International Journal of Lexicography, Vol. 30 No. 1, pp. 1–38
doi:10.1093/ijl/ecv048 Advance access publication 22 February 2016

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 Hajič: *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)*

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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[1] 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*[2] 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

© 2016 Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

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).

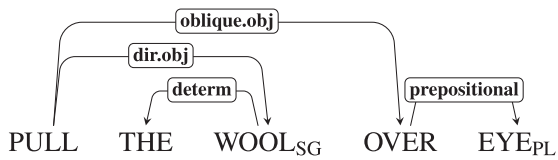
The need for a dictionary combining phraseological and valency information has long been recognised in Slavic linguistics and elsewhere, especially, in the work of Igor Mel'čuk and his colleagues starting in the 1960s and culminating in the development of the concept of an Explanatory Combinatorial Dictionary (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 [N_Y'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



Government pattern

X ⇔ I		Y ⇔ II
1. N	1. *of* N	THE EYES OF N (determ, attrib)
	2. N's	N's EYES (possessive)
	3. $A_{(poss)}$(N)	$A_{(poss)}$(N) EYES (determ)

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*.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 + 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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*

PDT-Vallex (Hajič *et al*. 2003, Hajič and Urešová 2003, Urešová 2009, 2011; 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, Lopatková *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 phraseological 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]

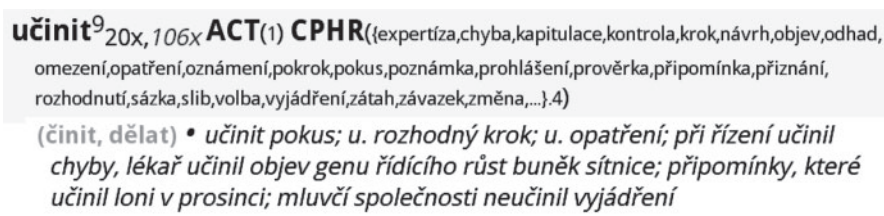
Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

(2) *PDT-Vallex*:[6,7]

In both dictionaries, the frame corresponds to the third meaning of BRÁT SI, as indicated by the boxed '3' in (1) and by the superscript '3' in (2).[8] In both, three arguments are postulated, labelled as ACT(or), PAT(ient) and DPHR (Dependent PHRaseme). The single digits that follow ACT and PAT indicate cases: 1 stands for the nominative and 4 – for the accusative, so the morpho-syntactic information for ACT and 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, *VALLEX* explicitly marks all arguments[9] as either obligatory (all arguments in this frame), optional or typical (Lopatková 2003), while *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 DPHR argument. *VALLEX* provides two strings that may occur in this position (corresponding to two different idioms): *na paškál* 'to task' (as in 'to take somebody to task') and *na mušku*, 'on aim', lit. 'on front sight', while *PDT-Vallex* notes only one of these, but also provides its internal structure: the head *na* 'on' (in its first meaning, i.e. as a preposition, hence 'na-1') and its dependent which is a form of MUŠKA 'front sight' in the singular ('S') and in the accusative case ('4') (i.e. the form *mušku*).

Another difference between the two dictionaries is that *PDT-Vallex*, but not *VALLEX*, makes use of another phraseological type of argument, namely, CPHR (Compound Phraseme) (i.e., roughly, the nominal element in a light-verb construction). For example, one of the frames for the verb UČINIT 'make', which may occur in many light-verb constructions, is given in (3) below. Within the CPHR argument, many nouns carrying the main meaning of the construction are listed, including the noun 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).[10]

(3) *PDT-Vallex*:[11,12]



(4) Učinil rozhodnutí. (Czech)
made.MASC.SG.PAST decision.ACC
'He made a decision.'

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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.[13] 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 (F), 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; Hajič 2004), 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':[14]

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

(7) Holčina dělá dlouhý nos na
girl.NOM.F.SG makes.F.SG.3.PRES long.ACC.SG.MASC nose.ACC.SG.MASC on
kolemjdoucí. (Czech)
passers-by.ACC.PL
'The girl thumbs her nose at passers-by.'

