Building Dataset of Continuous Multi-word Expressions in Czech

Zuzana Nevěřilová
(xpopelk@fi.muni.cz)

Natural Language Processing Centre
Masaryk University, Brno, Czech Republic

Abstract. Multi-word expressions frequently cause incorrect annotations in corpora, especially in cases they contain inter-lingual homographs or out-of-vocabulary words.

We created a dataset of Czech continuous multi-word expressions (MWEs). The candidates were discovered automatically from Czech web corpus considering their orthographic variability. The candidates were classified and annotated manually. Afterwards, the dataset was extended automatically by generating all word forms of those MWEs that were annotated as nouns. We used the dataset as positive examples, we filtered out negative examples from the MWE candidates. We trained a classifier with mean accuracy 92.7%.

We have shown that the combined approach slightly outperforms approaches concerning only association measures mainly on MWEs containing inter-lingual homographs and out-of-vocabulary words. The discovery methods can be applied to other languages with the possibility of orthographic variability.

1 Introduction

Multi-word expressions (MWEs) consist of several words but behave as a single word to some extent [4]. Their idiosyncracy causes (among others) problems in corpus annotation which is conventionally token-based.

MWEs do not form a homogeneous group. [4] point out four main characteristics of MWEs: syntactic anomaly, non-compositionality, non-substitutability, and ambiguity between MWE and non-MWE readings. It is nevertheless necessary to mention that not all MWEs have all four characteristics.

[15] provides a taxonomy of MWEs with two basic groups: lexicalized phrases and institutionalized phrases. The former group can be broken down into three subgroups: fixed expressions, semi-fixed expressions, and syntactically-flexible expressions. Fixed MWEs never change and are sometimes considered as a word with spaces, e.g. *ad hoc*. Semi-fixed expressions “undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection”. [15]. Syntactically-flexible expressions have a larger degree of syntactic variability including word order and possible gaps.

In this work, we focus on fixed and semi-fixed expressions and their discovery. The aim was to create an extensive list of MWEs that will be used for Czech corpora annotation. We show that traditional methods based on association measures do not work well for a considerable number of Czech MWEs. We therefore proposed a method based on discovery of orthographic variability. Using this method, we extracted 26,704 MWE candidates that were annotated manually by four annotators who marked about 5,800 of the candidates as MWEs.

We observed other features of the annotated data and we built a classifier that was later used for discovery of further MWEs.

The paper is organized as follows: Section 2 introduces MWE annotation, Section 3 references MWE discovery methods and MWE-annotated corpora. In Section 4, we describe in detail the construction of the MWE dataset. Section 5 summarizes results of this work. A concluding discussion is found in Section 6.

2 MWE Annotation

Annotation pipelines mostly consist of tokenization, morphological analysis, and tagging. A naive approach would be to create a large list and to treat MWEs as words with spaces, i.e. to tokenize sentences like *It is a priori impossible* to (*It*, *is*, *a priori*, *impossible*).

This approach is suitable for fixed, non-decomposable expressions, less suitable for semi-fixed expressions, and unsuitable for syntactically-flexible expressions.

For example, [15] shows that the verb-particle construction *look up* can have two meanings (to look upwards and to search) in some contexts and only one in others. Lists are not sufficiently *flexible* to cope with this ambiguity.

Moreover, MWE lists should be extensive yet not complete. This lack of generality is called *lexical proliferation problem* in [15].

The approaches such as [10,20] annotate MWEs both as single tokens and as MWEs. For example, in Wiki50, MWEs are split into tokens and annotated according to the Inside Outside Beginning (IOB) standard [1]. The corpus LexSem [17] distinguishes strong and weak multi-word groupings.

2.1 MWE Annotation in Czech Corpora

The pipelines for Czech corpus annotation do not consider MWEs at all, at least in case of the web corpus `czTenTen` [18] and SYNv6 corpus provided by the Czech National Corpus [7]. Instead, parts of the most frequent MWEs were included into the dictionaries used by morphological analyzers.

