ParaDi: Dictionary of Paraphrases of Czech Complex Predicates with Light Verbs

Petra Barančíková and **Václava Kettnerová**
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University,
Malostranské náměstí 25, 118 00, Praha, Czech Republic,
barancikova@ufal.mff.cuni.cz, kettnerova@ufal.mff.cuni.cz

Abstract

We present a new freely available dictionary of paraphrases of Czech complex predicates with light verbs, *ParaDi*. Candidates for single predicative paraphrases of selected complex predicates have been extracted automatically from large monolingual data using *word2vec*. They have been manually verified and further refined. We demonstrate one of many possible applications of *ParaDi* in an experiment with improving machine translation quality.

1 Introduction

Multiword expressions (MWEs) pose a serious challenge for both foreign speakers and many NLP tasks (Sag et al., 2002). From various 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 particular type of Czech multiword expressions – on complex predicates with light verbs (CPs). CPs consist of a light verb and another predicative element – a predicative noun, an adjective, an adverb or a verb; the pairs function as single predicative units. As such, most CPs have their single predicative counterparts by which they can be paraphrased, e.g. the CPs *dát polibek* and *dát pusu* 'give a kiss' can be both paraphrased by *políbit* 'to kiss'.

In contrast to their single predicative paraphrases, CPs manifest much greater flexibility in their modification, c.f. adjectival modifiers of the CP *dát polibek* 'give a kiss' and the corresponding adverbial modifiers of its single verb paraphrase *políbit* 'to kiss' in *dát vášnivý/něžný/letmý/manželský/májový/smrtící polibek* 'give a passionate/tender/fleeting/marriage/May/fatal kiss' vs. *vášnivě/něžně/letmo/*manželsky/*májově/*smrtelně políbit* 'kiss passionately/tenderly/fleetingly/*marriagely/*Mayly/?fatally'. Easier modification of CPs is usually considered as the main motivation for their widespread use (Brinton and Akimoto, 1999).

In this paper, we present *ParaDi*, a dictionary of single predicative verb paraphrases of Czech CPs. We restricted the dictionary only to CPs that consist of light verbs and predicative nouns, which represent the most frequent and central type of CPs in the Czech language.

ParaDi was built on a semi-automatic basis. First, candidates for single verb paraphrases of selected CPs have been automatically identified in large monolingual data using *word2vec*, a shallow neural network. The list of these candidates has been then manually checked and further refined. In many cases, if CPs are to be correctly paraphrased by the identified single predicative verbs, these verbs require certain semantic and/or syntactic modifications.

It has been widely acknowledged that many NLP applications – let us mention, e.g. information retrieval (Wallis, 1993), question answering, machine translation (Madnani and Dorr (2013); Callison-Burch et al. (2006); Marton et al. (2009)) or machine translation evaluation (Kauchak and Barzilay (2006); Zhou et al. (2006); Barančíková et al. (2014)) – can benefit from paraphrases.

Here we show how the dictionary providing high quality data can be integrated into an experiment with improving statistical machine translation quality. If translated separately, CPs often cause errors in machine translation. In our experiment, we use the dictionary to simplify Czech source sentences before translation by replacing CPs with their respective single predicative verb paraphrases. Human annotators have evaluated quality of the translated simplified sen-

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 1–10,
Valencia, Spain, April 4.

tences higher than of the original sentences contain CPs.

This paper is structured as follows. First, related work on CPs generally and on their paraphrases is introduced (Section 2). Second, the paraphrasing model for CPs is thoroughly described, especially the selection of CPs, an automatic extraction of candidates for their paraphrases and their manual evaluation (Section 3). Third, the resulting data and the structure of the lexical space of the dictionary are discussed (Section 4). Finally, in order to present one of many practical applications of this dictionary, a random sample of paraphrases from the *ParaDi* dictionary is used in a machine translation experiment (Section 5).

2 Related Work

A theoretical research on CPs with light verbs has a long history, which can be traced back to Jespersen (1965). An ample literature devoted to this language phenomenon so far is characterized by an enormous diversity in used terms and analyses, see esp. (Amberber et al., 2010) and (Alsina et al., 1997). Here we use the term *CP with the light verb* for a collocation within which the verb – not retaining its full semantic content – provides rather grammatical functions (incl. syntactic structure) and to which individual semantic properties are primarily contributed by the noun (Algeo, 1995).

