Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures

Eduard Bejček, Pavel Straňák, Pavel Pecina
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha 1, Czechia
{bejcek,stranak,pecina}@ufal.mff.cuni.cz

Abstract

We deal with syntactic identification of occurrences of multiword expression (MWE) from an existing dictionary in a text corpus. The MWEs we identify can be of arbitrary length and can be interrupted in the surface sentence. We analyse and compare three approaches based on linguistic analysis at a varying level, ranging from surface word order to deep syntax. The evaluation is conducted using two corpora: the Prague Dependency Treebank and Czech National Corpus. We use the dictionary of multiword expressions SemLex, that was compiled by annotating the Prague Dependency Treebank and includes deep syntactic dependency trees of all MWEs.

1 Introduction

Multiword expressions (MWEs) exist on the interface of syntax, semantics, and lexicon, yet they are almost completely absent from major syntactic theories and semantic formalisms. They also have interesting morphological properties and for all these reasons, they are important, but challenging for Natural Language Processing (NLP). Recent advances show that taking MWEs into account can improve NLP tasks such as dependency parsing (Nivre and Nilsson, 2004; Eryiğit et al., 2011), constituency parsing (Arun and Keller, 2005), text generation (Hogan et al., 2007), or machine translation (Carpuat and Diab, 2010).

The Prague Dependency Treebank (PDT) of Czech and the associated lexicon of MWEs SemLex[1] offer a unique opportunity for experimentation with MWEs. In this paper, we focus on identification of their syntactic structures in the treebank using various levels of linguistic analysis and matching algorithms.[2] We compare approaches operating on manually and automatically annotated data with various depth of annotation from two sources: the Prague Dependency Treebank and Czech National Corpus (CNC).

The remainder of the paper is organised as follows. Section 2 describes the state of the art of in acquisition and identification of MWEs. Section 3 explains what we consider a MWE. In Section 4 we describe the data used for our experiments. Section 5 gives the details of our experiments, and in Section 6 we analyse and discuss the results. Conclusions from the analysis are drawn in Section 7.

2 Processing of Multiword Expressions and Related Work

Automatic processing of multiword expressions includes two distinct (but interlinked) tasks. Most of the effort has been put into **acquisition** of MWEs appearing in a particular text corpus into a lexicon of MWEs (types) not necessarily linked with their occurrences (instances) in the text. The best-performing methods are usually based on lexical association measures that exploit statistical evidence of word occurrences and co-occurrences acquired from a corpus to determine degree of lexical association between words (Pecina, 2005). Expressions that consist of words with high association are then

[1] http://ufal.mff.cuni.cz/lexemann/mwe/semlex.zip

[2] We do not aim at disambiguating the occurrences as figurative or literal. We have not observed enough literal uses to substantiate working on this step. There are bigger improvements to be gained from better identification of syntactic occurrences.

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 106–115,
Atlanta, Georgia, 13-14 June 2013.

denoted as MWEs. Most of the current approaches are limited to bigrams despite the fact that higher-order MWEs are quite common.

The task of **identification** of MWE occurrences expects a list of MWEs as the input and identifies their occurrences (instances) in a corpus. This may seem to be a trivial problem. However, the complex nature of this phenomenon gives rise to problems on all linguistic levels of analysis: morphology, syntax, and semantics.

In *morphologically* complex languages, a single MWE can appear in a number of morphological variants, which differ in forms of their individual components; and at the same time, a sequence of words whose base forms match with base forms of components of a given MWE do not necessarily represent an instance of this MWE (*Pracoval dnem i nocí / He's been working day and night* vs. *Ti dva byli jako den a noc / Those two were as day and night*).

MWEs differ in the level of *syntactic* fixedness. On the one hand, certain MWEs can be modified by inserting words in between their components or by changing word order. Such expressions can only be identified by matching their syntactic structures, but only if a reliable syntactic information is available in both the lexion and text (*Po převratu padaly hlavy / After the coup, heads were rolling* vs. *Hlavy zkorumpovaných náměstků budou padat jedna za druhou / One head of a corrupt deputy will be rolling after the other*). On the other hand, some MWEs can appear only as fixed expressions with no modifications allowed. In that case, the syntactic matching approach can miss-indicate their instances because of an inserted word or altered word order (*Vyšší společnost / High society* vs. **Vyšší bohatší společnost / High rich society*).

