Lexicographic Description of Czech Complex Predicates: Between Lexicon and Grammar

Václava Kettnerová, Petra Barančíková, Markéta Lopatková

Charles University in Prague, Faculty of Mathematics and Physics

E-mail: kettnerova,barancikova,lopatkova@ufal.mff.cuni.cz

Abstract

In this paper, we propose a representation of Czech complex predicates with light verbs in a valency lexicon. We demonstrate that if such a representation is to be theoretically adequate and at the same time economical, the information on complex predicates should be divided between two components of the lexicon, a data component and a grammar component. The data component stores all the information necessary for the generation of well-formed deep and surface syntactic structures of complex predicates, namely valency frames of both light verbs and predicative nouns, links between these frames, mapping of verbal and nominal valency complementations, and mapping of the semantic participant Instigator onto a verbal complementation. The grammar component of the lexicon, representing a part of the overall grammar of the language, contains formal rules. These rules, which are instantiated by the information stored in the data component, allow users to obtain deep and surface structures of complex predicates. Finally, the proposed model is applied in the annotation of 1.215 Czech complex predicates selected from the Czech National Corpus on the basis of frequency and saliency. The resulting data forms a solid foundation for further survey into various semantic and syntactic aspects of Czech complex predicates.

Keywords: valency lexicon, complex predicates with light verbs, predicative nouns, syntactic structure formation

1 Introduction

Multiword expressions pose serious problems for foreign speakers as well as for automated language processing (Sag *et al.* 2002). The major challenges raised by multiword expressions can be overcome by a lexicographic representation allowing for their efficiently handling in both theoretical and computational linguistics. From various types of multiword expressions, those that involve verbs are of great significance as verbs represent the syntactic center of a sentence. In this paper, we focus on one type of verbal multiword expressions – on complex predicates with light verbs in Czech (CPs). We restrict our study to CPs within which the light verb is combined with a predicative noun expressed as direct object; these predicates represent the most frequent complex predicates in Czech. We propose a formal model of their representation in a valency lexicon. The aim of this representation is not only to provide an inventory of CPs in Czech but also to describe their syntactic structure. Although our proposal is primarily designed for the Valency Lexicon of Czech verbs, VALLEX, we suppose that its main tenets can be easily adopted by other lexical resources as well.

1.1 Related Work

CPs consist of two syntactic elements – a light verb and a predicative noun – which function as a single predicative unit. Light verbs, not having individual semantic properties, are usually characterized as (to some extent) semantically bleached. These verbs as verbal components of CPs fulfill mainly grammatical functions. The meaning of CPs is primarily expressed by predicative nouns representing their nominal components. The discrepancy between semantic and syntactic behavior represents one of the main characteristics of CPs (Algeo 1995): the syntactic center of CPs is formed by the light verb which governs a predicative noun; the semantic core of CPs is, however, represented by the predicative noun which selects a light verb.

As to syntactic structures, several formal mechanisms modeling the syntactic structure formation of CPs have been proposed, see e.g. the argument merger (Grimshaw & Mester 1988), argument fusion (Butt 2010) or argument composition (Hinrichs *et al.* 1998), and (Alonso Ramos 2007). However, to our best knowledge, none of these mechanisms have been verified on corpus data or integrated into a lexicographic representation so far. At present, the most sophisticated representation of CPs can be found in the Explanatory Combinatorial Dictionary of Modern Russian, representing the lexical component of the Meaning↔Text Theory (Mel'čuk & Zholkovsky 1984). In this dictionary, CPs are described by lexical functions (Mel'čuk 1996).

In the Czech linguistics, there are several works focused on lexical and syntactic aspects of Czech CPs, see e.g. (Macháčková 1994; Cinková 2009; Radimský 2010; Kolářová 2010). Although Czech as an inflected language, encoding syntactic relations via morphological cases, provides a great opportunity to study the syntactic structure formation of CPs, none of the theoretical works focusing on Czech CPs explicitly describe their syntactic formation.

From Czech lexical resources, let us introduce two valuable resources providing the information on Czech CPs. First, PDT-Vallex is a valency lexicon linked to the Prague Dependency Treebank 3.0 (PDT), see esp. (Urešová 2009). In this lexicon, the information on CPs is simply listed in two separate valency frames: in valency frames of light verbs and in valency frames of predicative nouns; however, no clear theoretical conception and methodology are adopted there. Second, *Slovník české frazeologie a idiomatiky – výrazy slovesné*, providing the information on combinatorial potentials of verbs and nouns, represents a valuable source of verbonominal collocations with light verbs (Čermák, Hronek & Machač 1994); however, this dictionary provides a simple list of collocations with light verbs without any further information. Moreover, no testable criteria for identifying collocations with light verbs were applied. The work presented in this paper attempts to fill the above mentioned gaps in the theoretical research into Czech CPs as well as in their lexicographic representation.

