*WDS'07 Proceedings of Contributed Papers, Part I, 131–136, 2007.* *ISBN 978-80-7378-023-4 © MATFYZPRESS*

# Czech Prefixed Verbs in a Valency Lexicon. Preliminary Study

K. Hrstková

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czech Republic.

**Abstract.** The paper reviews the specific features of prefixed verbs in Czech and approaches to their capture in a valency lexicon. We summarise means and functions of verbal prefixation. We illustrate a regular relation between base verbs and a derived verbs as to their syntactic and semantic properties. We show that the relation does not concern lexemes as wholes but specific lexical units only. We suggest describing such regularities in the form of rules and including them in a valency lexicon, namely in the Valency Lexicon of Czech Verbs VALLEX. To test the approach, we have formulated 21 rules in total for five prefixes. We used the rules to propose corrections of the manual annotations in VALLEX. We summarise results as well as problematic points. To handle those points in our future work, we consider corpus-based methods to be necessary. Therefore, we describe available corpora and methods used in related research.

## 1 Introduction

*Prefixation* – derivational process in which a prefix is attached to the front of a stem or existing base word resulting in a new lexical item – is the most productive type of verb derivation.

In terms of lexical derivation and verbal aspect, prefixed verbs have been traditionally studied by Czech linguists [e.g., *Trávníček*, 1923; *Kopečný*, 1962; *Dokulil*, 1962; *Šmilauer*, 1971; *Šlosar*, 1981]. Detailed monographic study, concerning different aspects of the issue, was provided by *Uher* [1987]. Current state of knowledge has been summarised in general grammars of Czech [*Petr et al.*, 1986; *Karlík et al.*, 1997]. Prefixed verbs have been studied also in other languages. As for the properties relevant for our task, particle verbs behave similarly to the prefixed verbs.[1] Intensive research into the particle verbs has been carried out in German, see Section 4.

This paper deals with derivational relations between lexical items covered in a valency lexicon of verbs. Therefore, it focuses on deverbative type of prefixation. The base word is a verb, referred to as a *base verb*. The product of prefixation is a verb as well, referred to as a *derived verb, prefixed verb* or simply *derivate*. Deadjective and desubstantive prefixation by verbs remains apart.[2]

The inventory of means deriving a new prefixed verb from a base verb is quite close. It can be listed as follows:[3]

- Twenty basic prefixes and their vocalised variants: *do-, na-, nad(e)-, o-, ob(e)-, od(e)-, po-, pod(e)-, pro-, pře-, před(e)-, při-, roz(e)-, s(e)-, u-, v(e)-, vy-, vz(e)-, z(e)-, za-*.
- Multiple prefixation – a) basic prefixes can be attached repeatedly and in combinations, e.g., *vy-z-po-vídat se* 'to confess one's sins'; b) groups consisting of many basic prefixes attached as a whole, e.g., *popo-jít*' 'to go a little further' is not derived from *po-jít* 'to perish' but directly from the base verb *jít* 'to go' [see *Hauser*, 1991].
- Combination with other types of derivation: with reflexivisation, e.g., *roz-plakat se* 'to burst into tears', or with suffixation, e.g., *po-křik-ova-t* 'to vociferate' [see *Uher*, 1987, 32].

The basic function of prefixation is to derive a new lexical item with lexical meaning different from that of the base word. As a result, the new verb may belong to different semantic class. The prefixation is usually

[1] The particle verbs consist of two parts: verb itself and a particle attached to its front. Unlike the prefix, the particle is separated in some syntactic constructions and given a different position in the clause.

[2] It might be considered in the future, as the nouns and adjectives may also be included in a valency lexicon.

[3] Prefix *ne-* and so-called prefixoides (terminology used by *Uher* [1987]) such as *spolu-, polo-* are quite similar to the listed means but they are usually not included in the inventory, since they lack some features characteristic for the prefixes.

accompanied by perfectivisation and it also may have impact on valency frame as to both valency complementations, and their morhological forms [see *Jirsová*, 1979; *Bémová*, 1979; 1981; 1991].

