Chapter 2

Paraphrases of verbal multiword expressions: the case of Czech light verbs and idioms

Petra Barančíková
Charles University

Václava Kettnerová
Charles University

In this chapter, we deal with two types of Czech verbal MWEs: light verb constructions and verbal idiomatic constructions. Many verbal MWEs are characterized by the possibility of being paraphrased by single words. We explore paraphrasability of Czech verbal MWEs by single verbs in a semiautomatic experiment using word embeddings. Further, we propose a lexicographic representation of the obtained paraphrases enriched with morphological, syntactic and semantic information. We demonstrate one of its practical application in a machine translation experiment.

1 Introduction

Multiword expressions (MWEs) are widely acknowledged as a serious challenge for both foreign speakers and many NLP tasks (Sag et al. 2002). Out of various MWEs, those that involve verbs are of great significance as verbs represent the syntactic center of a sentence. Baldwin & Kim (2010) distinguish the following four types of verbal MWEs:

- verb-particle constructions (also referred to as particle verbs, or phrasal verbs), e.g., *catch up*, *put on*, *swallow down*;

Petra Barančíková & Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: the case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 35–59. Berlin: Language Science Press. DOI:10.5281/zenodo.1469553

- prepositional verbs, e.g., *come across*, *refer to*;
- light-verb constructions (also referred to as verb-complement pairs, or support verb constructions), e.g., *do a report*, *give a kiss*, *make an attempt*;
- verb-noun idiomatic constructions (also referred to as VP idioms), e.g., *spill the beans*, *pull strings*, *shoot the breeze*.

In this chapter, we focus on two particular types of Czech verbal MWEs: light-verb constructions (LVCs) and idiomatic verbal constructions (IVCs) as they also represent MWEs in Czech in contrast to the first two types that are primarily expressed as single prefixed verbs.

We explore the possibility of expressing these two types of MWEs by single synonymous verbs, which is considered to be one of their prototypical features, see e.g. Chafe (1968) and Fillmore et al. (1988). The motivation for this work lies in the fact that paraphrases greatly assist in a wide range of NLP applications such as information retrieval (Wallis 1993), machine translation (Madnani & Dorr 2013; Callison-Burch et al. 2006; Marton et al. 2009) or machine translation evaluation (Kauchak & Barzilay 2006; Zhou et al. 2006; Barančíková et al. 2014).

The content of this chapter is an extended version of Barančíková & Kettnerová (2017). In addition, it is further explored with IVCs and linguistic properties of LVCs and IVCs relevant to the paraphrasing task are discussed in detail. The new version of the dictionary of paraphrases is larger and it provides a more elaborated set of morphological, syntactic and semantic features, including information on aspects and aspectual counterparts of verbs.

This chapter is structured as follows. First, linguistic properties of LVCs and IVCs are discussed (§2) and related work on their paraphrases is introduced. Second, a paraphrasing model is proposed, namely the selection of LVCs and IVCs, an automatic extraction of candidates for their paraphrases and their manual evaluation are described in detail (§3). Third, the resulting data and their representation in a dictionary of paraphrases are introduced (§4). Finally, in order to present one of the many practical applications of this dictionary, a random sample of paraphrases of LVCs is used in a machine translation experiment (§5).

2 Linguistic properties of LVCs and IVCs

Both LVCs and IVCs represent verbal multiword units: they are composed of separate words that, however, refer to an extralinguistic reality as a whole. Their

linguistic properties relevant for their paraphrasability by single verbs are introduced below.

2.1 Light-verb constructions

The theoretical research on light-verb constructions is characterized by an enormous diversity in terms and analyses used, see esp. Amberber et al. (2010) and Alsina et al. (1997). Here, we use the term LVC for a multiword unit within which the verb – not retaining its full semantic content – provides grammatical functions and to which the main predicative content is contributed by a noun; as a result, such a multiword unit serves as a single predicative unit, see e.g. Algeo (1995), Alsina et al. (1997) and Butt (2010).[1] In contrast to IVCs, predicative nouns in LVCs have the same meanings as in nominal structures, meanings of light verbs are rather impoverished when compared with their full verb counterparts, see §2.2.

In the Czech language, the central type of LVCs are represented by LVCs in which predicative nouns are expressed as a direct or indirect object of a light verb (e.g., *dostat strach* 'to get fear' ⟹ 'to become afraid', *vzdát úctu* 'to pay tribute', and *vyvolat pobouření* 'to provoke indignation' ⟹ 'to cause uproar'). The LVCs in which a predicative noun occupies an adverbial of the light verb, (e.g., *dát do pořádku* 'to put in order', *mít pod kontrolou* 'to have under control', *mít na starosti* 'to have on care' ⟹ 'to be responsible') are more syntactically and morphologically fixed than the central type of LVCs (Radimský 2010).

As single predicative units, most LVCs have their single predicative counterparts by which they can be paraphrased. A single verb paraphrase can be either morphologically related, or non-related with the predicative noun representing the nominal component of the paraphrased LVC. For example, the LVCs *dát polibek* and *dát pusu* 'give a kiss' can be both paraphrased by the verb *políbit* 'to kiss', which is morphologically related only with the nominal component of the first LVC. There is no synonymous verb morphologically related to the nominal component of the second LVC.