1.2 Valency Lexicon of Czech Verbs, VALLEX

Our description of Czech CPs is proposed for the purpose of valency description in the *Valency Lexicon of Czech verbs, VALLEX* (Lopatková *et al*. 2016). The *VALLEX* lexicon takes the Functional Generative Description as its theoretical background (Sgall, Hajičová & Panevová 1986). This lexicon provides the information on roughly 4.570 Czech verbs corresponding to almost 10.780 lexical units (selected on the verb frequency basis in the Czech National Corpus). The lexicon has been designed for both human users and NLP tasks. The main organizational concept is represented by a lexeme – an abstract two-fold unit associating a set of lexical forms of a verb with a set of its lexical units ('senses'). A valency characteristic is related to individual lexical units of a verb and it is captured in the form of a valency frame. The valency frame consists of valency slots: each slot corresponds to a single valency complementation characterized by a functor (a label indicating the

relation of the complementation to its governing verb), by obligatoriness and by morphemic forms. In addition, the lexicon describes other syntactic properties of individual lexical units as diatheses, reciprocity, reflexivity, control, semantic-syntactic class etc.

For the description of language phenomena at the lexicon-grammar interface, e.g., diatheses, reciprocity, reflexivity, the lexicon has been recently divided into a data component and a grammar component. The data component stores the information on unmarked valency frames of lexical units of verbs (corresponding to their active, unreciprocal and irreflexive uses). The grammar component provides formal rules allowing for deriving marked valency frames (describing their passive, reciprocal and reflexive uses) from the unmarked ones. As a result, the rules stored in the grammar component make it possible to obtain all possible surface syntactic manifestations of lexical units of verbs.

2 Complex Predicates in the Data Component of the Lexicon

In the data component of the lexicon, the information on CPs is provided by valency frames of predicative nouns (Section 2.1) and by valency frames of light verbs (Section 2.2). In addition, three special attributes, lvc, map and instig, are added to complete all the information necessary for the syntactic structure formation of CPs (Sections 2.3).

2.1 Valency Frames of Predicative Nouns

The valency frame of a predicative noun underlies its deep dependency structure. Morphemic forms recorded in the valency frame indicate a usage of the noun in both nominal structures and CPs. As in case of other predicative units, each valency complementation of a predicative noun corresponds to a semantic participant. For instance, the noun *příkaz* 'order' is described by the valency frame in (1). In this frame, ACTor is mapped onto Speaker, ADDRessee onto Recipient and PATient onto Information, see the mapping in (2) and example (3):

(1) *příkaz*$_{\text{PN}}$ 'order': ACT(gen,pos;obl) ADDR(dat;obl) PAT(k+dat,inf,že,aby;obl)

(2) *příkaz*$_{\text{PN}}$ 'order': ACT ↔ Speaker$_{\text{N}}$
ADDR ↔ Recipient$_{\text{N}}$
PAT ↔ Information$_{\text{N}}$

(3) *Kapitánův*$_{\text{Speaker.ACT}}$ *příkaz* *vojákům*$_{\text{Recipient.ADDR}}$ *k ústupu*$_{\text{Information.PAT}}$ *přišel pozdě.*
Captain's order to soldiers to retreat came late.

2.2 Valency Frames of Light Verbs

The valency frame of a light verb is typically identical with the valency frame of its predicative counterpart, i.e., the valency frame of a light verb typically consists of the same number and type of valency complementations (regarding their functors and obligatoriness) and of the same set of their morphemic forms as the valency frame of its predicative counterpart. The only difference represents the functor CPHR in the valency frame of light verbs which indicates the valency position of a predicative noun. See the valency frame of the predicative verb *dát* 'to give' in (4) and the valency frame of the light verb *dát* 'to give' in (5).

(4) *dát*$_{\text{PV}}$ 'to give': ACT(nom;obl) ADDR(dat;obl) PAT(acc;obl)

(5) *dát*$_{\text{LV}}$ 'to give': ACT(nom;obl) ADDR(dat;obl) CPHR(acc;obl)

However, there is a crucial difference between valency complementations of predicative verbs and

those of light verbs: while complementations of predicative verbs are mapped onto semantic participants, complementations of light verbs – being semantically bleached – do not typically correspond to any semantic participants. For instance, with the predicative verb *dát* 'to give', ACTor corresponds to Agent, ADDRessee to Recipient and PATient to Theme, see this correspondence in (6) and example illustrating this mapping in (7). In contrast, the ACTor and ADDRessee of the light verb *dát* 'to give' are not mapped to any semantic participants (CPHR is occupied by a predicative noun). To be semantically saturated, valency complementations of a light verb enter in a coreference relation with complementations of the respective predicative noun, as is shown Section 2.3.2.