The information on CPs is a part of several lexical resources containing manually annotated data. For instance, CPs are represented in syntactically rich annotated corpora from the family of the Prague Dependency Treebanks: the Prague Dependency Treebank 3.0 (PDT)[1] and the Prague Czech-English Dependency Treebank 2.0[2], see (Bejček et al., 2013) and (Hajič et al., 2012). Further, the PropBank[3] project has been recently enhanced with the information on CPs; the annotation scheme of CPs in PropBank is thoroughly described in (Hwang et al., 2010). Finally, the Hungarian corpus of CPs based on the data from the Szeged Treebank has been built (Vincze and Csirik, 2010).

At present, one of trending topics in NLP community is an automatic identification of CPs. In this task, various statistical measures often combined with information on syntactic and/or semantic properties of CPs are employed (e.g. Bannard (2007), Fazly et al. (2005)). The automatic detection benefits especially from parallel corpora representing valuable sources of data in which CPs can be automatically recognized via word alignment, see e.g. (Chen et al., 2015), (de Medeiros Caseli et al., 2010), (Sinha, 2009), (Zarrießand Kuhn, 2009).

Work on paraphrasing CPs is still not extensive. A paraphrasing model has been proposed within the Meaning↔Text Theory(Žolkovskij and Mel'čuk, 1965). Its representation of CPs by means of lexical functions and rules applied in the paraphrasing model are thoroughly described in (Alonso Ramos, 2007). Further, Fujita et al. (2004) present a paraphrasing model which takes advantage of semantic representation of CPs by lexical conceptual structures. Similarly as our proposed dictionary of paraphrases, this model also takes into account changes in the grammatical category of voice and changes in morphological cases of arguments, which have appeared to be highly relevant for the paraphrasing task.

3 Paraphrase Model

In this section, the process of paraphrase extraction is described in detail. First, we present the selection of CPs (Section 3.1). For their paraphrasing, we had initially intended to use some of existing sources of paraphrases, however, they turned out to be completely unsatisfactory for our task.[4]

Word2vec is a group of shallow neural networks generating representations of words in a continuous vector space depending on contexts they appear in (Mikolov et al., 2013). In line with distributional hypothesis (Harris, 1954), semantically

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

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

[3] https://verbs.colorado.edu/~mpalmer/projects/ace.html

[4] We used the *ParaPhrase DataBase* (PPDB), (Ganitkevitch and Callison-Burch, 2014; Ganitkevitch et al., 2013) the largest paraphrase database available for the Czech language. PPDB has been created automatically from large parallel data and it comes in several sizes ranging from S to XXL. However, the bigger its size, the bigger the amount of noise. We chose the size L as a reasonable trade-off between quality and quantity. We combined the phrasal paraphrases, many-to-one and one-to-many. We lemmatized and tagged the collection of PPDB using the state-of-the-art POS tagger *Morphodita* (Straková et al., 2014). Even though this collection contains almost 400k lemmatized paraphrases in total, it contained only 54 candidates for single predicative verb paraphrases of CP. Only 2 of these 45 candidates these candidates have been detected correctly, the rest was noise in PPDB. As a result, we chose not to use parallel data in our task but we have adopted another approach applying *word2vec*, a neural network based model to large monolingual data.

similar words are mapped close to each other (measured by the cosine similarity) so we can expect CPs and their single verb paraphrases to have similar vector space distribution.

Word2vec computes vectors for single tokens. As CPs represent MWEs, their preprocessing was necessary: CPs have to be first identified and connected into a single token (Section 3.2).

Particular settings of our model for an automatic extraction of candidates for single predicative verb paraphrases are presented in Section 3.3. Finally, a manual evaluation of the extracted candidates, including their further annotation with semantic and syntactic information, is described (Section 3.4).

3.1 CPs Selection

Two different datasets of CPs, containing together 2,257 unique CPs, have been used. As both these datasets have been manually created, they allow us to achieve the desired quality of the resulting data.