Downloaded from https://academic.oup.com/jl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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) Bral na lehkou váhu, že se mu
take.MASC.SG.3.PAST on light.ACC.F.SG weight.ACC.F.SG that REFL him
vysmívala. (Czech)
mocked
'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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

following frame of the verb UDĚLAT 'do', with a DPHR argument meaning 'turn of one hundred eighty degrees':

(12) **udělat** ACT(1) DPHR(obrat.S4[o-1[sto-2.S4[osmdesát.S4[stupeň.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) **žít** ACT(1) DPHR(z-1[ruka.S2],do-1[ústa.P2])

(14) Firma žije z ruky do úst. (Czech)
company.NOM.F.SG lives.F.SG.3.PRES from hand.GEN.SG to mouth.GEN.PL
'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) **být** ACT(1) DPHR(názor.S2[{jiný,stejný,podobný,opačný}.#];
názor.S2[že[.v]];názor.S2[ten.#,že[.v]])
?PAT(↓že)

(16) Byli toho názoru, že je to pravda. (Czech)
be.MASC.PL.3.PAST that.GEN.MASC.SG opinion.GEN.MASC.SG, that be.N.SG.3.PRES it.NOM.N.SG truth.NOM.F.SG
'They were of the opinion that it is true.'

(17) Jsme všichni stejného názoru. (Czech)
be.MASC.PL.1.PRES all.MASC.PL.NOM same.GEN.MASC.SG opinion.GEN.MASC.SG
'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,

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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*

Walenty[15] (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: *Świgra* (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 frames[17] 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 cp(że)).

(18) subj{np(str)}+obj{np(str); cp(że)}+{np(dat)}

(19) Niektórzy głoszą teorie o szkodliwości
some.NOM preach theories.ACC about harmfulness.LOC
przedszkoli i że najlepsze dla dziecka jest
kindergartens.GEN and that best for child is
przebywanie z matką... (Polish)
being with mother
'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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

nominal phrase (`np(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 `lex` symbol, as in the simplified valency frame for the verb PŁYNĄĆ 'flow' in (20), with an example of its use given in (21).

(20) `subj{lex(np(str),sg,'krew',ratr)} +`
`{lex(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 `lex`, i.e., apart from `np` and `prepnp` illustrated above, also `adjp`

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

(adjectival phrase), `prepadjp` (prepositional phrase with an obligatorily adjectival dependent), `advp` (adverbial 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(prepnp(w,loc),pl,'żyła',ratr({adjp(agr)}+{possp}))}`

Here, in case of both lexicalised arguments, the required dependents of KREW and ŻYŁA 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 *żyłach* '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 *żyłach* '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 żyłach. (Polish)
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 cierpnie skóra na myśl o czymś* 'something makes somebody's flesh creep', lit. 'somebody.DAT creeps skin.NOM on (the) thought.ACC about something.LOC'.[20] The part corresponding to *o czymś* 'about something' in the argument expressed here as *na myśl o czymś* 'on (the) thought of

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 cierpnie 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)' (`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)
```
{lex(prepnp(na,acc),sg,'myśl',
     ratr({prepnp(o,loc);cp(że);prepncp(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 cierpnie 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 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'.[21] 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)
```
{lex(prepnp(na,acc),sg,'myśl',
     ratr({prepnp(o,loc);cp(że);prepncp(o,loc,że)}
          +{adjp(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':[22]

(26)
```
subj{np(str)}+obj{np(str)}+
{xp(mod);
 lex(prepnp(z,inst),pl,XOR('ramię','ręka'),
   ratr1({lex(adjp(agr),agr,agr,pos,'otwarty',
     atr1({lex(advp(misc),pos,'szeroko',natr)}))}))}+
{lex(prepnp(pod,acc),_,'dach',
   atr({lex(adjp(agr),agr,agr,pos,
     OR('mój','nasz','swój','twój','wasz',
        'własny'),natr)}))}
```

Downloaded from https://academic.oup.com/jjl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 RAMIĘ '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 niezwykle 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 Hajičová 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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):

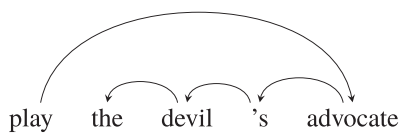
(27) *play the devil's advocate – PDT-Vallex*-style:
play ACT(1) DPHR(advocate.S['s[devil.S[the]]])

(28) *play the devil's advocate – Walenty*-style:[25]

```
play subj{np} +
     {lex(np,sg,'advocate',
       ratr1({lex(possp,'’s',
         ratr1({lex(np,'devil',
           ratr1({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:



illicit idiom:

Figure 2: An impossible idiom, according to O'Grady 1998

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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:

```
play    ACT(1) DPHR(.n['s[devil[the]]])
```

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

```
play subj{np}+
     {lex(np,_,_,
       ratr1({lex(possp,'s',
         ratr1({lex(np,'devil',
           ratr1({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_{NP} is after Y_{Numeral} 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':[26]

(31)

```
subj{np(str)}+{lex(np(inst),_,'tor',
ratr({np(gen)}+{adjp(agr)}))}
```

Downloaded from https://academic.oup.com/jjl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

The lexicalised argument headed by TOR 'track, course' in any number and in the instrumental case must be modified (*Coś biegło 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. (Polish)
association.GEN.PL
'His thoughts ran the course of strange associations.'