(6) *dát*$_{PV}$ 'to give':	ACT	↔	Agent$_V$	
	ADDR	↔	Recipient$_V$	
	PAT	↔	Theme$_V$	
(7) *Laban*$_{Agent.ACT}$	*dal*$_{PV}$	*velbloudům*$_{Recipient.ADDR}$		*slámu*$_{Theme.PAT}$.
Laban	gave	camels		straw.

2.3 Special Attributes

In this section, three special attributes, lvc (Section 2.3.1), map (Section 2.3.2) and instig (Section 2.3.3), are described in detail.

2.3.1 Attribute lvc

Individual valency frames of light verbs and predicative nouns that form CPs are linked by the attribute lvc, the value of which is a list of references to respective valency frames. This attribute is attached to valency frames of predicative nouns and (for user's convenience) to valency frames of light verbs as well. On the basis of references provided by the attribute lvc, individual combinations of light verbs and predicative nouns can be obtained.

2.3.2 Attribute map

One of the main characteristics of CPs is represented by the coreference between valency complementations. Light verbs being semantically bleached are typically not endowed with any semantic participants.[1] To acquire semantic specificity, they enter into the coreference with valency complementations of the predicative noun with which they form the CP. Pairs of coreferring complementations within CPs thus refer to the same semantic participants.

Let us demonstrate the mechanism of the semantic saturation of valency complementations of light verbs on the example of the CP *dát příkaz* 'to give an order'. The valency frame of the noun *příkaz* 'order' consists of three complementations, ACTor, ADDRessee and PATient, which are mapped onto Speaker, Recipient and Information, respectively, see the valency frame in (1) repeated below for convenience as (8) and the mapping in (2) repeated as (9). The valency frame of the light verb *dát* 'to give' has three complementations as well, see the valency frame in (5) repeated below as (10). CPHR is the valency position reserved for a predicative noun. The remaining two complementations, ACTor and ADDRessee, are not semantically specified (as they do not correspond to any semantic participants, see the mapping in (11)). Within the CP *dát příkaz* 'to give an order', the ACTor and the ADDRessee of the light verb corefer with the ACTor and the ADDRessee of the predicative noun *příkaz* 'order', respectively. Via the coreference, these verbal complementations acquire semantic

1 However, some light verbs contribute the participant Instigator to CPs, representing the only exception; we will address these cases below Section 2.3.3.

capacity as they refer to the same semantic participants, to Speaker and to Recipient, as the nominal complementations; see the scheme of the mapping of semantic participants in the CP *dát příkaz* 'to give an order' in (12) and illustrating example in (13).

(8) *příkaz*$_{PN}$ 'order': ACT(gen,pos;obl) ADDR(dat;obl) PAT(k+dat,inf,že,aby;obl)

(9) *příkaz*$_{PN}$ 'order': ACT ↔ Speaker$_N$
ADDR ↔ Recipient$_N$
PAT ↔ Information$_N$

(10) *dát*$_{LV}$ 'give': ACT(nom;obl) ADDR(dat;obl) CPHR(acc;obl)

(11) *dát*$_{LV}$ 'give': ACT ↔ θ
ADDR ↔ θ

(12) *dát*$_{LV}$ *příkaz*$_{PN}$ 'give an order': (ACT$_V$ → ACT$_N$) ↔ Speaker$_N$
(ADDR$_V$ → ADDR$_N$) ↔ Recipient$_N$
PAT$_N$ ↔ Information$_N$

(13) *Kapitán*$_{Speaker.V.ACT}$ *dal* *vojákům*$_{Recipient.V.ADDR}$ *příkaz*$_{V.CPHR}$ *k ústupu*$_{Information.N.PAT}$.
The captain gave soldiers the order to retreat.

As a result, valency complementations of the same light verb can be semantically specified by different semantic participants depending on the predicative noun with which it forms CPs. For instance, the light verb *dostat* 'get' forms the CPs with the predicative nouns *zpráva* 'message' and *trest* 'punishment'. Its valency complementations are then semantically saturated via the coreference with the respective nominal ones. See examples illustrating the changes in the correspondence of valency complementations of the given light verb with semantic participants depending on whether it combines with the predicative noun *zpráva* 'message' (17) or with the predicative noun *trest* 'punishment' (18) (see (14) and (15), respectively, for sets of valency complementations of the predicative nouns and their correspondence with semantic participants).