In contrast to their single predicative paraphrases, LVCs manifest greater flexibility in their modification, compare e.g. adjectival modifiers of the LVC *dát polibek* 'give a kiss' and the corresponding adverbial modifiers of its single verb paraphrase *políbit* 'to kiss': *dát vášnivý/něžný/letmý/manželský/májový/smrtící polibek* 'give a passionate/tender/fleeting/marriage/May/fatal kiss' vs. *vášnivě/*

[1] Besides predicative nouns, adjectives, adverbs and verbs can also serve as predicative elements. These cases are left aside here.

*něžně/letmo/*manželsky/*májově/*smrtelně políbit* 'to kiss passionately/tenderly/fleetingly/*marriagely/*Mayly/?fatally'. Easier modification of LVCs is often considered a motivation for their use (Brinton & Akimoto 1999).

Another motivation lies in the possibility to structure the expressed event in a more subtle way than what single verbs allow. For example, in Czech various combinations of the grammatical aspect of light verbs and the number of predicative nouns allow for the expression of several meanings that cannot be expressed with single verbs; these cases require lexical modification, see Table 1.

Finally, in many cases, the selection of different light verbs allows for per-

Table 1: Possible combinations of the grammatical aspect of the light verbs *dát*^{pf}, *dávat*^{pf} 'to give' and the number of the noun *polibek* 'kiss' and their paraphrasability by the perfective and imperfective single verbs *políbit*^{pf} and *líbat*^{impf} 'to kiss', respectively.

LVC	Single verb paraphrase	Lexical modification	Example[a]
sg & pf	pf	no	*Petr dal Janě polibek.* 'Peter gave a kiss to Jane.' ∼ *Petr Janu políbil.* 'Peter kissed Jane.'
pl & impf	impf	no	*Petr dával Janě polibky.* 'Peter gave kisses to Jane.' ∼ *Petr Janu líbal.* 'Peter was kissing Jane.'
pl & pf	pf	yes	*Petr dal Janě polibky.* 'Peter gave several kisses to Jane.' ∼ *Petr Janu několikrát políbil.* 'Peter kissed Jane several times.'
sg & impf	impf	yes	*Petr Janě dával polibek.* 'Peter was giving a kiss to Jane.' ∼ *Petr Janu právě líbal.* 'Peter was just kissing Jane.'

[a]Let us emphasize that the single verb paraphrases of the last two combinations require to be lexically modified – by the words *několikrát* 'several times' and *právě* 'just', respectively.

spectivization of the expressed event from the point of view of its different participants, see esp. Kettnerová & Lopatková (2015). For example, besides the light verb *dát* 'to give', the noun *polibek* 'kiss' can select the light verb *dostat* 'to get' as well. The LVC *dát polibek* 'to give a kiss' promotes a kisser in the subject position while the LVC *dostat polibek* 'to get a kiss' puts a kissee into this position. Both these LVCs are paraphrasable by a single verb *políbit* 'to kiss', however, with different values of the grammatical voice: the LVC *dát polibek* 'to give a kiss' can be paraphrased by the verb *políbit* 'to kiss' in the active voice (e.g., *Petr dal Janě polibek.* 'Peter gave a kiss to Jane.' ∼ *Petr Janu políbil.* 'Peter kissed Jane.') while the LVC *dostat polibek* 'to get a kiss' requires the passive voice of the verb *políbit* 'to kiss' (e.g., *Jana dostala od Petra polibek.* 'Jane got a kiss from Peter.' ∼ *Jana byla políbena od Petra.* 'Jane was kissed by Peter'.)

LVCs in NLP. One of the trending topics concerning LVCs in the NLP community is their automatic identification. In this task, various statistical measures often combined with information on syntactic and/or semantic properties of LVCs are employed, see e.g. Bannard (2007) and Fazly et al. (2005). The automatic detection benefits especially from parallel corpora representing valuable sources of data in which LVCs can be automatically recognized via word alignment, see e.g. Chen et al. (2015), de Medeiros Caseli et al. (2010), Sinha (2009), Zarrieß & Kuhn (2009). However, work on paraphrasing LVCs is still not extensive. For example, a paraphrasing model has been proposed within the Meaning↔Text Theory (Žolkovskij & Mel'čuk 1965); its representation of LVCs by means of lexical functions and rules applied in the paraphrasing model are thoroughly described in Alonso-Ramos (2007). Further, Fujita et al. (2004) presents a paraphrasing model which takes advantage of semantic representation of LVCs by lexical conceptual structures. As with our method proposed in §3, their model also takes into account several morphological and syntactic features of LVCs, which have turned out to be highly relevant for the paraphrasing task.

2.2 Idiomatic Verbal Constructions

Despite their low frequency, IVCs form a substantial part of a lexis, see e.g. Baldwin & Kim (2010), Sag et al. (2002) and Cowie (2001). Similarly to LVCs, definitions of idioms vary depending on diverse purposes of their description, see e.g. Healy (1968), Fraser (1970), van der Linden (1992) and Nunberg et al. (1994).