The first dataset resulted from the experiment examining the native speakers' agreement on the interpretation of light verbs (Kettnerová et al., 2013). CPs in this dataset consist of collocations of light verbs and predicative nouns expressed by a prepositionless case (e.g., *položit otázku* 'put a question'), by a simple prepositional case (e.g., *dát do pořádku* 'put in order'), and by a complex prepositional group (e.g., *přejít ze smíchu do pláče* 'go from laughing to crying').

The second dataset resulted from a project aiming to enhance the high coverage valency lexicon of Czech verbs, VALLEX,[5] with the information on CPs (Kettnerová et al., 2016). In this case, only the nominal collocates expressed in the prepositionless accusative were selected as they represent the central type of Czech CPs. As the frequency and saliency have been taken as the main criteria for their selection, the resulting set represents a valuable source of CPs for Czech.

The overall number of CPs in the datasets is presented in Table 1. The union of CPs from these datasets – 2,257 CPs in total – has been used in the paraphrase candidates extraction task.

3.2 Data Preprocessing

For *word2vec* training, only monolingual data – generally easily obtainable in a large amount – is necessary. We have used large lemmatized corpora

[5] http://ufal.mff.cuni.cz/vallex/3.0/

	CPs	Verbs	Nouns
First dataset	726	49	612
Second dataset	1640	126	699
Union	2257	154	1061

Table 1: The number of unique CPs, light verbs and predicative nouns from two datasets. Their union has been used in the paraphrase extraction task.

Corpus	Sentences	Tokens
CNK2000	2.78	121.81
CNK2005	7.95	122.99
CNK2010	8.18	122.48
Czeng 1.0	14.83	206.05
Czech Press	258.40	4018.89
Total	292.14	4592.22

Table 2: Basic statistics of datasets (numbers in millions of units).

of Czech texts: SYN2000 (Čermák et al., 2000), SYN2005 (Čermák et al., 2005), SYN2010 (Křen et al., 2010) and CzEng 1.0 (Bojar et al., 2011). As these four large corpora with almost 600 million tokens in total have turned out to be insufficient, they have been extended with the data from the Czech Press – a large collection of contemporary news texts containing more than 4,000 million tokens. The overall statistics on all datasets is presented in Table 2.

To generate CPs paraphrases, all the selected CPs (Section 3.1) had to be automatically identified in the given corpora. For the identification of the CPs, we proceeded from light verbs. First, all verbs in the corpora were detected. From these verbs, only those verbs that represent light verbs as parts of the selected CPs were further processed.

For each identified light verb, each noun phrase in the context $\pm$ 4 words from the given light verb was extracted in case the verb and the given noun phrase can combine in some of the selected CPs.

Further, as *word2vec* generates representations of single word units, every detected noun phrase was connected with its respective light verb into a single word unit. In case that some light verb could combine with more than one noun phrase into CPs, or in case that one noun phrase could be connected with more than one light verb, we have followed the principle that every verb should be connected to at least one candidate in order to maximize a number of identified CPs.

rank	CP	frequency
1.	*mít problém* 'have a problem'	319,791
2.	*mít možnost* 'have a possibility'	300,330
3.	*mít šanci* 'have a chance'	292,340
...	...	...
998.	*vznést žalobu* 'bring charges'	535
...	...	...
1775.	*vést k sebevyvrácení* 'lead to self-refutation'	1
1776.	*dojít k flagelantství* 'flagellation takes place'	1

Table 3: The ranking of the CPs identified in the corpora, based on their frequency.

For example, if there were two light verbs v_1 and v_2 in a sentence and v_1 had a candidate c_1, while v_2 had two candidates c_1 and c_2, v_1 was connected with c_1 and v_2 with c_2. In case this principle was not sufficient, the light verb was assigned the closest noun phrase on the basis of word order.

When each noun phrase was connected maximally with one light verb and each light verb was connected maximally with one noun phrase, we have joined the noun phrases to their respective light verbs into single word units with the underscore character and erase the noun phrases from their original position in sentences.