(14) *zpráva*$_{PN}$ 'message': ACT ↔ Speaker$_N$
ADDR ↔ Recipient$_N$
PAT ↔ Information$_N$

(15) *trest*$_{PN}$ 'punishment': ACT ↔ Punisher$_N$
ADDR ↔ Punishee$_N$
PAT ↔ Reason$_N$

(16) *dostat*$_{LV}$ 'get': ACT(nom;obl) CPHR(acc;obl) ORIG(od+gen;opt)

(17) *Hlídka*$_{Recipient.V.ACT}$ *nedostala*$_{LV}$ *od posádky*$_{Speaker.V.ORIG}$ *žádnou zprávu*$_{CPHR}$.
The guard did not get from the crew any message.

(18) *Brankář*$_{Punishee.V.ACT}$ *dostal*$_{LV}$ *od rozhodčího*$_{Punisher.V.ORIG}$ *trest*$_{CPHR}$.
The goalkeeper got from the referee a punishment.

The information on the coreference between valency complementations of light verbs and complementations of predicative nouns is provided in the attribute map. We attach this attribute to predicative nouns. The adopted solution is supported by two observations: (i) It is the predicative noun that selects an appropriate light verb. (ii) The coreferential relations between valency complementations within the CPs which are based on the same predicative noun can be rearranged depending on different light verbs, c.f. the mapping in the CP *dát příkaz* 'give an order' in (12) and in the CP *dostat příkaz* 'get an order' in (19).

(19) *dostat*$_{LV}$ *příkaz*$_{PN}$ 'get an order': (ORIG$_V$ → ACT$_N$) ↔ Speaker$_N$

$(ACT_V \rightarrow ADDR_N) \leftrightarrow Recipient_N$
$PAT_N \leftrightarrow Information_N$

The value of the map attribute is a list of pairs of coreferring complementations accompanied with references to relevant light verbs (see Section 2.3.1). Each predicative noun can be assigned by more than one attribute map reflecting different coreference relations (see above the observation (ii) illustrated in (12) and (19)).

2.3.3 Attribute instig

Within CPs, an event expressed by the predicative noun can be instigated by a participant external for the given event. This participant is provided by the respective light verb. We assign this participant the semantic role of Instigator. The information on the mapping of the Instigator onto a valency complementation of the light verb is recorded in the attribute instig. This attribute is accompanied with a list of references to predicative nouns which select light verbs with the Instigator (Section 2.3.1).

For instance, the event expressed by the predicative noun *možnost* 'opportunity' comprises two semantic participants, Agent and Situation. When this noun selects the light verb *dát* 'give', the event denoted by this noun is instigated by the Instigator contributed to the resulting CP by the given light verb. See the valency frame of the predicative noun *možnost* 'opportunity' in (20) and the mapping of the semantic participants provided by the noun onto nominal complementations in (21). Moreover, see the frame of the light verb in (5) repeated as (22) and the mapping of the Instigator in (23). The ACTor of the light verb (being mapped onto the Instigator) is semantically specified. The CPHR is occupied by the predicative noun *možnost* 'opportunity'. The only complementation of the light verb which remains semantically unspecified is the ADDRessee. To acquire semantic capacity, this ADDResse corefers with the ACTor of the noun, which corresponds to Agent. As a result, both the verbal ADDRessee and the nominal ACTor refer to the given participant. See the correspondence of the semantic participants with the valency complementations in the CP *dát možnost* 'to give an opportunity' in (24) and example illustrating this CP in (25).

(20) *možnost*$_{PN}$ 'opportunity': ACT(gen,pos;obl) PAT(gen,k+dat,inf,že,aby;obl)

(21) *možnost*$_{PN}$ 'opportunity': ACT $\leftrightarrow$ Agent$_N$
PAT $\leftrightarrow$ Situation$_N$

(22) *dát*$_{LV}$ 'give': ACT(nom;obl) ADDR(dat;obl) CPHR(acc;obl)

(23) *dát*$_{LV}$ 'give': ACT $\leftrightarrow$ Instigator$_V$

(24) *dát*$_{LV}$ *možnost*$_{PN}$ 'give opportunity': $ACT_V \leftrightarrow Instigator_V$
$(ADDR_V \rightarrow ACT_N) \leftrightarrow Agent_V$
$PAT_N \leftrightarrow Situation_N$

(25) Internet$_{Instigator.V.ACT}$ *dal* *lidem*$_{Agent.V.ADDR}$ nové možnosti$_{CPHR}$ publikovat.$_{Situation.N.PAT}$.
Internet gave people new opportunities to publish.

