Annotation of Multi-Word Expressions in Czech Texts

Zuzana Nevěřilová

Natural Language Processing Centre,
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xpopelk@fi.muni.cz

Abstract. Multi-word expressions (MWEs) are difficult to define and also difficult to annotate. Some of them cause serious errors in the traditional annotation pipeline tokenization – morphological analysis – morphological disambiguation. Many cases of incorrect annotation in Czech corpora are known. To narrow the research topic, we focus only in fixed MWEs – those with fixed word order and no ellidable components.
In this paper, we propose a corpus-based method that reveals fixed MWE candidates. From the web-based corpus of Czech, we extracted 25,091 expressions, 2,140 of them were identified as MWEs, 332 as probable MWEs, and 174 of them can be either MWEs or one single word.
Our method is based on corpus data observation that indicates that people are unsure when writing a MWE whether it is one word, a word with dashes, or several words. The result is a list of MWE candidates and also an application that classifies the input as MWE, probable MWE, or non-MWE.

Keywords: multi-word expressions, corpus, orthographical variants

1 Introduction

Most corpora are single-word tokenized, i.e. the input text is segmented on single tokens (surrounded by white spaces or punctuation). This common starting point sometimes causes problems in subsequent text analysis.

For example, the expression *a priori* is tokenized in two tokens: *a* and *priori*. The token *a* is ambiguous in Czech (it means *and* among others), and *priori* is not a Czech word. The expression is widely used in Czech (1.10 per million in Czech corpus `czTenTen`) but the tagging in `czTenTen` and `syn2010` (one part of the Czech National Corpus) is incorrect: *a* is tagged as conjunction, *priori* is tagged as adverb in both corpora.

To solve this problem, a large dictionary of multi-word expressions (MWEs) that are problematic for tagging would be useful. [7] define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)" and distinguish three classes: fixed MWEs, semi-fixed MWEs (such as compound nominals or proper names), and syntactically flexible expressions (such as idioms).

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 103–112, 2015.

Although many types of MWEs exist, in this work, we focus on two types of fixed MWEs:

- fixed MWEs (also called frozen MWEs or words with spaces): e.g. *a priori*, *křížem krážem* (meaning *(travel) the length and breadth*)
- *almost* fixed MWEs (e.g. those with inflectional component and no ellidable component): e.g. *hot dog* which has inflected forms such as *hot dogy*, *hot dogu* etc., *Karel IV.* (meaning the king Charles IV) with inflected forms such as *Karla IV.*, *Karlu IV.* etc.

Both types have fixed word order and no ellidable component. These criteria differ the studied MWEs from other types. For example, *New York Rangers* is a MWE but it has ellidable component since we can find *Rangers* in texts in the meaning of *New York Rangers*.

In order to reduce tagging errors, we need to distinguish problematic MWEs from all other n-grams. We did several measurements on MWEs found in Czech Wiktionary, n-grams with high associative measures, and a set of random n-grams. We propose a method to distinguish fixed MWEs from all other collocations. Such MWEs can be stored in a dictionary with the appropriate grammatical information (part-of-speech and grammatical categories).

We plan to use our method to re-annotate Czech corpus *cztenten* [8].

1.1 Current Tokenization and Tagging of Czech Corpora

Currently, Czech corpora developed at NLPC are annotated automatically using the `unitok` tokenizer and the morphological analyzer `majka` [10]. In case of words unknown to `majka`, the word tag guesser is used. Afterwards, all possible lemmata and morphological tags are disambiguated using the tagger `desamb` [9] based on inductive logical programming.

The most problematic part of this process is the tagging of unknown words and cross-lingual homonyms. In the former case, the guesser proposes possible lemmata and tags regardless of the context (the context is used only in the disambiguation part). In the latter case, Czech homonyms are used even when their frequency is very low. For this reason, the corpus czTenTen contains 17,261,404 imperatives (3,405.00 per million) which is in reality very unlikely. Rather than imperatives, a significant part of these tokens are foreign words. For example, the English word *top* is homonymous to the Czech imperative from *to drown*. Most expressions such as *Top 10* are therefore annotated as imperative followed by a number.