For example, the word *priori* (being part of the MWE *a priori*) is in morphological dictionaries annotated as an adverb. As a result, *a priori* is annotated as two tokens, one being (incorrectly) a conjunction, another being an adverb. Apart from the Latin preposition, *a* is also a Czech conjunction (meaning *and*).

Similarly, the fixed MWE *hot dog* is annotated as an unknown token *hot* and a noun *doga* (a dog breed *Great Dane*). In `czTenTen`, *hot* is annotated as an interjection.

These two examples show that problems in MWE annotation concern out-of-vocabulary words and inter-lingual homographs.

3 Related Work

Multi-word expressions appear in a wide range of NLP tasks such as machine translation (e.g. [16]), parsing [6], or lexicography (e.g. [11]). Therefore the number of works concerning MWEs is enormous. In the following text, we focus on MWE discovery as well as works concerning Czech.

3.1 MWE Discovery

Early discovery approaches worked with collocation measures. Association measures, namely point-wise mutual information (first proposed by [**?**]) are among the most popular discovery methods. Later, works as [13] show that adding linguistic information improves the results.

Another group of approaches is based on lexical fixedness. For example, [5] use K-means clustering algorithm with cosine similarity to build an unsupervised approach to MWEs discovery.

MWEs are also discovered employing semantic properties: e.g. [9] use latent semantic analysis to identify non-compositional MWEs. An example approach combining several information sources is [19]. In this work, authors use among others orthographic variations with hyphens.

3.2 MWE Annotated Corpora

Although there are shared tasks at SIGLEX and working groups in PARSEME, not many MWE annotated corpora exist. Examples of such corpora are the social web corpus [17], French corpus with annotated multiword nouns [10], or the corpus of 50 Wikipedia articles with annotated MWEs – Wiki50 [20], or parallel corpus of TED talks [8].

3.3 MWEs in Czech

Czech MWEs are studied mainly with respect to their syntactic structure. SemLex, the lexicon of Czech MWEs, is used for syntactic identification of MWE occurrences in text. This process is described in detail in [3]. SemLex was built by identifying MWEs in the Prague Dependency Treebank [2].

4 Building the Dataset for Czech MWE Annotation

In this Section, we describe the process of automatic discovery of MWE candidates, manual classification, comparison with association measures, and training. We explain different aspects of the manual annotation that influence selection of the training examples.

4.1 MWE Discovery Based on Orthographic Variability

We made several observations on Czech MWEs and we found fluctuating orthography of frozen MWEs in the Czech web corpus `czTenTen` [18]. In other words, people are sometimes unsure whether the correct Czech orthography for expressions such as *a priori* is *apriori* or even *a-priori* (the correct variants are *a priori* and *apriori*). Similar experience is mentioned in [10] and [19] for different languages.

Using a web corpus was a key decision: first, `czTenTen` is so far one of the largest Czech corpora, second, web corpus contains many kinds of (mostly unedited) texts. Our observations indicated that e.g. in discussion groups, orthographic variants appear frequently since people use a language between correct written Czech and spoken Czech and do not care much about the correctness.

The MWE discovery is described more in detail in [12]. The approach is straightforward: if a chunk exists in corpus in all three forms (several words, one word, several words with dashes) with a minimum frequency, we consider it a MWE candidate. This method discovered 26,704 MWE candidates.

4.2 Classification of MWE Candidates and Observation

The classification of MWEs was manual: four annotators had to decide whether a sequence of words is random (non-MWE), MWE with function of a noun, MWE with function of an adverb, unspecified English loanword or other foreign. The resulting collection contained 3,219 MWEs with function of a noun, 80 MWEs with function of an adverb, 2,325 English and 140 non-English foreign MWEs.

