3). In fact, each form may be described with a morphosyntactic tag consisting of 15 positions (https://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html; Hajicˇ2004), only some of which may be specified directly in the way just indicated. However, should the need arise, other positions may be constrained with the special notation of the form \$number<value>, where number is the position number and value is the expected value in this position in this phraseological expression. This mechanism is used in (5) to constrain the value of position 11 (i.e. negation) to A (i.e. affirmative). This way the phraseological argument must have the form na výbornou ‘excellently’ and not, say, na nevýbornou ‘not excellently’ (putatively).

It is also possible to specify – with the use of the hash symbol # – that a given form of the lemma must agree in number, gender and case with its governor, as illustrated below, where the form of the adjective DLOUHÝ ‘long’ must agree with the singular accusative form of the masculine inanimate noun NOS ‘nose’:

(6) dělat ACT(1) DPHR(nos.S4[dlouhý:#]) PAT(na + 4)

(7) Holčina dělá dlouhý nos na kolemjdoucí. (Czech)

‘The girl thumbs her nose at passers-by.’
Not only morphosyntactic features of a lexeme may be specified, but also its (sub)part of speech (POS), although such specifications occur rarely in case of DPHR arguments, as the lemma usually determines the (sub)part of speech. Such a POS indicator occurs at the beginning of the surface specification, as in the example below, where the adjective DOBRÝ 'good' occurs (in the dative, cf. 3), as indicated by a, rather than the noun DOBRÝ 'good – as a mark (grade) at school':

(8) změnit ACT(1) DPHR(k-1[dobrý.a3]) PAT(4)

Other possible POS indicators are n (noun), d (adverb), i (particle), u (possessive pronoun), v (verb) and j (complementiser), but also indicators of the more specific form of the dependent, such as f – restricting the dependent to infinitival forms, s – indicating the direct speech, or c – signalling an interrogative content clause.

Specifications of surface realisations of DPHR arguments are not limited to single words or head–dependent pairs – they may describe larger dependency trees, as in the following examples:

(9) brát ACT(1) DPHR(na-1[váha:4[lehký:#]]) PAT(4;↓že;↓c)
(10) běžet ACT(1) DPHR(jako[na-1[drátek.P6];jako[po-1[drátek.P6]])

In the frame given in (9), for the verb BRÁT ‘take’, the phraseological argument is specified as headed by the preposition NA ‘on’ which governs an accusative form of the noun VÁHA ‘weight’ which in turn governs an agreeing form of the adjective LEHKÝ ‘light’. A possible use of this frame is given below:

(11) Brál na lehkou váhu, že se mu vysmívala. (Czech)

‘He took it lightly that she mocked him.’

This frame also illustrates the possibility to specify a given argument disjunctively: PAT is specified as realised either by an accusative phrase, or a subordinate clause introduced by the complementiser ŽE ‘that’ (↓že), or a subordinate interrogative content clause (↓c). In the next frame, given in (10), it is the DPHR argument which is specified disjunctively: in the Czech expression for ‘run like clockwork’, the phraseological argument of BĚŽET ‘run’ may be realised as either jako na drátkách or jako po drátkách, literally meaning ‘as on/along wires’.

The phraseological expression with the longest dependency chain present in the version of PDT-Vallex published as Urešová 2011 may be found in the
following frame of the verb *udělat* ‘do’, with a *DPHR* argument meaning ‘turn of one hundred eighty degrees’:

(12) \[
\text{udělat ACT(1) DPHR(o-brat.S4[o-1[sto-2.S4[osm-de-sát.S4[stu-peň.P2]])])}
\]

Also, it is possible to specify multiple phraseological dependents; they are always listed within a single *DPHR* argument, as in the following frame for *žít* ‘live’, where the phraseological part is realised as two prepositional phrases (PPs) meaning ‘from hand’ and ‘to mouth’ (where the Czech for ‘mouth’ is a *plurale tantum* noun, hence the *P*):

(13) \[
\text{žít ACT(1) DPHR(z-1[ruka.S2],do-1[ústa.P2])}
\]

(14) Firma \[
\text{žije z ruky do úst. (Czech)}
\]

‘The company hardly makes ends meet.’

Such multiple dependents may also be specified at deeper levels of the surface specification, as in the – rather complex – example of one of many phraseological frames of *být* ‘be’ in (15), where the *DPHR* argument is given as a disjunction of three possible surface realisations (separated by semicolons), the last of which specifies two dependents of the genitive singular form of the noun *názor* ‘opinion, view’: an agreeing form of the pronoun *ten* ‘this’ and a subordinate clause introduced by the complementiser *že* ‘that’ and headed by a verb (cf. *že[.v]* – the verb is assumed to be a dependent of the complementiser). This third alternative is illustrated in (16).

(15) \[
\text{být ACT(1) DPHR(názor.S2[\{jiný,stejný,podobný,opačný\}.#];názor.S2[že[.v]];názor.S2[ten.#,že[.v]]})
\]

\[
?\text{PAT(že)}
\]

(16) Byli \[
\text{toho názoru, že je to pravda. (Czech)}
\]

‘They were of the opinion that it is true.’

(17) Jsme \[
\text{všichni stejněho názoru. (Czech)}
\]

‘We are all of the same opinion.’

This frame also employs one more notational convention of *PDT-Vallex*, namely, the possibility to succinctly specify a set of possible lemmata,
as in: \{jiný, stejný, podobný, opačný\} (where the four lemmata translate as ‘(an)other’, ‘same’, ‘similar’ and ‘opposite’), exactly one of which must be realised on the surface. This is illustrated in (17) above, where the adjective stejný ‘same’ is picked out from the set.

3. Polish: Walenty

Walenty¹⁵ (Przepiórkowski et al. 2014b, Hajnicz et al. 2015; see http://zil.ipipan.waw.pl/Walenty for the home page of the resource and http://walenty.ipipan.waw.pl/ for the user interface) is a Polish valency dictionary which is being employed by two parsers: Świgrá (an implementation of a Definite Clause Grammar description of fragments of Polish syntax;¹⁶ Woliński 2004) and POLFIE (an implementation of a Lexical Functional Grammar description of considerable fragments of Polish; Patejuk and Przepiórkowski 2012, 2015). As these parsers are based on two rather different linguistic approaches, the valency dictionary must be sufficiently expressive to accommodate for the needs of both – and perhaps others to come. At the same time, the dictionary is meant to be readable for qualified and motivated humans – mainly linguists and lexicographers – although this requires learning the notation employed in Walenty.

Each verb is assigned a number of valency frames¹⁷ and each frame 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 frame (18) for the verb głosić ‘preach, advocate’, as used in (19)¹⁸ involving a coordinated phrase in the object position, consisting of an NP (teorie o szkodliwości przedszkoli ‘theories of the harmfulness of kindergartens’) and a declarative clause introduced by the complementiser że ‘that’ (że najlepsze dla dziecka jest przebywanie z matką ‘that what is best for the child is staying with the mother’; marked here as \(\text{cp}(że)\)).

(18) \[
\text{subj}\{\text{np(str)}\} + \text{obj}\{\text{np(str)}; \text{cp}(że)\} + \{\text{np(dat)}\}
\]

(19) Niektoří głoszą teorie o szkodliwości przedszkoli i że najlepsze dla dziecka jest przebywanie z matką... (Polish)

‘Some preach theories of the harmfulness of kindergartens and that what’s best for the child is staying with the mother.’

There are three argument positions (separated by +)¹⁹ given in this frame: 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 \( np(\text{dat}) \). The subject is also described as a nominal phrase, but its case is specified as structural, i.e. depending on the syntactic context. In Polish, such subjects are normally nominative, they are genitive in case the head verb is nominalised and – according to some approaches (Przepiórkowski 1999, 2004) – they bear the accusative case when they are realised as numeral phrases of a certain type. Similarly, the nominal realisation of the object is specified as structural, as it normally occurs in the accusative, unless the verb is nominalised or the object is in the scope of verbal negation, in which case it bears the genitive case (on the so-called Genitive of Negation in Polish, see Przepiórkowski 2000 and references therein). Crucially, though, the object is specified here not just as an NP, but also alternatively (see the semicolon ;) as a clausal argument \( cp \), for complementiser phrase introduced by a specific complementiser. A parser may take this information into account and properly analyse a sentence with unlike coordination like the one involving głosić ‘preach, advocate’ in (19).

Other features of the formalism of Walenty worth mentioning here, and described in more detail in Przepiórkowski et al. 2014b, are: the representation of control and raising (cf. Landau 2013 and references therein), 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 control information and semantically defined \( xp \) arguments, see below), 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.

Phraseological arguments are specified with the use of the \( \text{lex} \) symbol, as in the simplified valency frame for the verb płynić ‘flow’ in (20), with an example of its use given in (21).