From the *semantic* point of view, MWEs are often characterized by more or less non-compositional (figurative) meaning. Their components, however, can also occur with the same syntax but compositional (literal) semantics, and therefore not acting as MWEs (*Jedinou branku dal až v poslední minutě zápasu / He scored his only goal in the last minute of the match.* vs. *Rozhodčí dal branku zpět na své místo / The referee put a goal back to its place*).

Automatic discrimination between figurative and literal meaning is a challenging task similar to word sense disambiguation which has been studied extensively: Katz and Giesbrecht (2006), Cook et al. (2007), Hashimoto and Kawahara (2008), Li and Sporleder (2009), and Fothergill and Baldwin (2011). Seretan (2010) includes MWE identification (based on a lexicon) in a syntactic parser and reports an improvement of parsing quality. As a by-product, the parser identified occurrences of MWEs from a lexicon. Similarly, Green et al. (2013) embed identification of some MWEs in a Tree Substitution Grammar and achieve improvement both in parsing quality and MWE identification effectiveness. None of these works, however, attempt to identify all MWEs, regardless their length or complexity, which is the main goal of this paper.

3 Definition of Multiword Expressions

We can use the rough definition of MWEs put forward by Sag et al. (2002): *"idiosyncratic interpretations that cross word boundaries (or spaces)"*. We can also start from their – or Bauer's (1983) – basic classification of MWEs as *lexicalised* or *institutionalised phrases*, where lexicalised phrases include some syntactic, semantic or lexical (i.e. word form) element, that is idiosyncratic. Institutionalised phrases are syntactically and semantically compositional, but still require a particular lexical choice, e.g. disallowing synonyms (mobile phone, but not *movable phone).

We need to make just one small adjustment to the above: "phrase" above must be understood as a subtree, i.e. it can have holes in the surface sentence, but not in terms of a dependency tree.

In reality there is no clear boundary, in particular between the institutional phrases and other collocations. Like many other traditional linguistic categories, cf. Manning (2003), this phenomenon seems to be more continuous than categorial.

For the purpose of this paper, however, it is not important at all. We simply try to find all instances of the expressions (subtrees) from a lexicon in a text, whatever form the expression may take in a sentence.

4 Data

In this work we use two datasets: Czech National Corpus (CNC), version SYN2006-PUB, and the

Prague Dependency Treebank (PDT), version 2.5. We run and compare results of our experiments on both manual annotation of PDT, and automatic analysis of both PDT and CNC (see Section 5.3). We also make use of SemLex, a lexicon of MWEs in the PDT featuring their dependency structures that is described in Section 4.3.

4.1 Corpora – Czech National Corpus and Prague Dependency Treebank

CNC is a large[3] corpus of Czech. Its released versions are automatically segmented and they contain automatic morphological tagging (Hajič, 2004).

PDT (Bejček et al., 2011) is a smaller news-domain corpus based on a subset of the news section of CNC. It contains approx. 0.8 million words that have three layers of annotation: morphological, analytical (surface syntax), and tectogrammatical (deep syntax).

Annotation of a sentence on the *morphological layer* consists of attaching morphological lemma and tag to the tokens. A sentence at the *analytical layer* is represented as a rooted ordered tree with labelled nodes. The dependency relation between two nodes is captured by an edge with a functional label. On the *tectogrammatical layer* only content words form nodes in a tree (t-nodes).[4] Auxiliary words are represented by various attributes of t-nodes, as they do not have their own lexical meaning, but rather modify the meaning of the content words. Each t-node has a t-lemma: an attribute whose value is the node's basic lexical form, and a dependency function that relates it to its parent. Figure 1 shows the relations between the neighbouring layers of PDT.

4.2 MWE in Prague Dependency Treebank 2.5

In the Functional Generative Description (Sgall et al., 1986, FGD)[5] the tectogrammatical layer is construed as a *layer of the linguistic meaning* of text. This meaning is composed by means of "deep" (tecto-grammatical) syntax from single-meaning-carrying units: monosemic lexemes.