Here, we define an IVC as a verbal multiword unit that exhibits strong lexical co-occurrence restrictions so that at least one of its parts cannot be used with

the same meaning outside the given multiword unit. The idiomatic meaning of individual components of IVCs is reflected in the fact that they are only rarely interchangeable with words of similar meanings. IVCs thus represent highly conventionalized multiword units, see e.g. Everaert et al. (2014), Granger & Meunier (2008) and Cowie (2001). IVCs can exhibit the following specific properties, see e.g. Burger et al. (2007), Čermák (2001) and Everaert et al. (2014):

- markedness at the syntactic and/or morphological level: e.g., *vzít za své* 'take as one's own' ⇒ 'to be no more' (syntactically marked as the reflexive adjective *své* does not modify any noun), and *nalít někomu čistého vína* 'to pour someone pure wine' ⇒ 'to tell someone the honest truth' (morphologically marked due to the partitive genitive of the noun *víno* 'wine', which is highly restricted in contemporary Czech);

- figuration: e.g., *vstát z mrtvých*, 'raise the dead' (as it involves a metaphor), *pověsit se někomu na krk* 'to hang around someone's neck' (as it involves a metonymy);

- fixedness at syntactic and/or morphological level: e.g., *postavit někoho na nohy* 'to put someone back on his feet' (syntactically fixed as it cannot be transformed into the passive structure), and *přijít na jiné myšlenky* 'to come to different ideas' ⇒ 'to find something else to think about' (morphologically fixed as the noun *myšlenka* 'idea' can have only the plural form);

- proverbiality: IVCs are typically used for recurrent socially significant situations, implying often their subjective evaluation (e.g., *vidět někomu do duše* 'to see right through someone');

- informality: IVCs are typically of informal register (e.g., *strčit si něco za klobouk* 'to put something behind a hat' ⇒ 'to stick it up one's jumper').

Some IVCs can be paraphrased by a single word verb, see e.g. the IVC *podat někomu pomocnou ruku* 'to give someone helping hand' and its single verb paraphrase *pomoci* 'to help'. However, many IVCs are paraphrasable rather by a whole syntactic structure, see e.g. the IVC *mít slovo* 'to have a word' ⇒ 'to be someone's turn to speak'.

IVCs in NLP. There is considerable work focused on automatic identification of idioms in the text and their extraction (Cook et al. 2007; Li & Sporleder 2009;

Muzny & Zettlemoyer 2013; Peng et al. 2015; Katz 2006). However, little attention has been paid to paraphrases of idioms. Let us introduce two works focused on paraphrases of idioms. First, Pershina et al. (2015) identifies synonymous idioms based on their dictionary definitions and their occurrences in tweets. Similarly, Liu & Hwa (2016) generate paraphrases of idioms using dictionary entries. However, there are no lexical resources available for NLP applications providing information on idioms in Czech.

3 Paraphrase model

In this section, the process of extracting paraphrases is described in detail. First, we present the selection of LVCs and IVCs (§3.1). For their paraphrasing, we had initially intended to use some of the existing resources, however, they turned out to be completely unsatisfactory for our task.

First, we used the *ParaPhrase DataBase* (PPDB) (Ganitkevitch & Callison-Burch 2014), the largest paraphrase database available for the Czech language. PPDB was created automatically from large parallel data. Unfortunately, there were only 54 candidates for single verb paraphrases of LVCs present. A manual analysis of these candidates showed that only 2 of them were detected correctly, the rest was noise in PPDB. Similarly for idioms, PPDB contained a correct single verb paraphrase for only 6 IVCs from our data (i.e. about 1%). As this number is clearly insufficient, we chose not to use parallel data for paraphrasing.

Therefore, we adopted another approach to the paraphrasing task applying *word2vec* (Mikolov et al. 2013), a neural network model. *Word2vec* is a group of shallow neural networks generating word embeddings, i.e. representations of words in a continuous vector space depending on the contexts in which they appear. In line with the distributional hypothesis (Harris 1954), semantically similar words are mapped close to each other (measured by the cosine similarity) so we can expect LVCs and IVCs to have similar vector space distribution to their single verb paraphrases.

Word2vec computes vectors for single tokens. As both LVCs and IVCs represent multiword units, their preprocessing was thus necessary: each LVC and IVC had to be first identified and connected into a single token (§3.2). Particular settings of our model for an automatic extraction of candidates for single verb paraphrases are described in §3.3.

The advantage of this approach is that only monolingual data – generally easily obtainable in a large amount – is necessary for word embeddings training. The disadvantage is that not only paraphrases can have similar word embeddings.

Antonyms and words with more specific or even different meaning can appear in similar contexts as well. Therefore, a manual evaluation of the extracted candidates is necessary (§3.4).

3.1 Data selection

3.1.1 LVCs selection

Three different datasets of LVCs – containing together 2,389 unique LVCs[2] – were used in our experiment. As all the datasets were 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). This dataset consists of both LVCs in which predicative nouns are expressed as a direct or indirect object by a prepositionless case (e.g. *položit otázku* 'put a question') and LVCs in which predicative nouns are expressed as an adverbial by a simple prepositional case (e.g., *dát do pořádku* 'put in order') or by a complex prepositional group (e.g., the verb *přejít* 'go' plus the complex prepositional group *ze smíchu do pláče* 'from laughing to crying').

The second dataset resulted from a project aiming to enhance the high coverage valency lexicon of Czech verbs VALLEX[3] with the information on LVCs (Kettnerová et al. 2016). In this case, only the predicative nouns expressed as the direct object by the prepositionless accusative were selected. For identification of LVCs, the modified test of coreference was applied (Kettnerová & Bejček 2016). As the frequency and saliency have been taken as the main criteria for their selection, the resulting set represents a valuable source of LVCs for Czech.

The third small dataset is represented by LVCs in which the predicative noun is expressed as an adverbial. These LVCs were obtained from the VALLEX lexicon as a result of manual analysis of verbal multiword units marked as idioms. As these multiword units were treated inconsistently in the annotation, including not only IVCs but sometimes also LVCs with predicative nouns in adverbial positions, the obtained dataset had to be manually selected.