(33) Sprawy biegły swoimi torami. (Polish)
matter.NOM.PL ran self.INST.PL track.INST.PL
'The matters ran their own course.'

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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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:

(37) ACT(1) DPHR(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).[28] So, for example, in the frame (6) (repeated below as (38) for convenience), neither the node specified as nos.S4, nor its dependent node dlouhý:#, 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.

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

(39) **být** ACT(1)
DPHR(názor.S2[{jiný,stejný,podobný,opačný}.#];
názor.S2[že[.v]];názor.S2[ten.#,že[.v]])
?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 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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).[29] Then, the relevant constraint on the phraseological argument could be expressed as follows (we omit the other argument here):

(40) DPHR(`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_{NP} is after Y_{Numeral} 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 general 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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) `{lex(np(inst),_,'tor',ratr({np(gen)}+{adjp(agr)}))}`

(42) `{lex(np(inst),_,'tor',`
`XOR(PLUS({adjp(agr)}),{np(gen)}+STAR({adjp(agr)})))}`

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

5.2. *Word order*

Neither of the two dictionaries has any mechanisms to specify linearisation constraints on phraseological (or any other) arguments.[30] 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-rhematic – 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) `subj{np(str)}+{lex(np(str),pl,'noga',natr)}+`
`{lex(prepnp(za,acc),sg,'pas',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,[31] 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 `{lex(np(str),pl,'noga',natr)}` immediately precedes the other phraseological argument, `{lex(prepnp(za,acc),sg,'pas',natr)}`.[32,33]

Another, perhaps even more clear, example of such a linear constraint is the idiom *odsyłać kogoś od Annasza do Kajfasza* 'send someone from pillar to post', lit. 'send someone from Annas to Caiaphas', where the two PPs *od Annasza* and *do Kajfasza* must be adjacent and in the order indicated above.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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) `subj{np(str)}+{lex(np(str),pl,'noga',NONE)} ‹ {lex(prepnp(za,acc),sg,'pas',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) ACT(1) DPHR(`noga.P4,za[pas.S4]`)
(46) ACT(1) DPHR(`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 BRAĆ should follow this verb,[35] such a linear constraint may be expressed in the two dictionaries as in (47)–(48):

(47) `subj{np(str)}+«{lex(np(str),pl,'noga',NONE)} ‹ {lex(prepnp(za,acc),sg,'pas',NONE)}`
(48) ACT(1) «DPHR(`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)
```
subj{np(str)}+obj{np(str)}+
{xp(mod);
 lex(prepnp(z,inst),pl,XOR('ramię','ręka'),
   »{lex(adjp(agr),agr,agr,pos,'otwarty',
```

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

```
      OPT({lex(advp(misc),pos,'szeroko',NONE)}))})} +
{lex(prepnp(pod,acc),_,'dach',
  STAR({lex(adjp(agr),agr,agr,pos,
    OR('mój','nasz','swój','twój','wasz',
       'własny'),NONE)}))}
```

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 ŻEBY-type complementiser (`{cp(żeby)}`; this class of complementisers contains *żeby*, *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,'dąb',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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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
due moving heaven.GEN.SG and earth.GEN.SG by
zrozpaczoną matkę. . . (Polish)
despaired mother
'. . .due to moving heaven and earth by the mother in despair. . .'

(52) Manifest Matthiasa Polityckiego nie poruszył
manifesto.NOM.MASC.SG Matthias.GEN Polityki.GEN NEG moved
nieba i ziemi. . . (Polish)
heaven.GEN.SG and earth.GEN.SG
'Matthias Polityckі's manifesto didn't move heaven and earth. . .'