[3]It contains 200 mil. words in SYN2000, 600 mil. in SYN2006-PUB; http://www.korpus.cz.

[4]with a few exceptions (personal pronouns or coord. heads)

[5]FGD is a framework for systematic description of a language, that the PDT project is based upon.

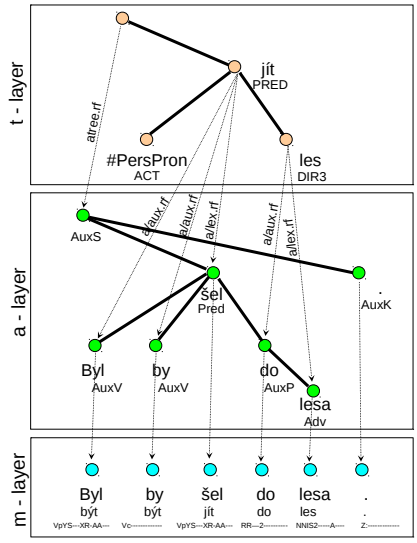

Figure 1: A visualisation of the annotation schema of PDT. Lit.: "[He] would have gone into forest."

In order to better facilitate this concept of t-layer, all multiword expressions in the release of PDT 2.5 (Bejček et al., 2011) have been annotated and they are by default displayed as single units, although their inner structure is still retained.

A lexicon of the MWEs has been compiled. A simple view of the result of this annotation is given in the Figure 2. A detailed description can be found in Bejček and Straňák (2010), and Straňák (2010). The MWEs in PDT 2.5 include both multiword lexemes (phrasemes, idioms) and named entities (NEs). In the present work we ignore the named entities, concentrating on the lexemes. Some NEs (names of persons, geographical entities) share characteristics of multiword lexemes, other NEs do not (addresses, bibliographic information).

We build on the PDT 2.5 data and MWE lexicon SemLex (Section 4.3) to evaluate the approach with various automatic methods for detection of MWEs.

4.3 Lexicon of MWEs – SemLex

SemLex is the lexicon of all the MWEs annotators identified during the preparation of PDT 2.5 t-layer. In the PDT 2.5 these instances of MWEs can then be displayed as single nodes and all the MWEs themselves are compiled in the SemLex lexicon. The lexicon itself is freely available. See `http://ufal.mff.cuni.cz/lexemann/mwe/`. Length (size)

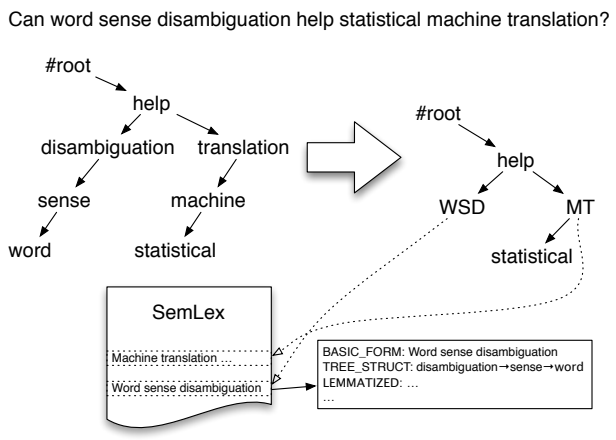

Figure 2: An illustration of changes in t-trees in PDT 2.5; every MWE forms a single node and has its lexicon entry

a) len	types	instances
2	7063	18914
3	1260	2449
4	305	448
5	100	141
6	42	42
7	16	15
8	4	5
9	4	3
11	1	0
12	2	2

b) len	types	instances
1[8]	148	534
2	7444	19490
3	843	1407
4	162	244
5	34	32
6	13	8
7	3	1
8	4	1
9	1	1
10	0	0

Table 1: Distribution of MWE length in terms of words (a) and t-nodes (b) in SemLex (types) and PDT (instances).

distribution of MWEs in PDT 2.5 is given in Table 1.

There are three attributes of SemLex entries crucial for our task:

BASIC_FORM – The basic form of a MWE. In many languages including Czech it often contains word forms in other than the basic form for the given word on its own. E.g. "vysoké učení" contains a neuter suffix of the adjective "vysoký" (high) because of the required agreement in gender with the noun, whereas the traditional lemma of adjectives in Czech is in the masculine form.