See the representation of the CPs *dát příkaz* 'to give an order' and *dát možnost* 'to give an opportunity' in figure 1:

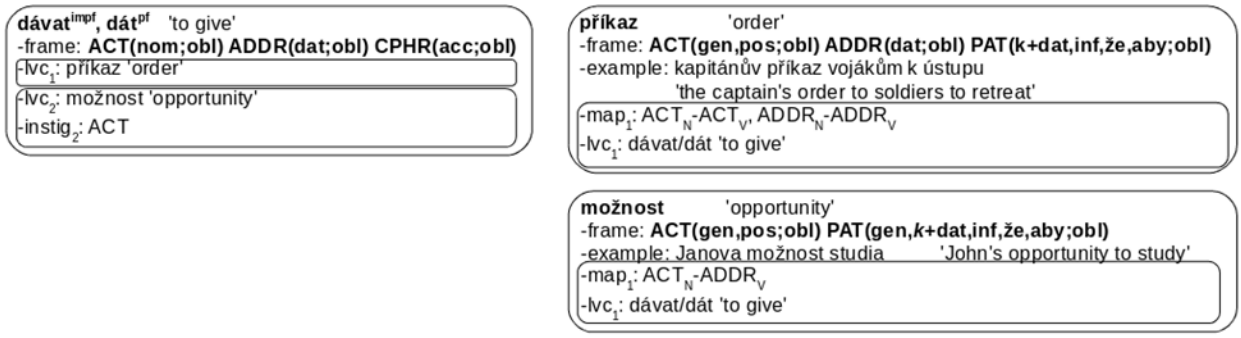

Figure 1: The lexical entry of the verb *dát, dávat* 'to give' and the predicative nouns *příkaz* 'order' and *možnost* 'opportunity' in the Valency lexicon of Czech Verbs, VALLEX.

3 Grammar Component

We can observe that the syntactic structure formation of Czech CPs is to a great extent a regular process, which can be described by formal rules. These rules – allowing for the generation of both deep and surface syntactic structures of CPs – are instantiated by the information stored in the data component of the lexicon.

3.1 Deep Syntactic Structure of Complex Predicates

For the generation of the deep syntactic structure of a CP, relevant valency frames of both the predicative noun (Section 2.1) and the light verb (Section 2.2) are identified by references provided by the attribute lvc (Section 2.3.1). Further, the information on coreference between valency complementations captured by the attribute map (Section 2.3.2) and the information on the mapping of Instigator (Section 2.3.3) is necessary, as was illustrated in previous sections.

3.2 Surface Syntactic Structure of Complex Predicates

Czech, encoding syntactic relations via morphemic forms, provides a solid basis for studying the distribution of valency complementations in the surface syntactic structure of Czech CPs. On the basis of morphemic forms of valency complementations, we can observe that this distribution is guided by the following principles:[2]

- From the valency frame of a light verb, all valency complementations are expressed on the surface, namely:

  (i) the valency complementation with the functor CPHR,
  (ii) the valency complementation corresponding to Instigator (if it is present in the frame) (the attribute instig)
  (iii) those complementations that corefer with any complementations of the predicative noun (the attribute map).

- From the valency frame of a predicative noun:

  (iv) only those valency complementations are expressed on the surface that do not corefer with any complementations of the light verb (the attribute map).

2 These principles were verified in the corpus data provided by the Prague Dependency Treebank with very satisfactory results, see (Kettnerová & Bejček 2016).

For instance, from the valency frames of the light verb *dát* 'to give' and from the frame of the predicative noun *příkaz* 'order' which form the CP *dát příkaz* 'to give an order', the following valency complementations are expressed in the surface structure: all the verbal valency complementations, namely CPHR reserved for the predicative noun (principle (i)), the verbal ACTor and ADDRessee, being in coreference with the nominal ACTor and ADDRessee (principle (iii)), and the nominal PATient, not being in coreference with any verbal complementation (principle (iv)). The nominal ACTor and ADDRessee, being in coreference with the verbal ones, remain unexpressed on the surface. See the valency frames of the light verb and of the predicative noun in (5) and (1), repeated as (26) and (27), respectively, and the information on coreference relations in (28); the surface structure of the given CP is illustrated in example (13) repeated here for convenience as (29).

(26) *dát*$_{\text{LV}}$ 'to give': ACT(nom;obl) ADDR(dat;obl) CPHR(acc;obl)
(27) *příkaz*$_{\text{PN}}$ 'order': ACT(gen,pos;obl) ADDR(dat;obl) PAT(k+dat,inf,že,aby;obl)
(28) *dát*$_{\text{LV}}$ *příkaz*$_{\text{PN}}$ 'give an order': $\text{ACT}_{\text{V}} \rightarrow \text{ACT}_{\text{N}}$, $\text{ADDR}_{\text{V}} \rightarrow \text{ADDR}_{\text{N}}$

(29) *Kapitán*$_{\text{V.ACT.nom}}$	*dal*	*vojákům*$_{\text{V.ADDR.dat}}$	*příkaz*$_{\text{V.CPHR.acc}}$	*k ústupu*$_{\text{N.PAT.k+dat}}$.
The captain	gave	soldiers	the order	to retreat.

