Distribution of Valency Complements in Czech Complex Predicates: Between Verb and Noun

Václava Kettnerová and Eduard Bejček
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
{kettnerova,bejcek}@ufal.mff.cuni.cz

Abstract

In this paper, we focus on Czech complex predicates formed by a light verb and a predicative noun expressed as the direct object. Although Czech – as an inflectional language encoding syntactic relations via morphological cases – provides an excellent opportunity to study the distribution of valency complements in the syntactic structure with complex predicates, this distribution has not been described so far. On the basis of a manual analysis of the richly annotated data from the Prague Dependency Treebank, we thus formulate principles governing this distribution. In an automatic experiment, we verify these principles on well-formed syntactic structures from the Prague Dependency Treebank and the Prague Czech-English Dependency Treebank with very satisfactory results: the distribution of 97% of valency complements in the surface structure is governed by the proposed principles. These results corroborate that the surface structure formation of complex predicates is a regular process.

Keywords: Czech complex predicates, distribution of valency complements, formal rules

1. Introduction

Multiword expressions (MWEs) have proven to be a serious challenge for NLP (Sag et al., 2002). From various types of MWEs, those that involve verbs are of great significance as the verb is the syntactic center of the sentence. In this paper, we focus on one type of Czech verbal MWEs – on complex predicates with light verbs (CPs). CPs consist of two syntactic elements – the light verb and a predicative noun, adjective, adverb or verb – which function together as a single predicative unit. With respect to the wide range of issues, we limit our study to a central type of Czech CPs – to those CPs in which the predicative noun is expressed as the direct object (e.g., *vést jednání* ‘to hold talks’, *mít potíže* ‘to have difficulties’, *udělat chybu* ‘to make a mistake’, *dostat příkaz* ‘to get an order’). We aim to formulate principles governing the surface structure of these CPs and to verify these principles on corpus data.

CPs exhibit a discrepancy between syntax and semantics. Their syntactic center is the light verb, which requires the predicative noun as one of its valency complements. Their semantic core is, however, the predicative noun, which selects the light verb (Algeo, 1995). Both the light verb and the predicative noun typically contribute their valency complements to the syntactic structure of CPs. See examples (1) and (2) with the verb *dostat* ‘to get’. Unlike the syntactic structure with the predicative verb *dostat* ‘to get’ in (1), the structure with the CP *dostat příkaz* in (2) is constructed of both valency complements of the light verb *dostat* ‘to get’ and complements of the predicative noun *příkaz* ‘order’, which adds the valency complement *střílet* ‘to shoot’ to the resulting structure.

(1) *Vojáci dostali od velitele zbraně.*
soldiers – got – from – the commander – guns
‘The soldiers got guns from the commander.’

(2) *Vojáci dostali od velitele rozkaz střílet.*
soldiers – **got** – from the commander – **the order** – to shoot
‘The soldiers got the order to shoot from the commander.’

The CP usually provides a redundant number of valency slots for the expression of participants of the action denoted by the given CP. As a result, some valency complements of the light verb and of the predicative noun within CPs typically corefer. As a consequence of this coreference, only several verbal and nominal complements are expressed in the surface structure of well formed sentences with CPs. To determine which valency complements of the light verb and which complements of the predicative noun should be present in the surface structure and which should be omitted from the surface, i.e., to determine the distribution of these valency complements, represents a substantial task of the current theoretical and computational linguistics.

Although Czech as an inflectional language gives us reliable clues in a form of morphological forms for determining whether a valency complement expressed on the surface belongs to the light verb or to the predicative noun, none of the works focused on Czech CPs, see esp. Radimský (2010) and Macháčková (1994), provides an explicit description of the distribution of verbal and nominal complements in the surface structure with CPs.

For other languages, several formal mechanisms describing the distribution of complements of the light verb and the predicative noun in the surface structure of CPs have been proposed: see e.g. argument merger (Grimshaw and Mester, 1988), argument fusion (Butt, 2010), argument composition (Hinrichs et al., 1998), and also Alonso Ramos (2007). However, to our best knowledge, none of these mechanisms have been verified on corpus data.