LEMMATIZED – "Lemmatised `BASIC_FORM`", i.e. take the basic form of an entry and substitute each form with its morphological lemma. This attribute is used for the identification of MWEs on the morphological layer. For more details see Section 5.

TREE_STRUCT (TS) – A simplified tectogrammatical dependency tree structure of an entry. Each node in this tree structure has only two attributes: its tectogrammatical lemma, and a reference to its effective parent.

4.4 Enhancing SemLex for the Experiments

SemLex contains all the information we use for the identification of MWEs on t-layer.[6] It also contains basic information we use for MWE identification on m-layer: the *basic form* and the *lemmatized form* of each entry. For the experiments with MWE identification on analytical (surface syntactic) layer we need to add some information about the surface syntactic structures of MWEs. Given the annotated occurrences of MWEs in the t-layer and links from t-layer to a-layer, the extraction is straightforward. Since one tectogrammatical TS can correspond to several analytical TSs that contain auxiliaries and use morphological lemmas, we add a list of a-layer TSs with their frequency in data to each SemLex entry (MWE). In reality the difference between t-layer and a-layer is unfortunately not as big as one could expect. Lemmas of t-nodes still often include even minute morphological variants, which goes against the vision of tectogrammatics, as described in Sgall et al. (1986).[7] Our methods would benefit from more unified t-lemmas, see also Section 6.2.

5 Methodology of Experiments

SemLex – with its almost 8,000 types of MWEs and their 22,000 instances identified in PDT – allows us to measure accuracy of MWE identification on various layers, since it is linked with the different layers of PDT 2.5. In this section, we present the method for identification of MWEs on t-layer in comparison with identification on a-layer and m-layer. The

[6] Automatic identification of MWES was, after all, one of the reasons for its construction.

[7] These variants are unified in FGD theory, but time consuming to annotate in practice. Therefore, this aspect was left out from the current version of PDT.

[8] Indeed, there are expressions that are multiword, but "single-node". E.g.: the preposition in *bez váhání* (without hesitation) does not have its own node on t-layer; the phrase *na správnou míru* (lit.: into correct scale) is already annotated as one phrasal node in PDT with the lemma "na_správnou_míru"; the verbal expression *umět si představit* (can imagine) has again only one node for reflexive verb "představit_si" plus an attribute for the ability (representing "umět" as explained in Section 4.1).

idea of using tectogrammatical TS for identification is that with a proper tectogrammatical layer (as it is proposed in FGD, i.e. with correct lemmatisation, added nodes in place of ellipses, etc.), this approach should have the highest Precision.

Our approach to identification of MWEs in this work is purely syntactic. We simply try to find MWEs from a lexicon in any form they may take (including partial ellipses in coordination, etc.). We do not try to exploit semantics, instead we want to put a solid baseline for future work which may do so, as mentioned in Section 2.

5.1 MWE Identification on t-layer

We assume that each occurrence of a given MWE has the same t-lemmas and the same t-layer structure anywhere in the text. During the manual construction of SemLex, these tectogrammatical *"tree structures"* (TSs) were extracted from PDT 2.5 and inserted into the lexicon. In general this approach works fine and for majority of MWEs only one TS was obtained. For the MWEs with more than one TS in data we used the most frequent one. These cases are due to some problems of t-layer, not deficiencies of the theoretical approach. See section 6.2 for the discussion of the problems.

These TSs are taken one by one and we try to find them in the tectogrammatical structures of the input sentences. Input files are processed in parallel. The criteria for matching are so far only t-lemmas and topology of the subtree.[9] Comparison of tree structures is done from the deepest node and we consider only perfect matches of structure and t-lemmata.

5.2 MWE Identification on a-layer and m-layer

We use identification of MWE occurrences on a-layer and m-layer mainly for comparison with our approach based on the t-layer.

