An Analysis of Annotation of Verb-Noun Idiomatic Combinations in a Parallel Dependency Corpus[†]

Zdenka Uresova and **Jana Sindlerova** and **Eva Fucikova** and **Jan Hajic**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics[*]
{uresova,sindlerova,fucikova,hajic}@ufal.mff.cuni.cz

Abstract

While working on valency lexicons for Czech and English, it was necessary to define treatment of multiword entities (MWEs) with the verb as the central lexical unit. Morphological, syntactic and semantic properties of such MWEs had to be formally specified in order to create lexicon entries and use them in treebank annotation. Such a formal specification has also been used for automated quality control of the annotation vs. the lexicon entries. We present a corpus-based study, concentrating on multilayer specification of verbal MWEs, their properties in Czech and English, and a comparison between the two languages using the parallel Czech-English Dependency Treebank (PCEDT). This comparison revealed interesting differences in the use of verbal MWEs in translation (discovering that such MWEs are actually rarely translated as MWEs, at least between Czech and English) as well as some inconsistencies in their annotation. Adding MWE-based checks should thus result in better quality control of future treebank/lexicon annotation. Since Czech and English are typologically different languages, we believe that our findings will also contribute to a better understanding of verbal MWEs and possibly their more unified treatment across languages.

1 Introduction: Valency and MWEs

Valency is a linguistic phenomenon which plays a crucial role in the majority of today's linguistic theories and may be considered a base for both lexicographical and grammatical work. After valency was first introduced into linguistics by L. Tesnière (1959), the study of valency was taken up by many scholars, with a wealth of material now available; cf. (Ágel et al., 2006). In the theoretical framework of Functional Generative Description (Sgall et al., 1986), the following researchers have substantially contributed to valency research: J. Panevová (1977; 1998); P. Sgall (1998), M. Lopatková (2010), V. Kettnerová (2012), Z. Urešová (2011a; 2011b).

In general, valency is understood as a specific ability of certain lexical units - primarily of verbs - to open "slots" to be filled in by other lexical units. By filling up these slots the core of the sentence structure is built. Valency is mostly approached syntactically, semantically or by combining these two perspectives. Valency terminology is not consistent (cf. valency, subcategorization, argument structure, etc.), however, valency as a verbal feature seems to be language universal (Goldberg, 1995).

MWEs are expressions which consist of more than a single word while having non-compositional meaning. They can be defined (Sag et al., 2002) as "idiosyncratic interpretations that cross word boundaries." As the MWE Workshop itself attests, MWEs form a complex issue, both theoretically and practically in various NLP tasks. Here, we will concentrate on certain types of verbal MWEs only.

Verbal MWEs can be divided into several groups

[†] This work has been supported by the Grant No. GPP406/13/03351P of the Grant Agency of the Czech Republic. The data used have been provided by the LINDAT/Clarin infrastructural project LM2010013 supported by the MSMT CR (http://lindat.cz).

[*] Authors' full address: Institute of Formal and Applied Linguistics, Charles University in Prague, Faculty of Mathematics and Physics, Malostranske nam. 25, 11800 Prague 1, Czech Republic

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 58–63,
Atlanta, Georgia, 13-14 June 2013. ©2013 Association for Computational Linguistics

(cf. Sect. 1.3.2 in (Baldwin and Kim, 2010)):

- verb-particle constructions (VPCs), such as *take off*, *play around*, or *cut short*,
- prepositional verbs (PVs), such as *refer to*, *look for*, or *come across*,
- light-verb constructions (LVCs or verb-complement pairs or support verb constructions, see e.g. (Calzolari et al., 2002)), such as *give a kiss*, *have a drink*, or *make an offer*,
- verb-noun idiomatic combinations (VNICs or VP idioms), such as the (in)famous *kick the bucket*, *spill the beans*, or *make a face*.

While (Baldwin and Kim, 2010) define VNICs as being "composed of a verb and noun in direct object position,"[1] we found that their syntax can be more diverse and thus we will include also constructions like *be at odds* or *make a mountain out of a molehill* into this class. Our goal is to look mainly at the surface syntactic representation of MWEs, therefore, we will follow the above described typology even though the exact classification might be more complex.