For example, after identifying the light verb *mít* 'have' in a sentence and the prepositionless noun phrase *problém* 'problem' in its context on the above principles, the given light verb and the given noun phrase have been connected into the resulting single word unit *mít_problém*; this whole unit then replaced the verb *mít* 'have' in the sentence, while the noun phrase *problém* 'problem' was deleted from the sentence.

On this basis, almost 8.5 million instances of CPs were identified in the corpora, 99,9% of them has frequency more than 100 occurrences in the corpora. However, only 1,776 unique CPs were detected – almost 500 CPs from the selected datasets (Section 3.1) did not occur even once. The rank and frequency of selected CPs identified in the corpora is presented in Table 3.

3.3 Word2vec Model

To the resulting data, we have applied *gensim*, a freely available *word2vec* implementation (Řehůřek and Sojka, 2010). In particular, we have used a model of vector size 500 with continuous bag of word (CBOW) training algorithm and negative sampling.

As it is impossible for the model to learn anything about a rarely seen word, we have set a minimum number of word occurrences to 100 in order to limit the size of the vocabulary to reasonable words. This requirement filtered also uncommonly used CPs from the identified CPs in the corpora: from 1,776 CPs only 1,486 CPs fulfilled the given limit.

After training the model, for each of 1,486 CPs we have extracted 30 words with the most similar vectors. From these 30 words, we have selected up to ten single verbs closest to the given CP. These verbs were taken as candidates for single predicative verb paraphrases of the given CP.

As a result, 8,921 verbs in total corresponding to 3,735 unique verb lemmas have been selected as candidates for single predicative verb paraphrases of the given 1,486 CPs.

3.4 Annotation Process

In this section, the annotation process of the extracted 8,921 candidates for single predicative verb paraphrases of CPs is thoroughly described. Manual processing of the extracted single verbs allowed us to evaluate the results of the adopted method.

Let us repeat that *word2vec* generates semantically similar words depending on their contexts they appear in. However, not only words having the same meaning can have similar space representation. Words with the opposite meaning (e.g. 'finish' vs 'start'), more specific meaning ('finish' vs. 'graduate') or even different meaning can be extracted as they can appear in similar contexts as well. Manual evaluation of the extracted candidates for single verb paraphrases is thus necessary.

In the manual evaluation, two annotators have been asked to indicate for each instance of the extracted candidates for single verb paraphrases of a CP whether it represents the paraphrase of the given CP, or not. For example, the single verbs *upřednostňovat* and *preferovat* 'to prefer' are the paraphrase of the CP *dávat přednost* 'to give a preference' while the verb *srazit* 'to run down'

not.

Moreover, single verbs antonymous with the respective CPs have been indicated as well as in particular context they can also function as a paraphrase. For example, depending on contexts both extracted single verbs *stoupnout* 'to rise' and *poklesnout* 'to drop' can function as paraphrases of the CP *zaznamenat propad* 'to experience a drop', while the first one has the meaning synonymous with the given CP, the meaning of the latter is antonymous.

Further, when the annotators have determined a certain candidate as the single verb paraphrase of a CP, they have taken the following three morphological, syntactic and semantic aspects into account.

First, they had to pay special attention to the morphological expression of arguments. Changes in their morphological expression reflect different syntactic perspectives from which the action denoted by the given CP and its single verb paraphrase is viewed. For example, the single verb *potrestat* 'to punish' can serve as the paraphrase of the CP *dostat trest* 'to get a punishment' in a sentence, however, the semantic roles of the subject and the object are switched.

Second, in some cases the reflexive morpheme *se/si*, reflecting the inchoative meaning, had to be added to single predicative verb paraphrases so that their meaning corresponds to the meaning of their respective CPs. For example, the CP *mít problém* 'have a problem' can be paraphrased by the verb *trápit* only on the condition that the reflexive morpheme is attached to the verb lemma *trápit se* 'to worry'.

Third, some single predicative verbs function as paraphrases of particular CPs only if nouns in these CPs have certain adjectival modifications. These paraphrases have been assigned the given adjectives during the annotation.

As the above given three features are not mutually exclusive, they can combine. For example, the verb *zaměstnat* 'to hire' is a paraphrase of the CP *nalézt uplatnění* 'to find an use' but both the reflexive morpheme *se* and a modification by the adverb *pracovní* 'working' is required.