Similarly, from the valency frames of the light verb *dát* 'to give', see above (26), and the frame of the predicative noun *možnost* 'opportunity', see (20) repeated as (30), the following valency complementations are realized in the surface structure: all the verbal complementations, namely CPHR occupied by the predicative noun (principle (i)), ACTor mapped onto Instigator (principle (ii)), ADDRessee which corefers with the nominal ACTor (principle (iii)), and the nominal PATient, which does not corefer with any verbal complementation (principle (iv)). See the valency frames in (26) and (30) and the information on coreference of complementations within the given CP provided in (31). The resulting surface structure of this CP is exemplified in (32).

(30) *možnost*$_{\text{PN}}$ 'opportunity': ACT(gen,pos;obl) PAT(gen,k+dat,inf,že,aby;obl)
(31) *dát*$_{\text{LV}}$ *možnost*$_{\text{PN}}$ 'give an opportunity': $\text{ADDR}_{\text{V}} \rightarrow \text{ACT}_{\text{N}}$

(32) *Učitel*$_{\text{V.ACT.nom}}$	*dal*	*žákům*$_{\text{V.ADDR.dat}}$	*možnost*$_{\text{V.CPHR.acc}}$	*opakovat*$_{\text{N.PAT.inf}}$	*zkoušku.*
The teacher	gave	the pupils	an opportunity	to repeat	the exam.

4 Annotation Scheme of Complex Predicates in VALLEX

In this section, the selection of the lexical stock (Section 4.1) and the annotation process (Section 4.2) are described in detail.

4.1 Selection of the Lexical Stock

The large amount of CPs in the Czech language is not easily manageable. Thus some selection criteria had to be determined at the beginning of the annotation process. The verb lemmas in VALLEX were selected on the basis of their frequency in the Czech National Corpus (CNC); at present, these lemmas cover almost 98% of verb occurrences in CNC. For the identification of CPs representing the lexical stock of the lexicon, we first had to identify an inventory of verb lemmas already stored in the VALLEX lexicon that can function as light verbs. For this purpose, we used the valency lexicon PDT-Vallex linked with the Prague Dependency Treebank (PDT).[3] In this lexicon,

3 Although PDT-Vallex contains CPs, they cannot be straightforwardly exploited as the lexical stock for two

valency frames of light verbs are indicated by the functor CPHR labeling the valency position of predicative nouns. On this basis, we automatically identified 124 verb lemmas that have at least one valency frame in PDT-Vallex with the CPHR functor. The intersection of verb lemmas obtained from PDT-Vallex and of verb lemmas contained in VALLEX was 105 in total. As VALLEX treats aspectual counterparts of verbs as a single verb lexeme, respective aspectual counterparts for selected verb lemmas have been added (if not already in the list). The resulting number of 133 verb lemmas has formed the inventory of verbs selected for the annotation process.

To identify the most frequent and most salient predicative nouns that form CPs with selected verbs, we used the Sketch Engine, a corpus tool allowing users to obtain summaries of verbs' grammatical and collocational properties; a balanced corpus of synchronous texts SYN2010 was used as the material base. In this way, the collocation lists of each selected verb lemma were obtained. From these lists, only the nominal collocates expressed as direct object in the accusative case (representing the central and most frequent case of light verb collocations in Czech) were selected: almost 3.050 nouns in total (23 nouns on average for a verb lemma).

A human annotator has been asked to indicate only those nouns in each list that represent predicative nouns forming CPs with the given verb. At the very beginning of the annotation, it was necessary to single out criteria for distinguishing collocates with light verbs from those with predicative verbs. As the main criterion, we have adopted the coreference test (Kettnerová *et al.* 2013). This test verifies whether any valency complementation of the verb is in coreference with the ACTor of the noun with which the verb forms the given collocation. The collocations that satisfy the coreference test are interpreted as CPs. The only exception is represented by the CPs that are formed by a light verb and a predicative noun with an empty valency frame, i.e., by valency frames with no valency complementation. The coreference test was satisfied by 1.215 collocations in total. These collocations represent the lexical stock of CPs prepared to be integrated into the VALLEX lexicon.

4.2 Annotation Process

The second stage of the annotation process consisted in the description of the selected CPs according to the above proposed principles (Section 2). The annotation process proceeded from the light verbs. At the very beginning, all the verb lemmas representing aspectual counterparts (or their orthographic variants) were joined under a single verb lexeme.