There were only 33 candidates that were annotated differently by the four annotators. We included all MWE candidates with a majority agreement in the dataset.

A more detailed observation of the classified data sample has shown that the positive entries (MWEs of any kind) do not contain evident non-MWEs. The high inter-annotator agreement was also caused by the similarity of the annotators: they shared the same field of study, interests, and age. On the other hand, observation of negative entries sample indicated that the dataset coverage is limited. We found two main reasons why the negative entries were rather noisy: First, many actual MWEs were annotated as non-MWE since the annotators did not understand the expression. Second, annotators sometimes did not annotate MWEs with non-standard spelling. Some MWEs are frequent with incorrect spelling, for example *á propos* has 876 occurrences in `czTenTen` while the correct spelling *à propos* has only 147 occurrences[1].

After observing the positive entries we could distinguish several types of Czech MWEs:

- **Czech fixed phrases** with syntactic anomalies (e.g. *stůj co stůj*, lit. imperative *go what go* meaning *by hook or crook*.

[1] The main reason in this case is that the character *á* is not in the Czech keyboard layout.

- **Non-English borrowings** which are not analyzable by most users of the Czech language (e.g. *faux pas*, *a priori*).
- **English calques** (loan translations) that are syntactically anomalous in Czech (e.g. *risk management* contains a noun modifier which precedes the head noun – such construction does not originally exist in Czech)
- **Proper names** (e.g. San Francisco, Air France). Although [15] show that location names in sport club names are ellidable (e.g. *the (San Francisco) 49ers*), proper names of sport clubs, companies, locations are still highly idiosyncratic. Personal names do not share this fixedness level, mainly for two reasons: first, they are not strictly continuous, since we can find constructions such as *Barack Obambi Obama*, second, most personal names (e.g. *John Smith*) are tuples formed from lists of first names (e.g. *John*, *Jane*) and last names (e.g. *Smith*). They are therefore decomposable and their components can be combined (e.g. *Jane Smith*).

The first two categories contain fixed MWEs whereas the third and fourth categories cover words that are often subject to inflection (semi-fixed MWEs). In some cases, the inflectional pattern is not clear to language users so they avoid inflection. For example, users prefer expressions such as *jít do obchodu Marks & Spencer* (*go to the store Marks & Spencer*) over (shorter) *jít do Markse & Spencera* (*go to Marks & Spencer's*).

The correct inflection requires gender assignment which is based mostly on the word ending. Roughly said, language users assign masculine inanimate gender to words ending with a consonant, feminine gender to words ending with *-a* or *-e* and neuter gender to words ending with *-o*. This rule is sometimes influenced by similar words that were adopted previously. For example, *after party* has the same gender as *party* (feminine) since *party* exists in Czech for decades. On the other hand, the gender of *Air France* is unclear and probably it is not assigned at all.

4.3 Automatic Extension of the Dataset

As mentioned above, semi-fixed MWEs are subject of inflection which is quite regular in Czech. We decided to extend the dataset with automatically generated word forms.

Within the manual annotations, we observed that annotators were often unsure whether English calques such as *news server* have to be annotated as nouns or unspecified English phrases. This indistinctness disappeared in case of inflected MWEs: e.g. if the annotator had to decide whether *news serveru* (genitive) is a noun or an unspecified English phrase, she was sure that it is a (Czech) noun. Therefore we decided to annotate all English calques such as *news server* as nouns if we found corpus evidence for their inflection, i.e. at least one inflected form.

Eventually, we modified the annotation for MWES with the same structure containing the same head noun. For example, we observed that *news server* is subject of inflection therefore we also inflected *mail server*, *DHCP server* etc. This modification was controlled manually since the rule is not generally applicable. For example, in case of *client server* the inflection makes no sense since the similarity in structure is only shallow (i.e. *client* is not modifying *server*).

