Influence of Treebank Design on Representation of Multiword Expressions

Eduard Bejček, Pavel Straňák, and Daniel Zeman

Charles University in Prague, Institute of Formal and Applied Linguistics
{bejcek,stranak,zeman}@ufal.mff.cuni.cz

Abstract. Multiword Expressions (MWEs) are important linguistic units that require special treatment in many NLP applications. It is thus desirable to be able to recognize them automatically. Semantically annotated corpora should mark MWEs in a clear way that facilitates development of automatic recognition tools. In the present paper we discuss various corpus design decisions from this perspective. We propose guidelines that should lead to MWE-friendly annotation and evaluate them on numerous sentence examples. Our experience of identifying MWEs in the Prague Dependency Treebank provides the base for the discussion and examples from other languages are added whenever appropriate.

1 Motivation

Grammatical theories have been thriving recently in computational linguistics. They describe phenomena of natural language in increasing detail with the purpose of creating a description that analyses and/or generates language as natural as possible.

Several treebanks have been developed during the past decade, new ones are still being created and the old ones are being enriched with additional annotations. A corpus is often designed and developed with the vision of further, deeper annotation, with the aim to add semantic information in future. Multiword expressions (MWEs; such as idioms, phrasemes, multiword named entities) are an important part of most natural languages. Usually they form a significant portion of vocabulary, particularly in special domains where terminology is in play, but not only there.

Although some grammatical theories have accounted for MWEs decades ago (see e.g. [1]), in treebanks, multiword expressions are one of the least developed phenomena. Recently, however, their processing started to attract attention, as they are proving to be important for information extraction, machine translation and other crucial tasks of NLP [2]. Therefore they should be an integral part of any serious semantic annotation.

In this paper, we discuss some decisions of a treebank design that have direct influence on representation of MWEs. A good treebank design can contribute to both more natural and more useful representation of MWEs, or even enable to capture certain rare forms of MWEs. We will also discuss the decisions that make the representation of MWEs harder or inefficient (see Section 3).

A. Gelbukh (Ed.): CICLing 2011, Part I, LNCS 6608, pp. 1–14, 2011.

We base the discussion on our experience with MWEs in the Prague Dependency Treebank 2.0 (PDT 2.0[1])[3]. Examples from other treebanks are presented for comparison. Examples that are not specifically marked are taken from PDT 2.0.

The rest of the paper is organized as follows: In Section 2, we provide some background on multiword expressions, and why they are important in NLP. In Section 3, we discuss the way MWEs are currently represented in selected treebanks, and what are the problems of these representations. Section 4 constitutes the core of the paper. We present a variety of linguistic phenomena and decisions of their representation that affect processing of MWEs to varying degree. We summarise our findings in Section 5.

2 Introduction to MWEs

Multiword expressions are a boundary phenomenon on the interface of grammar and lexicon. We understand them, in accordance with [4,5,6] and other authors, as phrases that contain some idiosyncratic elements that differentiates them from normal expressions. The idiosyncratic element can be morphological, syntactic, or semantic.[2] As a practical guideline we add that the idiomaticness must be significant enough to justify adding the given MWE into a lexicon.[3]

The idiosyncracy that defines the class of multiword expressions causes problems for various NLP applications.

- *Morphological* analysers have to analyse "words" that only exist in modern language as a part of an idiom (e.g. "criss" in criss-cross) in one fixed form. Even if it is a form of say singular, instrumental case, it does not fill such a morphological function.
- *Syntactic* analysers (and treebank designers before them) have to cope with analysis of idioms and other MWEs, in which the relations between the parts (words) do not have the meaning expressed by dependency relations or phrase structure types of the given grammar. The problem is usually solved by creating artificial annotation (grammar) rules with little to no linguistic motivation. Rules for analysis of named entities (NEs) like addresses or personal names can serve as good examples (see the relevant sections in [9]).
- *Semantic* idiosyncrasy limits the forms or even completely changes the translation equivalents of a MWE. One cannot translate "spill the beans" into a foreign language literally and keep its meaning. It is a big challenge for machine translation, especially in terminology (Supreme Court, Secretary of the Treasury, etc.).

The problems with handling MWEs in NLP applications are precisely why it is important to represent them correctly in treebanks. We will demonstrate that proper representation of MWEs can alleviate later problems with their treatment.

[1] `http://ufal.mff.cuni.cz/pdt2.0/`

[2] Some authors prefer still wider definition of MWEs and include also expressions that are fully regular and compositional on all layers of description, but are *statistically* significant. For instance the phrase "salt and pepper" is significantly more frequent than "pepper and salt" [7,6].