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

(53) DPHR(`i[niebo.S4 « ziemia.S4]`)

Note the use of the linear operator « introduced in Section 5.2 and expressing (not necessarily immediate)[38] 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 *ziemię i niebo*.[39] 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) DPHR(`i[(.n2;.u#) < oko.S7 « ucho.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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 *PDT-Vallex* to handle such cases by simply marking shared dependents with the special diacritic = and thus distinguishing them from direct conjuncts:

(55) DPHR(`i[(.n2;.u#)= < oko.S7 ≪ ucho.S7]`)

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 DPHR argument could be specified as: `i[(.n2;.u#)*= < oko.S7 ≪ 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):[40]

(56) DPHR(`i[.u#= < oko.S7 ≪ ucho.S7]; i[oko.S7 ≪ ucho.S7 < .n2=]`)

In case of *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 OR may be coordinated. This is best illustrated with the following current frame for the verb ZAMIENIAĆ SIĘ 'change (itself), metamorphose':

(57) `subj{np(str)}+{lex(prepnp(w,acc),sg,OR('proch';'pył'),natr)}`

This frame may be used to express that something has changed into ashes (Polish: a form of PROCH) or into dust (Polish: a form of PYŁ), but also into ashes and dust (*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 OR.

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

(58) `obj{lex(np(str),sg,AND('niebo';'ziemia')),NONE}`

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 *poruszyć niebo i ziemię* 'move heaven and earth',

Downloaded from https://academic.oup.com/jjl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

two singular instrumental nouns in *być czyimś 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[41] (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) Manifest Matthiasa Polityckiego nie poruszył
manifesto.NOM.MASC.SG Matthias.GEN Polityki.GEN NEG moved
ani nieba, ani ziemi... (Polish)
neither heaven.GEN.SG neither earth.GEN.SG
'Matthias Polityckiego'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) `{lex(prepnp(w,acc),sg,AND[i]('kotek';'myszka')),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)`:[42]

(62) a. `{lex(prepnp(pod,acc),pl,'strzecha',NONE)}`
b. `{lex(xp(adl[prepnp(pod,acc)]),pl,'strzecha',NONE)}`

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 `lex`, as usual:

(63) `{lex(np(inst),sg,AND[i]('oko';'ucho'),{possp})}`

In (63), expressing the phraseological argument in *być czyimś okiem i uchem* 'be somebody's eye and ear', the last parameter of `lex` is `{possp}`, which expresses a required (see Section 5.1 and note the lack of the optionality operator `OPT`) possessive dependent – either a possessive pronoun or a genitive NP, according to the definition of `possp` already assumed in *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 *Walenty* specification more fully analogous to that of *PDT-Vallex* given in (56) could then have the following form:[43]

(64) `{lex(np(inst),sg,AND[i]('oko';'ucho')),`
`XOR(«{np(gen)},»{adjp(agr)})}`

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.

(65) ...moim i mojej rodziny okiem i uchem... (Polish)
my.INST and my.GEN family.GEN eye.INST and ear.INST
'...the eyes and ears of my family and myself...'

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 *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 *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 ZDARZYĆ SIĘ 'happen', as in the attested[44] (66):

(66) Wiele się może zdarzyć między ustami a brzegiem
much REFL may happen between mouth.INST and edge.INST

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

pucharu. (Polish)
goblet.GEN
'Much may happen between a decision and its fulfilment.'

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) Kraje znajdują się między Scyllą autorytaryzmu
countries find REFL between Scylla.INST authoritarianism.GEN
a Charybdą oligarchii. (Polish)
and Charybdis.INST oligarchy.GEN
'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) Cała Kolumbia chucha i dmucha na Falcao. (Polish)
whole.NOM Colombia.NOM puffs and blows on Falcao.ACC
'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'.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

Another possibility, at least in *PDT-Vallex*, would be to represent this idiom under CHUCHAĆ, with I DMUCHAĆ treated as a dependent (e.g. headed by I). However, then a new mechanism would be necessary to express the fact that the verbal part of the dependent (i.e. the form of DMUCHAĆ) must be inflected the same way as the head verb CHUCHAĆ. 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 DWOIĆ SIĘ 'to duplicate oneself' and TROIĆ SIĘ 'to triplicate oneself' obligatorily share the reflexive marker SIĘ,[48] *coś kogoś ani ziębi, 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. 'sombody.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 attested[49] (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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

phraseological frames of the Czech verb NECHAT 'leave', given in (70) and exemplified by the attested[50] (71), requires the verb to be negated:

(70) **nechat** ACT(1) DPHR(kámen.S4,na-1[kámen.S6]) PAT(z+2) −(.∼)

(71) V roce 1997, kdy Jobs nenechal v Apple kámen
in year 1997 when Jobs not left in Apple stone.ACC.SG
na kameni...
on stone.LOC.SG (Czech)
'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 BAĆ 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:

(72) subj{np(str)}+{lex(np(gen),sg,'bóg',natr)}

(73) Ten, kto rozsypał azbest, chyba Boga się nie boi...(Polish)
that who spilled asbestos perhaps god.GEN.SG REFL NEG fears
'The one that spilled asbestos must have no fear of God...'