(20) \( \text{subj} \{ \text{lex}(\text{np(str),sg,'krew',ratr}) + \} \{ \text{lex}(\text{prepnp(w,loc),pl,'żyła',ratr}) \} \)

(21) Gorąca krew płynie w jego żyłach. (Polish)
hot.NOM.F.SG blood.NOM.F.SG flows in his vein.LOC.PL
‘Hot blood runs in his veins.’

There are two phraseological arguments in (20). The subject is a structurally-cased NP, as usual, but necessarily headed by krew ‘blood’ in the singular, and the NP must contain further dependents (as indicated by ratr, explained in detail below). The second argument is a prepositional phrase (PP) headed by the preposition w ‘in’ combining with a locative NP in the plural. This locative NP must be headed by a form of żyła ‘vein’, namely, by the locative plural form żyłach, and also must contain a dependent (cf. ratr again). In general, any type of phrase assumed in Walenty may be specified as lexicalised with the use of \( \text{lex} \), i.e., apart from np and prepnp illustrated above, also adjp
(adjectival phrase), prepadjp (prepositional phrase with an obligatorily adjectival dependent), advp (adverbal phrase), xp (semantically defined phrase; see below), cp (complementiser phrase, i.e. a subordinate phrase), infp (infinitival phrase), etc.

The frame in (20) is simplified: whenever a lemma is specified as requiring a dependent, the nature of this dependent should be indicated, as in (22):

(22) subj{lex(np(str),sg,'krew',ratr({adjp(agr)})+{possp})}+
{lex(prepp(np(w,loc),pl,'żyła',ratr({adjp(agr)})+{possp}))}

Here, in case of both lexicalised arguments, the required dependents of krew and żyla are specified as {adjp(agr)}+{possp}, i.e. as two possibilities: an agreeing adjectival phrase and a possessive phrase. In (21), the former possibility is realised by the adjective gorąca ‘hot’ agreeing with the form krew ‘blood’, and the latter – by the possessive pronoun jego ‘his’ modifying the nominal form żylach ‘veins’. This specification should be understood inclusively: both types of phrases may occur and each may occur in principle any number of times, as illustrated in (23), where krew ‘blood’ is modified by two adjectival forms – ta ‘this’ and gorąca ‘hot’ – and the possessive form ojca ‘father’, and the noun żylach ‘veins’ is modified by the possessive pronoun jego ‘his’ and the adjective młodych ‘young’:

(23) Ta gorąca krew ojca płynie teraz
this.NOM.F.SG hot.NOM.F.SG blood.NOM.F.SG father.GEN flows now
w jego młodych żylach.
in his young.LOC.PL vein.LOC.PL

‘This hot blood of his father runs now in his young veins.’

Apart from ratr (which stands for ‘required attribute’), other specifications of additional dependents may be used: ratr1 – exactly one dependent required, so ratr1({adjp(agr)})+{possp}) would be understood as the exclusive requirement of exactly one occurrence of exactly one of the two phrase types: adjp(agr) or possp; atr – optional dependents, i.e. as ratr but with the possibility of omitting the dependent; atr1 – at most one dependent, i.e. as ratr1 but with the possibility of omitting the dependent; natr (without any further specification) – no dependents allowed.

Note that morphosyntactic specifications of possible or required dependents 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 na myśl o czymś ‘something makes somebody’s flesh creep’, lit. ‘somebody.DAT creeps skin.NOM on (the) thought.ACC about something.LOC’. The part corresponding to o czymś ‘about something’ in the argument expressed here as na myśl o czymś ‘on (the) thought of
something’ may be realised in at least three ways: as a prepositional phrase (as just illustrated; prepnp(o, loc)), as a subordinate clause introduced by the complementiser że ‘that’ (cp(że); e.g. komuś cierpię skóra na myśl, że (to się stalo) 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 stalo) lit. ‘on (the) thought about this.LOC that (this happened.REFL)’ (prepnCP(o, loc, że)). Such a disjunctive specification of dependents may be expressed as follows (with the line broken for typographic reasons and indented for readability):

(24) \[ \{ \text{lex}(\text{prepnp}(\text{na}, \text{acc}), \text{sg}, \text{myśl}), \text{ratr}((\text{prepnp}(\text{o}, \text{loc}); \text{cp(że)}); \text{prepnCP}(\text{o}, \text{loc}, \text{że})) \} \]

This specification is still incomplete: the noun myśl may also be modified by an adjectival form, e.g. the adjectival pronoun te/a/skóra mi cierpnie na te/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 dependent type. But the status of this dependent type is different than the three dependent 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 cierpnie 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 dependents are analogous to two different arguments of a predicate occupying different syntactic positions, and the same notation could be used to specify them, with the + symbol (just as in (22) above):

(25) \[ \{ \text{lex}(\text{prepnp}(\text{na}, \text{acc}), \text{sg}, \text{myśl}), \text{ratr}((\text{prepnp}(\text{o}, \text{loc}); \text{cp(że)}); \text{prepnCP}(\text{o}, \text{loc}, \text{że}) + \{ \text{adjP}(\text{agr}) \}) \} \]

Let us finish this presentation of Walenty with a more complex frame, for the verb PRZYJMOwać ‘accept, welcome’, as in ‘somebody welcomes somebody under somebody’s roof with arms wide open’.

(26) subj\{np(str)\} + obj\{np(str)\} +
{xp(mod);
 lex(\text{prepnp}(\text{z}, \text{inst}), \text{pl}, \text{XOR('ramię', 'ręka')},
 ratr1((\text{lex}(\text{adjP}(\text{agr}), \text{agr}, \text{agr}, \text{pos,'otwarty'},
 atrl((\text{lex}(\text{advP(misc)}, \text{pos,'szeroko'}, \text{natr}))))) +
{\text{lex}(\text{prepnp}(\text{pod}, \text{acc}), _, \text{dach')},
 atrl((\text{lex}(\text{adjP}(\text{agr}), \text{agr}, \text{agr}, \text{pos},
 \text{OR('mój', 'nasz', 'swój', 'różni', 'wasz',
 'własny'), natr})))))}
There are four arguments mentioned in this frame: the usual NP subject, the usual NP object, and two (possibly) lexicalised arguments. The last argument must be a PP headed by the preposition POD ‘under’ taking an accusative dependent. This dependent is characterised as a form of DACH ‘roof’ – either singular or plural, as indicated by the underscore character signalling underspecification in the position indicating grammatical number. The form of DACH may in turn be modified by an adjectival phrase in agreeing (i.e. accusative) case (cf. adjp(agr)), agreeing number and gender (the next two occurrences of agr), and positive degree (pos), as long as this adjectival phrase is headed by one of the adjectival possessive pronouns: MÓJ ‘my’, NASZ ‘our’, SWÓJ ‘self’s’, TWÓJ ‘your.sg’, WASZ ‘your.pl’, WŁASNY ‘own’. Actually, these adjectival phrases may only contain such forms (cf. natr, signalling no further dependents, as explained above), but any number of such forms is allowed, as indicated by the inclusive OR. This correctly allows for the sequences such as MÓJ WŁASNY ‘my own’, but it also obviously overgenerates by allowing for sequences impossible in this context, such as NASZ TWÓJ ‘our your’.23

Finally, the penultimate argument in (26) may be realised either by a phraseological phrase to be discussed presently or by any manner phrase (xp(mod)); other such semantically defined phrases – with separate lists of their possible surface realisations – are, inter alia, xp(abl) (ablative), xp(adl) (adlative), xp(perl) (perlative), xp(locat) (locative), xp(temp) (temporal), xp(dur) (durative). The phraseological alternative to any manner phrase is a PP headed by the preposition Z ‘with’ which combines with the instrumental case. The NP dependent of this preposition must be in the plural and must be headed by either a form of RAMIE ‘arm’ or a form of RĘKA ‘hand’ (note the exclusive XOR). This form must have a single dependent (cf. ratr1) headed by the agreeing positive form of the adjective OTWARTY ‘open’, which in turn may be modified by a single (cf. atr1) adverb SZEROKO ‘widely’ in the positive degree. This accounts for the possibility of z OTWARTymi RĘKAMI, lit. ‘with open hands’, and z SZEROKO OTWARTymi RĘKAMI, lit. ‘with widely open hands’, and the impossibility of z RĘKAMI, ‘with hands’, or of Z NIEZWYKŁE SZEROKO OTWARTymi RĘKAMI, lit. ‘with unusually widely open hands’.

See Przepiórkowski et al. 2014a for some other features of the phraseological subformalism of Walenty, less relevant for the ensuing discussion.

4. Comparison

The two formalisms for expressing phraseological constructions in valency dictionaries were developed independently: while that of PDT-Vallex was proposed much earlier than that of Walenty, the developers of Walenty were at that time unaware of PDT-Vallex as a dictionary separate from VALLEX, and the latter does not contain information about the syntactic structure of
phraseological arguments. Nevertheless, the two dictionaries share a number of features – and display interesting differences.