In agreement with *Uher* [1987], we point out that not all the verbs with prefixes – we use his term *prefixal verbs* – are direct products of the prefixation. The prefixal verbs can be typically divided into two groups with respect to the verbal aspect and to the derivational process:

a) *Prefixed verbs* (derived by first grade prefixation) – a prefix is attached to an existing verb, which is mostly imperfective, e.g., *kreslit* 'to draw, impf.' → *vy-kreslit* 'to depict, pf'.
b) So-called *secondary imperfectives* – a new verb is derived from a prefixed verb by other derivation processes than prefixation, typically suffixation. The rest of the word after the prefix (so-called *postprefixal component* [see *Uher*, 1987, 16]) may not even exist as a separate verb, e.g., *vy-kresl-ova-t* 'to depict, impf'.

Nevertheless, the real situation is more complicated and cannot be addressed here in full detail. For illustration, some cases that do not clearly fit into the outlined division are, e.g., (1) a prefixed verb is perfective but its postprefixal component does not exist as a separate verb: *vy-běhnout* 'to run out, pf.'; (2) prefixation is accompanied by phonetical changes in the postprefixal component: *vléct* 'to drag, impf.' → *ob-léct* 'to dress, pf.'; *jíst* 'to eat, impf.' → *s-níst* 'to eat up, pf.'; *kráčet* 'to step, impf.' → *vy-kročit* 'to step forward, pf'.

With regard to character of our task, we focus first only on derivates of an existing verb that does not undergo any phonetic changes. We discuss the possibility how to describe them in a valency lexicon in such a way that the relation between a base verb and its derivate is sufficiently captured. This approach might not only result in decreasing redundancy of the valency lexicon data but also in better understanding of the lexical derivation system and in a method for automatic prediction of syntactic and semantic properties of non-annotated prefixed verbs.

The present paper is structured as follows: in Section 2, we show the specific properties of prefixed verbs and published approaches to their representation in a valency lexicon. Section 3 provides the results of our attempt to describe the relation between a base verb and a prefixed verb in the form of exact rules. In Section 4, we present methods that were applied in related works and that we consider as an inspiration for our future work. Finally, we conclude our findings and open questions.

## 2 Prefixed Verbs in a Valency Lexicon

### 2.1 Regular changes resulting from verbal prefixation

In all the main functions of prefixation, see Section 1, regularities can be observed within groups of verbs.

**Table 1.** Regular changes of verb properties in case of derivation by a prefix *při-* in the group of motion verbs.

| **Property** | **Base verb** | **Derived verb** |
|---|---|---|
| Lexical unit | *jít* 'to walk' | *při-jít* 'to come' |
| Valency frame[4] | ACTor(1;obl)<br>INTTention(k+3, na+4, inf;opt)<br>MANNer(;typ)<br>↑DIRection(;typ) | ACTor(1;obl)<br>↑DIRection3(;obl)<br>INTTention(k+3, na+4, inf;opt) |
| Meaning | 'to move by walking' | 'to reach a place by walking' |
| Aspect | imperfective | perfective |
| Syntactico-semantic class | motion | motion |
| Other verbs in the group | *jet* 'to drive', *letět* 'to fly', *plavat* 'to swim', … | *při-jet* 'to arrive by driving', *při-letět* 'to arrive by flying', *při-plavat* 'to arrive by swimming', … |

Table 1 documents the case of motion verbs with a prefix *při-*. However, both base verbs and prefixed verbs are usually polysemous. For the presented example, the other meanings of the verb *jít* are 'to walk a distance', 'to

[4] Information on valency frame and syntactico-semantic class of the verbs is taken from VALLEX 2.0 (http://ufal.mff.cuni.cz/vallex/2.0/doc/home.html). The labels of the frame elements were extended to the full words from abbreviations (capitalised) used in VALLEX. For detailed explanation of the notation, see *Lopatková et al.* [2006].