An annotation tool for efficiently handling all the necessary information was designed. For each selected CP, this tool allowed an annotator to select a valency frame of the predicative counterpart of the light verb (from the list of valency frames belonging to the given verb lexeme in the VALLEX lexicon) as a candidate for a valency frame of the given light verb. As valency frames of light verbs are identical with the frames of their predicative counterparts (see Section 2.2), we hypothesize that the selected frames will need only slight adjustments. The main adjustment is the change of a functor of the valency complementation which is occupied by the predicative noun into CPHR. In this way, 118 valency frames of verbs from VALLEX were selected – they represent candidates for valency frames of light verbs.

Further, for each selected CP, the annotator had to indicate an appropriate valency frame of the predicative noun which forms the given CP (see Section 2.1). As the VALLEX lexicon does not contain any predicative nouns, these valency frames were selected from the PDT-Vallex lexicon.

---

reasons: (i) the information on CPs in PDT is not complete (annotators indicated only those CPs that can be paraphrased by single predicating verbs) and (ii) PDT is not a balanced corpus – it contains only journalistic texts.

Although the annotation of predicative nouns in PDT-Vallex is not complete (only predicative nouns that occurred in the PDT are listed there) and suffers from many inconsistencies, it can serve as a solid starting point for further annotation of valency characteristics of predicative nouns. The annotator indicated 595 valency frames of predicative nouns in total.
The selected valency frames were automatically interlinked by references (see Section 2.3.1). After selecting appropriate valency frames for the light verb and for the predicative noun within the given CP, the annotator had to indicate the mapping of the Instigator – if it is relevant for the given CP – onto a verbal valency complementation. In most CPs (1.049 CPs), the Instigator is not present. It was indicated only with 166 CPs. This participant was mapped onto either the verbal ACTor (156 CPs), or the verbal ORIGin (10 CPs). Further, for each selected CP the mapping between verbal and nominal valency complementations which characterizes the given CP had to be given (see Section 2.3.2). In most cases (848 CPs), a single pair of a verbal and a nominal complementation was determined. Further, two pairs of coreferring verbal and nominal complementations are common as well (367 CPs). Three pairs were not found in the annotated data at all. In total, 19 types of the mapping between nominal and verbal complementations in the selected CPs were determined (see table 1).

Instig	Map	Num	Ex
	ACT-ACT	689	*vést diskusi* 'hold a discussion'
	ACT-ACT, ADDR-ADDR	107	*vydat povolení* 'grant permission'
	ACT-ACT, PAT-ADDR	71	*poskytovat útěchu* 'offer a comfort'
	ACT-ACT, PAT-DIR3	4	*klást důraz* 'lay emphasis'
	ACT-ACT, PAT-LOC	2	*najít inspiraci* 'find inspiration'
	ACT-ACT, ADDR-DIR3	1	věnovat pohled 'give a look'
	ACT-ACT, DIR3-DIR3	1	*vrhat pohled* 'cast a look'
	ACT-LOC, PAT-ACT	92	*najít pochopení* 'find an understanding'
	ACT-LOC, ORIG-ACT	3	*vyvolat pocit* 'arouse a feeling'
	ACT-LOC, ADDR-ACT	2	*budit podezření* 'raise a suspicion'
	ACT-ORIG, ADDR-ACT	35	*vzít si úplatek* 'take a bribe'
	ACT-ORIG, PAT-ACT	19	*přijmout křest* 'get a baptism'
	ACT-ADDR, PAT-ACT	22	*činit obtíže* 'pose difficulties'
	ACT-ADDR, LOC-ACT	1	*udělit audienci* 'give an audience'
ACT	ACT-ADDR	76	*přinést úleva* 'bring a relief'
ACT	ACT-LOC	42	*vyvolat protest* 'raise protest'
ACT	ACT-ORIG	17	*převzít úkol* 'take over a task'
ACT	ACT-BEN	14	*zvedat sebevědomí* 'raise self-confidence'
ACT	ACT-LOC, PAT-ADDR	7	*vynést nominaci* 'bring a nomination'
ORIG	ACT-ACT	10	dostat přednost 'get priority'

Table 1: Quantitative properties of the mappings between verbal and nominal complementations and of the occurrences of the Instigator within Czech CPs; in pairs of coreferring valency complementations, the first complementation is the nominal one, the second complementation is the verbal one.

5 Conclusion

We have proposed a novel representation of Czech complex predicates with light verbs. This representation has been proposed for the Valency Lexicon of Czech Verbs, VALLEX; however, it can

be applied in other lexical resources as well. We have demonstrated that if this representation should be adequate and economical, the information on complex predicates is to be divided into the data and grammar component of the lexicon. The information on the valency frames of both light verbs and predicative nouns provided by the data component captures a core of the deep syntactic structure of complex predicates. Moreover, the information on the correspondence of verbal and nominal valency complementations and on the mapping of the participant Instigator together with valency frames serves as the input for deriving surface syntactic structures of complex predicates. This derivation is specified by general rules captured in the grammar component of the lexicon.