Both dictionaries are heavily corpus-based, as is to be expected in the era of corpus-based lexicography, but in different ways. *PDT-Vallex* is strongly coupled with the Prague Dependency Treebank: frames were added to *PDT-Vallex* as they were encountered during the annotation of *PDT*, and relevant *PDT* nodes contain pointers to *PDT-Vallex* frames which are realised in the dependency subtrees rooted in these nodes. *Walenty* is less strongly coupled with the National Corpus of Polish (*NKJP*; Przepiórkowski *et al.* 2010, 2012; http://nkjp.pl/), but all frames must be documented with attested examples, preferably from *NKJP*. Moreover, information is present which of the examples provided for a given frame contain which realisations of which of the possible arguments. On the other hand, the construction of lexical entries is more similar to that of *VALLEX*: once a lexical entry is added, an attempt is made to describe all of its possible frames and find corpus examples for all their realisations.

Another similarity is that both dictionaries do not rely on syntactic obligatoriness as test for argumenthood: given that in both languages it is possible to omit almost any argument in the right context, and that both are so-called pro-drop languages, with verbs happy to occur without overt subjects, frames based on syntactic obligatoriness would often be empty and always inadequate. Instead, what counts as an argument in *PDT-Vallex* is determined by the dialogue test presented (after Sgall and Hajicˇova´ 1970) by Jarmila Panevová (1974: 17–19). This test may be illustrated on the basis of the verb *ARRIVE* and used to decide whether the possible ablative (from where) and adlative (where to) dependents are semantically obligatory (and, hence, arguments in the sense used in this paper), even though both are syntactically optional. Let us imagine that A said ‘John arrived.’ If the dialogue continues by B asking ‘Where from?’ and A answering ‘I don’t know’, there is nothing particular about the dialogue. However, if B asks ‘Where?’ and A answers ‘I don’t know’, there is something funny about it: how could have A said ‘John arrived’ if he cannot answer the question where John arrived? A different verb should have been used by A. Hence, according to Panevová 1974, the adlative dependent, unlike the ablative dependent, is semantically obligatory and should be mentioned in the valency dictionary. On the other hand, there are cases where this test does not give clear results (Urešová 2006: 95) and, in general, as discussed in Lopatková and Panevová 2006, there are principled difficulties in classifying some dependents as ‘arguments’ or ‘adjuncts’.24

This test was also discussed at the initial stages of the development of *Walenty*, but it was decided that it is too difficult to apply it in too many cases to make it the sole criterion for determining argumenthood. Instead, *Walenty* lexicographers rely on their intuition which, just as in case of a vast majority of valency dictionaries for various languages, is not supported by any
clear tests. However, when in doubt, they are told to include the doubtful dependent type as an argument. Hence, Walenty frames possibly contain ‘arguments’ which would in various theories be classified as ‘adjuncts’, but they should not omit any ‘true arguments’.

Despite very different notation, the expressive power of both formalisms is also rather similar. Lexicalised dependency trees may be represented with no limitation on the levels of embedding, as assumed already in the work of Igor Mel’čuk (cf. Figure 1 above). O’Grady 1998: 284 claims that [a]n idiom’s [lexically specified] component parts must form a chain and gives play the devil’s X as an example of a potential idiom prohibited by this constraint (see Figure 2) – the dependency chain would have to include the non-lexical variable component X here. Such lexicalised chains are easy to express in both formalisms; for example, the representation of the English idiom play the devil’s advocate could have the following PDT-Vallex-style and Walenty-style representations (provided the possessive ‘s is treated as a separate token):

(27) play the devil’s advocate – PDT-Vallex-style:

\[
\text{play} \quad \text{ACT}(1) \quad \text{DPHR}(\text{advocate.s[devil.s[the]]})
\]

(28) play the devil’s advocate – Walenty-style:25

\[
\text{play} \quad \text{subj} \{\text{np} + \\
\{\text{lex(np,sg,‘advocate’,} \\
\text{ratr1(\{\text{lex(possp,‘s’,} \\
\text{ratr1(\{\text{lex(np,’devil’,} \\
\text{ratr1(\{\text{lex(det,’the’,natr)}))))))))\)}}\}
\]

On the other hand, the formalism of PDT-Vallex also makes it possible to represent putative idioms violating O’Grady’s generalisation. For example, play the devil’s X could be represented as in (29), with .n signalling any noun.

Similarly, while currently Walenty only allows underspecification at the level of morphosyntactic categories (see the first underscore in (30), meaning ‘any permissible idiom:

\[
\text{play the devil’s advocate}
\]

illicit idiom:

\[
\text{play the devil’s X } \leftarrow \text{chain broken here since X is not part of the idiom}
\]

Figure 2: An impossible idiom, according to O’Grady 1998
grammatical number’), it would be a minimal extension to allow for underspecified head lemma, as in (30) below (see the second underscore there):

(29)  *play the devil’s X – PDT-Vallex-style:

\[\text{play} \quad \text{ACT(1) DPHR(.n[‘s[devil[the]]])}\]

(30)  *play the devil’s X – Walenty-style:

\[\text{play} \quad \text{subj{np}+}
\{\text{lex(np,’_’),}
\text{ratr1({\text{lex(possp,’s’,}}}
\text{ratr1({\text{lex(np,’devil’,}}}
\text{ratr1({\text{lex(det,’the’,natr}}))))))}\]

This shows that PDT-Vallex already is – and Walenty may easily be – more expressive than needed given the generalisation in O’Grady 1998. However, this extra expressive power may actually be useful, as – at least in Polish – there seem to exist expressions which violate this generalisation. For example, Bogusławski and Danielewiczowa 2005: 100 mention the following idiomatic expression of Polish: *X jest po Y głębszych ‘somebody is tipsy’, lit. ‘X<sub>NP</sub> is after Y<sub>Numeral</sub> deeper ones (glasses of liquor)*. In this expression, according to the usual analysis of Polish numeral phrases as headed by the numeral (Saloni and Świdziński 1985, Przepiórkowski 1999), głębszych ‘deeper ones’ must be analysed as a dependent of the numeral Y (possibly an indirect dependent, via an elided noun meaning ‘glass(es)’). Moreover, Y may be expressed by a noun representing quantity, such as TUZIN ‘dozen’, where it is completely uncontroversial that it is the syntactic governor of głębszych ‘deeper ones’ in *On jest po tuzinie głębszych*, lit. ‘He is after a dozen of deeper ones.’ Hence, a uniform representation of this idiom in a dictionary would have to violate O’Grady’s generalisation.

After demonstrating some similarities between the two dictionaries, and before moving to limitations common to both, let us discuss some differences. The one that is immediately clear from (27)–(28) and previous examples concerns conciseness and readability: in PDT-Vallex even relatively complex phraseological frames can be encoded concisely in a way readable to human after some training, while Walenty frames quickly become difficult to read even with considerable training. While a slightly more readable format is offered by the web interface to Walenty, at http://walenty.ipipan.waw.pl/, clearly a more robust visualisation component is needed.

Only to some extent does this difference in readability reflect a real difference in expressiveness of the two formalisms, which will be illustrated with the following frame from Walenty, for the verb BIEC ‘run’, as in the idiom biec swoim torem ‘run its course’:

(31)  *subj{np(str)} +{\text{lex(np(inst),’_’,’tor’,}}
\text{ratr({np(gen})+{adjp(agr}}))))}
The lexicalised argument headed by *tor* ‘track, course’ in any number and in the instrumental case must be modified (Cos bieg torem ‘Something ran course’ is not phraseological), as indicated by *ratr*, but it may be specified either by a genitive NP, as illustrated in (32), or by an agreeing adjectival phrase, as in (33), or by – in principle – any number of such modifications, as illustrated in (34), where *torami* ‘tracks’ has three dependents (two agreeing adjectival and one genitive nominal):

(32) Jego myśli biegły torami dziwnych
    his thought.NOM.PL ran track.INST.PL strange.GEN.PL
    skojarzeń. association.GEN.PL

    ‘His thoughts ran the course of strange associations.’

(33) Sprawy biegły swoimi torami.
    matter.NOM.PL ran self.INST.PL track.INST.PL

    ‘The matters ran their own course.’

(34) Jego myśli biegły swoimi najzwyklejszymi w świecie torami dziwnych skojarzeń.
    his thought.NOM.PL ran self.INST.PL common.INST.PL.SUP in
    world track.INST.PL strange.GEN.PL association.GEN.PL

    ‘His thoughts ran their own most common in the world course of strange associations.’