From nominative singular forms, we generated genitive, dative, accusative, locative, and instrumental. We also generated plural forms for those MWEs that are not named entities. We did not generate vocative since it is used only for animate nouns. Also, the plural forms of named entities are rather rare, so we did not generate them. As a result, we obtained a dataset of 24,807 word forms.

4.4 Preparing Training Data

Next step in the task was to examine the relationship between traditional association measures and the annotated data. According to [4], association measures are straightforward for two-word expressions but rather complicated for longer ones. In the dataset, 24,360 entries are two-word expressions, 423 entries are three-word expressions, and 24 are longer. We decided not to take longer entries into account since they make only 1,8% of the data.

In Section 4.2, we described the nature of positive (MWEs) and negative (non-MWEs) annotations. Although the majority of the candidates mentioned in Section 4.1 were annotated as non-MWEs, we could not simply use them as negative examples. We filtered out named entities automatically and made further manual cleaning.

T-score: $\frac{f_{xy} - \frac{f_x f_y}{N}}{\sqrt{f_{xy}}}$ MI-score: $\log_2 \frac{f_{xy} N}{f_x f_y}$ MI^3-score: $\log_2 \frac{f_{xy}^3 N}{f_x f_y}$

min. sensitivity: $\min(\frac{f_{xy}}{f_y}, \frac{f_{xy}}{f_x})$ logDice: $14 + \log_2 \frac{2 \cdot f_{xy}}{f_x + f_y}$

log-likelihood: $2 \cdot (xlx(f_{xy}) + xlx(f_x - f_{xy}) + xlx(f_y - f_{xy}) + xlx(N) + xlx(N + f_{xy} - f_x - f_y) - xlx(f_y) - xlx(N - f_x) - xlx(N - f_y))$, where $xlx = f \ln(f)$

Fig. 1. Different association measures are based on frequency f_x of the token x, frequency f_y of the token y, frequency f_{xy} of the bigram (x, y) and the corpus size N.

In order to avoid skewed classes, we selected randomly the same number of positive and negative examples.

We computed association measures T-score, MI-score, MI^3-score, minimum sensitivity, logDice, and log-likelihood as described in [14]. The formulas are shown in Figure 4.4. The data were rescaled to $[-1, 1]$ using the function $r = \frac{(2x - max - min)}{(max - min)}$.

We are aware that the association measures are not independent features, therefore we employed the recursive feature elimination (RFE) in order to suppress less effective features. The best results (i.e. matching the largest number of positive examples and not matching the largest number of negative examples) were provided by MI-score, MI^3 and logDice. Roughly said, if the MI-score is a large number, the bigram is likely to be a MWE.

In Section 1, we mentioned that association measures do not work well for a large group of MWEs. For MWEs containing inter-lingual homographs, the MI-score is somewhat lower than for MWEs without inter-lingual homographs but still higher than for random bigrams. However, the minimum sensitivity is significantly lower for MWEs with homographs than for MWEs without homographs. This discrepancy which can be seen in Figure 4.4 causes problems in MWE discovery. MWEs containing inter-lingual homographs are not discovered by means of association measures.