As in the VALLEX lexicon, information on aspectual counterparts of the given verbs is available, we have used it to expand these datasets by adding missing aspectual counterparts. The overall number of LVCs in the datasets is presented below in Table 2. The union of LVCs from these datasets has been used in the paraphrase candidates extraction task.

[2]When counting aspectual counterparts separately, the number increases to 3,509 unique LVCs
[3]http://ufal.mff.cuni.cz/vallex/3.0/

3.1.2 IVCs selection

The dataset of IVCs was extracted from the VALLEX lexicon after the manual filtering of LVCs with predicative nouns in adverbial positions, see the third dataset in §3.1.1. From the obtained IVCs, those IVCs that include the highly polysemous pronoun *to* 'it' were removed as their automatic identification could be unreliable. The final set consists of 595 IVCs (counting aspectual counterparts separately 621 IVCs), see the statistics provided in Table 2.

Table 2: The number of LVCs and IVCs, verbs and nominal components in the three datasets described in §3.1.1, before (first number) and after (second number) the aspectual counterparts expansion.

Dataset	LVCs	IVCs	Verbs	Nominal components
First	726/1,167	0/0	49/84	612
Second	1,640/2,366	0/0	126/131	699
Third	104/106	595/621	310/324	324
Union[a]	2,389/3,509	595/621	417/446	1444

[a]The numbers do not add up due to a small overlap among the datasets.

3.2 Data preprocessing

We used four large lemmatized and POS-tagged corpora 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). These corpora were further extended with the data from the Czech Press – a large collection of contemporary news texts containing more than 2,000 million lemmatized and POS-tagged tokens. The overall statistics on all datasets is presented in Table 3.

To generate LVCs and IVCs paraphrases, all the selected LVCs and IVCs (§3.1) had to be automatically identified in the given corpora. For their identification, we started with verbs. First, all verbs in the corpora were detected. From these verbs, only those verbs that represent parts of the selected LVCs and IVCs were further processed. For each selected verb, each noun phrase in the context ± 4 words from the given verb was identified based on POS tags and extracted in case the verb and the given noun phrase can combine in some of the selected LVCs or IVCs.

Further, as word embeddings are generated for single words, each detected noun phrase was connected with its respective verb into a single word unit. In

Table 3: Basic statistics of datasets (numbers in millions of units).

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	57.03	2447.68
Total	90.77	3021.01

cases where some verb could combine with more than one noun phrase into LVCs or IVCs, or in cases where a particular noun phrase could be connected with more than one verb, we followed the principle that every verb should be connected to at least one noun phrase in order to maximize the number of identified LVCs and IVCs. For example, if there were two verbs v_1 and v_2 in a sentence and v_1 had a candidate noun phrase c_1, while v_2 had two candidate noun phrases 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, a verb was assigned the closest noun phrase on the basis of word order. When each noun phrase was connected maximally with one verb and each verb was connected maximally with one noun phrase, we have joined the noun phrases to their respective verbs into single word units with the underscore character and deleted the noun phrases from their original positions in sentences.

Further, to compensate sparsity of LVCs and IVCs in the data, after identifying a verb from the selected LVCs and IVCs in the data, its aspectual counterpart – if relevant – has been automatically added. For example, after detecting the imperfective verb *vcházet*$^{\text{impf}}$ 'enter' in the data and the prepositional noun phrase *do dějin* 'to history' in its context, not only the given imperfective verb, but also its perfective counterpart *vejít*$^{\text{pf}}$ have been connected with the given noun phrase into the resulting unit *vcházet_vejít_do_dějin*. We refer to such an artificially constructed unit as an *abstract unit* from now on. The abstract unit *vcházet_vejít_do_dějin* then replaced the verb *vcházet* in the sentence, while the noun phrase *do dějin* was deleted from the sentence. Each LVC and IVC identified in the data is thus represented by a single abstract unit representing also its relevant aspectual counterparts.

On this basis, almost 7 million instances of LVC and IVC abstract units were generated in the corpora, see Table 4. The rank and frequency of the most and the least common ones are presented in Table 5.

Table 4: The number of LVCs and IVCs detected in the data. The first row shows the total number of LVC and IVC abstract units identified in the data. The second row represents the number of their unique instances. The third row provides the number of those unique units with higher frequency than 100 occurrences. The last row shows the number of unique LVCs and IVCs without aspectual counterparts expansion, i.e. after splitting the generated abstract units back to a single verb–a single noun phrase pairs.

	LVCs	IVCs
abstract units	6,541,394	374,493
unique abstract units	1,776	211
unique abstract units > 100	1,361	153
unique MWEs	2,954	353

Table 5: The ranking of LVC and IVC abstract units identified in the data, based on their frequency.

rank	type	abstract unit	frequency
1.	LVC	*mít_problém* 'have a problem'	211,296
2.	LVC	*mít_možnost* 'have a possibility'	207,330
⋮	⋮	⋮	⋮
29.	IVC	*mít_na_mysli* 'have in mind'	43,521
⋮	⋮	⋮	⋮
1986.	IVC	*chytnout_chytat_chytit_za_špatný_konec* 'get hold of the wrong end of the stick'	1
1987.	LVC	*přechodit_přecházet_přejít_ze_smíchu_do_pláče* 'go from laughing to crying'	1

3.3 Word2vec model

To the resulting data, we applied *gensim*, a freely available *word2vec* implementation (Řehůřek & Sojka 2010). In particular, we 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 set a minimum number of word occurrences to 100 in order to limit the size of the vocabulary to reasonable words. Even though we increased frequencies of LVCs and IVCs by the unified representation for their aspectual counterparts, this limit still filtered more than 300 rarely used LVC and 50 IVC abstract units; the resulting number is provided in the third row of Table 4.