[9]It is not sufficient, though. Auxiliary words that are ignored on t-layer are occasionally necessary for distinguishing MWE from similar group of nodes. (E.g. *"v tomto směru"* (*"in this regard"*) is an MWE whereas *"o tomto směru"* (*"about this direction"*) is not.) There are also attributes in t-layer that are—although rarely—important for distinguishing the meaning. (E.g. words typeset in bold in *"Leonardo **dal** svým **gólem** signál."* (*"Leonardo signalled by his goal."*) compose exactly the same structure as in *"Leonardo **dal gól**."* (*"Leonardo scored a goal."*). I.e., the dependency relation is "*dal* governs *gól*" in both cases. The difference is in the dependency function of *gól*: it is either MEANS or DIRECT_OBJECT (CPHR).)

We enhance SemLex with a-tree structures as explained in Section 4.4, and then **a-layer** is processed in the same manner as t-layer: analytical TS is taken from the SemLex and the algorithm tries to match it to all a-trees. Again, if more than one TS is offered in lexicon, only the most frequent one is used for searching.

MWE identification on the **m-layer** is based on matching lemmas (which is the only morphological information we use). The process is parametrised by a width of a window which restricts the maximum distance (in a sentence) of MWE components to span (irrespective of their order) measured in the surface word order. However, in the setting which does not miss any MWE in a sentence (100% Recall), this parameter is set to the whole sentence and the maximum distance is not restricted at all.

The algorithm processes each sentence at a time, and tries to find all lemmas the MWE consists of, running in a cycle over all MWEs in SemLex. This method naturally over-generates – it correctly finds all MWEs that have all their words present in the surface sentence with correct lemmatisation (high Recall), but it also marks words as parts of some MWE even if they appear at the opposite ends of the sentence by complete coincidence (false positives, low Precision).

In other experiments, the window width varies from two to ten and MWE is searched for within a limited context.

5.3 Automatic Analysis of Data Sets

The three MWE identification methods are applied on three corpora:

- **manually annotated PDT:** This is the same data, from which the lexicon was created. Results evaluated on the same data can be seen only as numbers representing the maximum that can be obtained.
- **automatically annotated PDT:** These are the same texts (PDT), but their analysis (morphological, analytical as well as tectogrammatical) started from scratch. Results can be still biased – first, there are no new lexemes that did not appear during annotation (that is as if we had a complete lexicon); second, it should be evaluated only on eval part of the data – see discussion in Section 6.1.
- **automatically annotated CNC:** Automatic analysis from scratch on different sentences. The

layer/span	PDT/man	PDT/auto	CNC/auto
tecto	61.99 / 95.95 / 75.32	63.40 / 86.32 / 73.11	44.44 / 58.00 / 50.33
analytical	66.11 / 88.67 / 75.75	66.09 / 81.96 / 73.18	45.22 / 60.00 / 51.58
morpho / 2	67.76 / 79.96 / 73.36	67.77 / 79.26 / 73.07	51.85 / 56.00 / 53.85
3	62.65 / 90.50 / 74.05	62.73 / 89.80 / 73.86	46.99 / 60.00 / 52.70
4	58.84 / 92.03 / 71.78	58.97 / 91.29 / 71.65	42.83 / 61.33 / 50.48
5	56.46 / 92.94 / 70.25	56.59 / 92.16 / 70.12	40.09 / 61.33 / 48.49
6	54.40 / 93.29 / 68.81	54.64 / 92.51 / 68.70	38.27 / 61.33 / 47.13
7	52.85 / 93.42 / 67.51	53.01 / 92.64 / 67.43	36.99 / 61.33 / 46.15
8	51.39 / 93.46 / 66.32	51.57 / 92.68 / 66.27	35.59 / 61.33 / 45.04
9	50.00 / 93.46 / 65.15	50.18 / 92.68 / 65.11	34.67 / 61.33 / 44.30
10	48.57 / 93.46 / 63.92	48.71 / 92.68 / 63.86	33.84 / 61.33 / 43.64
∞	35.12 / 93.51 / 51.06	35.16 / 92.72 / 50.99	22.70 / 62.00 / 33.24
	P / R / F	P / R / F	P / R / F

Table 2: Evaluation of all our experiments in terms of Precision (P), Recall (R) and F_1 score (F) in percent. Experiments on the m-layer are shown for different widths of window (see Section 5.2).

disadvantage here is the absence of gold data. Manual evaluation of results has to be accomplished.

