Czech MWE Database

Karel Pala, Lukáš Svoboda, Pavel Šmerk

Faculty of Informatics
Masaryk University Brno
{pala,xsvobod4,smerk}@fi.muni.cz

Abstract

In this paper we deal with a recently developed large Czech MWE database containing at the moment 160 000 MWEs (treated as lexical units). It was compiled from various resources such as encyclopedias and dictionaries, public databases of proper names and toponyms, collocations obtained from Czech WordNet, lists of botanical and zoological terms and others. We describe the structure of the database and give basic types of MWEs according to domains they belong to. We compare the built MWEs database with the corpus data from Czech National Corpus (approx. 100 mil. tokens) and present results of this comparison in the paper. These MWEs have not been obtained from the corpus since their frequencies in it are rather low. To obtain a more complete list of MWEs we propose and use a technique exploiting the Word Sketch Engine, which allows us to work with statistical parameters such as frequency of MWEs and their components as well as with the salience for the whole MWEs. We also discuss exploitation of the database for working out a more adequate tagging and lemmatization. The final goal is to be able to recognize MWEs in corpus text and lemmatize them as complete lexical units, i. e. to make tagging and lemmatization more adequate.

1. Introduction

Effective collocation searching is one of basic processes in almost all computational text processing. For many applications we need a module that is able to recognize MWEs (collocations) in a raw text and process them appropriately, i.e. as correctly as possible for the purpose of lemmatization, tagging or parsing.

One possibility how to handle (recognize) MWEs is to use statistical techniques. We use them when it makes sense but for some tasks like semantic categorization they offer rather approximate solutions. The second way is to compile a large MWE database (partly automatically) and then classify its items semantically. In our view, a reliable semantic classification of the MWEs can hardly be achieved with statistical techniques only. In this sense we prefer here rather rule-based techniques.

In this paper we present large MWE database recently built for Czech that at the moment contains more than 160 000 items. It was compiled from various resources such as encyclopedia headwords categorized by their meaning, public databases of the toponyms, collocations obtained from Czech WordNet (Pala and Smrž, 2004), lists of abbreviations, names of institutions and others. Thus we have built a database of MWEs first and now we use it for recognizing MWEs appearing in an input (corpus) text.

Another possibility is to try to build a MWE database using Word Sketch Engine (Kilgarriff et al., 2004) and exploit statistical parameters such as MI- or T-score, salience and others. The obtained results have to be cleaned and corrected manually but we are looking for the rules that would make the task easier.

2. Motivation

One good reason for having MWE database and recognizing MWEs in texts is to achieve more adequate tagging and lemmatizing. According to our experience it is rather difficult to judge the quality of tagging the corpus texts if we do not know how correctly the MWEs are tagged (if they are tagged at all). If we are told that a tagger's precision is 96 % and nothing is said (typically) about handling the MWEs in corpus text this 'good' result is rather disputable.

Many examples from Czech grammatically tagged corpus SYN2000 (Čermák et al., 2000) can be adduced that demonstrate how problematic is the existing tagging in some relevant aspects. Many (if not all) collocations containing proper names like *Václav Havel* are tagged as two independent items in the corpus SYN2000 though it is one whole representing a proper name. The frequency of this collocation in SYN2000 is 6565 (from 100 mil. positions). The same case of this sort is proper name *Václav Klaus* with frequency 7206 in SYN2000. It is lemmatized as two units with two distinct grammatical tags. Toponym *Nové Město na Moravě (New Town at Moravia)* displays frequency 65 but it is lemmatized as four lexical units with four tags, respectively. Such tagging and lemmatizing offers a distorted picture of the language (we will say more about this below) and the authors of the taggers used are usually silent about this state. In this case it would be useful to speak about a 'conventional' way of tagging which does not reflect structure of the language adequately.

The second good reason are many NLP applications in which, however, the results of the above mentioned tagging are used without any further explanation. In fact, the Czech MWE database was originally built for a NLP application representing the QA system UIO (Svoboda, 2004) but obviously many other NLP applications require more adequate MWEs recognition as well.

3. Description of the Czech MWE database

We define MWEs as sequences of words representing one lexical unit (semantic whole) or as Sag et al. (2002) say idiosyncratic interpretations that cross word boundaries. As we indicated above the information about domains, i.e. a semantic categorization and tagging of MWEs, is useful both for morphological and sense disambiguation and (partial) parsing.