The main goal of this paper is thus twofold: (i) to formulate principles governing the distribution of valency complements of the light verb and the predicative noun in surface structures of Czech CPs and (ii) to verify these principles on well-formed structures of CPs.

Our study takes advantage of the theoretical results formulated in Kettnerová and Lopatková (2015) and applies these results to corpus data. In our experiment, we make use of the language data from the Prague Dependency Treebank 3.0 (Bejček et al., 2013, PDT),[1] which is linked with the valency lexicon PDT-Vallex (Urešová et al., 2014).[2] The rich annotation of the PDT allows us to observe valency complements of the light verb and the predicative noun and coreferential relations between them. However, the number of CPs in this corpus is limited. We thus supplemented the data from the PDT with the data from the Czech part of the Prague Czech-English Dependency Treebank 2.0 (Hajič et al., 2012).[3] Both these treebanks use the same theoretical background, the Functional Generative Description (FGD), see esp. (Sgall et al., 1986), and their annotation guidelines are thus very similar. However, the annotation of CPs in the PCEDT is not complete. For this reason, we formulated principles for the distribution of valency complements in the surface structure of CPs on the data from the PDT. Then we verified the proposed principles on both the data from the PDT and the data from the PCEDT. However, with respect to the incomplete information on CPs in the PCEDT, the results from this corpus had to go through a manual correction.

The paper is structured as follows. First, we introduce the annotation principles of CPs adopted in the PDT, see Section 2. Second, on the basis of the manual analysis of the data from the PDT, we propose principles governing the distribution of valency complements in syntactic structures of Czech CPs and we evaluate cases where the formulated principles incorrectly determine this distribution. On the basis of this evaluation, we slightly modify the proposed principles, see Section 3. Finally, we verify these principles by applying them to the data from the PDT and from the Czech part of the PCEDT, see Section 4.

2. Complex Predicates in the PDT

On the tectogrammatical layer of the PDT (representing the deep syntactic structure of a sentence in the form of a tree with labeled nodes and edges), a CP is represented by two nodes: by a node representing the light verb and by a node representing the predicative noun. The node representing the light verb is assigned a functor (a syntactic-semantic label indicating the relation of a dependent node to its governing node) according to the function of the CP in the sentence structure, typically PRED for a predicate. The node of the dependent predicative noun – which is represented as a direct daughter of the light verb – is given the functor CPHR (compound phraseme), see below Figure 1.

2.1. Valency frames

In the first step of the annotation of CPs on the tectogrammatical layer, both the light verb and the predicative noun within CPs were assigned their respective valency frames from the PDT-Vallex lexicon. In PDT-Vallex, the deep syntactic structure of a CP is described by a valency frame of the light verb and by a valency frame of the predicative noun. These frames are modeled as sequences of valency slots: each slot stands for one valency complement. It consists of a functor and a surface syntactic form (specifying surface dependency relation and morphological form). Moreover, the information whether the complement is obligatory or optional is specified in the frame.

- The *valency frame of the light verb* typically corresponds to the valency frame of its predicative verb counterpart. The only exception is the functor CPHR labeling the predicative noun. The surface form of the CPHR is in the lexicon represented by a list of predicative nouns and their morphological form. See the simplified valency frame of the predicative verb (PV) *dostat* 'to get' in (3) and the light verb (LV) *dostat* 'to get' in (5) and examples illustrating these frames in (4) and (6), respectively:[4]

(3) $dostat_{PV}$: ACT$_{1}$ PAT$_{4}$?ORIG$_{\text{od+2,z+2}}$
'to get'

(4) *Jana*$_{\text{ACT:1}}$ *dostala od otce*$_{\text{ORIG:od+2}}$ *nové kolo*$_{\text{PAT:4}}$.
Jane$_{\text{ACT}}$ – got – from father$_{\text{ORIG}}$ – new bicycle$_{\text{PAT}}$
'Jane got a new bicycle from her father.'