be going to', 'to be possible' etc.; and of the verb *přijít* 'to be received', 'to happen' etc. There are different causes for polysemy of derived verbs: Firstly, polysemy of a base verb might cause polysemy of the derived verb, e.g., *balit* 'to pack' → *s-balit* 'to pack up', vs. *balit* 'to flirt' → *s-balit* 'to pick up (a partner)'. Secondly, a derived verb may obtain its polysemous meanings as a result of polysemous meanings of the prefix, e.g., *vy-nést* 'to bring sth out from somewhere', vs. 'to take sth up' [cf. *Lopatková – Panevová*, 2007]. Moreover, there may be also homographic relations between prefixes with the same spelling, e.g., *do-mluvit* 'to finish speaking', vs. 'to admonish'.[5] Finally, derived verbs may be a subject of lexicalisation, metaphorical, or idiomatical shifts, thus their lexical meaning differs from the derivational meaning, i.e., it cannot be simply explained as a combination of the meaning of the base verb with the meaning of the prefix, e.g., *nést* 'to carry' → *do-nést* 'to peach'. Such a verb, whose meaning is *opaque*, is called *non-compositional*, in contrast to *compositional* verbs with *transparent meaning*.

Because of polysemy, it is necessary to understand prefixation only as a relation between a base verb and a derived verb in one of their meanings – lexical units[6] –, and not between lemmas. We propose to speak about a *base lexical unit* and *prefixed lexical unit*.[7]

### 2.2 Valency lexicon

In this research, we work with VALLEX, a valency lexicon of Czech verbs [see *Lopatková et al.*, 2006]. VALLEX provides information on valency structure of verbs and on their syntactic properties as verbal aspect, reciprocity, control, and reflexive forms. The key information on valency structure is captured in valency frames. According to the theoretical background of VALLEX, given by the Functional Generative Description [see *Sgall et al.*, 1986], valency frame consists of arguments (inner participants) and obligatory free modifications (adjuncts). In addition, quasi-valency complementations and some non-obligatory free modifications are captured as well as morphological forms of all valency complementations. An entry in VALLEX corresponds to a verb lexeme, in the sense of an abstract unit that associates all the lexical forms and lexical units (LUs). LU can be roughly understood as 'the given word in the given meaning'. Current version, VALLEX 2.0, [8] covers roughly 2,730 (4,250) [9] lexemes, containing 6,460 LUs.

Imperfective and perfective counterparts are treated as one lexeme in VALLEX. It entails that a prefixed verb and a secondary imperfective are to be described simultaneously.

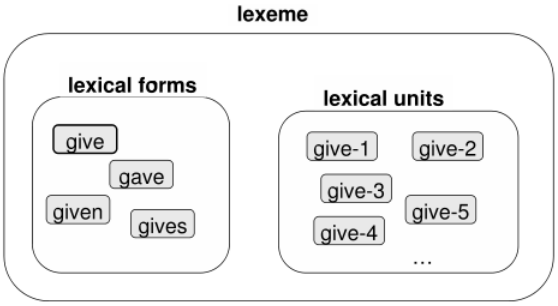


**Figure 1.** Illustration of the notions of *lexeme, lexical form,* and *lexical unit* (illustration from *Žabokrtský* [2005], 35).

*Žabokrtský* [2005] proposes two possible approaches to the annotation of the prefixed verbs:

I) Interlinking – General approach to the prefixed verbs is not "to merge a base verb and all its derived prefixed verbs into one lexeme" but "to interlink the base with its prefixed derivates in the lexicon", "ideally interlink LUs in different lexemes and not the whole lexemes" [*Žabokrtský*, 2005, 42-43].

[5] Apparently, it may be often difficult to distinguish polysemy and homography of the prefixes.
[6] We use the term *lexical unit* as it is defined in VALLEX and we will further explain the notion in Subsection 2.2.
[7] These should not be confused with *basic lexical unit* and *derived lexical unit* as used by *Žabokrtský* [2005].
[8] http://ufal.mff.cuni.cz/vallex/2.0/doc/home.html
[9] If the perfective and imperfective aspectual counterparts are counted separately.