If this was a matter of the choice between the genitive NP and the agreeing adjective, it would be possible to represent (31) in *PDT-Vallex*, as shown in (35):\(^{27}\)

(35) ACT(1) DPHR(*tor.7[.n2;.a#]*)

However, (35) does not allow for the simultaneous realisation of the genitive NP and the agreeing adjectival phrase (AdjP), so it would have to be extended at least to (36):

(36) ACT(1) DPHR(*tor.7[.n2;.a#]*;tor.7[.n2,.a#]*

This would probably be statistically satisfactory, in the sense that it would cover the vast majority of the textual occurrences of this frame, but it is linguistically unsatisfactory – not only because this disjunctive notation misses the generalisation that the head of this construction is always a form of *tor* and the variation occurs within its dependents, but mainly because there is no way to express in *PDT-Vallex* that any number of adjectival dependents are allowed here. In short, in *PDT-Vallex*, there is no mechanism equivalent to the Kleene star used in regular expressions, while such an equivalent is present in *Walenty* in the form of the *ratr* and *atr* operators.
An attempt to solve this problem, at the cost of some overgeneration, would be to simplify the specification to (37), with the intended semantics that the phraseological component may be realised by just any grammatical tree rooted in the instrumental form of TOR:

\[
\text{(37) } \text{ACT(1) DPHR(\text{tor.7})}
\]

However, this solution would be at odds with the current interpretation of such specifications in *PDT-Vallex*. In short, the current interpretation is that nodes specified by particular lemmata cannot be extended by dependents, while nodes specified only with grammatical information can be freely extended (but in concord with the principles of the grammar, including other valency information). So, for example, in the frame (6) (repeated below as (38) for convenience), neither the node specified as nos.S4, nor its dependent node dlouhýː#, allow any further dependents, while in the frame (15) (repeated below as (39)), the node .v, i.e. specified only as verbal, may – and usually will – have various dependents.

\[
\text{(38) } \text{dělat ACT(1) DPHR(nos.S4[dlouhýː#]) PAT(na+4)}
\]

\[
\text{(39) } \text{být ACT(1)}
\]

\[
\text{DPHR(názor.S2[\{jiný,stejný,podobný,opačný\}.#]; názor.S2[že[.v]];názor.S2[ten.#,že[.v]])}
\]

\[
?\text{PAT(že)}
\]

Given this convention, the phraseological argument in (37) must be understood as involving the single form of tor, without any dependents.

Let us finish by saying that the lack of an equivalent of the Kleene star would be a non-negligible problem for *Walenty*, as – out of 9001 frames containing at least one phraseological argument in the version of 25 May 2015 – 3135 (almost 35%) contain a ‘Kleene star operator’: atr or ratr (as opposed to natr, atr1 or ratr1).

5. Limitations and perspectives

5.1. Regular operators

The frame in (31), describing the variation in the idiom *biec swoim torem* ‘run its course’ and given there to show the greater expressive power of *Walenty* than that of *PDT-Vallex*, may also be used to illustrate one of the limitations of the former. As explained above, the specification \( \text{ratr(\{np(gen)\} + (adjp(agr)))} \) means that at least one dependent is necessary (hence ratr rather than atr) and that any number of genitive NPs and agreeing adjectival phrases may occur. This specification overgenerates, as – while any number of adjectival dependents are allowed in principle – in fact there may only occur at
most one genitive NP in this idiom (see (32)–(34) above). Rather, the right
generalisation – not expressible in the current formalism of Walenty (or
PDT-Vallex) – is that an instrumental form of tor must be modified by
either a genitive NP or an agreeing AdjP, and may additionally be modified
by any number of agreeing AdjPs. Fortunately, both formalisms may be easily
extended to express such constraints. Let us start with PDT-Vallex.

The most straightforward extension of PDT-Vallex would consist in adding
the usual regular operators: * (Kleene star, indicating zero or more), + (Kleene
plus, indicating one or more) and ? (indicating optionality). Then, the rele-
vant constraint on the phraseological argument could be expressed as follows
(we omit the other argument here):

(40) \text{DPHR} (\text{tor.7[.a#+;n2,.a#*]})

We assume here that the comma, expressing a conjoint requirement, binds
more strongly than the semicolon, expressing a disjoint requirement. Hence,
(40) is saying that the phraseological argument should be headed by an
instrumental form of tor with either one or more adjectival dependents (.a#+)
or an obligatory genitive nominal dependent (.n2) and any number (including
zero) of adjectival dependents (.a#*). Note that we follow here the implicit
convention that nodes specified lexically cannot be extended beyond what is
said in the specification, while nodes specified only grammatically may be so
extended. That is, the form of tor is expected not to have any dependents
beyond those specified in (40), while the nominal node and the adjectival
nodes mentioned there may (and often will) have their own dependents.
This convention is unproblematic if the generalisation postulated in O'Grady 1998
(and discussed above) is true, but may become problematic
in case of idioms such as X jest po Y głębszych ‘somebody is tipsy’, lit. ‘X is
after Y Numerals deeper ones (glasses of liquor)’, which (again, as discussed
above) seem to contain an inner node specified grammatically rather than lexically. Namely, the problem is that the grammatically specified numeral
node expects exactly one dependent (an appropriately cased adjectival form),
while the current interpretation of such non-lexically specified numeral nodes
would allow for additional nominal dependents, in accordance with the gen-
eral rules of the grammar.

More far-reaching changes seem to be needed in case of Walenty, whose
format is already baroque and should rather be simplified than added more
complexity. For this reason we propose to get rid of the specialised and perhaps
confusingly-named operators atr, ratr, etc., and instead extend the logical
operators OR and XOR (see (26) above) to AND and the regular operators STAR
(Kleene star), PLUS (Kleene plus) and OPT (optionality), as well as the explicit
NONE indicating no further dependents. Such verbose names of common regular
operators are needed in order not to overload the + operator, which already has
a different meaning in *Walenty*. So, the phraseological argument of biec ‘run’ 
given in (31) (and repeated in (41) below without the other argument) should 
rather be represented as in (42):

\[(41) \{\text{lex}(\text{np}(\text{inst}),\_,'\text{tor}',\text{ratr}(\{\text{np}(\text{gen})\} + (\text{adjp}(\text{agr}))))\}\]

\[(42) \{\text{lex}(\text{np}(\text{inst}),\_,'\text{tor}',
XOR(PLUS(\{\text{adjp}(\text{agr})\}),\{\text{np}(\text{gen})\} + \text{STAR}(\{\text{adjp}(\text{agr})\}))\})\}\]

The latter specification correctly constrains the surface form of the phraseo-
logical argument to instrumental NPs headed by a form of *tor* ‘track, course’ 
(of any grammatical number) with the only allowed dependents being either (cf. 
\text{XOR}) one or more agreeing adjectival dependents (cf. PLUS(\{\text{adjp}(\text{agr})\})) or 
an obligatory genitive nominal dependent (cf. \{\text{np}(\text{gen})\}) and (+) any number 
(including zero) of adjectival dependents (cf. \text{STAR}(\{\text{adjp}(\text{agr})\})).

### 5.2. Word order

Neither of the two dictionaries has any mechanisms to specify linearisation con-
straints on phraseological (or any other) arguments. This is understood, as both 
Czech and Polish are so-called free word order languages, where linear position 
of phrases is often regulated by the information structure of the sentence (i.e.,
using other terminologies, its thematic-rhetorical – or functional – structure).
However, some phraseological expressions that should be described in a valency 
dictionary are linearly constrained beyond the general word order principles. One 
example is the Polish idiom *brać nogi za pas* ‘take to one’s heels, leg it, run away’,
lit. ‘take legs behind (the) belt’, currently described in *Walenty* as shown in (43):

\[(43) \quad \text{subj}\{\text{np}(\text{str})\} + \{\text{lex}(\text{np}(\text{str}),\text{pl},'\text{noga}',\text{natr})\} + \\
\{\text{lex}(\text{preppnp}(\text{za},\text{acc}),\text{sg},'\text{pas}',\text{natr})\}\]

Apart from the usual subject, there are two phraseological arguments in this 
frame: the plural of *noga* ‘leg’, normally in the accusative case (*nogi*), but in the 
genitive (*nóg*) when in the scope of negation or when the verb is nominalised, 
and a PP consisting of the preposition *za* ‘behind’ and the accusative singular 
form of *pas* (i.e. the form *pas*). Given the head verb *brać* ‘take.IMPERF’ or *wziąć* 
‘take.PERF’ and the three arguments, $4! = 24$ different word orders should be 
possible. However, the results of relevant corpus queries and native intuitions 
suggest that only $3! = 6$ of these are possible, because the argument described 
as \{\text{lex}(\text{np}(\text{str}),\text{pl},'\text{noga}',\text{natr})\} immediately precedes the other 
phraseological argument, \{\text{lex}(\text{preppnp}(\text{za},\text{acc}),\text{sg},'\text{pas}',\text{natr})\}.32,33