After training the model, for each of 1,361 LVC and 153 IVC abstract units with more than 100 occurrences we extracted 30 words with the most similar vectors. From these 30 words, we selected up to 15 single verbs closest to a given LVC or IVC abstract unit. These verbs were taken as candidates for single verb paraphrases of LVCs or IVCs in that abstract unit. On average, there were 7 candidates for each LVC abstract unit and 10 candidates for each IVC abstract unit.

Before the manual evaluation of the candidates, the abstract units were divided back to individual IVCs or LVCs and their paraphrase candidates were again enriched with their aspectual counterparts from the VALLEX lexicon. This way, annotators could select a paraphrase with a proper aspect for each verbal MWE.

3.4 Annotation process

In this section, the annotation process of the candidates for single verb paraphrases of LVCs and IVCs is thoroughly described. Let us repeat that *word2vec* generates semantically similar words depending on the context in which they appear. However, not only words having the same meaning can have similar space representations, but words with an opposite meaning, more specific meaning or even different meaning can be extracted as they can appear in similar contexts as well. Manual processing of the extracted single verbs was thus necessary for evaluating the results of the adopted method.

In the manual evaluation, two annotators were asked to indicate for each instance of the unique paraphrase candidates of an LVC or IVC whether it represents a single verb paraphrase of the given LVC or IVC, or not. For example, the single word verbs *upřednostňovat* and *preferovat* 'to prefer' were indicated as paraphrases of the LVC *dávat přednost* 'to give a preference'. Similarly, for the IVC *prásknout do bot* 'to bang to the shoes' ⇒ 'to take to one's heels', the single verbs *utéci* 'to run away' and *zdrhnout* 'to make off' among others were chosen as paraphrases.

Moreover, single verbs antonymous to LVCs or IVCs were marked as well since they can also function as paraphrases in a modified context. For example, for the LVC *vypovídat pravdu* 'to tell the truth' the antonymous verb *lhát* 'to lie' was selected as well, as the sentence *Nevypovídá pravdu.* 'He is not telling the truth.' can be paraphrased as *Lže.* 'He is lying.'.

Further, when the annotators determined a certain candidate as a single verb paraphrase of an LVC or IVC, they took into account the following four morphological, syntactic and semantic aspects.

First, they had to pay special attention to the morphosyntactic expression of arguments. As Czech encodes syntactic relations via morphological forms, changes in the morphological expression of arguments reflect different perspectives from which the event denoted by an LVC or IVC on the one hand and its single verb paraphrase on the other hand is viewed. For example, the single verb *potrestat* 'to punish' paraphrases the LVC *dostat trest* 'to get a punishment', however, the morphological forms of the punisher and the punishee, two semantic roles evoked by the given LVC and the single verb, differ. In the LVC *dostat trest* 'to get punishment', the punishee (*Petr* 'Peter') is expressed by the nominative and the punisher (*otec* 'father') has the form of the prepositional group *od*+genitive (e.g., $Petr_{nom}$ *dostal od* $otce_{od+gen}$ *trest.* 'Peter got punishment from his father.'), while with its single verb paraphrase *potrestat* 'to punish' the nominative encodes the punisher and the accusative expresses the punishee (e.g., $Otec_{nom}$ $Petra_{acc}$ *potrestal.* 'Father punished Peter.').

Second, the annotators had to take into account differences between the syntactic structure of a sentence created by an LVC or IVC and by its respective paraphrase. Particularly, the difference between sentences with a subject and subjectless sentences had to be indicated. For example, the LVC *dojít k oddělení* 'to happen to the separation' ⇒ 'the separation happens' is paraphrasable by the single verb *oddělit se* 'to separate', although the LVC forms a subjectless structure, the syntactic structure of its single verb paraphrase needs a subject.

Third, in some cases the reflexive morpheme *se/si*, marking usually intransitive verbs, has to be added to a single verb paraphrase so that its meaning corresponds to a meaning of its respective multiword counterpart. For example, the IVC *vejít do dějin* 'to come into history' ⇒ 'to go down in history' can be paraphrased by the verb *proslavit* only on the condition that the reflexive morpheme *se* is attached to the verb lemma *proslavit se* 'to achieve fame'.

Lastly, some verbs function as paraphrases of particular LVCs or IVCs only if nouns in these LVCs or IVCs have certain adjectival modifications. These paraphrases were paired with appropriate adjectives during the annotation. For ex-

ample, if the LVC *provozovat praxi* 'to run a practice' is to be paraphrased by the single verb *ordinovat* 'to see patients', the adjective *lékařský* 'medical' has to modify the noun *praxe* 'practice'.

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

To summarize, for each identified single verb paraphrase *v* of an LVC or IVC *l*, the annotators have chosen from the following options:

- *v* is a paraphrase of *l*
  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 morphosyntactic 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 change in a sentence structure is required
  e.g., *dojít k poruše* 'to happen to the failure' ⟹ 'the failure happens' and *porouchat se* 'to breakdown';
- *v* is a paraphrase of *l* but the modification of the verb lemma by the reflexive morpheme *se*/*si* is necessary
  e.g., *nést název* 'bear a name' and *nazývat se* 'to be called';
- *v* is a paraphrase of *l* only if a noun component of *l* is modified by a particular adjectival modification
  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*.