II) Alternation-based model – *Žabokrtský* [2005, 73-80] describes some alternation[10] types in Czech and proposes a minimal lexicon version containing "only basic lexical units with associated lists of applicable alternations", which allow to generate an expanded lexicon by resolving the alternation rules. He shows that a base verb and its derivate may be involved in an inter-lexeme alternation.

| | |
|---|---|
| *Marie malovala obrázky na stěnu.* | *Marie pomalovala stěnu obrázky.* |
| *Mary drew pictures on the wall.* | *Mary drew the wall with pictures.* |
| | |
| *Marie cestuje do Prahy.* | *Marie se do Prahy nacestuje hodně.* |
| *Mary travels to Prague.* | *Mary often travels (will often travel) to Prague.* |

Representation of German particle verbs in different types of lexicon is discussed by *Lüdeling* [2000]. She proposes not to create lexicon entries for the particle verbs at all. Instead, she supposes to treat the verb and particle within separate lexicon entries, the particle as the regular part-of-speech, e.g., preposition. As the only exception, she expects to cover non-compositional[11] particle verbs as idioms in a semantic lexicon.

Based on the observation as that of presented in Table 1, we believe exact rules of some kind can be formulated to allow application of the alternation-based model. However, we estimate there always remain prefixed verbs that need to be covered in separate lexicon entries. At least, the meaning of the non-compositional verbs cannot be derived by any rules; just the interlinking approach can be applied to them. The approach of *Lüdeling* [2000] shows a similar tendency.

## 3 Preliminary Experiments and Recognised Problems: Manual Description of Rules

In this section, we present results of a first attempt to describe relations between base and prefixed verbs in a form of rules that allow to predict valency frames, meaning, verbal aspect or syntactico-semantic class of the derivates.

### 3.1 Rules for changes resulting from verbal prefixation

The observation was based on the data in VALLEX 1.0[12] and in *Slovník spisovné češtiny* [2003][13]. Besides, we used the findings provided by related works [*Karlík et al.*, 1997; *Petr et al.*, 1986; *Bémová*, 1979; 1981]. We dealt with the most common meanings of five prefixes, namely *do-, na-, nad-, o-* and *ob-*. We looked for the corresponding base LU and prefixed LU, then compared their properties and described regularities shared by a group of verbs. As a result, we obtained 21 rules such as the sample presented in Table 2. For delimitation of the group of verbs affected by a particular rule, it seems to be convenient to use the information on syntactico-semantic class as assigned in VALLEX. [14]

In the next step, we used the relations between base LUs and prefixed LUs for checking the data consistency of VALLEX 1.0 and of the working version VALLEX-n. In total, we have proposed 170 corrections. 61 based directly on the rules and 109 based on comparison of base LUs and prefixed LUs, while 44% of the former and 53% of the latter have been reflected in VALLEX 2.0.

### 3.2 Recognised Problematic Points

Apart from the difficulties resulting from "irregularities" in the derivation process, as they were discussed in Section 1 and Subsection 2.1, we have encountered also some other problematic points:

**Small number of prefixed verbs covered in VALLEX 1.0.** Since VALLEX 1.0 was the only source of the information on valency structure of observed verbs, the research was limited by a small number of the prefixed verbs contained. For the five prefixes we focused on in the respective rules, there are only 62 pairs of a base verb and a prefixed verb that both are described in VALLEX 1.0. The situation has now improved by VALLEX 2.0 where the coverage of the prefixed verbs is higher, with 212 such pairs.

**Unclear boundaries between different meanings.** We encountered the question of the most appropriate granularity in word sense disambiguation.

[10] The term 'alternation' is used in accordance with *Levin* [1993].
[11] See Section 2.1.
[12] http://ufal.mff.cuni.cz/~zabokrtsky/vallex/1.0/
[13] Monolingual descriptive lexicon of current standard Czech language.
[14] Information on the syntactico-semantic class is asigned to 44.9% of LUs in VALLEX 2.0, and the coverage was very similar in VALLEX 1.0.