2 Verbal Valency and MWEs in Dependency Treebanks

In the Prague Dependency Treebank family of projects (PDT(s)) annotated using the *Tectogrammatical Repesentation* of deep syntax and semantics (Böhmová et al., 2005), valency information is stored in valency lexicons. Each verb token in PDTs is marked by an ID (i.e., linked to) of the appropriate valency frame in the valency lexicon. For Czech, both the PDT (Hajič et al., 2012a) and the Czech part of the PCEDT 2.0 (Hajič et al., 2012b)[2] use PDT-Vallex[3]; for English (the English part of PCEDT, i.e. the texts from the Wall Street Journal portion of the Penn Treebank (WSJ/PTB), cf. (Marcus et al., 1993)) we use EngVallex,[4] which follows the same principles, including entry structure, labeling of arguments etc.

Here is an example of a valency lexicon entry (for the base sense of *to give*, simplified):

```
give ACT(sb) PAT(dobj) ADDR(dobj2)
```

The verb lemma (`give`) is associated with its arguments, labeled by *functors*: `ACT` for actor (deep subject), `PAT` for Patient (deep object), and `ADDR` for addressee.[5]

In the valency lexicon entries, two more argument labels can be used: effect (`EFF`) and origin (`ORIG`). In addition, if a free modifier (e.g. adverbial, prepositional phrase, etc.) is so tightly associated to be deemed *obligatory* for the given verb sense, it is also explicitly put into the list of arguments. The P(CE)DT use about 35 free modifications (such as `LOC`, `DIR1`, `TWHEN`, `TTILL`, `CAUS`, `AIM`, ...), most of which can be marked as obligatory with certain verbs (verb senses).

At each valency slot, requirements on surface syntactic structure and inflectional properties of the arguments may be given. This is much more complex in inflective languages but it is used in English too, often as a 'code' assigned to a verb sense, e.g. in OALDCE (Crowther, 1998).

For details of surface-syntactic structural and morphological requirements related to Czech valency and subcategorization in Czech, see e.g. Urešová (2011a; 2011b).

For the annotation of (general) MWEs (Bejček and Straňák, 2010) in the P(CE)DT, the following principle have been chosen: each MWE is represented by a single node in the deep dependency tree. This accords with our principles that "deep" representation should abstract from (the peculiarities and idiosyncrasies of) surface syntax and represent "meaning."[6] The syntactic (and related morphological) representation of MWEs is annotated at a "lower", purely syntactic dependency layer (here, each word token is represented by its own node).

[1] (Baldwin and Kim, 2010), Sect. 1.3.2.4

[2] Also available from LDC, Catalog No. LDC2012T08.

[3] `http://ufal.mff.cuni.cz/lindat/PDT-Vallex`

[4] `http://ufal.mff.cuni.cz/lindat/EngVallex`; since it was created for the WSJ/PTB annotation, the starting point was PropBank (Palmer et al., 2005) to which it is also linked.

[5] We say that a verb has (zero or more) *valency slots*; the verb `give` as presented here has three.

[6] Under this assumption, each node in such a dependency tree should ideally represent a single unit of meaning, and the "meaning" of the tree - typically representing a sentence - should be derived compositionally from the meanings of the individual nodes and their (labeled, dependency) relations (i.e. functors, as they are called in the PDT-style treebanks).

Subsequently, the two representations are linked.

However, here arises a problem with modifiable MWEs (such as *lose his/my/their/... head*): if the whole MWE is represented as a single node, the modifier relation to the MWE would be ambiguous if put simply as the dependent of the MWE (i.e., which part of the MWE does it modify?). Therefore, a rather technical, but unambiguous solution was adopted: the verb as the head of the verbal MWE is represented by a node, and the "rest" of the MWE gets its own appropriately marked node (technically dependent on the verb node). Such a relation is labeled with the DPHR functor ("Dependent part of a PHRase"). The modifier of the MWE can thus be unambiguously attached as either the dependent node of the verb (if it modifies the whole MWE, such as a temporal adverbial in *hit the books on Sunday*), or to the DPHR node (if it modifies only that part of the MWE, such as in *hit the history books*).[7] We believe that this solution which allows the flexibility of considering also modifiable verbal VNICs to be annotated formally in the same way as fully fixed VNICs is original in the PDT family of treebanks, since we have not seen it neither in the Penn Treebank nor in other treebanks, including dependency ones.