[3] For a description of a lexicon of MWEs see for instance [8].

For the purpose of this paper the problem whether a particular expression is a MWE or not is not crucial. What is important, is an agreement that MWEs exist, and that in representing them the linguistic phenomena discussed below have to be tackled.

3 Representation of MWEs

A handful of corpora provide MWE annotation on the layer of tokenisation. That means a MWE is actually converted to one-word expression. Not only is it an indivisible meaning from the perspective of deep syntax; it is also one token from the point of view of morphology and surface syntax. Even if the treebank does not have a dedicated deep-syntactic layer of annotation, the idiomaticness of the MWE can be captured by the annotation; the price is that it is no more possible to describe the inner structure of the MWE as well, should one desire that. Tokenisation-based annotation is typically limited to contiguous MWEs (otherwise, one would have to reorder tokens, apart from joining them). The CoNLL Shared Task Treebanks ([10], [11]) of Portuguese, Spanish and Catalan belong to this class. For instance, consider the following Spanish sentence:

(1) *sentence:* Durante la presentación del libro " **La_prosperidad_por_-**
lit.: During the presentation of-the book " **The_prosperity_through_-**

_medio_de_la_investigación_._La_investigación_básica_en_EEUU " ,
_means_of_the_research_._The_research_basic_in_U.S. " ,

editado por la **Comunidad_de_Madrid** , el secretario general de la
edited for the **Community_of_Madrid** , the secretary general of the

Confederación_Empresarial_de_Madrid-CEOE – CEIM – ,
Confederation_of-Company_of_Madrid-CEOE – CEIM – ,

Alejandro_Couceiro , abogó por la formación de los investigadores en
Alejandro_Couceiro , advocated for the formation of the investigators in

temas de innovación tecnológica .
themes of innovation of-technology .

trans.: During the presentation of the book "Prosperity through Research. The Basic Research in the U.S.", edited for the Community of Madrid, the Secretary General of the Confederación Empresarial de Madrid (CEOE), Alejandro Couceiro, advocated for the training of researchers in the field of technological innovation.

We believe that MWEs should be viewed as single units, but not on the morphological layer, as in the above mentioned Iberian treebanks. Even in terms of surface syntax, it is usually possible[4] to view MWEs as relations between words. It is the layer of the meaning of a sentence, i.e. deep syntax, where it is natural to tackle MWEs as single units, because units of this layer are supposed to be "meanings" [1,12,13]. In the PDT 2.0, the deep syntactic layer is called the *tectogrammatical layer* [9] and we demonstrate (mainly in Section 4) that it is the layer of description most suitable to represent MWEs.

[4] Even though sometimes quite awkward.

The same is true for other treebanks that already include some deep structures: in the beginning of treebanking, all treebanks (including PDT 1.0) were based only on the surface syntax. Most of them, however, have been accepting some deep syntactic features. These include PropBank [14] and NomBank [15] for the Penn Treebank, Chinese PropBank [16] and annotation of some named entities integrated in the in recent Chinese Treebank (see Figure 1), Salsa project [17] for the German Tiger treebank, and several others. The main problem of most of these annotation projects is, however, that the deep structures are annotated without any relation to the (surface) syntax, thus often ending up in conflict with it.

An illustration of this problem is given in Figure 1. The NEs, as well as coreference, were annotated on plain text and are stored separately from the syntactic annotation of the Chinese Treebank. This results in many cases in a unit of coreference annotation or a NE that does not form a phrase and thus cannot be represented in the tree. This points towards an error of either syntactic or deeper annotation, because any unit that is a member of a coreference relation or that forms a named entity should also form a phrase in a phrase structure tree.

3.1 List Structures

Some MWEs really have no internal syntactic structure in the given language. For instance embedded passages in a foreign language cannot be analysed using the grammar of the "main" language of the treebank.