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, KIWNĄĆ 'nod' as used in *nie kiwnąć 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 KIWNĄĆ – it may be expressed elsewhere in the sentence, e.g. on the higher verb, as in (75):

(74) subj{np(str)}+{lex(np(inst),sg,'palec',natr)}

(75) Nikt nie chciał nawet kiwnąć palcem. (Polish)
nobody.NOM.MASC NEG wanted.MASC even nod.INF finger
'Nobody even wanted to lift a finger.'

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

However, *Walenty* currently lacks a more general mechanism of specifying arbitrary conditions on the morphosyntactic properties of the head of a given

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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 *gadał*, a form of GADAĆ '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 ZAĆPAĆ SIĘ (where ĆPAĆ means 'take drugs') or ZABEŁKOTAĆ SIĘ (where BEŁKOTAĆ means 'mumble, babble').

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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).

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

2 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).

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

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

5 https://ufal.mff.cuni.cz/vallex/2.6/data/html/generated/lexeme-entries/brat-3.html

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

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

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

9 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).

10 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.

11 http://lindat.mff.cuni.cz/services/PDT-Vallex/PDT-Vallex.html?block=U&verb=u%C4%8Dinit

12 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 (Hajič *et al*. 2012; 106 occurrences). On phraseological expressions in the latter, see also Dušek *et al*. 2014.

13 On the other hand, CPHR arguments are hardly ever – with the notable exception of the copulas BÝT and BÝVAT – more complex than in (3).

14 The specification `na+4` of the PAT argument is a shorthand for `na-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.

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

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

17 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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

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)):