Another, perhaps even more clear, example of such a linear constraint is the 
idiom *odsylać kogoś od Annasa do Kaifasa* ‘send someone from pillar to post’,
lit. ‘send someone from Annas to Caiaaphas’, where the two PPs *od Annasa* and 
do *Kaifasa* must be adjacent and in the order indicated above.
It is an empirical question whether there are idiomatic expressions where more complex linearisation constraints are needed, for example constraints on three or more arguments (e.g. ‘A should precede both B and C, but B and C may occur in any order’). If so, a more general linearisation component should be developed for the two dictionaries. If not, there is a simple and conservative solution requiring the introduction of two linearisation operators: « (Unicode symbol U+00AB), for expressing linear precedence, and < (Unicode symbol U+2039), for expressing immediate precedence. The usual symbol combining two dependents (+ in Walenty and , in PDT-Vallex) may then be replaced with either of the linearisation operators, as needed. In case of Walenty, (43) above would be replaced with (44) below, where the last + is substituted by <:34

\[
(44) \quad \text{subj}\{\text{np(str)}\} + \{\text{lex}(\text{np(str)}, \text{pl}, 'noga', \text{NONE})\} < \\
(\text{lex} (\text{preppnp}(\text{za, acc}), \text{sg}, 'pas', \text{NONE}))
\]

In case of PDT-Vallex, the putative representation in (45) would be replaced with (46), where the comma separating the two phraseological dependents is substituted by <:

\[
(45) \quad \text{ACT}(1) \text{DPHR} (\text{noga.P4, za[pas.S4]}) \\
(46) \quad \text{ACT}(1) \text{DPHR} (\text{noga.P4 < za[pas.S4]})
\]

It is also possible to extend this notation to encode linear relations between a head and a dependent. This can be done by prefixing the dependent with one of the four linearisation operators: « and < already introduced above, as well as » (Unicode symbol U+00BB) for expressing linear consequence and > (Unicode symbol U+203A) for expressing immediate consequence. Assuming that the phraseological argument of BRAC should follow this verb,35 such a linear constraint may be expressed in the two dictionaries as in (47)–(48):

\[
(47) \quad \text{subj}\{\text{np(str)}\} + \langle\{\text{lex}(\text{np(str)}, \text{pl}, 'noga', \text{NONE})\} < \\
(\text{lex} (\text{preppnp}(\text{za, acc}), \text{sg}, 'pas', \text{NONE}))
\]

\[
(48) \quad \text{ACT}(1) \langle\text{DPHR} (\text{noga.P4 < za[pas.S4]})
\]

Similarly, given that in the expression ‘z otwartymi rękami ‘with open hands’ discussed above the adjective otwartymi must precede the governing noun rękami, and that this is not strictly required by the general rules of the Polish grammar, which also allow adjectival modifiers to follow governing nouns, the specification given above in (26) should be further constrained as follows (note the »):36

\[
(49) \quad \text{subj}\{\text{np(str)}\} + \text{obj}\{\text{np(str)}\} + \\
(\text{xp (mod);} \\
(\text{lex} (\text{preppnp}(\text{z, inst}), \text{pl}, \text{XOR('ramię', 'ręka'))}, \\
\langle(\text{lex} (\text{adjp (agr)}), \text{agr, agr, pos, 'otwarty')},
\]

Note that the use of » (rather than ›) implies that otwartymi must precede rękami, but not necessarily immediately precede it. This is because there is no linearisation constraint on the realisation of the optional adverbial modifier szeroko ‘widely’, so both sequences are possible: z szeroko otwartymi rękami (immediate precedence) and z otwartymi szeroko rękami (not immediate precedence). On the other hand, the fact that the preposition z is initial in this expression follows from the general rule that, in Polish, adpositions are prepositions (with just a couple of well-defined exceptions), so it does not have to be stated in the lexicon.37

5.3. Coordination within arguments

Neither dictionary handles coordination properly. Let us first consider the Polish idiom poruszyć niebo i ziemię (żeby coś zrobić) ‘move heaven and earth (to do something)’. Its current representation in Walenty is given in (50):

(50) subj{np(str)}+obj{fixed(np(str),'niebo i ziemię')}+{cp(żeby)}

Apart from the usual subject (subj{np(str)}) and a subordinate clause introduced by a żeb-type complementiser ({cp(żeby)}; this class of complementisers contains żeb, aby and by), there is a phraseological object described with the use of the symbol fixed, not explained so far. Typical uses of fixed are concerned with lexicalised arguments which are morphologically unusual. An example would be stanąć dęba ‘rear, jib’ (of a horse) or ‘stand on end’ (of hair), lit. ‘stand oak.Gen.SG’. The problem is that, in contemporary Polish, the singular genitive form of DĄB ‘oak’ is dębu, not dęba. Hence, this part of the phraseological expression cannot be described with the construct lex(np(gen),sg,'daśb',NONE), as a Polish parser taking advantage of this description would expect stanąć dębu, and a generator would produce this string instead of the correct stanąć dęba. Moreover, the intuition is that dęba behaves here as an adverb rather than as a noun, so the relevant argument of STANĄĆ ‘stand’ is described as {fixed(advp(misc),'dęba')}. In contrast to dęba, there is nothing unusual about the string niebo i ziemię – it is a simple coordination (with i ‘and’) of singular accusative forms of NIEBO ‘sky, heaven’ and ZIEMIA ‘earth’. Obviously, fixed was used here only because lexicographers found no other way to describe a phraseological expression involving coordination. In fact, this description is simply wrong: it rightly
marks the coordination as $np(str)$, but it wrongly fixes the form to the accusative *niebo i ziemię*. When the verb is nominalised or in the immediate scope of negation, this expression should be *nieba i ziemi*, with the genitive forms of *niebo* and *ziemia*, as the following examples from *NKJP* testify:

(51) ...
dzięki poruszeniu nieba i ziemi przez zrozpaczoną matką...

(52) Manifest Matthiasa Polityckiego nie poruszył nieba i ziemi...

*PDT-Vallex* seems to fare better here, as the coordinated phraseological argument could be simply represented as in (53):

(53) $\text{DPHR}(i[\text{niebo.S4} \ « \text{ziemia.S4}])$

Note the use of the linear operator « introduced in Section 5.2 and expressing (not necessarily immediate) linear precedence: this, together with general rules placing the conjunction before the last conjunct, would ensure the surface realisation *niebo i ziemię*, as opposed to much less frequent and perhaps not phraseological *ziemia i niebo*. Moreover, appropriately sophisticated grammar rules could then interpret the accusative specification 4 as genitive in the right contexts.

However, it would be more difficult to model in *PDT-Vallex* a phraseological construction involving coordination and a modifier shared by the conjuncts, as in the Polish idiom *być czyimś okiem i uchem* ‘be somebody’s informant’, lit. ‘be somebody’s eye and ear’. According to the representation of coordination in *PDT* (see e.g. Popel et al. 2013), the shared modifier, which can be expressed by a possessive pronoun or a genitive NP, is represented as yet another dependent of the conjunct. In *PDT-Vallex*, this could be represented as in (54) (with parentheses delimiting the disjunctive specification of the first dependent):

(54) $\text{DPHR}(i[.(n2;.u#) « \text{oko.S7} « \text{uchy.S7}])$

In the full *PDT*, such shared dependents of conjuncts are distinguished from the true conjuncts with the use of appropriate dependency labels. But the formalism of *PDT-Vallex* does not use dependency labels at the level of surface realisation, so the frame in (54) does not distinguish between a coordination of
three elements and a coordination of two elements with a shared modifier. Clearly, additional mechanisms are needed in both formalisms to handle phraseological expressions containing coordination properly.

Again, it is easy to extend the formalism of \textit{PDT-Vallex} to handle such cases by simply marking shared dependents with the special diacritic = and thus distinguishing them from direct conjuncts:

\begin{equation}
\text{(55) DPHR}(i[(.n2;.u#)= \text{oko}.S7 \text{ ucho}.S7])
\end{equation}

We may additionally assume that this special diacritic combines with diacritics expressing regular expressions and occurs after them. For example, if – contrary to fact – any number of shared dependents were allowed in this idiomatic expression, the \text{DPHR} argument could be specified as: \text{i[(.n2;.u#)\*= \text{oko}.S7 \text{ ucho}.S7]}. This frame assumes that the possessive modifier must occur immediately before the first conjunct. A better approximation would be that, in case the modifier is a possessive pronoun, it must occur at the beginning of the coordination, and when it is a genitive noun – it should be found at the end of the construction. Hence, a perhaps more precise specification that the possessive pronominal modification normally precedes the coordination and the genitive NP normally follows it, may look as in (56):

\begin{equation}
\text{(56) DPHR}(i[.u#= \text{oko}.S7 \text{ ucho}.S7]; \text{i[oko}.S7 \text{ ucho}.S7 \text{ .n2=}])
\end{equation}

In case of \textit{Walenty}, there is already a mechanism related to coordination, which has not been introduced so far, namely the possibility to specify that elements introduced by \text{OR} may be coordinated. This is best illustrated with the following current frame for the verb \text{ZAMIENIAĆ SIĘ} ‘change (itself), metamorphose’:

\begin{equation}
\text{(57) subj\{np(str)\} + \{lex\{prepnp(w,acc),sg,\text{OR('proch';'pył'),natri}\}}
\end{equation}

This frame may be used to express that something has changed into ashes (Polish: a form of \text{PROCH}) or into dust (Polish: a form of \text{PYL}), but also into ashes and dust (\text{w proch i pył}); the possibility of using not just one of the alternatives but also their coordination is signalled by the use of semicolon ; instead of the comma , within \text{OR}.