(5) $dostat_{LV}$: ACT$_{1}$ CPHR$_{\{\text{příkaz},\ldots\}4}$?ORIG$_{\text{od+2,z+2}}$
'to get'

(6) *Jana*$_{\text{ACT:1}}$ *dostala od otce*$_{\text{ORIG:od+2}}$ *příkaz*$_{\text{CPHR:4}}$ *pohlídat mladšího bratra.*
Jane$_{\text{ACT}}$ – **got** – from father$_{\text{ORIG}}$ – **order**$_{\text{CPHR}}$ – to watch – younger – brother
'Jane was ordered by her father to watch her younger brother.'

- The *valency frame of the predicative noun* (PN) underlies the usage of the noun in nominal structures. See the valency frame of the predicative noun *příkaz* 'order' (7) and example documenting this frame in (8):

(7) $příkaz_{PN}$: ACT$_{2,u}$ ADDR$_{3}$ PAT$_{\text{k+3,f,aby,ať,že,c}}$
'order'

(8) *Otcův*$_{\text{ACT:u}}$ *příkaz Janě*$_{\text{ADDR:3}}$, *(aby pohlídala mladšího bratra)*$_{\text{PAT:aby}}$, *nebyl spravedlivý.*
'Father's$_{\text{ACT}}$ order to Jane$_{\text{ADDR}}$ (to watch her younger brother)$_{\text{PAT}}$ was not fair.'

2.2. Distinguishing Verbal and Nominal Complements

In the second step of the annotation of CPs on the tectogrammatical layer, annotators had to determine whether a valency complement expressed in the surface structure of the given sentence belongs to the light verb or to the predicative noun. In most instances, morphological forms of valency complements unambiguously identify to which valency frame a certain complement falls. See the valency frame of the light verb *dostat* 'to get' repeated in (9) and the frame of the predicative noun *příkaz* repeated in (10), which form the tectogrammatical structure of the CP *dostat příkaz* 'to get an order'. The following repeated ex-

[1] http://ufal.mff.cuni.cz/pdt3.0

[2] http://lindat.mff.cuni.cz/services/PDT-Vallex/

[3] http://ufal.mff.cuni.cz/pcedt2.0/en/index.html

[4] The question mark in valency frames indicates optional valency complements. The numbers refer to respective morphological cases. f indicates infinitive, u possessive pronoun or adjective and c dependent content clause introduced by a relative pronoun or adverb. The dependent content clauses introduced by conjunctions are represented by the respective conjunctions.

ample (11) illustrates this CP. The nominative case identifies ACT, the prepositional group od+2 identifies ORIG, and the accusative identifies CPHR from the valency frame of the light verb. The infinitive indicates PAT from the valency frame of the predicative noun.

(9) *dostat*$_{LV}$: ACT$_{1}$ CPHR$_{\{\text{příkaz},\ldots\}4}$?ORIG$_{\text{od}+2,\text{z}+2}$
‘to get’

(10) *příkaz*$_{PN}$: ACT$_{2,\text{u}}$ ADDR$_{3}$ PAT$_{\text{k}+3,\text{f},\text{aby},\text{ať},\text{že},\text{c}}$
‘order’

(11) *Jana*$_{\text{V:ACT:1}}$ *dostala od otce*$_{\text{V:ORIG:od+2}}$ *příkaz*$_{\text{V:CPHR:4}}$ *pohlídat*$_{\text{N:PAT:f}}$ *mladšího bratra.*
Jane$_{\text{V:ACT}}$ – **got** – from father$_{\text{V:ORIG}}$ – **order**$_{\text{V:CPHR}}$ – to watch$_{\text{N:PAT}}$ – younger – brother
‘Jane was ordered by her father to watch her younger brother.’

In certain cases, however, a valency complement expressed on the surface has the same form in the valency frame of both the light verb and the predicative noun. In this case, the complement was treated as a complement of the light verb. Although these cases require further investigation, several indications (esp. diatheses and word order) justify the adopted convention.