For the automatic analysis we use the modular NLP workflow system Treex (Popel and Žabokrtský, 2010). Both datasets were analysed by the standard Treex scenario "Analysis of Czech" that includes the following major blocks:

1) standard rule-based Treex segmentation and tokenisation
2) morphology (Hajič, 2004) and Featurama tagger (Spousta, 2011) trained on the *train* part of the PDT
3) MST Parser with an improved set of features by Novák and Žabokrtský (2007)
4) and t-trees structure provided by standard rule-based Treex block.

6 Results

Effectiveness of our methods of identification of MWE occurrences is presented in Table 2. Numbers are given as percentages of Precision and Recall The first two columns show the results of the evaluation against gold data in PDT 2.5, the third column reflects the manual evaluation on 546 sentences. The results obtained for PDT (the first two columns) are also visualised in Figure 3.

The important issue to be decided when evaluating MWE identification is whether partial match between automatic identification and gold data MWE is to be counted. Because of cases containing ellipses (see Section 6.2), it can happen that longer MWE is used for annotation of its subset in text.[10] We do not want to penalise automatic identification (either performing this behaviour or confronted with it in the gold data), so we treated subset as a match.

Another decision is that although the MWEs cannot be nested in gold data, we accept it for automatic identification. Since one word can belong to several MWEs, the Recall rises, while Precision declines.[11]

6.1 Discussion of Results

The automatically parsed part of the CNC consists of 546 sentences. Thus the third column in Table 2 represents evaluation on a much smaller data set. During manual annotation of this data carried out by one annotator (different from those who annotated PDT data, but using the same methodology and a tool), 163 occurences of MWEs were found. Out

[10] Let us say, only elliptic term *Ministry of Industry* is seen in the data (instead of the full name *Ministry of Industry and Trade*) annotated by the full-term lexicon entry. Whenever *Ministry of Industry and Trade* is spotted in the test data, its first part is identified. Should that be qualified as a mistake when confronted with the gold annotation of the whole term? The assigned lexicon entry is the same – only the extent is different.

[11] For example, annotator had to choose only one MWE to annotate in *vládní návrh zákona o dani z příjmu* (lit.: government proposal of the Law on Income Tax), while it is allowed to automatically identify *vládní návrh zákona*, *zákon o dani* and *daň z příjmu* together with the whole phrase. Recall for this example is 1, whereas Precision is 0.25.

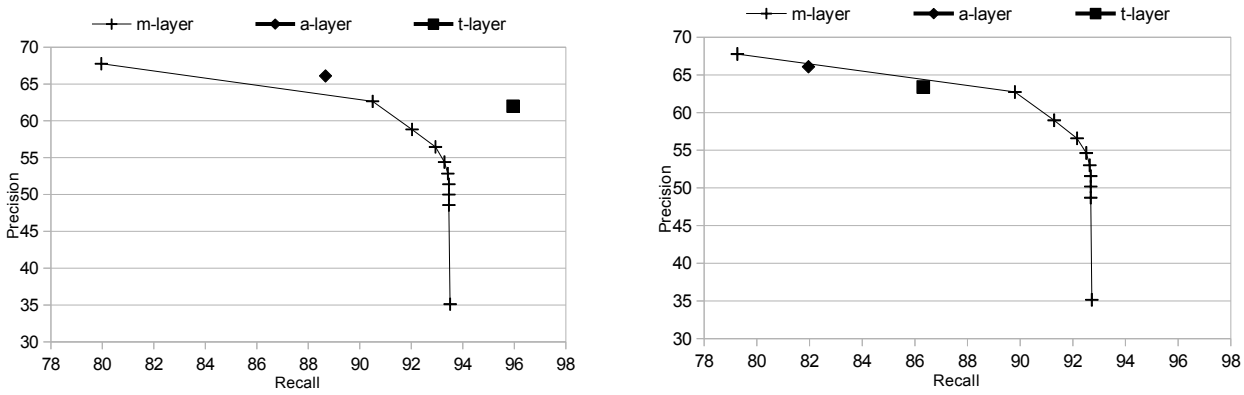

Figure 3: Precision–Recall scores of identification of MWE structures on manually/automatically annotated PDT.

of them, 46 MWEs were out-of-vocabulary expressions: they could not be found by automatic precedure using the original SemLex lexicon.