(i)
```
atr({lex(adjp(agr),agr,agr,pos,XOR('mój','nasz','swój',
'twój','wasz'),natr)}+{lex(adjp(agr),agr,agr,pos,
'własny',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

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

poetry, one is stylised for old Polish, and two are from Usenet groups and also felt to be attempts at stylisation.

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.

41 http://dorzeczy.pl/id,1149/W-kolejce-do-tronu.html, accessed on 4th June 2015.

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)`.

44 http://www.proszynski.pl/Miedzy_ustami_a_brzegiem_pucharu-p-1255-1800-.html, accessed on 11 July 2015.

45 http://www.demoseuropa.eu/index.php?option=com_content&view=article&id=1474%3Aukraiskie-panta-rhei&catid=148%3A2014kom&Itemid=174&lang=pl, accessed on 11 July 2015; here simplified.

46 http://ekstraklasa.net/cala-kolumbia-chucha-i-dmucha-na-falcao-zawodnika-odwiedzil-sam-prezydent-wideo,artykul.html?material_id=52e56545b564da7113f64d85, accessed on 4th June 2015.

47 http://info.wyborcza.pl/temat/wyborcza/chucha, accessed on 4th June 2015.

48 See Kupść 1999 on the haplology of SIĘ in Polish.

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

50 http://www.appliste.cz/jonathan-ive-jako-pani-columbova/, accessed on 11 July 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.

Downloaded from https://academic.oup.com/jjl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

52 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.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

References

Al-Haj, H., A. Itai and S. Wintner (2013). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography* 27.2: 130–38.

Bejček, E., V. Kettnerová and M. Lopatková (2014). Automatic mapping lexical resources: A lexical unit as the keystone. In Calzolari *et al*. (2014), 2826–2832.

Bogusławski, A. and M. Danielewiczowa (2005). *Verba polona abscondita. Sonda słownikowa III*. Warsaw: Uniwersytet Warszawski, Katedra Lingwistyki Formalnej.

Böhmová, A., J. Hajič, E. Hajičová and B. Hladká (2003). The Prague Dependency Treebank: Three-level annotation scenario. In Abeillé A. (ed.), *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*. Dordrecht: Kluwer, 103–127.

Calzolari, N., K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds). (2014). *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland. ELRA.

Crystal, D. (1997). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell, 4th edition.

Czerepowicka, M. and I. Kosek (2011). Problemy opisu związków frazeologicznych w formalizmie „Multifleks" (na przykładzie rodzaju wyrażeń frazeologicznych). In Bańko M. and D. Kopcińska (eds), *Różne formy, różne treści: Tom ofiarowany Profesorowi Markowi Świdzińskiemu*. Warsaw: Uniwersytet Warszawski, Wydział Polonistyki, 117–126.

Dušek, O., J. Hajič and Z. Urešová (2014). Verbal valency frame detection and selection in Czech and English. In *Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Baltimore, MD. 6–11.

Gardent, C., B. Guillaume, G. Perrier and I. Falk (2005). Maurice Gross' grammar lexicon and Natural Language Processing. In Z. Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference*, Poznań, Poland.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: Chicago University Press.

Gross, M. (1984). Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING 1984)*, Stanford. 275–282.

Hajič, J. (2004). *Disambiguation of Rich Inflection*. Prague: Karolinum.

Hajič, J. and Z. Urešová (2003). Linguistic annotation: from links to cross-layer lexicons. In Nivre and Hinrichs (2003).

Hajič, J., J. Panevová, Z. Urešová, A. Bémová, V. Kolářová and P. Pajas (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In Nivre and Hinrichs (2003).

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková Razímová and Z. Urešová (2006). Prague Dependency Treebank 2.0 (PDT 2.0).

Hajič, J., E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová and Z. Žabokrtský (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In LREC (2012).

Hajnicz, E., B. Nitoń, A. Patejuk, A. Przepiórkowski and M. Woliński (2015). Internetowy słownik walencyjny języka polskiego oparty na danych korpusowych. *Prace Filologiczne* LXV: 95–110.

Herbst, T. and I. Roe (1996). How obligatory are obligatory complements? An alternative approach to the categorization of subjects and other complements in valency grammar. *English Studies* 2: 179–199.

Herbst, T., D. Heath, I. F. Roe and D. Götz (eds). (2004). *A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Mouton de Gruter.

Hlaváčková, D. and A. Horák (2005). VerbaLex – new comprehensive lexicon of verb valencies for Czech. In R. Garabík (ed.), *Computer Treatment of Slavic and East European Languages: Proceedings of the Third International Seminar, Bratislava, Slovakia, 10–12 November 2005*, Bratislava. VEDA: Vydavatel'stvo Slovenskej akadéme vied, 107–115.

Karolak, S. (1984). Składnia wyrażeń predykatywnych. In Topolińska Z. (ed.), *Gramatyka współczesnego języka polskiego: Składnia*. Warsaw: Wydawnictwo Naukowe PWN, 11–211.

Kosek, I. (2008). *Fleksja i składnia nieciągłych imiennych jednostek leksykalnych*. Olsztyn: Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego.

Kosek, I. (2013). Paradygmaty zwrotów frazeologicznych – problemy opisu leksykograficznego. In Dziamska-Lenart G. and J. Liberek (eds), *Perspektywy współczesnej frazeologii polskiej. Między teorią a praktyką leksykograficzną*, Poznań. Wydawnictwo Naukowe UAM, 51–61.

Kupść, A. (1999). Haplology of the Polish reflexive marker. In Borsley R. D. and A. Przepiórkowski (eds), *Slavic in Head-Driven Phrase Structure Grammar*. Stanford, CA: CSLI Publications, 91–124.