Since DPHR is technically a dependent node, it can then be formally included as a slot in the valency dictionary, adding the surface syntactic and/or morphological representation in the form of an encoded surface dependency representation, such as in the following example of an English VNIC:

```
make DPHR(mountain.Obj.sg[a],
          out[of,molehill.Adv.sg[a])
```

In Czech, the formal means are extended, e.g. for the required case (1 - nominative, 6- locative):[8]

```
běhat DPHR(mráz.S1,po[záda.P6])
```

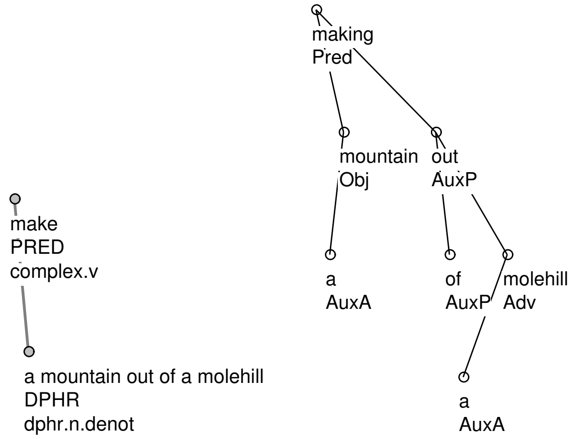

Figure 1: Verbal MWE: tectogrammatical (left) and syntactic (right) annotation of a VNIC

In Fig. 1, the phrase *making a mountain out of a mole* is syntactically annotated in the following way:

- *mountain* is annotated as the syntactic direct object of *making*,
- *out of a molehill* is annotated as a prepositional phrase (with the preposition as the head)

On the tectogrammatical layer of annotation, the verb is the head and the defining part of the MWE gets a separate node (marked by DPHR).

In the corpus-based analysis of verbal MWEs in the valency lexicons and the treebanks presented here, we concentrate mainly on VNICs (see Sect. 1) and briefly mention LVCs, since the boundary between them is often a bit grayish. In the P(CE)DT treebanks, LVCs are always represented as two nodes: the (light) verb node and the noun complement node. Formally, the representing structure is the same for both mentioned groups of MWEs, but it differs in the labels of the verb arguments: CPHR (Compound PHRase) for LVCs vs. DPHR for VNICs. Whereas lexical units marked as DPHRs are mostly limited to a fixed number of words and therefore are listed in the lexicon, lexical units marked as CPHRs are often not limited in their number and therefore it does not make sense to list them all in the lexicon.

[7] One can argue that in very complex MWEs, this simple split into two nodes might not be enough; in the treebanks we have explored no such multiple dependent modifiers exist.

[8] The repertoire of possible syntactic and morphological constraints, which can be used for the description of possible forms of the fixed part of the idiomatic expression, covers all aspects of Czech word formation: case, number, grammatical gender, possessive gender and number, degree of comparison, negation, short/long form of certain adjectives, analytical dependency function etc.

A possible solution to the problem of automatic *identification* of (general) MWEs in texts using the annotation found in the PDT, which is related to the topic described in this paper but goes beyond its scope, can be found in (Bejcek et al., 2013).

3 Corpus Analysis

To compare annotation and use of VNICs in Czech and English, we have used the PCEDT. The PCEDT contains alignment information, thus it was easy to extract all cases where a VNIC was annotated (i.e. where the `DPHR` functor occurs).[9]

We found a total of 92890 occurrences of aligned (non-auxiliary) verbs. Czech VNICs were aligned with English counterparts not annotated as a VNIC in 570 cases, and there were 278 occurrences of English VNICs aligned with Czech non-VNICs, and only 88 occurrences of VNICs annotated on both sides were aligned.[10] These figures are surprisingly small (less than 1.5% of verbs are marked as VNICs), however, (a) it is only the VNIC type (e.g., phrasal verbs would account for far more), and (b) the annotator guidelines asked for "conservativeness" in creating new VNIC-type verb senses.[11]

Ideally (for NLP), VNICs would be translated as VNICs. However, as stated above, this occurred only in a 88 cases only (a few examples are shown below).