3.1. Types of MWEs

Here we present the MWE types that were originally introduced in the Czech MWE database, however, when we started working with the corpus data the changes in the classification of types appeared necessary:

- toponyms — names of the states, cities, streets, railway and coach stations, etc., — *San Francisco, Kurilské ostrovy (Kuril Islands), Nové Město na Moravě (New Town at Moravia), Brno hl. n. (Brno, Main Station)*,
- names of the famous people — *Albert Einstein, Karel IV, Tomáš Garrigue Masaryk, Václav Havel*,
- names of culture events or institutions such as movies, cinemas, galeries, TV stations, radio stations — *Moravská galerie (Moravian Gallery), ČT 1 (Channel 1 of Czech Television), Matrix Reloaded, Radio Country, Obsluhoval jsem anglického krále (I served to English King)*,
- botanical and zoological terms (*chrpa polní (bluebottle), kočka domácí (true cat)*),
- other smaller domains such as currencies (*americký dolar (US dollar)*), physical, mathematical and other units (*námořní míle (imperial nautical mile), metrický cent (quintal), čtvereční metr (square meter), krychlový metr (cubic meter)*),
- unsorted idiomatic collocations (expressions) — obtained from synsets in the Czech WordNet, e.g. *vysoká škola (university), lovecký pes (hunting dog)* and other resources such as the Dictionary of Czech Phraseologisms and Idioms (Čermák et al., 1994), Wikipedia, encyclopedia Diderot and varia.

It has to be remarked that most of the given types of MWEs can be also characterized as named entities, however, we are treating them here as collocate expressions only. There is an obvious overlap between the two but we are not convinced that we would gain more here if we try to examine them from the N-E side. Of course, more detailed semantic classification obtained in this way can be helpful for tagging, lemmatizing and parsing.

3.2. MWE database — basic statistics

In the Table 1 given below we offer a basic statistics showing the types of MWEs and their numbers in the Czech MWE database. The types mentioned above are slightly different in some cases but the classification remains basically the same. The Table 2 shows for each domain three most frequent MWEs in SYN2000 corpus.
The Table 1 makes it possible to compare the MWE data from the database with the data obtained from the Czech National Corpus (CNC, its version SYN2000).

- the column **# MWEs** contains a number of the MWEs from a given domain in our MWE database.
- in the column **# Occs** we find a number of MWE occurences from a given domain in the CNC.
- **# Unique** is a number of the individual MWEs from a given domain which occur in the CNC at least once.
- **% of all** is a percent of MWEs occuring in the CNC in comparison with all MWEs from a given domain.
- **# HL** denotes "hapax legomena", i.e. MWEs with only one occurrence in the CNC.
- **# not in corpus** is a number of the MWEs, which did not occur in the CNC.

Average length of the found MWEs is 2.12 word, i.e. found MWE occurrences cover 3.25 % of the Syn2000 (containing approx. 100 M words). It is possible that some found MWE occurrences may not, in fact, be MWEs but only the words appearing in the syntactic configurations randomly, but we have observed only very few examples of it and therefore we abstract from this possibility for the moment. We also assume that the found MWEs are not overlapping. The numbers in the Table 1 show that the intersection between the MWEs obtained from CNC and MWEs in our database represents 3,25 %, i.e. some MWEs from the Czech database are not frequent in CNC (80,5 % is a complement to the % of all). This is caused by the fact that the MWE database was originally designed for a QA system. Pragmatically speaking, this means that there is a sort of conflict between the data from corpus and from the MWE database because database data are obtained from the text that do not exist in corpora. In our view, it still makes sense to have MWE database as large as possible.

Botanics, zoology	
vlašský ořech *(wallnut)*	148
německý ovčák *(Alsatian)*	142
réva vinná *(grapevine)*	87
Culture	
Snídaně s Novou *(name of a teleview)*	534
Městské divadlo *(name of a theatre)*	508
Stavovské divadlo *(name of a theatre)*	471
Toponyms	
Hradec Králové *(city)*	19501
Ústí nad Labem *(city)*	3945
České Budějovice *(city)*	3879
Proper names (people)	
Václav Klaus *(current president of CR)*	11638
Václav Havel *(previous president of CR)*	10409
Miloš Zeman *(previous premier of CR)*	4252
Unsorted and smaller domains	
Česká republika *(Czech Republic)*	29107
životní prostředí *((living) environment)*	10314
Evropská unie *(European Union)*	9927