2.3. Ellipsis of Valency Complements

As the final step of the annotation of CPs on the tectogrammatical layer, all obligatory valency complements not present on the surface (but present in the valency frames, i.e., in the deep syntactic structure) were added. Valency complements omitted from the surface structure with CPs can be esp. of the following types:

- #QCor: an actant[5] omitted from the surface structure due to grammatical ellipsis: the omitted actant is in grammatical coreference (so called quasi control) with a coreferred element. This coreference reflects the fact that certain valency complements of light verbs and predicative nouns within CPs are referentially identical (see Figure 1).

- #PersPron: an actant omitted from the surface structure due to textual ellipsis: the omitted actant is in textual coreference with a coreferred element. In addition, it represents the null subject of the governing verb of a clause which is omitted on the surface due to Czech being a pro-drop language. Moreover, it can substitute personal or possessive pronoun.

- #Gen: an actant omitted from the surface which is subject to systemic grammatical ellipsis. This actant refers to entities usual or typical in the given situation which can be determined from a broad context. This actant is not in any type of coreference.

- #Oblfm: an omitted obligatory free modification which is subject to different types of ellipsis.

For example, ACT and ADDR from the valency frame of the predicative noun *příkaz* ‘order’, although not expressed on the surface, are added to the tectogrammatical structure with the CP *dostat příkaz* ‘to get an order’ given in (11). See the valency frame of the light verb *dostat* in (9) and the frame of the predicative noun *příkaz* in (10) in Section 2.2. The simplified tectogrammatical tree of this sentence is displayed in Figure 1.

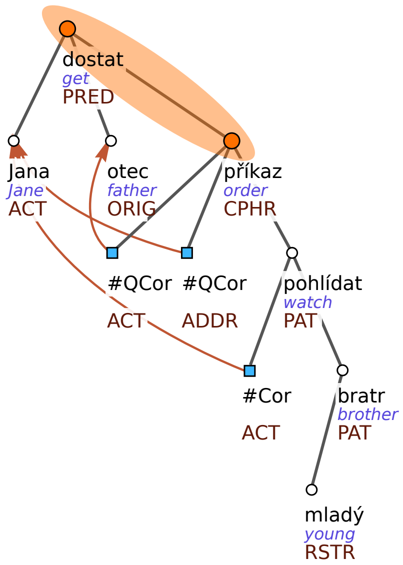

Figure 1: The simplified tectogrammatical structure of the sentence in example (11) (an orange oval shows the respective CP, blue squares mark the nodes that are not present on the surface and brown arrows indicate grammatical coreference).

2.4. Basic Statistics

The data from the PDT with its rich annotation provides a solid basis for the study of such a complex language phenomenon as CPs are. However, the number of CPs in the PDT is limited to 2,778 instances of CPs in 2,558 sentences. We restricted our experiment to the CPs in which the light verb has a finite active form and the predicative noun is expressed as its direct object in prepositionless accusative as these CPs represent the most frequent and central type of Czech CPs.

In addition, these CPs have to be realized in the same clause; this requirement filters out those instances in which the predicative noun is relativized. In these cases, the functor CPHR is typically assigned to a relative pronoun referring to the predicative noun. As valency frames are not assigned to pronouns, these cases are not relevant for our experiment.

In the PDT, 1,695 instances of CPs satisfy the given conditions. This number was determined by means of the PML-TQ search engine, a tool for processing complex queries over the data in the PDT and in the PDT-Vallex lexicon (Pajas and Štěpánek, 2009), see Figure 2 displaying an example of such a complex query.

We divided the selected instances of CPs into two portions: the first portion (659 CPs, i.e., 40% of the given instances) were subject to a manual analysis and the second portion

[5] In accordance with the FGD, valency complements are divided into actants (roughly corresponding to arguments) and free modifications (corresponding to adjuncts) (Panevová, 1994).