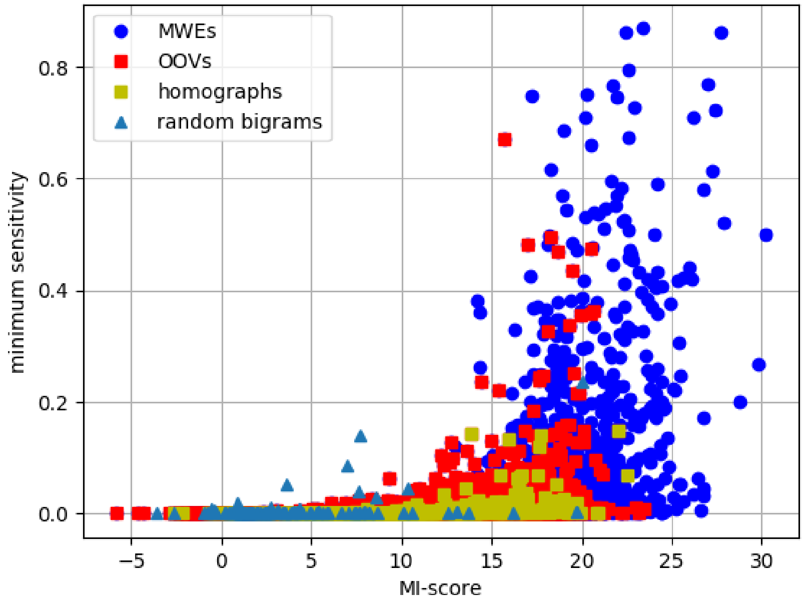

Fig. 2. MWEs have often high MI-score and higher minimum sensitivity than non-MWEs. However, MWEs containing inter-lingual homographs and OOVs have lower MI-scores and minimum sensitivity than other MWEs.

We decided to include the information about inter-lingual homography as a feature using a list of 251 Czech-English homographs. Similarly, we added the information whether the bigram contains OOV words.

4.5 Classifier Training

We used logistic regression on 5,792 examples. Using 4-fold cross-validation, the classifier had 91.4% mean accuracy without using information about OOV and homography as a feature. The information about homography increased the mean accuracy to 91.8%. Information about OOV increased the mean accuracy up to 92.7%. Eventually, information about OOV proved to be more useful.

5 Results

We obtained two results: a dataset of fixed and semi-fixed MWEs, and a classifier for MWEs.

Currently, the resource contains 4,731 MWE lemmata (24,807 word forms). The Table 5 shows different categories of the entries. Most of the MWEs have function of a noun, 634 are indeclinable.

We compared our dataset to SemLex [3]. The overlap in both resources is very small, only 40 lemmata are in both resources. The reason can be different source data: Prague Dependency Treebank contains mostly correct Czech sentences, while web corpora often contain non-standard language. In our resource, many entries are non-standard, however, some of them are frequent.

category	# of entries	# number of lemmata
foreign	2,221	2,221
nouns	22,459	2,422
adjectives	42	3
adverbs	81	81
particles	4	4

Table 1. Overview of Czech MWEs dataset

The classifier was trained on 75% of the example data and tested on 25%. The resulting mean accuracy is 92.7%, precision 93.5%, recall 95.9%, and F1-score 94.7.

6 Conclusion and Future Work

Frozen continuous MWEs are in many cases incorrectly annotated in Czech corpora. It is caused by the idiosyncratic nature of such MWEs: they often contain rare of foreign words and they sometimes evince syntactic anomalies. The aim of this work is to discover MWEs.

The paper presents a new dataset of Czech fixed and semi-fixed MWEs. We described the acquisition of the data and the annotation process. The annotated data were automatically extended since many MWEs are subject to inflection. Finally, we used the data for classifier training.

To our knowledge the dataset is the largest list of fixed and semi-fixed MWEs. Another dataset for Czech, SemLex contains all types of MWEs including syntactically-flexible ones. The overlap with SemLex [3], is insignificant: only 40 MWE lemmata occur in both resources. The results are difficult to compare with other works: some deal with syntactically-flexible MWEs which are more difficult to discover, some work for languages with weak inflection.

We plan to use the dataset and the classifier for identifying MWEs in new version of the Czech web corpus `czTenTen`. This application can provide extrinsic evaluation if a measure of quality of corpus annotation will be defined.

7 Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781.