Table 2: Three Most Frequent MWEs for Each Domain

3.3. MWE recognition

In Czech we have to deal with rich inflection so the full morphological analysis has to be performed. First, the morphological analyzer ajka (Sedláček and Smrž, 2001) is used

Domain	# MWEs	# Occs	# Unique	(% of all)	# HL	# not in corpus
Botanics, zoology	48608	10472	2820	(5.9)	1280	45248
Culture	27278	72190	3425	(12.6)	708	23853
Toponyms	21471	178726	4126	(19.2)	1164	17345
Proper names (people)	19190	197565	6758	(35.2)	1387	12342
Unsorted and smaller domains	48632	1076065	15057	(31.0)	2302	33575
Total	164639	1535018	32186	(19.5)	6841	132453

Table 1: Statistics of Czech MWE Database

for this purpose and input text is tagged. The words not recognized by ajka are handled by a simple guesser which determines a lemma and tag as well. The result is that every word in text now has its lemma. At this moment we start searching MWE database for sequences of lemmata that represent the particular MWEs. The found candidates are checked for different properties, e. g. whether they display the particular values of the case or number where it is necessary, etc.

4. Acquiring MWEs using Word Sketch Engine

Compiling MWE database from the various lists is one way how to obtain some types of MWEs with a relatively little effort. However, there are other MWE types that are difficult to recognize and they cannot be found in the lists of proper names or toponyms etc. This becomes obvious when we try to find MWEs from the indicated lists in corpora, even in the big ones. If we find them at all they display very low frequencies. On the other hand it can be observed that there are other types of MWEs, quite frequent in corpora, for instance *vzhledem k (with regard to)*, which usually do not occur in the lists mentioned above. If we want to find them we have to turn to corpora and use statistical techniques – one of them enables us to explore expressions as they occur in their typical contexts. This statistically based tool is called Word Sketch Engine (Kilgarriff etal., 2004) and we have used it for obtaining candidates of MWEs from corpora.

4.1. Word Sketch Engine

Word sketches are one-page automatic, corpus-based tables capturing the word's grammatical and collocational behaviour. The Sketch Engine then is a corpus tool which takes as an input a tagged corpus text of any language and corresponding grammar patterns and generates word sketches for the words of that language. It also generates a thesaurus (semantic cluster) and sketch differences, which specify similarities and differences between near-synonyms.

The word sketch also provides a well-founded salience statistics for collocations. The necessary condition is a grammatically tagged and lemmatized corpus and grammar patterns. Rather than looking at an arbitrary window of text around the headword, we look, in turn, for each grammatical relation that the word participates in. For English about 27 grammatical relations are used, for Czech 23 ones. The word sketch then provides one list of collocates for each grammatical relation the word participates in. For a verb, the subject, the objects, the conjoined verbs (drink and smoke), modifying adverbs, prepositions and prepositional objects, are all presented in different lists.

An example of the word sketch table is presented in Figure 1 and it shows what WSE offers for the noun *school*. The underlined numbers denote frequencies in the British National Corpus, clicking on them a user gets the respective concordancy lists. The non-underlined number represents a respective value of the salience parameter computed for the individual contexts. If we have a look at the columns with the respective labels (modifier, object of, subject of, etc.) it can be seen that approx. 90 % bigrams in the Figure 1 are either MWEs or their very good candidates. The examples from the word sketch table demonstrate this very convincingly, take, for instance, bigrams such as *public school*, *grammar school*, *high school* or *medical school*.

For the purpose of further analysis we have generated the list of bigrams and n-grams from the whole Czech National Corpus (Čermák et al., 2000) using the WS Engine. The number of bigrams with frequency higher than one hundred is 27362, number of n-grams (n>2) with frequency higher than 100 is 20148. The basic evaluation of all bigrams shows that from the first 300 bigrams (according to the salience) only 8 cannot be accepted as MWEs (2,6 %). 130 MWEs from the rest are already in our MWE database and the remaining 162 are new with 143084 occurrencies in corpus SYN2000. Similarly, from the first 100 tri/tetragrams we get 63 MWEs. Only 7 are already in our MWE database and 56 are new and represent 148735 occurrencies in corpus. Together with bigrams it is more than 0.7 % of the corpus text (counts are multiplied by three/four or two respectively, as we count tri/tetragrams or bigrams). It can be also seen that with all n-grams when salience decreases more noise can be observed. That is why it is necessary to look for some rules that would make the task easier - we already have performed some experiments with the rules based on POS structures which are, in fact, included in WSE. This can also reduce a considerable amout of the manual work necessary for cleaning the lists. In our view these numbers represent a good starting point for a more detailed evaluation which will come in the next phase.