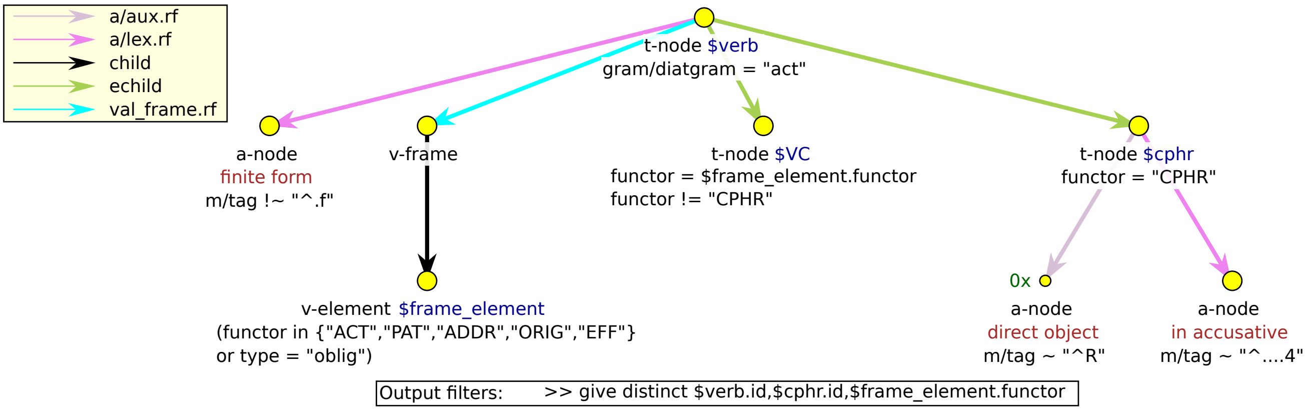

Figure 2: An example of a PML-TQ query in its graphical form. The query finds light verbs (the top node marked as `$verb`) expressed in a finite active form having a dependent predicative noun (`$cphr`) realized as direct object in prepositionless accusative. As output this query gives all verbal complements (`$VC`) (except for CPHR (other than `$cphr`)) that are in the valency frame of the `$verb` in the PDT-Vallex.

(1,036 CPs, i.e., 60% of the given instances) were used in an automatic experiment.

First, we manually analyzed the distribution of 2,034 valency complements in the surface structure belonging to 659 CPs from the first portion of the selected CPs in order to establish principles governing their distribution in the surface structure, see Section 3.

Second, 3,264 valency complements of 1,036 CPs from the second portion of the selected data were used in the automatic experiment. To extend the data for the experiment, we made use of the Czech part of the PCEDT, containing a manually annotated Czech translation of the Penn Treebank-Wall Street Journal texts. The PCEDT contains 2,116 CPs of the above given type which have 2,649 valency complements in total, see Section 4.

Let us stress that from both the analysis and the experiment, we excluded valency complements of light verbs labeled with the CPHR functor because the expression of this valency complement is already given by our criteria for the selection of CPs (we made use only those CPs in which CPHR is expressed as the direct object in prepositionless accusative).

3. Manual Analysis

In the manual analysis, we have identified that the main role in the distribution of valency complements in the surface structure of CPs is played by the grammatical coreference of *quasi control*. This type of coreference is specific to CPs within which pairs of verbal and nominal valency complements typically share the same reference, i.e., they refer to the same entity, see Section 2.3.

For example, in the sentence with the CP *dostat příkaz* ‘to get an order’ in (11) in Section 2.2., both the verbal ACT and the nominal ADDR refer to a person to whom the order is given (*Jana* ‘Jane’). At the same time, both the verbal ORIG and the nominal ACT refer to a person who gives the order (*otec* ‘father’). We can observe that from the coreferring pairs of valency complements only the verbal ones, ACT and ORIG, are expressed in the surface structure of the sentence. See Figure 1 in Section 2.3. in which the verbal ACT and ORIG, which are expressed on the surface, are represented by their respective lemmas, *Jana* ‘Jane’ and *otec* ‘father’, respectively, whereas the nominal ACT and ADDR, which are not expressed on the surface, are represented by the lemma #QCor. From the nominal complements, only PAT, which does not corefer with any verbal complement, is expressed on the surface. See Figure 1 in which the PAT is represented by the lemma *pohlídat* ‘to watch’.