Given this convention, the most natural extension for expressing coordination would involve introducing \text{AND} (in fact, already introduced in Section 5.1) with possible conjuncts separated by the semicolon:

\begin{equation}
\text{(58) obj\{lex\{np(str),sg,\text{AND('niebo';'ziemia')},NONE\}}
\end{equation}

This conservative extension is based on the assumption that only single words with the same inflectional characteristics may be coordinated, e.g. two singular structurally-cased nouns in \text{poruszyć niebo i ziemię} ‘move heaven and earth’,
two singular instrumental nouns in być czymś okiem i uchem ‘be somebody’s eye and ear’, or two plural instrumental nouns in wciskać się drzwiami i oknami ‘try to get in (of a large number of people)’, lit. ‘squeeze in (through) doors and windows’. Note that the conjunction is not explicitly specified in (58); in such cases we may assume any non-contrastive conjunctive (as opposed to disjunctive) conjunction: not only i ‘and’, but also oraz ‘and, as well as’ and ani... ani... ‘neither... nor...’ under negation (but not the contrastive A ‘and’ or the disjunctive LUB ‘or’). The attested\(^4\) (59) and the constructed (60) (based on (52) above) illustrate this variability of conjunction:

(59) William obiecał Kate, że... poruszy niebo
William.NOM promised Kate.DAT that move.FUT heaven.ACC.SG
oraz ziemię... (Polish)
as well as earth.ACC.SG
‘William promised Kate that... he will move heaven and earth...’

(60) Manifesto Matthiasa Polityckiego nie poruszy
manifesto.NOM.MASC.SG Matthias.Gen Politycki.GEN NEG moved
ani nieba, ani ziemię... (Polish)
neither heaven.GEN.SG neither earth.GEN.SG
‘Matthias Politycki’s manifesto moved neither heaven nor earth...’

On the other hand, there are idiomatic constructions where only one conjunction may be used, as in the Polish bawić się w kotka i myszkę ‘play hide and seek’, lit. ‘play reFl in cat.ACC.SG and mouse.ACC.SG’ (bawić się w kotka oraz myszkę, i.e. with oraz replacing i, is not phraseological). The phraseological argument could be represented here as follows:

(61) \{\{\text{prepnp}(w,\text{acc}),\text{sg,\text{AND}[i]}('kotek';'myszka')),\text{NONE}\}\}

The use of square brackets after AND is only a moderate extension of the current formalism of Walenty, which already allows for such an optional further specification elsewhere, for example in the frame for the Polish verb docierać ‘reach’, as in docierać z czymś pod strzechy ‘get the message through to ordinary folk’, lit. ‘reach with something under thatches’. The simplest specification of the argument pod strzechy ‘under thatches’ would be as in (62)(a), where STRZECHA is the Polish lemma for ‘thatch’, but it would miss the point that this is a kind of an adlative phrase, normally represented in Walenty as xp(adl) (cf. Section 3). For this reason, a combined representation illustrated in (62)(b) was devised for such cases, with the main type of phrase given as xp(adl), where the adl symbol is further specified – with the help of the square brackets – as prepnp(pod,acc):\(^4\)

(62) a. \{\{\text{prepnp}(\text{pod},\text{acc}),\text{pl,}'strzechy',\text{NONE})\}\}

b. \{\{\text{xp}(adl[\text{prepnp}(\text{pod},\text{acc})]),\text{pl,}'strzechy',\text{NONE})\}\}
The advantage of such a representation of phraseological coordinated arguments is that it does not suffer from the problem of shared modification of conjuncts, which was problematic in case of PDT-Vallex: the shared modification is simply given as the last parameter of \texttt{lex}, as usual:

\begin{equation}
(63) \{\text{\texttt{lex(np(inst),sg,AND[i]('oko';'uco'),{possp})}}\}
\end{equation}

In (63), expressing the phraseological argument in \textit{być czyimś okiem i uchem} ‘be somebody’s eye and ear’, the last parameter of \texttt{lex} is \{\texttt{possp}\}, which expresses a required (see Section 5.1 and note the lack of the optionality operator \texttt{OPT}) possessive dependent – either a possessive pronoun or a genitive NP, according to the definition of \texttt{possp} already assumed in \textit{Walenty}.

Note that the order of the modifier with respect to the coordination is not specified here. Just as in case of PDT-Vallex, further extensions are needed in case lexical specification of linear constraints is desirable, for example, by prefixing the specification of the modifier with either « (coordination precedes modification) or » (coordination follows modification). A \textit{Walenty} specification more fully analogous to that of PDT-Vallex given in (56) could then have the following form:

\begin{equation}
(64) \{\text{\texttt{lex(np(inst),sg,AND[i]('oko';'uco')),}}
\text{\texttt{XOR(«{np(gen)},»{adjp(agr)})}}\}
\end{equation}

However, the above specification includes adjectives which are not possessive pronouns and excludes coordination of possessives of different kinds, as in (65), so the specification in (63) seems more adequate.

\begin{equation}
(65) \ldots\text{moim i mojej rodziny okiem i uchem\ldots} \quad \text{(Polish)}
\end{equation}

\text{my.INST and my.GEN family.GEN eye.INST and ear.INST}

\text{‘\ldots the eyes and ears of my family and myself\ldots’}

Let us finally note that – while the initial assumption that only single words may be coordinated in Polish phraseological expressions seems to be true about a vast majority of cases – there are potential exceptions. One such expression is \textit{między ustami a brzegiem pucharu}, lit. ‘between mouth.INST and edge.INST goblet.GEN’, meaning roughly ‘between intention and its execution’ or ‘between a decision and its fulfilment’, where the second conjunct is a 2-word NP \textit{brzegiem pucharu}. Currently, there is no phraseological frame requiring this expression as an argument, but it is imaginable that this expression could be treated as a dependent of the verb \textit{ZDARZYĆ SIĘ} ‘happen’, as in the attested\textsuperscript{44} (66):

\begin{equation}
(66) \text{Wiele się może zdarzyć między ustami a brzegiem much REFL may happen between mouth.INST and edge.INST}
\end{equation}
Another potential counterexample to the single-word-coordination assumption is the idiom *znajdować się między Scyllą a Charybdą* ‘find oneself between Scylla and Charybdis’, which sometimes occurs in texts as in the attested\(^{45}\) (67), i.e. with genitive modifiers:

\[
(67) \text{Kraje znajdują się między Scyllą autorytarzyzmu a Charybdą oligarchii.} \quad (\text{Polish})
\]

‘The countries find themselves between the Scylla of authoritarianism and the Charybdis of oligarchy.’

Again, if such uses where to be described in *Walenty*, a more general mechanism than the simple solution proposed here would be necessary.

### 5.4. Coordination within predicates

While phraseological coordinated arguments pose a problem that requires some extensions of the discussed dictionaries, another kind of coordination – exemplified below\(^{46}\) – poses more fundamental problems:

\[
(68) \text{Cała Kolumbia chucha i dmucha na Falcao.} \quad (\text{Polish})
\]

‘The whole Colombia dotes on Falcao.’

The head of this example consists of a coordination of verbs *chuchać* ‘puff’ and *dmuchać* ‘blow’, which has a phraseological meaning ‘dote’ and opens two valency position: for the usual NP subject and for a PP complement headed by the preposition *na* ‘on’ (taking an accusative NP).

Given the organisation of both dictionaries, by the head lemma, it would be necessary to postulate lexical entries headed by *chuchać i dmuchać*, in effect treating it as a single verb. The fact that the lemma and all forms of this ‘verb’ would contain spaces is only a minor problem. Also the missing generalisation that parts of this ‘verb’ would conjugate just as two existing verbs *chuchać* and *dmuchać* is perhaps not really a showstopper. The real problem is that, given the relative free word order of Polish, it is possible to linearly realise one of the arguments *within* this ‘verb’, as in the attested\(^{47}\) *chucha na nich i dmucha* ‘(S)he dotes on them’, lit. ‘puffs on them and blows’.
Another possibility, at least in *PDT-Vallex*, would be to represent this idiom under CHUCHAC’, with DMUCHAC’ treated as a dependent (e.g. headed by 1). However, then a new mechanism would be necessary to express the fact that the verbal part of the dependent (i.e. the form of DMUCHAC’) must be inflected the same way as the head verb CHUCHAC’. For example, while both chucham i dmucham ‘puff.sg.1 and blow.sg.1’ and chuchasz i dmuchasz ‘puff.sg.2 and blow.sg.2’ are fine, chucham i dmuchasz ‘puff.sg.1 and blow.sg.2’ is at best non-phraseological.