Moreover, based on the information provided in the lexicon, predicative nouns can be easily classified according to their combinatorial potentials with light verbs and further on the basis of the correspondence of their valency complementations with complementations of light verbs. Similar observations can be made also for light verbs. Moreover, in case of light verbs, the links between valency frames of light verbs and valency frames of their predicative counterparts provide users with possibility to further study the process of semantic bleaching. To sum, all the information on complex predicates stored in the data component of the lexicon forms a solid basis for further exploration of those semantic aspects of complex predicates that determine their syntactic behavior.

References

Algeo, J. (1995). Having a Look at the Expanded Predicate. In B. Aarts, Ch. F. Meyer (eds.) *The Verb in Contemporary English: Theory and Description.* Cambridge: Cambridge University Press, pp. 203-217.

Alonso Ramos, M. (2007). Towards the Synthesis of Support Verbs Constructions. In L. Wanner (ed.) *Selected Lexical and Grammatical Issues on the Meaning Text Theory*. Amsterdam: John Benjamins Publishing Company, pp. 97-137.

Butt, M. (2010). The Light Verb Jungle: Still Hacking Away. In M. Amberber, M. Harvey & B. Baker (eds.) *Complex Predicates in Cross-Linguistic Perspective*. New York: Cambridge University Press, pp. 48-78.

Czech National Corpus. Accessed at: http://www.korpus.cz [01/05/2016].

Cinková, S. (2009). Words that Matter. Towards a Swedish-Czech Colligational Lexicon of Basic Verbs. Praha: Ústav formální a aplikované lingvistiky.

Čermák, F., Hronek, J. & Machač, J. (1994). *Slovník české frazeologie a idiomatiky – výrazy slovesné*. Praha: Academia.

Grimshaw, J., Mester, A. (1988). Light Verbs and Θ-Marking. In *Linguistic Inquiry*, 19(2), pp. 205-232.

Hinrichs, E., Kathol, A. & Nakazawa, T. (eds.) (1998). *Complex Predicates in Nonderivational Syntax, Syntax and Semantics*. San Diego: Academic Press.

Kettnerová, V., Bejček, E. (2016). Distribution of Valency Complements in Czech Complex Predicates: Between Verb and Noun. In *Proceedings of the 10th Conference on Language Resources and Evaluation*, Portorož.

Kettnerová, V., Lopatková, M. (2015). At the Lexicon-Grammar Interface: The Case of Complex Predicates in the Functional Generative Description. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015).* Uppsala: Uppsala University, pp. 191-200.

Kettnerová, V., Lopatková, M., Bejček, E., Vernerová, A. & Podobová, M. (2013). Corpus Based

Identification of Czech Light Verbs. In K. Gajdošová, K., A. Žáková (eds.) *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*. Lüdenscheid: RAM-Verlag, pp. 118-128.

Kolářová, V. (2010). Valence deverbativních substantiv v češtině (na materiálu substantiv s dativní valencí). Praha: Karolinum.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A. & Žabokrtský, Z. (2016). *Valenční slovník českých sloves*. Praha: Karolinum.

Macháčková, E. (1994). Constructions with Verbs and Abstract Nouns in Czech (Analytical Predicates). In S. Čmejrková, F. Štícha (eds.) *The syntax of sentence and text*. Amsterdam, Philadelphia: John Benjamins Publishing Company, pp. 365-374.

Mel'čuk, I.A. (1996). Lexical Functions. A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural language Processing*. Amsterdam, Philadelphia: John Benjamins Publishing Company, pp. 37-105.

Mel'čuk, I.A., Zholkovsky, A.K. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna: Wiener Slawistischer Almanach.

PDT-Vallex. Accessed at: http://lindat.mff.cuni.cz/services/PDT-Vallex [01/05/2016].

Prague Dependency Treebank. Accessed at: http://ufal.mff.cuni.cz/pdt3.0 [01/05/2016].

Radimský, J. (2010). *Verbonominální predikát s kategoriálním slovesem*. České Budějovice: Editio Universitatis Bohemiae Meridionalis.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics.* Mexico City, pp. 1-15.

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

Urešová Z. (2009). Building the PDT-VALLEX Valency Lexicon. In *On-line Proceedings of the Fifth Corpus Linguistics Conference*. University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009/

VALLEX 3.0. Accessed at: http://ufal.mff.cuni.cz/vallex/3.0 [01/05/2016].

Acknowledgements

The work on this project has been supported by the grant GAČR No. GA15-09979S and partially by the LINDAT/CLARIN project No. LM2015071. This work has been using language resources distributed by the LINDAT/CLARIN project of the MŠMT No. LM2015071.