3.1. Principles of the Distribution

As the same principles of the distribution of verbal and nominal complements in the surface structure hold for most manually analyzed CPs from the PDT, we formulated two basic principles:

(i) *from the valency frame of the light verb*, only those complements are expressed on the surface that *are* in the coreference of quasi control with any nominal ones,

(ii) *from the valency frame of the predicative noun*, only those complements are expressed on the surface that *are not* in the coreference of quasi control with any verbal ones.

These principles apply only to those valency complements of light verbs and predicative nouns that belong to their valency frames, i.e., to those complements that form the syntactic core of well-formed sentences. Optional free modifications expressing circumstances such as time, place, direction, manner, etc. were left aside. Recall that the complement with the functor CPHR was excluded from the analysis as its surface realization is given by conditions on CPs, see Section 2.4.

The manual analysis has shown that these principles correctly determine the distribution of 92.1% of valency complements in the surface structure (the distribution of 1,873 out of 2,034 valency complements in total complies with the proposed principles), see Table 1. The principles were violated by 7.9% of valency complements (the distribution of 161 complements in total does not conform the above given principles). See Table 2 summarizing the results of the manual analysis.

	PDT
CPs	659
All complements	2,034
Verbal complements	796
Nominal complements	1,238

Table 1: Basic statistics on the manual analysis.

Distribution	Correct	Incorrect
All complements	1,873	161
	92.1%	7.9%
Verbal complements	653	143
	32.1%	7.0%
Nominal complements	1,220	18
	60.0%	0.9%

Table 2: Statistics on the distribution of valency complements in the surface structure of CPs governed by the proposed principles.

Let us remark that valency complements subject to other types of ellipsis than that brought about by the coreference of quasi control, see Section 2.3., were treated as expressed on the surface. This treatment is supported by the observation that unlike the valency complements subject to the coreference of quasi control these valency complements can be easily added to the surface sentence with CPs. Moreover, if we restricted our experiment only to sentences with CPs where ellipsis as a consequence of the coreference of quasi control is present, the number of sentences would be rather low.

3.2. Modification of the Principles

In the final step of the manual analysis, we analyzed valency complements whose distribution in the surface structure does not comply with the proposed principles (161 cases, see Table 2 in Section 3.1.). A certain portion of these cases arises from annotation errors (e.g., missing valency complements, incorrect coreference relations). However, there is one type of cases which is not governed by the above given principles despite not being brought about by annotation errors.

We observe that principle (i) incorrectly determines the distribution of such verbal ACT that can be characterized as an instigator of the action denoted by a given CP (64 cases out of 161 incorrectly distributed complements). This ACT is expressed on the surface although it is not in the coreference of quasi control with any nominal complement. See example (12) and its tectogrammatical structure displayed in Figure 3:

(12) *Televize*$_{\text{V:Instigator:ACT}}$ *dává i další příležitosti k podnikání.*
'The television$_{\text{V:Instigator:ACT}}$ gives even more opportunities for business.'

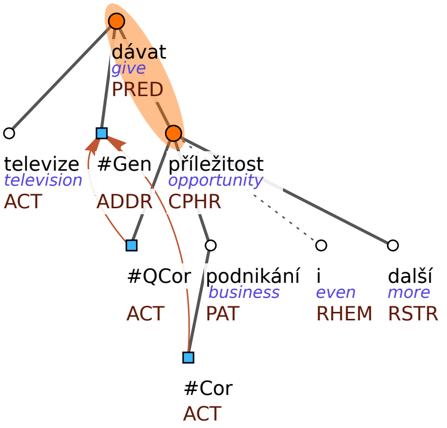

Figure 3: The simplified tectogrammatical structure of the sentence in example (12).