1.2 Paper Outline

In the Section 2, we describe the MWE fixedness issue in order to define MWE as most precisely as possible, Section 3 shows our preliminary research. In Section 4, we present the new MWE discovery method, Section 5 shows the results, and Section 6 discusses them. Section 7 proposes future work.

2 Related Work

Extraction of MWEs has been widely studied in many languages and several techniques exist. Statistical techniques are based on measuring co-occurence of tokens by various measures, e.g. t-score or pointwise mutual information [2], logDice [6] or mean and variance [3]. All techniques generally lead to a list of MWE candidates. However, finding fixed or semi-fixed candidates requires testing of basic properties of MWEs.

Usually, MWEs are identified by three criteria:

- non-compositionality: the meaning of the MWE is not perceivable by meanings of its components
- non-modifiability: the MWE can be used e.g. only in singular, inflectional languages distinguish between absolute non-modifiability and limited modifiability
- non-substitutability: the components of the MWE cannot be substituted by their synonyms

[4, p. 24] adds another criterion – the asymmetric association: "lexical association between components is much stronger from one component to another than vice versa". [4] shows that asymmetry is very important in case of noun phrases.

Testing the appropriateness of a MWE candidate includes testing the criteria, e.g. [1] tested non-compositionality, [5] tested non-substitutability of MWE candidates.

Multi-word tokenization has been deeply studied by [4].

3 A Preliminary Study

In our approach, we narrow the wide set of MWEs only to those that have problematic tagging, i.e. contain a non-Czech word or are a fixed sentence (a routine formula or a named entity). The hypothesis is that such MWEs are strongly fixed: either they are frozen (such as *křížem krážem* meaning *criss-cross*), or almost-frozen: they have fixed word order and only some components are subject of inflection (such as *hot dog*).

In the preliminary study, we did not focus much on MWE extraction. First, we only examined collocations from Wiktionary page concerning word expressions[1]. We observed the modifiability and asymmetry of the MWEs extracted from Czech Wiktionary. We also noticed that the orthography of MWE is sometimes difficult. Language users are probably unsure whether a frozen MWE is one word or several words. Therefore, many frozen MWEs occur in corpora as one word or as one word with dashes. Our observations have shown that this feature discriminates frozen MWE that are often subject of incorrect annotation.

[1] `http://cs.wiktionary.org/wiki/Kategorie:Česká_slovní_spojení`

Table 1. MWEs that were found in the corpus as one word or one word with dashes. Significant asymmetry or significant number of occurrences as one word or one word with dashes are marked with bold.

expression	freq	asymmetry	one word	dash
a to	1000000	1.0000	**3669**	**23**
Abú Dhabí	625	0.1025	1	**9**
Abú Zabí	1869	0.2335	3	**4**
ad acta	454	**0.0075**	**22**	3
ad hoc	5593	**0.0773**	**251**	**1266**
ano i	3731	1.8137	**54**	**12**
čáry máry	539	**0.0328**	**80**	**51**
český jazyk	8096	0.5621	3	**49**
chtě nechtě	10846	1.1990	**76**	**799**
dejme tomu	28626	**30.2847**	**370**	**13**
domino efekt	171	**67.3155**	**12**	**21**
ex officio	202	**0.0080**	2	**9**
ex offo	848	**0.0317**	**6**	**136**
fata morgána	387	0.3592	**123**	**9**
faux pas	3091	**13.1264**	**30**	**624**
hned tak	19743	1.0000	**223**	**7**
Hradec Králové	58366	1.8396	**18**	3
i když	1000000	1.0000	**74121**	**82**
IP adresa	6205	0.5918	**26**	**103**
IQ tykve	849	**0.0647**	**13**	**20**
Kanárské ostrovy	3945	11.8394	2	**4**
Karlovy Vary	43242	0.6917	**37**	**18**
křížem krážem	4633	0.2343	**44**	**55**
lážo plážo	735	1.0676	**69**	**138**
mírnix týrnix	36	1.1222	**12**	**10**
mírnyx týrnyx	41	1.5405	**11**	**5**
na shledanou	5379	**0.0086**	**12784**	3
New York	61917	0.2749	**175**	**43**
nomen omen	479	1.1692	**8**	**41**
obchodní dům	5819	1.2479	**8**	**169**
Pán Bůh	14501	3.4631	**11539**	**4**
po o	1166	1.0000	**908**	**49**
po spa	14	**0.0037**	**777**	**4**
s to	18286	1.0000	**223647**	**4**
San Francisco	5889	0.0989	**12**	**5**
San Marino	2305	**0.0379**	**7**	3
Srí Lanka	2782	0.3019	**20**	**23**
techtle mechtle	263	0.9644	**28**	**52**
to do	96790	1.0000	**1278**	**347**
volky nevolky	378	0.9077	2	**16**
zatím co	13675	1.0000	**667751**	**52**
zuby nehty	7283	0.3530	**76**	**217**