To summarize, for each identified single predicative verb paraphrase *v* of a CP *l*, the annotators have chosen from the following options:

- *v* is a synonymous paraphrase of *l* (without any modification of the context)
  e.g., *mít zájem* 'to be interested' and *chtít* 'to want'

- *v* is an antonym of *l* (the modification of the context is necessary)
  e.g., *zaznamenat propad* 'to experience a drop' and *stoupnout* 'to rise'

- *v* is a paraphrase of *l* but changes in the morphological expression of arguments are necessary
  e.g., *dostat nabídku* 'to get an offer' and *nabídnout* 'to offer'

- *v* is a paraphrase of *l* but the reflexive morpheme *se/si* has to be added (the modification of verb lemma is necessary)
  e.g., *nést název* 'to be called' and *nazývat se* 'to be called'

- *v* is a paraphrase of *l* with a particular adjectival modification (the adjective modifier of the noun should be present)
  e.g., *podat oznámení* 'to make an announcement' can be paraphrased as *žalovat* 'to sue' only if the noun *oznámení* is modified with the adjective *trestní* 'criminal'

- *v* is a not a paraphrase of *l*

	synonyms	antonyms
no constrains	1607	51
+ reflexive morpheme	353	2
+ voice change	173	5
+ an adjective	53	–
total	2177	58

Table 4: The basic statistics on the annotation. The *synonyms* column does not add up as the conditions are not mutually exclusive as mentioned earlier.

As a result of the annotation process, the total number of the indicated single verb paraphrases of CPs was 2,177. For 999 CPs at least one single verb paraphrase has been found. The highest number of single verb paraphrases indicated for one CP has been eight; it has been the CP *vznést dotaz* 'to ask a question'. Figure 1 shows the number of paraphrases per CPs.

Table 4 presents more detailed results of the annotation. It shows frequency of additional morphological, syntactic and semantic features.

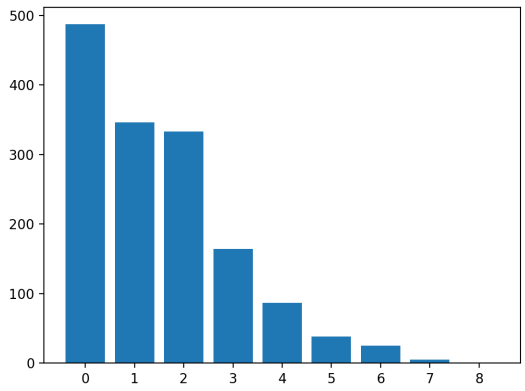

Figure 1: The number of single predicative verb paraphrases and antonymous verbs per CPs in the *ParaDi* dictionary.

4 Dictionary of Paraphrases

2,235 single predicative verbs indicated by the annotators as synonymous or antonymous verbs of 999 CPs (Section 3.4) form the lexical stock of *ParaDi*, a dictionary of single verb paraphrases of Czech CPs. The format of the *ParaDi* dictionary has been designed with respect to both human and machine readability. The dictionary is represented in JSON, as it is flexible and language-independent data format.

The lexical entries in the dictionary describe individual light verbs. Under light verb keys, all predicative nouns constituting CPs with the given light verb are listed. The predicative nouns are lemmatized; the information on their morphology is included under their *morph* keys the value of which are prepositionless and prepositional cases.

Each CP in the lexical entry might be assigned one or two lists of single predicative verbs: one for synonymous paraphrases and the other for antonymous verbs. Paraphrases in the lists are sorted based on the distance from their respective LVC in the vector space. Moreover, each verb may be assigned one or more following features:

- *voice_change* – indicating changes in the morphosyntactic expression of arguments,
- *adjective* – indicating necessary adjectival modification,
- *reflexive* – indicating that reflexive morpheme is necessary,

```
'lverb': 'zaznamenat',
[{'noun': 'propad',
'morph': '4',
'synonyms': [
{'lemma': 'poklesnout'},
{'lemma': 'klesnout'},
{'lemma': 'propadnout',
 'reflexive': 'se'}
],
'antonyms': [
{'lemma': 'stoupnout'}
],
...
]
}
```

Figure 2: The lexical representation of the CP *zaznamenat propad* 'to record a slump'.