**Table 2.** A rule for derivation of *do-* prefixed verbs from the verbs with meaning of 'particular human activity'.

| **Prefix DO-, rule no. 1** | **Base verb** | **Derived verb** |
|---|---|---|
| Meaning | 'particular human activity' | 'to finish the activity gradually', 'to reach the very end of the activity, or a certain limit'; four sub-meanings were distinguished, depending on the semantics of base verbs |
| Syntactico-semantic class | different syntactico-semantic classes, but 'production' is the most typical one | the same as of the base verb |
| Valency frame | any | without any changes, except for the verbs with meaning 'add to an appropriate extent, and thus reach the completeness', whose derivates bear the information on the limit in their valency frame as a typical modifier: EXTent (do+2, po+4;typ) or MANNer (;typ) |
| Aspect | imperfective; alternatively, perfective prefixed verb in case of multiple prefixation | perfective |
| Examples | *dělat* 'to make, impf.'; *kreslit* 'to draw, impf.'; *stavět* 'to build, impf.'; *vy-světlit* 'to explain, pf.' | *do-dělat* 'to finish, pf.'; *do-kreslit* 'to complete a drawing, pf.'; *do-stavět* 'to finish building, pf.'; *do-vy-světlit* 'to complete explaining, pf.' |

## 4 Outlines of our Future Work

To overcome the problematic points, we consider larger data and corpus-based methods to be necessary.

**Larger data.** Apart from the valency lexicon VALLEX, we assume to get improvement by increasing the exploited data in the future investigation. Therefore, we consider using available corpora for Czech: the Czech National Corpus (CNC)[15] and the Prague Dependency Treebank (PDT 2.0). [16]

**Methods.** There has not been any empirical corpus-based research into the relation between base and prefixed in Czech so far. At this place, we summarise the work of *Aldinger* [2004] and *Schulte im Walde* [2004; 2005], since their methods used for German particle verbs might be worth reimplementing for Czech.

*Aldinger* [2004] focuses on syntactic argument structure changes[17] by German separable complex verbs in a broad sense (*preverb-verb constructions*), including particle verbs. She examines their syntactic argument structures in relation to that of the base verbs manually, using the following data: (a) lists of particles, derived verbs, and corresponding base verbs extracted automatically from a corpus of 200 million words, (b) an automatically generated lexicon of verb subcategorisation frames. The results show that the effects of the German particles on the argument structure are regular even for derivates with opaque meaning.

*Schulte im Walde* [2004] and *Schulte im Walde* [2005] aim to predict compositionality and semantic class of the particle verbs automatically. Based on corpus data, each verb is associated with following quantitative information: (a) frequency and probability distribution over 38 frame types with/without specification of prepositional phrases; (b) frequencies for nominal lexical heads with/without respect to frame-slot combinations. *Schulte im Walde* [2004] observes a significant agreement of the properties (a) and (b) between compositional particle verbs and corresponding base verbs, and significant disagreement between non-compositional particle verbs and corresponding base verb. She concludes that the quantified similarity to the base verbs indicates the degree of transparency, and a list of verbs with the most similar values of these properties indicates the semantic class of the particle verb. *Schulte im Walde* [2005] realises that nominal preferences – property (a) – give significantly better results than subcategorisation frames – property (b). The interpretation is that the compositional particle verbs are semantically similar to their base verbs, but their subcategorisation frames may differ not only from the base verb but also from the other verbs in the same semantic class. On the other hand, for the non-compositional particle verbs, it is assumed that their syntactic behaviour is similar to that of their semantic neighbours, but they do not share the meaning with their base verbs. Both papers propose statistical measures for quantification.

[15] http://ucnk.ff.cuni.cz/

[16] http://ufal.mff.cuni.cz/pdt2.0/

[17] Argument structures correspond to valency frames in our terminology.