The observations can be seen in Table 3. It shows bigrams from Czech Wiktionary that were also found as one word in the corpus `czTenTen` (with frequency min. 3) and as one word with a dash. It can be seen that asymmetry is not a determining feature in this case and in further research, we did not take it into account.

It can also be seen that some of the bigrams (český jazyk, Karlovy Vary, Kanárské ostrovy, obchodní dům) are noun phrases formed by an adjective and a noun in grammatical agreement. Such cases are not problematic to annotate. Thus, for further research, we probably have to add this syntactic criterion as well.

4 Methods

From the preliminary observations, we concluded that fixed MWEs are characterized by orthographic variability. They can be written as two words, one word, or one word with dashes.

The second step was to find such occurrences in Czech corpora and to examine whether or not they are fixed MWEs. We tried to find more MWEs again in the corpus `czTenTen`. We started with words with dashes. We found 6,296,839 occurrences of words with dashes, 2,029,715 unique occurrences, and 388,364 appearing more than once, and 205,708 occurrences appearing more than twice.

We categorized the occurrences of words with dashes in several categories:

1. compound adjectives (such as *česko-německý*, meaning Czech-German). These words are recognized correctly by the morphological analyser `majka`
2. abbreviations (such as *KDU-ČSL*, *DVB-T*, *CD-ROM*)
3. proper names (such as *Aix-en-Provence*, *Mercedes-Benz*, *Saint-Exupéry*, *Müller-Thurgau*)
4. chemical nomenclature (such as *beta-karoten*, *L-karnitin*, *B-komplex*)
5. originally English words occurring ordinarily in Czech, sometimes even with inflection (*e-mail*, *play-off*, *sci-fi*, *know-how*, *pop-music*, *set-top-box*, *line-up*). Some of them are recognized by the morphological analyser `majka`, some of them are not in its database.
6. other words of foreign origin (*kung-fu*, *au-pair*, *tee-pee*, *laissez-faire*)
7. nicknames (such as *Margaret-ka*, *babča-helča*, *mam-ča*)
8. frozen expressions formed by Czech words (such as *více-méně*, *sem-tam*, *jakž-takž*, *vepřo-knedlo-zelo*)
9. URLs
10. tokenization error where more words should be output instead of one (such as *Brno-Praha*, *po-pá* meaning *Monday-Friday*, *voda-vzduch* meaning *water-air*)
11. gender neutral variants[2], such as *chtěl-a bys potkat někoho zajímavého?* (*do you want to meet someone interesting?*), *obráběč-ka kovů* (*machinist*)

[2] see `https://en.wikipedia.org/wiki/Gender_neutrality_in_languages_with_grammatical_gender`

The category 1 should be easily detected by the morphological analyser, categories 2, 3, and 4 could be detected by a proper gazetteer (list of abbreviations, list of proper names, list of chemical names).

Categories 5 and 6 can be recognized by searching corpora for other languages (and we can expect a high number of occurrences of non-English words in English corpora too).

Category 7 is difficult to recognize and it deserves a deep linguistic research.

Category 8 is somewhat similar to categories 5 and 6 but the difference is that occurrences in English corpora will be rare. Similarly to categories 5 and 6, category 8 will often contain words not recognizable by the morphological analyser. Moreover, some of the MWEs can be found in the MWE list available in NLPC[3] created in 2013.

Category 9 is surprisingly not detected by current tools. Nevertheless, dashes in URLs mean often the same thing as spaces between words. Therefore, we can take dashes in URLs into account.

Category 10 will mostly contain words recognizable by the morphological analyser or found in appropriate gazetteer (such as list of place names).