An illustrative example of the lexical representation of paraphrases in *ParaDi* is presented in Figure 2. It displays the lexical entry of the CP *zaznamenat propad* 'to record a slump'. Under the light verb *zaznamenat* 'to record', there is a list of nouns that combine with this light verb into CPs. In case of the noun *propad* 'slump', the noun is expressed by the prepositionless accusative. This CP has three single verb paraphrases (*poklesnout* 'to decrease', *klesnout* 'to drop', *propadnout se* 'to slump') and one antonymous verb (*stoupnout* 'to increase'). The paraphrase *propadnout* 'to slump' needs to have the reflexive morpheme *se*.

ParaDi is freely available at the following URL: `http://hdl.handle.net/11234/1-1969`

5 Machine Translation Experiment

We have taken advantage of the *ParaDi* dictionary in a machine translation experiment in order to verify its benefit for one of key NLP tasks. We have selected 50 random CPs from the dictionary. For each of them, we have randomly extracted one sentence from our data containing the given CP. This set of sentences is referred to as BEFORE. By substituting a CP for its first (i.e. closest in the vector space) paraphrase on the basis of the dictionary, we have created a new dataset AFTER.

We have translated both these datasets – BEFORE and AFTER – using two freely avail-

Source	Moses	GT
BEFORE	30%	33%
AFTER	45%	44%
TIE	25%	23%

Table 5: Results of the experiment. First column shows a source of better ranked sentence from the pairwise comparison or whether they tied.

able MT systems – *Google Translate*[6] (GT) and *Moses*[7] in the Czech to English setting.

We have used crowdsourcing for evaluation of the resulting translations. Both options were presented in a randomized order and the annotators were instructed to choose whether one translation is better or they have the same quality.

We have collected almost 300 comparisons. We measured inter-annotator agreement using Krippendorff's alpha (Krippendorff, 2007), a reliability coefficient developed to measure the agreement between judges. The inter-annotator agreement has achieved 0.58, i.e. moderate agreement.

The results (see Table 5) are very promising: in most cases the annotators clearly preferred translations of AFTER (i.e. with single predicative verbs) to BEFORE (i.e. with CPs). The results are consistent for both translation systems.

However, it is clear from the example in Table 6 that even though the change in the source sentence was minimal, the translations changed substantially as both the translation models are phrase-based. Based on this fact, we can expect that not only difference in quality between translations of CPs and their respective synonymous verbs was evaluated. This low quality translation inevitably reflected in lower inter-annotator agreement, typical for machine translation evaluation (Bojar et al., 2013).

6 Conclusion

We have presented *ParaDi*, a semiautomatically created dictionary of single verb paraphrases of Czech complex predicates with light verbs. We have shown that such paraphrases are automatically obtainable from large monolingual data with a manual verification. *ParaDi* represents a core of such dictionary, which can be further enriched. We have demonstrated one of its possible applications, namely an experiment with improving machine translation quality. However, the dictionary can be used in many other NLP tasks (text simplification, information retrieval, etc.) and can be similarly created for other languages.

[6] http://translate.google.com

[7] http://quest.ms.mff.cuni.cz/moses/demo.php

Acknowledgments

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. GA15-09979S. 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.

References

John Algeo. 1995. Having a look at the expanded predicate. In B. Aarts and Ch. F. Meyer, editors, *The Verb in Contemporary English: Theory and Description*, pages 203–217. Cambridge University Press, Cambridge.

Margarita Alonso Ramos. 2007. Towards the synthesis of support verb constructions: Distribution of syntactic actants between the verb and the noun. In L. Wanner and I. A. Mel'čuk, editors, *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*, pages 97–137. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Alex Alsina, Joan Bresnan, and Peter Sells, editors. 1997. *Complex Predicates*. CSLI Publications, Stanford.