As a result of the evaluation, we modified the principles governing the distribution of verbal complements proposed in Section 3.1. in the following way:

(i) *from the valency frame of the light verb*, the following valency complements are expressed on the surface:
  - ACT, regardless of its coreference of quasi control,
  - all the remaining complements that *are* in the coreference of quasi control with any nominal ones,

(ii) *from the valency frame of the predicative noun*, the following valency complements are expressed on the surface:
  - all complements that *are not* in the coreference of quasi control with any verbal ones.

Although principle (ii) underlying the distribution of nominal complements does not require any modification, the manual analysis revealed rare cases in which the nominal complement ACT is expressed on the surface despite its coreference with the verbal ACT. For example, in (13) the nominal ACT corefers with the verbal ACT and it is still expressed on the surface. These cases, however, do not violate principle (ii) as this coreference is not the coreference of quasi control and such complements are thus treated in accordance with principle (ii) as expressed on the surface.

(13) *Nikdo*$_{\text{V:ACT}}$ *nevěnuje lásce celý svůj*$_{\text{N:ACT}}$ *čas.*
'Nobody$_{\text{V:ACT}}$ gives love all their$_{\text{N:ACT}}$ time.'

4. Experiment

To verify the proposed principles, we have applied them to 3,264 valency complements pertaining to the second portion of the data selected from the PDT comprising 1,036 CPs in total, see Table 3. First, we identified the light verb and the predicative noun for each CP on the basis of the annotation, see Section 2. Then all valency complements of the light verb (except for CPHR) and all complements of the predicative noun within the given CP were automatically extracted from their valency frames stored in the PDT-Vallex. Moreover, the information on the coreference of quasi control between verbal and nominal complements within the given CPs recorded in the PDT was used. Second, for each extracted valency complement, we have applied the above given principles to determine whether it

	PDT	PCEDT
CPs	1,036	2,116
All complements	3,264	N/A
Complements of verbs	1,293	2,649
Complements of nouns	1,971	N/A

Table 3: Basic statistics on the automatic experiment.

should be or whether it should not be expressed in the surface structure. As a result, each valency complement of the light verb and of the predicative noun within the given CP was assigned a binary value.
Finally, in s̃entence with the given CP in the PDT, we automatically verified whether each assigned value is correct or not. As a result, we found out that the distribution of 97.0% of valency complements follows the proposed principles. The distribution of valency complements was incorrectly determined only in 97 cases. These results show that the proposed principles are quite reliable in predicting the distribution of complements in the surface structure with CPs, see Table 4.

Distribution	Correct	Incorrect
All complements	3,167	97
	97.0%	3.0%
Verbal complements	1,206	87
	36.9%	2.7%
Nominal complements	1,961	10
	60.1%	0.3%

Table 4: Statistics on the distribution of valency complements in the surface structure of CPs from the PDT governed by the proposed rules.

To obtain more data, we made use of the Czech part of the PCEDT, see above Table 3. However, the annotation of CPs in the PCEDT is not as rich as in the PDT. The main difference lies in the annotation of predicative nouns: only a part of deverbal predicative nouns were assigned all valency complements from their valency frames; the remaining predicative nouns – even if the noun is part of a CP – were assigned only those complements that are expressed in the surface structure. As a result, information on the coreference of quasi control is missing. For these reasons, the principle determining the distribution of nominal complements in the surface structure of CPs was not verifiable on the PCEDT data and the principles for the distribution of verbal complements could not be tested using the coreference of quasi control, see Section 3.2. These principles were applied in the experiment even though we were aware that results would necessarily be negatively biased, see Table 5. The results then had to be manually checked as the information on the coreference of quasi control was not available. One half of the incorrectly distributed valency complements were processed by a human annotator. The annotator had to indicate whether these verbal complements really violate the proposed principles or whether the perceived non-adherence to the principles is only due to the missing information on the coreference of quasi control. The manually corrected results show that the proposed principles determine the distribution of almost 95% of verbal complements, which corresponds to the results obtained from more precise but small data of the PDT, see Table 6.