(1) (wsj0062) točit[*turn*] se[*oneself-acc.*]
zády[*back-Noun-sg-instr.*]:
thumb(ing) its nose

(2) (wsj0989) podřezávat[*saw down*]
si[*oneself-dat.*] pod[*under*]
sebou[*oneself-instr.*]
větev[*branch-Noun-sg-acc.*]:
bit(ing) the hand that feeds them

Manual inspection of these alignments revealed (except for a few gray-area cases) no errors. We have thus concentrated on the asymmetric cases by manually exploring 200 such cases on each side. The results are summarized in Tab. 1.

Direction / Annotated as (by type)	VNIC in En, not Cz	VNIC in Cz, not En	Examples
Correctly annotated (as non-VNIC)			
LVC	26	4	lámat[*break*] rekordy: set records
non-MWE	138	124	přerušit [*interrupt*]: cut short
Annotation Error (should have been VNIC)			
LVC	7	17	držet[*hold*] krok[*step*]: keep abreast
non-MWE	28	52	zlomit (mu) srdce: break sb's heart
other error	1	3	

Table 1: Breakdown of VNICs linked to non-VNICs

3.1 English VNICs Linked to Non-VNIC Czech

The first column of counts in Tab. 1 refers to cases where the verb in the English original has been annotated as VNIC, but the Czech translation has been marked as a non-VNIC. We have counted cases, where we believe that the annotation is correct, even if it is not annotated as a VNIC (164 in total) and cases which should have been in fact annotated as a VNIC (35 cases). Within these two groups, we separately counted cases where the translation has not been annotated as a VNIC, but at least as a LVC, another MWE type (total of 33 such cases). The proportion of errors (approx. 18%) is higher than the 5.5% rate reported for semantic relation annotation (Štěpánek, 2006). Typically, the error would be corrected by adding a separate idiomatic verb sense into the valency lexicon and adjusting the annotation (verb sense and the `DPHR` label) accordingly.

[9] The alignment is automatic, the Czech and English tectogrammatical annotation (including verb sense/frame assignment) is manual.

[10] The total number of Czech VNICs in the PCEDT (1300) is higher than the sum of extracted alignments (570+88=658). The difference is due to many of the Czech VNICs being aligned to a node which does not correspond to a verb, or which is not linked to an English node, or where the alignment is wrong.

[11] By "conservative" approach we mean that splitting of verb senses into new ones has been discouraged in the annotation guidelines.

3.2 Czech VNICs Linked to Non-VNIC English

The second column of counts in Tab. 1 shows the same breakdown as described in the previous section, but in the opposite direction: Czech VNICs which in the English original have been annotated differently. The first difference is in the number of erroneously annotated tokens, which is visibly higher (approx. twice as high) than in the opposite direction both for LVCs (17) and for constructions which have not been marked as MWEs at all (52). This suggests that the authors of the English valency lexicon and the annotators of the English deep structure have been even more "conservative" than their Czech colleagues by not creating many VNIC-typed verb senses.[12] Second, there are only 4 cases of VNICs translated into and correctly annotated as LVCs, compared to the English → Czech direction (26 cases).

4 Conclusions

We have described the treatment of (an enriched set of) verb-noun idiomatic combinations (and briefly other types of MWEs) in the PDT style treebanks and in the associated valency lexicons. We have explored the PCEDT to find interesting correspondences between the annotation and lexicon entries in the English and Czech annotation schemes.

We have found that VNICs, as one of the types of MWEs, are translated in different ways. A translation of a VNIC as a VNIC is rare, even if we take into account the annotation errors (88+7+17+28+52=192 cases of the 936 extracted). By far the most common case of translating a VNIC in both directions is the usage of a completely non-MWE phrase. There is also a substantial amount of errors in each direction, higher in cases where the Czech translation was annotated as a VNIC and the English original was not. While the low overall number of VNICs found in the parallel corpus can be explained by not considering standard phrasal verbs for this study and by the required conservatism in marking a phrase as a true VNIC, we can only speculate why only a small proportion of VNICs are translated as VNICs in(to) the other language: manual inspection of several cases suggested (but without a statistically significant conclusions) that this does not seem to be caused by the specific nature or genre of the Wall Street Journal texts, but rather by the fact that the two languages explored, Czech and English, went generally through different developments under different circumstances and contexts throughout the years they evolved separately.