5. The role of MWEs in tagging and lemmatization

If we want to search a corpus for some word forms, either alone or in contexts, typically we would not be interested in their occurrence in a MWE because such con-

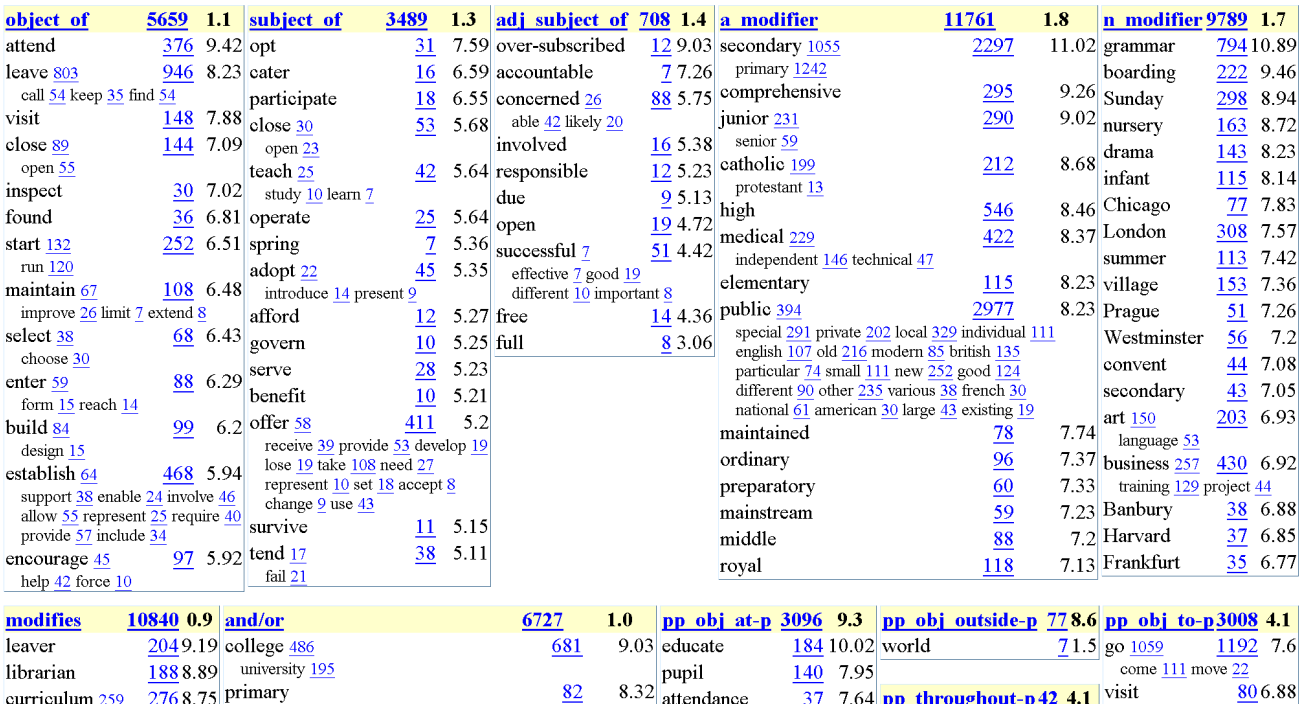

Figure 1: Word Sketch of English noun 'school'

texts do not yield any useful information about syntactic, semantic or pragmatic behaviour of those word forms in language. For instance, occurrences of the word forms *lesy (forests)*, *černými (black)* or *Moravě (Moravia)* in the toponyms *Kostelec nad Černými lesy (Kostelec on Black Forests)* or *Nové Město na Moravě (New Town at Moravia)* are not relevant with regard to the most of the queries containing words *černý (black)* or *Morava (Moravia)*.