Distribution	Correct	Incorrect
Verbal complements	2,116	533
	79.9%	20.1%

Table 5: Statistics on the distribution of valency complements in the surface structure of CPs from the PCEDT governed by the proposed rules.

Distribution	Correct	Incorrect
Verbal complements	1,252	68
	94.8%	5.2%

Table 6: Statistics on the manually corrected distribution of valency complements in the surface structure of CPs from the PCEDT governed by the proposed rules.

5. Conclusion

We have formulated principles for the distribution of valency complements in the surface structure of Czech complex predicates which provide a solid basis for study of this distribution. We have verified these principles on the data from the PDT and the PCEDT with very satisfactory results. From a theoretical point of view, the result of our experiment supports the assumption that the surface structure formation of CPs is regular enough to be described on the rule basis. From a computational point of view, the proposed principles can greatly assist in generation of well-formed structures with complex predicates. Our experiment further shows that the information on valency complements of the light verb and complements of the predicative noun and at the same time the information on coreferential relations between them is beneficial for this task. In the future, this experiment should be extended to other languages to verify if the proposed principles based on the given types of information hold cross-linguistically.

6. Acknowledgements

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. GA15-09979S, and partially by the Ministry of Education, Youth and Sports of the Czech Republic, grant No. LD14117. This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2015071.

7. Bibliographical References

Algeo, J. (1995). Having a look at the expanded predicate. In B. Aarts et al., editors, *The Verb in Contemporary English: Theory and Description*, pages 203–217. Cambridge University Press, Cambridge.

Butt, M. (2010). The light verb jungle: Still hacking away. In Mengistu Amberber, et al., editors, *Complex Predicates in Cross-Linguistic Perspective*, pages 48–78. Cambridge University Press, Cambridge.

Grimshaw, J. and Mester, A. (1988). Light verbs and θ-marking. *Linguistic inquiry*, 19(2):205–232.

Hinrichs, E., Kathol, A., and Nakazawa, T. (1998). *Complex Predicates in Nonderivational Syntax. Syntax and Semantics 30.* Academic Press, San Diego.

Kettnerová, V. and Lopatková, M. (2015). At the lexicon-grammar interface: The case of complex predicates in the functional generative description. In Eva Hajičová et al., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 191–200, Uppsala, Sweden. Uppsala University.

Macháčková, E. (1994). Constructions with verbs and abstract nouns in Czech (analytical predicates). In Světla Čmejrková et al., editors, *The Syntax of Sentence and Text: A Festschrift for František Daneš*, volume 42 of *Linguistic and Literary Studies in Eastern Europe*, pages 365–374. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Pajas, P. and Štěpánek, J. (2009). System for querying syntactically annotated corpora. In Gary Lee et al., editors, *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 33–36, Suntec, Singapore. Association for Computational Linguistics.

Panevová, J. (1994). Valency frames and the meaning of the sentence. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Radimský, J. (2010). *Verbo-nominální predikát s kategoriálním slovesem*. Editio Universitatis Bohemiae Meridionalis, České Budějovice.

Ramos, M. A. (2007). Towards the synthesis of support verb constructions: Distribution of syntactic actants between the verb and the noun. In L. Wanner et al., editors, *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*, pages 97–137. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg.

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

8. Language Resource References

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). *Prague Dependency Treebank 3.0.* Charles University in Prague, MFF, ÚFAL, URL: http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). *Prague Czech-English Dependency Treebank 2.0.* Charles University in Prague, MFF, ÚFAL, URL: http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4.

Urešová, Z., Štěpánek, J., Hajič, J., Panevová, J., and Mikulová, M. (2014). *PDT-Vallex: Czech Valency lexicon linked to treebanks*. Charles University in Prague, MFF, ÚFAL, 2.0, URL: http://hdl.handle.net/11858/00-097C-0000-0023-4338-F.