Note that results obtained using automatically parsed PDT are very close to those for manual data on all layers (see Table 2). The reasons need to be analysed in more detail. Our hypotheses are:

- M-layer identification reaches the same results on both data. It is caused by the fact that the accuracy of morphological tagging is comparable to manual morphological annotation: 95.68% (Spoustová, 2008).
- Both a- and t-parsers have problems mostly in complex constructions such as coordinations, that very rarely appear inside MWEs.

There are generally two issues that hurt our accuracy and that we want to improve to get better results. First, better data can help. Second, the method can always be improved. In our case, all data are annotated—we do nothing on plain text—and it can be expected that with a better parser, but also possibly a better manual annotation we can do better, too. The room for improvement is bigger as we go deeper into the syntax: data are not perfect on the a-layer (both automatically parsed and gold data) and on the significantly more complex t-layer it gets even worse. By contrast, the complexity of methods and therefore possible improvements go in the opposite direction. The complexity of tectogrammatic annotation results in a tree with rich, complex attributes of t-nodes, but simple topology and generalised lemmas. Since we only use tree topology and lemmas, the t-layer method can be really simple. It is slightly more complex on the a-layer (with auxiliary nodes, for example); and finally on the m-layer there is virtually unlimited space for experiments and a lot of literature on that problem. As we can see, these two issues (improving data and improving the method) complement each other with changing ratio on individual layers.

It is not quite clear from Table 2 that MWE identification should be done on the t-layer, because it is currently far from our ideal. It is also not clear that it should be done on the m-layer, because it seems that the syntax is necessary for this task.

6.2 Error Analysis and Possible Improvements

There are several reasons, why the t-layer results are not clearly better:

1. our representation of tree structures proved a bit too simple,
2. there are some deficiencies in the current t-layer parser, and
3. t-layer in PDT has some limitations relative to the ideal tectogrammatical layer.

Ad 1. We thought the current SemLex implementation of simple tree structures would be sufficient for our purpose, but it is clear now that it is *too simple* and results in ambiguities. At least auxiliary words and some further syntactico-semantic information (such as tectogrammatical functions) should be added to all nodes in these TSs.

Ad 2. Current tectogrammatical parser does not do several things we would like to use. E.g. it cannot

properly generate t-nodes for elided parts of coordinated MWEs that we need in order to have the same TS of all MWE occurrences (see below).

Ad 3. The total of 771 out of 8,816 SemLex entries, i.e. 8.75%, have been used with more than one tectogrammatical tree structure in the PDT 2.5. That argues against our hypothesis (stated in Section 5.1) and cause false negatives in the output, since we currently search for only one TS. In this part we analyze two of the most important sources of these inconsistent t-trees and possible improvements:

• *Gender opposites, diminutives and lemma variations.* These are currently represented by variations of t-lemma. We believe that they should rather be represented by attributes of t-nodes that could be roughly equivalent to some of the lexical functions in the Meaning-text theory (see Mel'čuk (1996)). This should be tackled in some future version of PDT. Once resolved it would allow us to identify following (and many similar) cases automatically.

- *obchodní ředitel* vs. *obchodní ředitelka* (lit.: managing director-man vs. managing director-woman)
- *rodinný dům* vs. *rodinný domek* (lit.: family house vs. family little-house; but the diminutive *domek* does not indicate that the house is small)
- *občanský zákon* vs. *občanský zákoník* (lit.: citizen law vs. citizen law-codex, meaning the same thing in modern Czech)

These cases were annotated as instances of the same MWE, with a vision of future t-lemmas disregarding this variation. Until that happens, however, we cannot identify the MWEs with these variations automatically using the most frequent TS only.

• *Elided parts of MWEs in coordinations.* Although t-layer contains many newly established t-nodes in place of elided words, not all t-nodes needed for easy MWE annotation were there. This decision resulted in the situation, when some MWEs in coordinations cannot be correctly annotated, esp. in case of coordination of several multiword lexemes like *inženýrská, montážní a stavební společnost* (engineering, assembling and building company), there is only one t-node for *company*. Thus the MWE *inženýrská společnost / engineering company* is not in PDT 2.5 data and cannot be found by the t-layer identification method. It can, however, be found by the m-layer surface method, provided the window is large enough and MWEs can overlap.