From this point of view it is then useful to mark each token to indicate whether it is or is not a part of a MWE so a user can query only positions in corpus that are not parts of a MWE (or specify that the searched position can be a part of a MWE, as a default we may assume that user is not interested in such contexts).

Similarly, if a corpus user wants to search for an evidence related to syntactic relations the occurrences of the unmarked MWEs are problematic since they violate sentence structure and consequently user's expectations. In some cases MWEs are not formed according to the actual syntactic rules but they are 'inherited' from the past (toponyms) or they can be completely agramatical (e.g. names of movies or books).

In such cases it is not enough to mark positions that are parts of a MWE but it is also necessary for a corpus manager to be able to represent the whole MWE as one position with a morphological tag capturing properties of the head of the phrase which represents MWE in the sentence. Of course, the possibility of querying particular positions in a MWE has to be retained, thus consequently, at the same time the corpus manager has to allow for representing the same corpus as having more levels where the particular levels differ in number of tokens (positions). The corpus manager Manatee/Bonito developed at NLP Centre FI MU (Rychlý, 2000) allows for this.

Exactly the same, i.e. treating whole MWEs as one token (position) with an appropriate grammatical tag is very useful for morphological disambiguation (Šmerk, 2007) and partial parsing (Žáčková, 2002) algorithms because not recognized MWEs complicate or even disrupt a sentence structure.

Since we have the large Czech MWE database we can solve all the above mentioned tasks in the following way: from the possible MWEs corresponding to a certain sequence of the word forms in the input text we always choose the longest one. Obviously, it is possible to construct artificial counterexamples for which this approach will not give the correct results but we have not observed such cases in the real texts so far. Also overlapping of MWEs is theoretically possible but we have not found convincing enough evidence that would confirm it.

6. Conclusions

We describe a recently prepared Czech database containing more than 160 000 MWEs (treated as lexical units). According to our results the obtained MWEs cover approx. 3.25 % of the CNC (100 M) which appears to be a new result for Czech.

To obtain a more reliable and complete list of MWEs we have proposed and used a technique exploiting the Word Sketch Engine, which allows us to work with the frequency of MWEs and their components and also to exploit statistical parameters such as salience or MI-score for finding MWEs as whole units. The list of bigrams and n-grams obtained via Word Sketch Engine was analyzed and compared with the MWE database mentioned above. In our view presented data and their analysis also show that corpora larger than 100 mil. are necessary, obviously the frequency of some MWEs is either low or they are hapax

legomena. This can be also observed when Word Sketch Engine is exploited – there are many contexts (bigrams) whose frequencies are low though they can be considered as very good candidates of MWEs. The development goes in this direction – there are already corpora containing more than one billion tokens (cf. OEC, itWAC, deWAC, see e.g. http://www.lexmasterclass.com/), and a compilation of a Czech corpus with similar size is presently going on as well. The issues discussed in the paper also lead to the conclusion that it is inevitable to work out a more adequate tagging, lemmatization and parsing. As we argue in the paper this cannot be done without MWEs or, more concretely, without the database containing them.

7. Acknowledgements

This work has been partly supported by the Academy of Sciences of Czech Republic under the project 1ET200610406, by the Ministry of Education of CR within the Center of Computational Linguistics LC536, and in the National Research Programme II project 2C06009.

8. References

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex*, pages 105–116.

Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*.

Pavel Rychlý. 2000. *Corpus Managers and Their Effective Implementation*. Ph.D. thesis, Masaryk University, Brno, Czech Republic.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Anne Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*.

Radek Sedláček and Pavel Smrž. 2001. A New Czech Morphological Analyser ajka. *Proceedings of the 4th International Conference Text, Speech and Dialogue*.

Lukáš Svoboda. 2004. Processing of Natural Language Multiword Expressions. *Proceedings of Znalosti*.

František Čermák et al. 1994. *Slovník české frazeologie a idiomatiky*. Academia, Prague, Czech Republic.

František Čermák et al. 2000. *The Czech National Corpus — SYN2000*. The Institute of the Czech National Corpus, Charles University, Prague, Czech Republic.

Pavel Šmerk. 2007. *Towards morphological disambiguation of Czech*. Ph.D. thesis proposals, Masaryk University, Brno, Czech Republic (in Czech).

Eva Žáčková. 2002. *Partial syntactic analysis of Czech*. Ph.D. thesis, Masaryk University, Brno, Czech Republic (in Czech).