Mengistu Amberber, Brett Baker, and Mark Harvey, editors. 2010. *Complex Predicates in Cross-Linguistic Perspective*. Cambridge University Press, Cambridge.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petra Barančíková, Rudolf Rosa, and Aleš Tamchyna. 2014. Improving Evaluation of English-Czech MT through Paraphrasing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 596–601, Reykjavík, Iceland. European Language Resources Association.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko,

Source	BEFORE	Fotbalisté Budějovic opět **nedali branku** Football players Budějovice again did not give gate Football players of Budějovice didn't make a goal again
	AFTER	Fotbalisté Budějovic opět **neskórovali** Football players Budějovice again did not score Football players of Budějovice didn't score again
GT	BEFORE	Footballers Budejovice again not given goal
	AFTER	Footballers did not score again Budejovice
Moses	BEFORE	Footballers Budějovice again gave the gate
	AFTER	Footballers Budějovice score again

Table 6: An example of the translated sentences. The judges unanimously agreed that AFTER translations are better than BEFORE. Moses translated the CP *dát branku* literally word by word and the meaning of this translation is far from the meaning of the source sentence.

Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague dependency treebank 3.0.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011. Czech-english parallel corpus 1.0 (CzEng 1.0). LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Laurel Brinton and Minoji Akimoto, editors. 1999. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

František Čermák, Renata Blatná, Jaroslava Hlaváčová, Jan Kocek, Marie Kopřivová, Michal Křen, Vladimír Petkevič, and Michal Schmiedtová, Věra Šulc. 2000. SYN2000: balanced corpus of written Czech. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

František Čermák, Jaroslava Hlaváčová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír Petkevič, Věra Schmiedtová, Hana Skoumalová, Johanka Spoustová, Michal Šulc, and Zdeněk Velíšek. 2005. SYN2005: balanced corpus of written Czech. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Wei-Te Chen, Claire Bonial, and Martha Palmer. 2015. English light verb construction identification using lexical knowledge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2375–2381. AAAI Press.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Afsaneh Fazly, Ryan North, and Suzanne Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, DeepLA '05, pages 38–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atsushi Fujita, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto, and Koichi Takeuchi. 2004. Paraphrasing of japanese light-verb constructions based on lexical conceptual structure. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The Multilingual Paraphrase Database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

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. 2012. 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.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden, July. Association for Computational Linguistics.

Otto Jespersen. 1965. *A Modern English Grammar on Historical Principles VI., Morphology*. A Modern English Grammar on Historical Principles. George Allen & Unwin Ltd., London.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics.

Václava Kettnerová, Markéta Lopatková, Eduard Bejček, Anna Vernerová, and Marie Podobová. 2013. Corpus Based Identification of Czech Light Verbs. In Katarína Gajdošová and Adriána Žáková, editors, *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, pages 118–128, Lüdenscheid, Germany. Slovak National Corpus, L'. Štúr Institute of Linguistics, Slovak Academy of Sciences, RAM-Verlag.

Václava Kettnerová, Petra Barančíková, and Markéta Lopatková. 2016. Lexicographic Description of Czech Complex Predicates: Between Lexicon and Grammar. In *Proceedings of the XVII EURALEX International Congress*.

Michal Křen, Tomáš Bartoň, Václav Cvrček, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Renata Novotná, Vladimír Petkevič, Pavel Procházka, Věra Schmiedtová, and Hana Skoumalová. 2010. SYN2010: balanced corpus of written Czech. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Klaus Krippendorff. 2007. Computing Krippendorff's Alpha Reliability. Technical report, University of Pennsylvania, Annenberg School for Communication.

Nitin Madnani and Bonnie J. Dorr. 2013. Generating Targeted Paraphrases for Improved Translation. *ACM Trans. Intell. Syst. Technol.*, 4(3):40:1–40:25, July.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 381–390, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

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

R. Mahesh K. Sinha. 2009. Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 40–46, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Veronika Vincze and János Csirik. 2010. Hungarian Corpus of Light Verb Constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1110–1118, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander K. Žolkovskij and Igor A. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija*, (6).

Peter Wallis. 1993. Information Retrieval based on Paraphrase.

Sina Zarrießand Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 77–84, Stroudsburg, PA, USA. Association for Computational Linguistics.