While this paper describes only an initial analysis of multiword expressions (of the verb-noun idiomatic combination type) in parallel treebanks, we plan to apply the same classification and checks as described here to the whole corpus (perhaps automatically to a certain extent), to discover (presumably) even more discrepancies and also more correspondence types. These will again be classified and corrections in the data will be made. Eventually, we will be able to get a more reliable material for a thorough study of the use of MWEs in translation, with the aim of improving identification and analysis of MWEs (e.g., by enriching the approach taken by and described in (Bejcek et al., 2013)). We would also like to improve machine translation results by identifying relevant features of MWEs (including but not limited to VNICs) and using the associated information stored in the valency lexicons in order to learn translation correspondences involving MWEs.

Acknowledgments

The authors would like to thank the four reviewers, especially reviewer #4, for helpful comments which hopefully have lead to a clearer version of this paper. Also, we would like to thank to all the annotators and technical support staff who made our research possible by creating the treebanks and lexicons which we can now build upon.

References

Vilmos Ágel, Ludwig M. Eichinger, Hans-Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, Henning Lobin, and Guta Rau. 2006. *Dependenz und Valenz*. Walter de Gruyter, Berlin & New York.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

[12] None of the annotators of the English side of the parallel treebank was a fully native English speaker, which might also explain this "conservatism."

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Eduard Bejcek, Pavel Pecina, and Pavel Stranak. 2013. Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In *Workshop on Multiword Expressions (NAACL 2013, this volume)*, New Jersey. Association for Computational Linguistics.

Alena Böhmová, Silvie Cinková, and Eva Hajičová. 2005. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *LREC*.

Jonathan Crowther. 1998. *Oxford Advanced Learner's Dictionary*. Cornelsen & Oxford, 5th edition.

A.E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Jan Hajič, Eduard Bejček, Jarmila Panevová, Jiří Mírovský, Johanka Spoustová, Jan Štěpánek, Pavel Straňák, Pavel Šidák, Pavlína Vimmrová, Eva Šťastná, Magda Ševčíková, Lenka Smejkalová, Petr Homola, Jan Popelka, Markéta Lopatková, Lucie Hrabalová, Natalia Klyueva, and Zdeněk Žabokrtský. 2012a. Prague Dependency Treebank 2.5. https://ufal-point.mff.cuni.cz/xmlui/handle/11858/00-097C-0000-0006-DB11-8.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012b. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Václava Kettnerová. 2012. *Lexikálně-sémantické konverze ve valenčním slovníku*. Ph.D. thesis, Charles University, Prague, Czech Republic.

Markéta Lopatková. 2010. *Valency Lexicon of Czech Verbs: Towards Formal Description of Valency and Its Modeling in an Electronic Language Resource*. Prague.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Jarmila Panevová. 1998. Ještě k teorii valence. *Slovo a slovesnost*, 59(1):1–14.

Jarmila Panevová. 1977. Verbal Frames Revisited. *The Prague Bulletin of Mathematical Linguistics*, (28):55–72.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, pages 1–15.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia.

Petr Sgall. 1998. Teorie valence a její formální zpracování. *Slovo a slovesnost*, 59(1):15–29.

Jan Štěpánek. 2006. Post-annotation Checking of Prague Dependency Treebank 2.0 Data. In *Lecture Notes in Artificial Intelligence, Text, Speech and Dialogue. 9th International Conference, TSD 2006, Brno, Czech Republic, September 11–15, 2006*, volume 4188 of *Lecture Notes in Computer Science*, pages 277–284, Berlin / Heidelberg. Springer.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Editions Klincksieck, Paris.

Zdeňka Urešová. 2011a. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Prague.

Zdeňka Urešová. 2011b. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Prague.