Landau, I. (2013). *Control in Generative Grammar: A Research Companion*. Cambridge: Cambridge University Press.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the valency lexicon. *The Prague Bulletin of Mathematical Linguistics* 79–80: 37–60.

Lopatková, M. and J. Panevová (2006). Recent developments in the theory of valency in the light of the Prague Dependency Treebank. In Šimková (2006), 83–92.

Lopatková, M., Z. Žabokrtský and V. Kettnerová (2008). *Valenční slovník českých sloves*. Prague: Karolinum.

LREC (2012). *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey. ELRA.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3.1: 31–56.

Mel'čuk, I. and A. Zholkovsky (1984). *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary*. Vienna: Wiener Slawistischer Almanach.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja and A. Lessard (1984). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques I*. Montreal: Les Presses de l'Université de Montréal.

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Dagenais, L. Elnitsky, L. Iordanskaja, M.-N. Lefebvre and S. Mantha (1988). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques II*. Montreal: Les Presses de l'Université de Montréal.

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Iordanskaja and S. Mantha (1992). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques III*. Montreal: Les Presses de l'Université de Montréal.

Mel'čuk, I., N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha and A. Polguère (1999). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-semantiques IV*. Montreal: Les Presses de l'Université de Montréal.

New Oxford Style Manual (2012). *New Oxford Style Manual*. Oxford: Oxford University Press.

Nivre, J. and E. Hinrichs (eds). (2003). *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway.

O'Grady, W. (1998). The syntax of idioms. *Natural Language and Linguistic Theory* 16: 279–312.

Panevová, J. (1974). On verbal frames in Functional Generative Description. Part 1. *The Prague Bulletin of Mathematical Linguistics* 22: 3–40.

Patejuk, A. and A. Przepiórkowski (2012). Towards an LFG parser for Polish: An exercise in parasitic grammar development. In LREC (2012), 3849–3852.

Patejuk, A. and A. Przepiórkowski (2015). Parallel development of linguistic resources: Towards a structure bank of Polish. *Prace Filologiczne* LXV: 255–270.

Popel, M., D. Mareček, J. Štěpánek, D. Zeman and Z. Žabokrtský (2013). Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria. 517–527.

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

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

Przepiórkowski, A. (2004). O wartości przypadka podmiotów liczebnikowych. *Biuletyn Polskiego Towarzystwa Językoznawczego* LX: 133–143.

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

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

Przepiórkowski, A., E. Hajnicz, A. Patejuk and M. Woliński (2014a). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, Dublin, Ireland. Association for Computational Linguistics and Dublin City University, 83–91.

Przepiórkowski, A., E. Hajnicz, A. Patejuk, M. Woliński, F. Skwarski and M. Świdziński (2014b). Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari *et al*. (2014), 2785–2792.

Downloaded from https://academic.oup.com/ijl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020

Saloni, Z. and M. Świdziński (1985). *Składnia współczesnego języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN, 2nd (changed) edition.

Savary, A. (2008). Computational inflection of multi-word units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology* 1.2: 1–53.

Schumacher, H., J. Kubczak, R. Schmidt and V. de Ruiter (2004). *VALBU – Valenzwörterbuch deutscher Verben*, volume 31 of *Studien zur deutschen Sprache. Forschungen des Instituts für Deutsche Sprache*. Tübingen: Narr.

Sgall, P. and E. Hajičová (1970). A "functional" generative description. *The Prague Bulletin of Mathematical Linguistics* 14: 9–37.

Sgall, P., L. Nebeský, A. Goralčíková and E. Hajičová (1969). *A Functional Approach to Syntax in Generative Description of Language*. New York: Elsevier Science B.V.

Sgall, P., E. Hajičová and J. Panevová (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel.

Šimková, M. (ed.). (2006). *Insight into Slovak and Czech Corpus Linguistics*. Bratislava: Veda.

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

Świdziński, M. and S. Szpakowicz (1994). Sentence schemata for the universal basic dictionary of contemporary Polish. *International Journal of Lexicography* 7.1: 1–30.

Tolone, E. and B. Sagot (2011). Using lexicon-grammar tables for French verbs in a large-coverage parser. In Z. Vetulani (ed.), *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6–8, 2009, Revised Selected Papers*, volume 6562 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, 183–191.

Trask, R. L. (1993). *A Dictionary of Grammatical Terms in Linguistics*. London: Routledge.

Urešová, Z. (2006). Verbal valency in the Prague Dependency Treebank from the annotator's viewpoint. In Šimková (2006), 93–112.

Urešová, Z. (2009). Building the PDT-Vallex valency lexicon. In *On-line Proceedings of the fifth Corpus Linguistics Conference*. University of Liverpool.

Urešová, Z. (2011). *Valenční slovník Pražského závislostního (PDT-Vallex)*. Prague: Ústav formální a aplikované lingvistiky.

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

Žabokrtský, Z. and M. Lopatková (2004). Valency frames of Czech verbs in VALLEX 1.0. In A. Meyers (ed.), *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Boston, Massachusetts, USA. Association for Computational Linguistics, 70–77.

Žabokrtský, Z. and M. Lopatková (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics* 87: 41–60.

Downloaded from https://academic.oup.com/jjl/article/30/1/1/2555471 by MFF CUNI user on 09 December 2020