As a result of the annotation, for 1,421 of 2,954 LVCs identified in the data (48,1%) and for 200 of 353 IVCs (56,6%) at least one single verb paraphrase was found. The highest number of single verb paraphrases indicated for one multiword unit was nine and that was the LVC *provést řez* 'to make an incision' and the LVC *dát do pořádku* 'to put in order'. The total number of the indicated single

verb paraphrases of LVCs and IVCs was 2,912 and 498, respectively, see Table 6 providing results of the annotation including the frequency of the additional morphological, syntactic and semantic features used in the annotation.

Table 6: The basic statistics on the annotation.

	LVC	IVC
no constraints	2063	336
+ antonymous	115	47
+ reflexive morpheme	473	85
+ morphosyntactic change	270	38
+ syntactic change	43	0
+ an adjective	30	1
total[4]	2912	498

4 Dictionary of paraphrases

3,410 single verbs indicated by the annotators as paraphrases or antonyms of 1,421 LVCs and 200 IVCs (§3.4) form the lexical stock of *ParaDi* 2.0, a dictionary of single verb paraphrases of Czech multiword units of the selected types.[5]

The format of *ParaDi* 2.0 has been designed with respect to both human and machine readability. The dictionary is thus represented as a plain table in the TSV format, as it is a flexible and language-independent data format.

Each lexical entry in the dictionary describes an individual LVC or IVC, providing the following information:

(i) *type* – the type of the given verbal multiword expression with the following three possible values: LVC (indicating an LVC with the predicative noun in the direct or indirect object position), ILVC (representing an LVC with the predicative noun in the adverbial position), or IVC;

(ii) *verb* – a lemma of the verbal component of the given multiword unit;

(iii) *reflexive* – the reflexive morpheme of the lemma, if relevant;

[4] The columns do not add up as the features are not mutually exclusive as mentioned earlier.
[5] *ParaDi* 2.0 is freely available at the following URL: http://hdl.handle.net/11234/1-2377.

(iv) *aspect* – a value of the grammatical aspect of the verb;

(v) *aspectual counterpart* – the aspectual counterpart of the verb, if relevant;

(vi) *noun phrase* – the nominal component of the given multiword unit;

(vii) *morphology* – the morphemic form of the given noun phrase;

(viii) *lemmatized noun phrase* – a lemma representing the noun phrase;

(ix) *synonyms* – a list of synonymous single verb paraphrases;

(x) *antonyms* – a list of antonymous single verbs;

(xi) *adj-modification* – a list of single verb paraphrases and adjectival modifications of the nominal component of the LVC or IVC;

(xii) *structural_change* – a list of single verb paraphrases requiring a change in their sentence structure;

(xiii) *voice_change* – a list of single verb paraphrases requiring changes in the morphosyntactic expression of arguments.

While the information provided in the columns (i)-(viii) concerns multiword units, the information given in (ix)-(xiii) is relevant for their single verb paraphrases. A single verb paraphrase can appear in several columns if it is relevant. For example, the verb paraphrase *zalíbit se* 'to find appealing' of the LVC *nalézt zalíbení* 'to find a delight' ⇒ 'to find appealing' is present in both columns *reflexive* and *voice_change* as it represents the verb paraphrase, which requires both adding the reflexive morpheme *se* to the verb lemma and changes in the morphosyntactic expression of its arguments.

5 Machine translation experiment

In this section, we show how the dictionary providing high quality data can be integrated into an experiment with improving statistical machine translation quality. If translated separately, multiword expressions often cause errors in machine translation. For example, IVCs have been reported to negatively affect statistical machine translation systems which might achieve only half of the BLEU score (Papineni et al. 2002) on the sentences containing IVCs compared to those that do not (Salton et al. 2014).

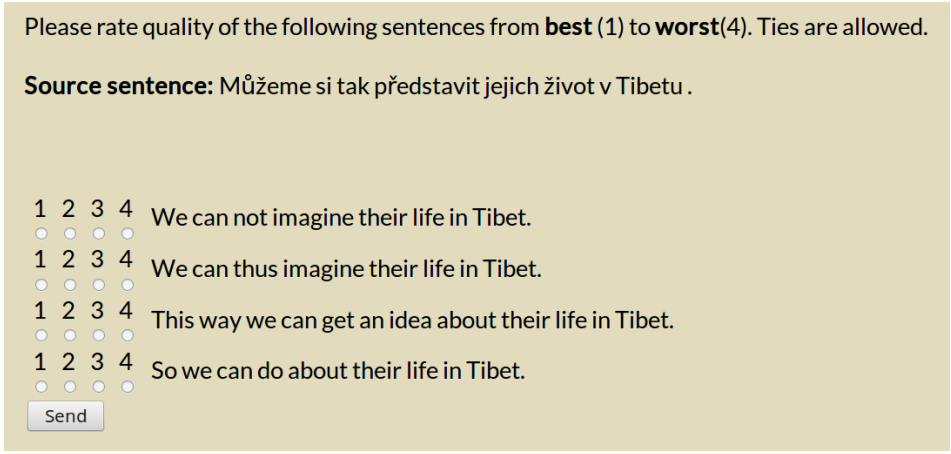

Figure 1: Example of the annotation interface for the MT experiment.

We took advantage of the *ParaDi* dictionary in a machine translation experiment in order to verify its benefit for one of the key NLP tasks. We experimented only with LVCs as we expected quality of LVC translations higher than those of IVCs due to their weaker lexical markedness and their more common use as their higher frequencies in the data suggested (see Table 4).