Category 11 could be recognized as a noun or adjective and an ending in the same case. The word without the dash would make no sense.

In further research, we concentrate on categories 5, 6, and 8. The hypotheses are as follows:

- For categories 5 and 6, we could find many occurrences in an English corpus.
- In Czech corpus, we could find enough occurrences of the one word variant (such as *aupair*) and for multiple words variant (such as *au pair*). This applies for all categories 5, 6, and 8 (and probably also for categories 1, 2, 3, and 4).
- Some components of the words in the categories 5 and 8 are subject of inflection.

4.1 Processing the List of Words with Dashes

We filtered out all compound adjectives, i.e. all compound words that are recognizable as a whole by the morphological analyser `majka` or their components are adverb and adjective recognizable by `majka`. In our data, we found 17,777 such occurrences.

We then filtered out all proper names, abbreviations and chemical names found in our gazetteers:

- Czech surnames (provided by Czech ministry of the interior[4]), 8,647 surnames contain a dash
- Czech first names (provided by Czech ministry of the interior[5]), 10,173 names contain a dash

[3] `/corpora/dicts/mwe/xsmerk_mwe.txt`

[4] `http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx`

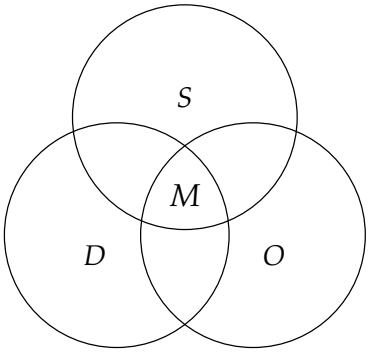
[5] `http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx`

– place names from Geonames[6], 232,231 names contain a dash
– chemical names from Czech Wikipedia (category Chemistry), 28 words contain a dash
– list of abbreviations provided by Seznam.cz, 82 abbreviations contain a dash

Applying the gazetteers reduced the list of words with dashes to 182,237 occurrences. This step should eliminate many members of categories 1 to 4.

Afterwards, we searched for the variants one word (i.e. after removing the dashes) and the variants with spaces (i.e. replacing dashes with spaces). Figure 4.1 shows how many words were in the intersections of all three sets.

It is necessary to mention that not all occurrences are only orthographic variants of a MWE. For example, we can found *A-Beat* (which is a name of a band), *a beat* (which is either a part of an English phrase or even a part of a Czech phrase).



set	# of occurrences
word with dashes D	182,237
$D \cap$ several words S	65,185
$D \cap$ one word O	46,724
MWE candidates $M = D \cap S \cap O$	26,704

Fig. 1. Number of orthographic variants of MWE candidates.

We identified 26,704 MWE candidates, 5,530 of them contain a one-letter word. MWE candidates with one-letter words can be the most problematic part for disambiguation.

4.2 Selection of Multi-Word Expressions

From the intersection of datasets D, S, and O it can be seen that by far not all words with dashes are good candidates for MWEs.

In this phase, we employed the frequencies computed on the corpus `cztenten12`. For performance reasons, we limited the queries to 1,000.

Let s be the number of occurrences as several words (e.g. *a priori*). Let d be the number of occurrences as one word with dashes (e.g. *a-priori*), let o be the number of occurrences as one word (e.g. *apriori*).We then applied the following decision procedure:

[6] `www.geonames.org`

- if $s > 100 \wedge d > 100$ we found a probable MWE
- if $d < 10 \wedge s > 100o$ we found a normal bigram (no MWE)
- if $d = 1000 \wedge s = 1000$ we found a word with both variants (probably the expression is ambiguous)
- if $s > d \wedge s > o$ and all the conditions above did not apply, we found a MWE

The decision process classified 25,091 expressions, 2,140 of them as MWEs, 332 as probable MWEs, and 174 both MWEs and one word.

5 Results

In Tables 5, we present the most frequent MWEs, probable MWEs, and expressions that can be MWEs and one word at the same time, according to our decision procedure.

It can be seen that still some expressions could be filtered out since they are named entities or noun phrases. We did not employ any language analysis which should be the next step. However, we can see that many MWEs were captured by this procedure.