## 5 Conclusion

The present paper reports on syntactic and semantic relations between prefixed verbs and their base verbs. Prefixation is the most productive type of verb derivation. We describe changes it entails, i.e., change of the lexical meaning, valency frame, verbal aspect, and possibly also of the syntactico-semantic class. We have shown some regularities of the changes. Furthermore, we summarise related works proposing how to exploit the relation in a valency lexicon. In the next part, we report on our first attempt to capture the relations in a form of rules. Finally, we summarise the results and difficulties that we encountered and outline our future work.

**Aknowledgements.** The present work is supported by the Charles University Grant Agency under the grant no. GA UK 7982/2007. We thank to Markéta Lopatková, Václava Kettnerová-Benešová and two anonymous WDS reviewers for constructive comments, which greatly improved this article.

## References

Aldinger, N., Towards a dynamic lexicon: Predicting the syntactic argument structure of complex verbs. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal, May 2004, pp. 427-430.

Bémová, A., Syntaktické vlastnosti prefigovaných sloves. In *Explicite Beschreibung der Sprache und automatische Textbearbeitung*, (V. Transducing components of functional generative description 2), 1979.

Bémová, A., Slovesná prefixace z hlediska intence. *Slovo a slovesnost*, 42, 1981, pp. 143-148.

Bémová, A., Vztah prefixace a syntaktických vlastností slovesa. Ph.D. thesis, Prague: Charles University, 1991.

Dokulil, M., *Tvoření slov v češtině*. Prague, 1962.

Hauser, P., Zdvojené slovesné předpony *nana-*, *popo-*. *Acta Universitatis Palackianae Olomucensis Facultas Paedagogica, Studia Philologica*, XI, 1991, pp. 28-32.

Jirsová, A., Prefixace sloves a slovesná vazba. *Naše řeč*, 62, 1979, pp. 1-7.

Karlík, P. et al., *Příruční mluvnice češtiny*. Prague, 1997.

Kopečný, F., *Slovesný vid v češtině*, Prague, 1962.

Levin, B. C., *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago, IL: University of Chicago Press, 1993.

Lopatková, M. et al., *Valency Lexicon of Czech Verbs VALLEX 2.0.* ÚFAL Technical Report TR-2006-34, Prague: Charles University, 2006.

Lopatková, M., J. Panevová, Valence vybraných sloves pohybu v češtine (antonyma, nebo synonyma?). In Piper, P (ed.) *Sbornik Matice srpske za slavistiku*, No 71-72/2007, Novi Sad, Serbia and Monte Negro, 2007, pp. 105-115.

Lüdeling, A., Particle Verbs in NLP lexicons. In *Proceedings of the 9th EURALEX International Congress, EURALEX 2000*, IMS, Stuttgart, pp. 625-630.

Petr, J. et al., *Mluvnice češtiny I.* Prague, 1986.

Schulte im Walde, S., Identification, Quantitative Description, and Preliminary Distributional Analysis of German Particle Verbs. In *Proceedings of COLING Workshop on Enhancing and Using Electronic Dictionaries*. Geneva, Switzerland, August 2004.

Schulte im Walde, S., Exploring Features to Identify Semantic Nearest Neighbours: A Case Study on German Particle Verbs. In *Proceedings of the International Conference on Recent Advances in NLP.* Borovets, Bulgaria, September 2005.

Sgall, P. et al., *The meaning of the sentence in its semantic and pragmatic aspects.* Praha: Academia, Dodrecht: D. Riedel, 1986.

*Slovník spisovné češtiny.* Prague, 2003.

Šlosar, D., *Slovotvorný vývoj českého slovesa*. Brno, 1981.

Šmilauer, V., *Novočeské tvoření slov*. Prague, 1971.

Trávníček, F., *Studie o českém vidu slovesném*. Prague, 1923.

Uher, F., *Slovesné předpony*. Brno: Univerzita J. E. Purkyně, 1987.

Žabokrtský, Z. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Prague: Charles University, 2005.