We selected 50 random LVCs from the dictionary. For each of them, we randomly extracted one sentence from our data containing the given LVC. This set of sentences is referred to as BEFORE. By substituting the LVC for its first paraphrase, i.e. the closest paraphrase in the vector space, we have created a new dataset, referred to as AFTER. We have translated both these datasets – BEFORE and AFTER – to English using two freely available MT systems – *Google Translate*[6] (GT) and *Moses*.[7]

We used crowdsourcing for evaluation of the resulting translations. Six annotators were presented randomly a Czech source sentence either from the dataset BEFORE or from AFTER and their English translations in a randomized order. The annotation interface is displayed in Figure 1. For each translated sentence, the annotators had to indicate its quality, allowing for the same ranking of more than one translated sentences.

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

[6]http://translate.google.com
[7]http://quest.ms.mff.cuni.cz/moses/demo.php

ate agreement. The results of replacing the selected verbal MWEs by their single verb paraphrases in machine translation are very promising: annotators clearly preferred translations of AFTER (i.e. the translations with single verbs) to BEFORE (i.e. with LVCs), in 45% of cases for Moses and in 44% of cases for Google Translate. The results are consistent for both translation systems, see Table 7.

Table 7: Results of the manual evaluation of the MT experiment. The first column shows the source of the better ranked sentence in the pairwise comparison within one translation model or whether they tied.

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

However, the example in Table 8 illustrates that even minimal change in a source sentence can substantially change its translations as both the translation models are phrase-based.[8] Based on this fact, we can expect that the evaluation of the translations was not affected only by differences between translations of LVCs and their respective single verb paraphrases but by overall low quality of the translations, which is inevitably reflected in the lower inter-annotator agreement, typical of machine translation evaluation (Bojar et al. 2013). The judges unanimously agreed that the translations of the AFTER source sentence are better than the translations of the BEFORE source sentence. Both systems exhibited a tendency to translate the LVC *dát branku* literally word by word, resulting in incorrect translations of the BEFORE source sentence.

6 Conclusion

We have explored the paraphrasability of Czech light-verb constructions and idiomatic verbal constructions. We have shown that their single verb paraphrases are automatically obtainable from large monolingual data with a manual verification in a significantly larger scale than from paraphrase tables generated from parallel data. Our semiautomatic experiment further revealed that although these verbal multiword units exhibit different linguistic properties, the possibility to

[8]The translations were performed on 9th July 2016, i.e. before a massive expansion of neural translation systems.

Table 8: An example of translated sentences.

Source	BEFORE	*Fotbalisté Budějovic opět* ***nedali*** ***branku*** Footballers Budějovice again did.not.give gate 'Footballers of Budějovice didn't make a goal again.'
	AFTER	*Fotbalisté Budějovic opět* ***neskórovali*** Footballers Budějovice again did.not.score 'Footballers 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

paraphrase them is very similar; for about one half of the selected light-verb constructions and idiomatic verbal constructions single verb paraphrases have been detected.

The results of our experiment form the lexical stock of a new version of the freely available *ParaDi* dictionary. 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.). We have used largely language independent methods, a similar dictionary can be thus created for other languages as well.

Acknowledgments

The research reported in this chapter was supported by the Czech Science Foundation GA ČR, grant No. GA15-09979S. This work used 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.

Abbreviations

GT	Google Translate	LVCS	light-verb constructions
IVCS	idiomatic verbal constructions	MT	Machine Translation

References

Algeo, John. 1995. Having a look at the expanded predicate. In Bas Aarts & Charles F. Meyer (eds.), *The verb in contemporary English: Theory and description*, 203–217. Cambridge: Cambridge University Press.

Alonso-Ramos, Margarita. 2007. Towards the synthesis of support verb constructions: Distribution of syntactic actants between the verb and the noun. In Leo Wanner (ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, 97–137. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Alsina, Alex, Joan Bresnan & Peter Sells (eds.). 1997. *Complex predicates*. Stanford: CSLI Publications.

Amberber, Mengistu, Brett Baker & Mark Harvey (eds.). 2010. *Complex predicates in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing, Second edition*, 267–292. Boca Raton: CRC Press.

Bannard, Colin James. 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), 1–8. Association for Computational Linguistics.

Barančíková, Petra & Václava Kettnerová. 2017. ParaDi: Dictionary of paraphrases of Czech complex predicates with light verbs. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 1–10. Association for Computational Linguistics. April 4, 2017.

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

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, 1–44. Association for Computational Linguistics.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra GaluščÁková, Martin Majliš, David MareČek, Jiří Maršík, Michal Novák, Martin Popel & 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.

Brinton, Laurel & Minoji Akimoto (eds.). 1999. *Collocational and idiomatic aspects of composite predicates in the history of English*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Burger, Harald, Dmitrij O. Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.). 2007. *Phraseologie / phraseology: Ein internationales Handbuch zeitgenössischer Forschung / an international handbook of contemporary research: Volume 1*. Berlin/ New York: Walter de Gruyter.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker & Mark Harvey (eds.), *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge: Cambridge University Press.

Callison-Burch, Chris, Philipp Koehn & 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), 17–24. Association for Computational Linguistics.

Čermák, František. 2001. Substance of idioms: Perennial problems, lack of data or theory? *International Journal of Lexicography* 14(1). 1–20.