Such idioms with a coordinated verbal head and open valency positions are not exceptional in language. Some other examples from Polish are ktoś dwoi się i troi ‘somebody.NOM acts with lots of zeal and energy, somebody.NOM gets out of his way (to do something)’, lit. ‘somebody duplicates REFL and triplicates’, where the two verbs DWOIC’ SIE ‘to duplicate oneself’ and TROIC’ SIE ‘to triplicate oneself’ obligatorily share the reflexive marker SIE, cos’ kogoś ani ziebi, ani grzeje ‘something leaves somebody indifferent’, lit. ‘something.NOM somebody.ACC neither cools down nor warms up’, or ktoś chce i boi się ‘somebody wavers’, lit. ‘somebody.NOM wants and fears’.

An interesting variation of this difficulty is presented by the idiom bić i patrzeć, czy równo puchnie ‘keep beating (somebody) black and blue’, lit. ‘beat and watch whether evenly swells’. While this idiom often occurs in the infinitival form, usually as nic tylko bić i patrzeć, czy równo puchnie ‘one should keep beating (somebody contextually salient) black and blue’, lit. ‘nothing but beat.INF and watch.INF whether evenly swells’, it may occur with the subject and an object, as in the attested49 (69), with the pro-dropped first person feminine subject:

(69) Biłam ją i patrzyłam czy równo beat.F.SG.PAST she.ACC and watch.F.SG.PAST whether evenly puchnie... (Polish)
swells
‘I kept beating her to a pulp.’

What is interesting about this example is that one argument (the subject) is clearly shared by both verbs, while the other argument in this idiom is required only by the first verb, bić ‘beat’. Again, we see no way of representing such constructions in the two dictionaries without fundamental changes in their formalisms.

5.5. Paradigmatic constraints

Some expressions have a phraseological meaning only when additional conditions on the form of their verbal head are met. For example, one of the
phraseological frames of the Czech verb NECHAT ‘leave’, given in (70) and exemplified by the attested\textsuperscript{50} (71), requires the verb to be negated:

\begin{equation}
\text{(70)} \ \text{nechat} \ \text{ACT}(1) \ \text{DPHR}(\text{kámen}.S4, \text{na-1}[\text{kámen}.S6]) \ \text{PAT}(z+2)-(.)
\end{equation}

\begin{equation}
\text{(71)} \ \text{V roce 1997, kdy Jobs nenechal v Apple kámen} \ \text{na kameni}... \ \text{(Czech)}
\end{equation}

‘In 1997, when Jobs rearranged everything in Apple...’

This requirement is expressed by the final ‘argument’ –(.), where – refers to the head lemma and the specification in the following parentheses describes its possible morphosyntactic forms. In this case, .~ specifies that the verb must be negated.

A similar Polish example involves the inherently reflexive verb BAC’ SIĘ ‘fear’ with the current phraseological frame (72), as used in (73), where the literal meaning of ‘not fearing God’ refers to acting immorally and without fear of punishment:

\begin{equation}
\text{(72)} \ \text{subj}\{\text{np(str)}\} + \{\text{lex}(\text{np(gen), sg, ‘bóg’, natr})\}
\end{equation}

\begin{equation}
\text{(73)} \ \text{Ten, kto rozsypał azbest, chyba Boga się nie boi}...(\text{Polish})
\end{equation}

‘The one that spilled asbestos perhaps god.GEN.SG REFL NEG fears’

\textit{Walenty} can deal with such cases as each frame has a negation flag saying whether this frame only occurs in negated contexts (very rare: well below 1% of frames), only in affirmative contexts (extremely rare), or whether it is insensitive to polarity. In fact, this flag may be interpreted not as a direct requirement on the head verb, but rather as a requirement on the general context. For example, KIWNAČ ‘nod’ as used in \textit{nie kiwnać palcem} (see the current frame (74)), lit. ‘not nod finger.INST’, means ‘not lift a finger’ and it is a negative polarity item, just as its English equivalent, but the negation does not have to be expressed on the verb KIWNAČ – it may be expressed elsewhere in the sentence, e.g. on the higher verb, as in (75):

\begin{equation}
\text{(74)} \ \text{subj}\{\text{np(str)}\} + \{\text{lex}(\text{np(inst), sg, ‘palec’, natr})\}
\end{equation}

\begin{equation}
\text{(75)} \ \text{Nikt \ nie chciał nawet kiwnać palcem.} \ \text{(Polish)}
\end{equation}

‘Nobody even wanted to lift a finger.’

Hence, the negation flag of \textit{Walenty} has a broader sense than the specification of negation in (70) and has no equivalent in PDT-Vallex.

However, \textit{Walenty} currently lacks a more general mechanism of specifying arbitrary conditions on the morphosyntactic properties of the head of a given
frame, similar to the mechanism of PDT-Vallex exemplified in (70). It is clear that such a mechanism is needed; as shown in Kosek 2008, 2013 (see also Czerepowicka and Kosek 2011), paradigmatic restrictions may refer to various morphosyntactic properties of the head, not just to polarity. For example, the expression utopić kogoś w łyżce wody ‘to hate somebody’, lit. ‘drown somebody in spoon water..gen’, with the open subject position, is limited to the conditional and infinitival contexts, and the attempt to use this expression in, say, the past tense triggers a literal interpretation. Similarly, urwać komuś głowę ‘to scold somebody, to tell somebody off’, lit. ‘tear-off somebody.dat head.acc’, with two argument positions, cannot be used in the past tense. Note that these constraints pertain to the whole idiomatic expressions, not to their head verbs, which – used in other contexts – enjoy full paradigms.

In general, such morphological and morphosyntactic properties of phraseological expressions have been studied much more extensively than their valency properties; see e.g. Savary 2008, Al-Haj et al. 2013 and references therein. Clearly, further work is needed on finding a natural way to combine valency and paradigmatic constraints on particular phraseological expressions.51

5.6. Constructional valency

Current valency dictionaries, the two Slavic dictionaries included, do not make a distinction between basic valency and what might be called constructional valency. In the latter, arguments are added to the basic valency frames via certain productive or semi-productive processes, as in the famous Pat sneezed the napkin off the table, where the basically intransitive verb SNEEZE receives two additional arguments (here: the napkin and off the table).52 This is not a technical problem, as such derived valency frames may be added to the lexicon next to basic frames, but the resulting description certainly misses a linguistic generalisation.

An interesting Polish example of phraseological construction of this kind is noted in Bogusławski and Danielewiczowa 2005: 266–267 and may be presented as ktoś za-V się na śmierć, lit. ‘somebody za-V refl to death’, where ‘V’ is – in principle – any activity verb. The meaning of this construction is that somebody died or is at the brink of death as a result of (excessive) V-ing. For example, ktoś zagadał się na śmierć, where V is gadal, a form of GADAC ‘talk, babble’, means that somebody talked to the point of complete exhaustion (or, indeed, death). Again, neither of the discussed (or any other, to the best of our knowledge) valency dictionaries is able to describe this phenomenon in a way that does not miss the generalisation; instead, Walenty contains a number of relevant frames with the phraseological argument na śmierć ‘to death’ for verbs such as ZACPAĆ SIĘ (where ĆPAĆ means ‘take drugs’) or ZABELKOTAĆ SIĘ (where BELKOTAĆ means ‘mumble, babble’).
6. Conclusion

This paper constitutes the first journal presentation of phraseological components of two wide-coverage valency dictionaries: PDT-Vallex for Czech and Walenty for Polish. The comparison of these components demonstrates their similar expressive power despite very different notations employed in them. The slightly greater – due to the use of regular operators – expressive power of Walenty is achieved at the cost of much more complex (indeed, baroque) notation; PDT-Vallex arguably achieves a better compromise between readability and expressiveness.

On the other hand, the expressive power of both formalisms is too limited to truthfully and precisely represent the surface structure of some phraseological expressions. In some cases (Sections 5.1–5.3), it is possible to extend the formalisms in a conservative way to handle such problematic idioms; in other cases (Sections 5.4–5.6), more fundamental modifications to the underlying structure of these dictionaries seem to be needed.

To the best of our knowledge, PDT-Vallex and Walenty are unparalleled with respect to the scope and depth of their descriptions of surface realisations of phraseological arguments, and may only be compared to the less formalised and less extensively applied methodology of Mel’čuk and Zholkovsky 1984 and the empirically extensive but syntactically shallow approach of Gross 1984. We hope that this paper will not only provide an insight into the two dictionaries, but will also help the future developers of similar dictionaries for other languages to design a formalism capable of handling the whole spectrum of phraseological expressions.