On the other hand, it seems that the condition the word must occur in all three forms is too strict. Many known MWEs that cause problems with annotation have no occurrence as a word with dashes.

6 Discussion

In some cases, several interpretations of a bigram are possible. For example, *to do* is a strong collocation with English origin that can be found separately in Czech texts. At the same time *to do* are two very common Czech words that can appear in a completely Czech sentence such as *Jsou to do víkendu dva dny* (Two days remaining to the end of the week). Similar case is the fixed expression *po o* meaning *after lunch*, only in the context of kindergartens. One can easily find a Czech sentence where *po o* are two subsequent prepositions.

From these two examples, it can be seen that reckless annotation of all occurrences as frozen MWEs would lead to disaster. This work should therefore be seen as preliminary leading to future re-annotation of the Czech corpus. Next step will consist of context-based filtering of MWE candidates.

7 Conclusion and Future work

In this paper, we present a promising corpus-based tool for recognition of fixed multi-word expressions. The complete list of all expressions including their classification is available at `https://nlp.fi.muni.cz/projekty/mwes/mwe_count.txt`. A web demo that decides whether the input is a MWE or not is available at `https://nlp.fi.muni.cz/projekty/mwes/index.py`.

Table 2. Most frequent MWEs and probable MWEs.

MWEs	probable MWEs	both MWE and one word
reality show	IP adresy	play off
last minute	IP adresu	sem tam
IP adresa	MS Windows	jakž takž
pro rodinné	ne já	open source
i té	PET lahví	Jo jo
IP adres	science fiction	kde kdo
plus mínus	PCI Express	pro aktivní
pole position	in vitro	jo jo
křížem krážem	Sebastian Vettel	ne ne
power play	nahoru dolů	Ne ne
LED diody	SD kartu	pop music
raz dva	Rolls Royce	On Line
shora dolů	in situ	Ski areál
já ty	Pentium M	jel a
MS Word	jakous takous	M ČR
Open Source	Kung Fu	r u
SIM kartu	one man	n i
PS PČR	PEN klubu	po užití
set top	one man show	off topic
LED diod	ready made	P ČR
LCD TV	Obchodní dům	R A
Land Rover	LDL cholesterolu	jisto jistě
MS Excel	No Limit	jakého si

A sound evaluation has not been made yet. We plan to compare lists of MWEs output by our tool with other lists, as well as manual annotation.

As the future work, we plan to implement language analysis in the process, namely morphological analysis and probably simplified syntactic analysis.

Finally, we plan to implement this tool into the annotation pipeline and examine if the corpus annotation improves.

Acknowledgments This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2015* (MUNI/A/1165/2014) and by the Ministry of Education of CR within the Czech-Norwegian Research Pogramme in the HaBiT Project 7F14047.

References

1. Bu, F., Zhu, X., Li, M.: Measuring the non-compositionality of multiword expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 116–124. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), `http://dl.acm.org/citation.cfm?id=1873781.1873795`

2. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. 16(1), 22–29 (Mar 1990), http://dl.acm.org/citation.cfm?id=89086.89095
3. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
4. Michelbacher, L.: Multi-word tokenization for natural language processing. Ph.D. thesis, Universität Stuttgart (2013)
5. Pearce, D.: Synonymy in collocation extraction. In: Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations. CMU (2001), http://www.cogs.susx.ac.uk/users/darrenp/academic/dphil/publications/data/Conferences/naacl2001/paper.pdf
6. Rychlý, P.: A lexicographer-friendly association score. In: RASLAN 2008. pp. 6–9. Masarykova Univerzita, Brno (2008)
7. Sag, I., 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
8. Suchomel, V.: Recent czech web corpora. In: Aleš Horák, P.R. (ed.) 6th Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 77–83. Tribun EU, Brno (2012)
9. Šmerk, P.: Unsupervised Learning of Rules for Morphological Disambiguation. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD. Lecture Notes in Computer Science, vol. 3206, pp. 211–216. Springer (2004), http://dblp.uni-trier.de/db/conf/tsd/tsd2004.html#Smerk04
10. Šmerk, P.: K morfologické desambiguaci češtiny [Towards morphological disambiguation of Czech]. thesis proposal, Masaryk University (2008), http://is.muni.cz/th/3880/fi_r/