Čermák, František, Renata Blatná, Jaroslava HlaváČová, Jan Kocek, Marie Kopřivová, Michal Křen, Vladimír PetkeviČ, Věra Schmiedtová & Michal šulc. 2000. *SYN2000: Balanced corpus of written Czech*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

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

Chafe, Wallace L. 1968. Idiomaticity as an anomaly in the chomskyan paradigm. *Foundations of Language* 4(2). 109–127.

Chen, Wei-Te, Claire Bonial & Martha Palmer. 2015. English light verb construction identification using lexical knowledge. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (AAAI'15), 2375–2381. AAAI Press.

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on*

Multiword Expressions (MWE '07), 41–48. Association for Computational Linguistics.

Cowie, Anthony (ed.). 2001. *Phraseology: Theory, analysis, and applications*. Oxford, UK: Oxford University Press.

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

Everaert, Martin, Erik-Jan van der Linden, André Schenk & Rob Schreuder. 2014. *Idioms: Structural and psychological perspectives*. New York, USA & East Sussex, UK: Psychology Press.

Fazly, Afsaneh, Ryan North & 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), 38–47. Association for Computational Linguistics.

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connore. 1988. Regularity and idiomaticity in grammatical constructions: The case of *Let Alone*. *Language* 64(3). 501–538.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language* 6(1). 22–42. http://www.jstor.org/stable/25000426.

Fujita, Atsushi, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto & 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 2004)* (MWE '04), 9–16. Association for Computational Linguistics.

Ganitkevitch, Juri & Chris Callison-Burch. 2014. The multilingual paraphrase database. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Joseph Mariani Bente Maegaard, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014). European Language Resources Association (ELRA).

Granger, Sylviane & Fanny Meunier (eds.). 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam / Philadelphia: John Benjamins.

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

Healy, Adam. 1968. English idioms. *Kivung (Journal of the Linguistic Society of the University of Papua New Guinea* 1(2). 71–108.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-*

06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE '06), 12–19. Association for Computational Linguistics.

Kauchak, David & 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), 455–462. Association for Computational Linguistics.

Kettnerová, Václava, Petra Barančíková & Markéta Lopatková. 2016. Lexicographic description of Czech complex predicates: Between lexicon and grammar. In George Meladze Tinatin Margalitadze (ed.), *Proceedings of the 17th EURALEX international congress*. Tbilisi, Georgia: Ivane Javakhishvili Tbilisi University Press. September 6-10, 2016.

Kettnerová, Václava & Eduard Bejček. 2016. Distribution of valency complements in Czech complex predicates: Between verb and noun. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 515–521. Paris, France: European Language Resources Association.

Kettnerová, Václava & Markéta Lopatková. 2015. At the lexicon-grammar interface: The case of complex predicates in the functional generative description. In Eva Hajičová & Joakim Nivre (eds.), *Proceedings of depling 2015*, 191–200. Uppsala, Sweden: Uppsala University.

Kettnerová, Václava, Markéta Lopatková, Eduard Bejček, Anna Vernerová & Marie Podobová. 2013. Corpus based identification of Czech light verbs. In Katarína Gajdošová & Adriána Žáková (eds.), *Proceedings of the 7th International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, 118–128. Lüdenscheid, Germany: RAM-Verlag.

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

Krippendorff, Klaus. 2007. *Computing krippendorff's alpha reliability*. Tech. rep. University of Pennsylvania, Annenberg School for Communication. http://repository.upenn.edu/asc_papers/43.

Li, Linlin & Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09), 315–323. Singapore: Association for Computational Linguistics.

Liu, Changsheng & Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *HLT-NAACL*, 363–373.

Madnani, Nitin & Bonnie J. Dorr. 2013. Generating targeted paraphrases for improved translation. *ACM Transactions on Intelligent Systems and Technology* 4(3). 40:1–40:25.

Marton, Yuval, Chris Callison-Burch & 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), 381–390. Association for Computational Linguistics.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. arXiv preprint arXiv:1301.3781.

Muzny, Grace & Luke S. Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013), 1417–1421. Association for Computational Linguistics.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (ACL '02), 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI:10.3115/1073083.1073135

Peng, Jing, Anna Feldman & Hamza Jazmati. 2015. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of recent advances in natural language processing*, 507–511. Association for Computational Linguistics.

Pershina, Maria, Yifan He & Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Workshop on linking models of lexical, sentential and discourse-level semantics* (LSDSem), 76–82. Association for Computational Linguistics.

Radimský, Jan. 2010. *Verbonominální predikáty s kategoriálním slovesem*. České Budějovice: Editio Universitatis Bohemiae Meridionalis.

Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for*

NLP Frameworks, 45–50. European Language Resources Association (ELRA). May 22, 2010.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Salton, Giancarlo, Robert J. Ross & John D. Kelleher. 2014. An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation, HyTra@EACL 2014, April 27, 2014, Gothenburg, Sweden*, 36–41. http://aclweb.org/anthology/W/W14/W14-1007.pdf.

Sinha, R. Mahesh K. 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), 40–46. Association for Computational Linguistics.

van der Linden, Erik-Jan. 1992. Incremental processing and the hierarchical lexicon. *Computational Linguistics* 18(2). 219–238.

Wallis, Peter. 1993. Information retrieval based on paraphrase. In *PACLING '93, 1st Pacific Association for Computational Linguistics Conference](formerly JAJSNLP, the Japan-Australia Joint Symposia on Natural Language Processing).*

Zarrieß, Sina & 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). Association for Computational Linguistics.

Zhou, Liang, Chin-Yew Lin & 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), 77–84. Association for Computational Linguistics.

Žolkovskij, Alexander K. & Igor A. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija* 6. 23–28.