Acknowledgements

We would like to thank Iwona Kosek, Agnieszka Patejuk and Agata Savary for comments on (parts of) a previous version of this paper. Also, detailed comments of one of the reviewers led to multiple improvements in the form and content of this article. Shuly Wintner helped with LaTeX formatting for OUP. Work reported here has been partially financed by the Polish Ministry of Science and Higher Education within the CLARIN ERIC programme 2015–2016 (http://clarin.eu/), by the grant GP13-03351P of the Grant Agency of the Czech Republic, by the project LM2010013 of the MEYS of the Czech Republic and by the IC 1207 COST Action PARSEME (http://www.parseme.eu/).

Notes

1 We adopt here the terminology of Igor Mel’čuk (e.g. Mel’čuk 2012).
In this paper we consistently use the variant ‘valency’, rather than ‘valence’, as the former seems to be more widespread in linguistics; in particular, linguistic dictionaries usually contain the former term and not the latter (e.g. Trask 1993: 296) and Crystal 1997: 407). Also, ‘valency’ seems to be chiefly British, while ‘valence’ is chiefly American (New Oxford Style Manual 2012: 796).

However, an attempt is currently being made at linking entries between VALLEX and PDT-Vallex; Bejcˇek et al. 2014.

That is, control information is missing in PDT-Vallex itself; however, grammatical control is fully annotated in PDT.

http://lindat.mff.cuni.cz/services/PDT-Vallex/PDT-Vallex.html?block=B&verb=br%C3%A1t+si

The lower part of such entries provides examples of use of the frame given in the upper part.

This is accidental; in general, there is no correspondence between meanings of lexemes in the two dictionaries.

Throughout this paper we try to simplify and unify the terminology used to describe the two dictionaries. In particular, argument is understood here as any element of a valency frame. Note that this use differs from the terminology of Functional Generative Description, where arguments are understood as only those valency elements which are labelled with one of the five core roles: ACT, PAT, ADDR(essee), ORIG(in) or EFF(ect). In FGD, all valency elements are called valency complementations, or sometimes valency frame members (Urešová 2009: 8) or valency frame slots (Urešová 2006: 99).

We do not attempt to provide all morphosyntactic features in word-by-word translations of such examples, only those which seem relevant and helpful for the understanding of the structure of the example. The abbreviations mostly adhere to those recommended in Leipzig Glossing Rules (http://www.eva.mpg.de/lingua/resources/glossing-rules.php). While both Czech and Polish distinguish between different masculine (sub)genders, we simplify here and annotate masculine forms as MASC.

The subscript numbers indicate the frequency of the verb with this frame in the Prague Dependency Treebank mentioned above (20 occurrences) and in the Prague Czech-English Dependency Treebank (Hajicˇ et al. 2012; 106 occurrences). On phraseological expressions in the latter, see also Dušek et al. 2014.

On the other hand, CPHR arguments are hardly ever – with the notable exception of the copulas BY´T and BY´VAT – more complex than in (3).

The specification nā+ 4 of the PAT argument is a shorthand for nā−1[.4], i.e. the preposition NA and its accusative dependent. Note also that, for the purpose of this article, the full stop , and the colon : may be considered as synonyms used interchangeably to signal the following surface specification.

This description is based on Przepiórkowski et al. 2014a and updates it to some extent.

The grammar itself is presented in Świdziński 1992; see also Świdziński and Szpakowicz 1994 for related work.

In Walenty publications, the term valency schema is used for the syntactic level of valency, while the term valency frame is reserved for the semantic level; here we break
with this convention for the sake of terminological uniformity with the description of PDT-Vallex.

18 Taken in a slightly simplified form from the National Corpus of Polish.

19 Whitespace characters are not meaningful around + or anywhere else in this notation and are only provided to enhance readability.

20 Arguably, this is a conflation of two phraseological expressions: komuś cierpnie skóra (z jakiegoś powodu) 'somebody.DAT creeps skin.NOM (for some reason)' and na myśl o czymś 'on (the) thought.ACC about something.LOC', which may be combined with a large number of verbs.

21 On the other hand, the realisation of adjp(agr) is lexically limited in such cases, possibly to the single adjective SAM 'sheer, alone', so a more subtle and precise description should probably replace (25) in the final version of the lexicon.

22 Again, this is probably a conflation of two separate idioms meaning ‘welcome somebody with arms wide open’ and ‘welcome somebody under one’s roof’.

23 This overgeneration can be avoided by the following atr specification (replacing the last atr in (26)):

\[
\text{(i) atr} \left\{ \text{lex(adjp(agr),agr,agr,pos,XOR('mój','nasz','swój','twój','wraz'),natr)} \right\} + \text{lex(adjp(agr),agr,agr,pos,}
\text{'wlasny',natr})
\]

24 Scare quotes here and in the next paragraph reflect our doubts regarding the reality of the argument/adjunct distinction. Obviously, the notion of semantic obligatoriness inherent in Panevová’s test is only one of many possible understandings of obligatoriness of a dependent, as discussed at length in Herbst and Roe 1996, as well as – in the Polish “semantic syntax” tradition – in Karolak 1984: Section 4.2 and 8.3.

25 We assume that this idiom is not passivisable, hence the lack of the obj specification on the lexicalised argument.

26 The complex dependence of the grammatical number of the phraseological argument and the NP subject is not expressed here.

27 We assume that the lack of specification of a given grammatical category (here: number) means that any value of this category is possible.

28 However, this was not completely checked at the time PDT was published.

29 If need be, round parentheses () could be added for grouping.

30 We are grateful to Urszula Andrejewicz for providing us with some of the Polish examples used in this and subsequent subsections.

31 Such variation in the surface realisation of a lexicalised argument is the main reason for specifying the argument via a lemma and morphosyntactic description, rather than providing a specific inflected form of the argument. Note that morphosyntactic description is independently needed in many cases in order to model agreement between the lexicalised element and its agreeing dependents, if any.

32 The first intuition is that these two phraseological arguments must immediately follow the head verb, but impeccable examples of the verb following them and of the verb separated from them may be found in corpora.

33 In the whole NKJP, six examples are found with the query "[Nn]ogi " []{1,5} za pas within s (find Nogi or nogi followed by 1 to 5 tokens followed by the two tokens za pas, all within a single sentence; see http://nkjp.pl/poliqarp/help/en.html for a short tutorial on the relevant query language), which seem to contradict this generalisation, but two of these are clear cases of ‘playing with words’, one is an attempt at
34 The proposed representations take into account the new notation for regular operators introduced in the preceding subsection.

35 There is certainly a statistical tendency to this effect, but not a hard constraint that should actually be encoded in a valency dictionary.

36 Note also that what follows » is the single operand of the old operator ratr1. Given the extensions proposed in the previous subsection, the ratr1 operator may simply be dropped now.

37 As noted by the reviewer, this paragraph suggests that we assume a strong divide between the grammar and the lexicon, contrary to various arguments put forward within Construction Grammar and elsewhere. This is not so: we only note that some linguistic knowledge pertains to particular lexemes (certainly including phraseological information and at least some valency information) and that there are also some general grammatical rules which do not refer to any specific lexemes. But we leave it open whether these two kinds of knowledge represent two very different components of the language, or whether they are only two extreme points on a continuous scale. See also Section 5.6.

38 Note that the operator of the immediate linear precedence  < could not be used here, as the two dependents are separated by the conjunction i. 

39 While this is not clear in case of this idiom, many other idiomatic expressions involving coordination have a strictly fixed order of conjuncts.

40 On the other hand, this arguably follows from the general rules of the grammar, so it probably does not need to be specified in the lexicon.


42 In general, the content of square brackets is a list of possible realisation of a given phrase type; here the list is of length 1.

43 Since Walenty does not have a natural notion of an “agreeing possessive pronoun” (cf. u# in PDT-Vallex), although this notion may be approximated by listing relevant lemmata in a lex(adjp(agr),agr,agr,pos,XOR(...),NONE) specification, we simplify here by generalising possessive pronouns – most of which are adjectival in Polish – to adjp(agr).


48 See Kupś 1999 on the haplology of sie in Polish.

49 In the Google snippet found with the query ‘biłam ja i patrzyłam czy równo puchnie’ on 4th June 2015.


51 This point is also made – from the opposite direction of a morphosyntactic dictionary of Multiword Expressions – in Al-Haj et al. 2013: Section 6.8, esp. fn. 10.
The term *constructional valency* is justified by the fact that, in Construction Grammar (CxG), such additional “argument roles” are contributed to the basic arguments (“participant roles”) of the verb by an appropriate construction – here, by the “caused-motion” construction; see Goldberg 1995: 54–55, 152–179. As noted by the reviewer, this term may be a little misleading, since the argument structure resulting from the fusion of participant roles (of a verb) and argument roles (of a construction) is not considered a valency frame of the verb in CxG.

References