References

1. Baldwin, B.: Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO+. https://lingpipe-blog.com/2009/10/14/ (October 2009), https://lingpipe-blog.com/2009/10/14/coding-chunkers-as-taggers-io-bio-bmewo-and-bmewo/, [accessed 2017-09-28]
2. Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: Prague Dependency Treebank 3.0 (2013), http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University
3. Bejček, E., Straňák, P., Pecina, P.: Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In: Proceedings of the 9th Workshop on Multiword Expressions. pp. 106–115. Association for Computational Linguistics, Atlanta, Georgia, USA (2013)
4. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword Expression Processing: A Survey. Computational Linguistics 0(ja), 1–92 (2017), https://doi.org/10.1162/COLI_a_00302
5. Van de Cruys, T., Moirón, B.n.V.: Semantics-based Multiword Expression Extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. pp. 25–32. MWE '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=1613704.1613708
6. Eryiğit, G., İlbay, T., Can, O.A.: Multiword Expressions in Statistical Dependency Parsing. In: Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages. pp. 45–55. SPMRL '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=2206359.2206365
7. Hnátková, M., Křen, M., Procházka, P., Skoumalová, H.: The syn-series corpora of written czech. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H.,

Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)

8. Johanna Monti, Federico Sangati, M.A.: Ted-mwe: a bilingual parallel corpus with mwe annotation: Towards a methodology for annotating mwes in parallel multilingual corpora. In: Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015. Accademia University Press, Torino
9. Katz, G., Giesbrecht, E.: Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. pp. 12–19. MWE '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=1613692.1613696
10. Laporte, E., Nakamura, T., Voyatzi, S.: A French Corpus Annotated for Multiword Nouns. In: Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions. pp. 27–30. Marrakech, Morocco (2008), https://halshs.archives-ouvertes.fr/halshs-00286552
11. Loukachevitch, N., Lashevich, G.: Multiword expressions in Russian thesauri RuThes and RuWordnet. In: 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL). pp. 1–6 (Nov 2016)
12. Nevěřilová, Z.: Annotation of Multi-Word Expressions in Czech Texts. In: Aleš Horák, P.R., Rambousek, A. (eds.) Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 103–112. Tribun EU, Brno (2015)
13. Ramisch, C., Schreiner, P., Idiart, M., Villavicencio, A.: An Evaluation of Methods for the Extraction of Multiword Expressions. In: In Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions MWE 2008. pp. 50–53. Marrakech, Morocco (2008)
14. Rychlý, P.: A Lexicographer-Friendly Association Score. In: RASLAN 2008. pp. 6–9. Masarykova Univerzita, Brno (2008)
15. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, vol. 2276, pp. 1–15. Springer Berlin Heidelberg (2002), http://dx.doi.org/10.1007/3-540-45715-1_1
16. Sakamoto, S., Ogawa, Y., Nakamura, M., Ohno, T., Toyama, K.: Utilization of Multi-word Expressions to Improve Statistical Machine Translation of Statutory Sentences. In: Otake, M., Kurahashi, S., Ota, Y., Satoh, K., Bekki, D. (eds.) New Frontiers in Artificial Intelligence: JSAI-isAI 2015 Workshops, LENLS, JURISIN, AAA, HAT-MASH, TSDAA, ASD-HR, and SKL, Kanagawa, Japan, November 16-18, 2015, Revised Selected Papers. pp. 249–264. Springer International Publishing, Cham (2017), https://doi.org/10.1007/978-3-319-50953-2_18
17. Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M.T., Conrad, H., Smith, N.A.: Comprehensive Annotation of Multiword Expressions in a Social Web Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation. pp. 455–461. ELRA, Reykjavík, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf

18. Suchomel, V.i.t.: Recent Czech Web Corpora. In: Aleˇ s Hor´ ak, P.R.y. (ed.) 6th Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 77–83. Tribun EU, Brno (2012)
19. Tsvetkov, Y., Wintner, S.: Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 836–845. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), `http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=2145432.2145525`
20. Vincze, V., Nagy T., I., Berend, G.: Multiword Expressions and Named Entities in the Wiki50 Corpus. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. pp. 289–295. Association for Computational Linguistics (2011), `http://aclanthology.coli.uni-saarland.de/pdf/R/R11/R11-1040.pdf`