7 Conclusions

Identification of occurrences of multiword expressions in text has not been extensively studied yet although it is very important for a lot of NLP applications. Our lexicon SemLex is a unique resource with almost 9 thousand MWEs, each of them with a tree-structure extracted from data. We use this resource to evaluate methods for automatic identification of MWE occurrences in text based on matching syntactic tree structures (tectogrammatical – deep-syntactic, and analytical – surface-syntactic trees) and sequences of lemmas in the surface sentence.

The theoretically ideal approach based on tectogrammatical layer turned out not to perform better, mainly due to the imperfectness of the t-layer implemented in PDT and also due to the low accuracy of automatic tectogrammatical parser. It still shows very high Recall, as expected – due to simple topology of the trees – however Precision is not ideal. Morphology-based MWE identification guarantees high Recall (especially when no limits are put on the MWE component distance) but Precision of this approach is rather low. On the other hand, if the maximum distance is set to 4–5 words we get a very interesting trade-off between Precision and Recall. Using analytical layer (and thus introducing surface syntax to the solution) might be a good approach for many applications, too. It provides high Precision as well as reasonable Recall.

Acknowledgements

This research was supported by the Czech Science Foundation (grant n. P103/12/G084 and P406/2010/0875). This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). We want to thank to our colleagues Michal Novák, Martin Popel and Ondřej Dušek for providing the automatic annotation of the PDT and CNC data.

References

Abhishek Arun and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 306–313, Ann Arbor, Michigan.

Laurie Bauer. 1983. *English Word-formation*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, (44):7–21.

Eduard Bejček, Jarmila Panevová, Jan Popelka, Lenka Smejkalová, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, Josef Toman, Zdeněk Žabokrtský, and Jan Hajič. 2011. Prague dependency treebank 2.5. `http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8`. Data.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 242–245, Stroudsburg, PA, USA.

Paul Cook, Afsaneh Fazly, and 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, pages 41–48.

Gülşen Eryiğit, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, SPMRL '11, pages 45–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Fothergill and Timothy Baldwin. 2011. Fleshing it out: A supervised approach to MWE-token and MWE-type classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 911–919, Chiang Mai, Thailand.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 992–1001.

Deirdre Hogan, Conor Cafferkey, Aoife Cahill, and Josef van Genabith. 2007. Exploiting multi-word units in history-based probabilistic generation. In *EMNLP-CoNLL*, pages 267–276. ACL.

Graham Katz and Eugenie Giesbrecht. 2006. 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*, MWE '06, pages 12–19.

Linlin Li and 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: Volume 1 - Volume 1*, EMNLP '09, pages 315–323.

Christopher D. Manning, 2003. *Probabilistic Linguistics*, chapter Probabilistic Syntax, pages 289–341. MIT Press, Cambridge, MA.

Igor Mel'čuk. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Language Companion Series*, pages 37–102. John Benjamins.

Joachim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Dias, G., Lopes, J. G. P. and Vintar, S. (eds.) MEMURA 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications, Workshop at LREC 2004*, pages 39–46, Lisbon, Portugal.

Václav Novák and Zdeněk Žabokrtský. 2007. Feature engineering in maximum spanning tree dependency parser. In Václav Matoušek and Pavel Mautner, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 92–98, Berlin / Heidelberg. Springer.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *LNCS*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword

expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing*, volume 2276/2002 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg.

Violeta Seretan. 2010. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Academia/Reidel Publ. Comp., Praha/Dordrecht.

Miroslav Spousta. 2011. Featurama. `http://sourceforge.net/projects/featurama/`. Software.

Drahomíra "johanka" Spoustová. 2008. Combining statistical and rule-based approaches to morphological tagging of Czech texts. *The Prague Bulletin of Mathematical Linguistics*, 89:23–40.

Pavel Straňák. 2010. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. Ph.D. thesis, Charles University in Prague.