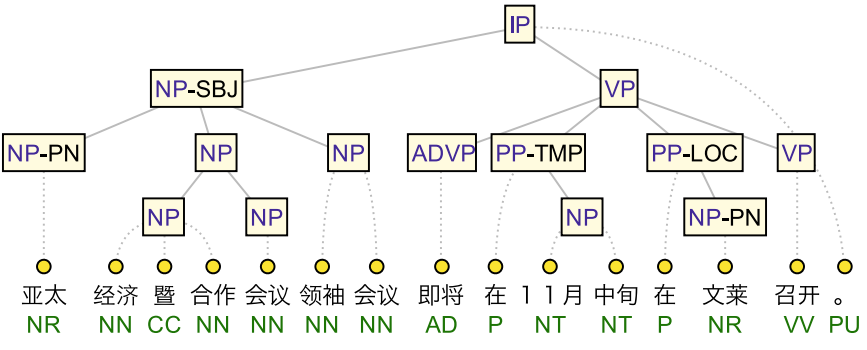

Fig. 1. A sentence from the Chinese Treebank 7, romanised *yàtài jīngjì jì hézuò huìyì lǐngxiù huìyì jíjiāng zài 11yuè zhōngxún zài wénlái zhàokāi*, meaning *"Asia-Pacific Economic Cooperation [APEC] Summit will be held in mid-November in Brunei."* lit. *"Asia-Pacific economy and cooperation conference leader meeting upcoming in November mid in Brunei hold."* The first five terminal nodes together constitute a named entity (MWE) that is the Chinese translation of *APEC*. However, the syntactic annotation does not contain any nonterminal spanning just this expression. The `NP-SBJ` span includes two additional terminals and describes an event (meeting of APEC leaders) rather than the institution. On the other hand, its second child `NP` fails to cover the node of *Asia-Pacific*. Thus the MWE cannot be properly marked without changing the parse tree first.

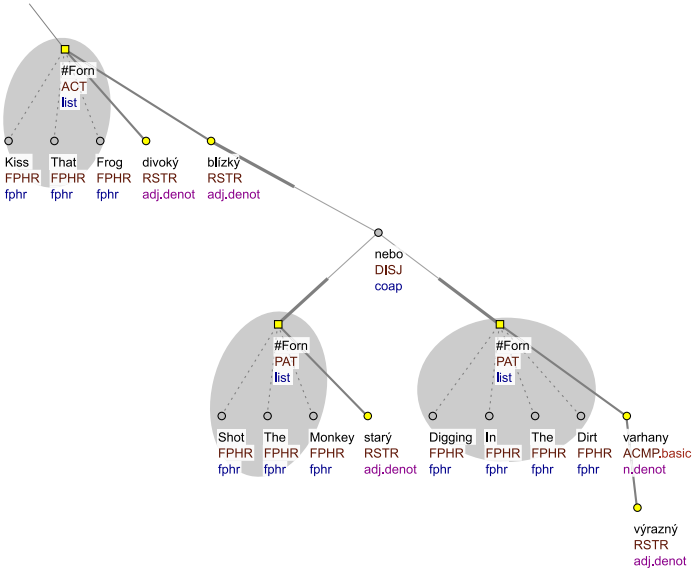

Fig. 2. *. . . divoké Kiss That Frog blízké staršímu Shot The Monkey nebo Digging In The Dirt s výraznými varhanami.*
trans.: . . . wild Kiss That Frog similar to older Shot The Monkey or Digging In The Dirt with striking organ.
Foreign expressions (English in the Czech sentence) represented as lists. The first MWE is modified by an attribute "similar (to)" and a coordination of the other two MWEs that are also further modified. (Example from the PDT 2.0.)

In PDT 2.0 these constructions are represented as lists of words with a generated root node that has a t-lemma[5] substitute specifying the type of the list (an Idiomatic Phrase, or a Foreign Phrase).

The list members (words in a list structure) cannot have children, since the whole point of creating these list structures is to specify either that there are no syntactic relations inside these objects, or that we cannot describe them. The whole structure can of course have children (e.g. attributes). Such children are represented as brothers of the members of list structures, and are distinguished by their tectogrammatical function, as seen in Figure 2.

We believe that creating list structures with artificial rigid and flat structures serves no point. Lemmas of the parts of such structures are foreign morphological forms (e.g. "shot"), and the dependency edges do not really represent any dependency relations. Thus we believe that non-analysable idioms and foreign phrases should be represented just as a single node.[6]

[5] A lemma of a node of a tectogrammatical tree, i.e. a tree on the tectogrammatical layer.

[6] One may also want to annotate the original structure according to the foreign grammar in parallel to the one-node representation assigned to the phrase once it entered the host language and became a MWE.

4 Linguistic Phenomena Reflected by Treebank Features

We present an overview of common linguistic phenomena that complicate capturing the MWEs. Every phenomenon is described and exemplified, the problem is discussed and a potential solution in the dependency treebank is proposed.

Two principles are to be borne in mind while making decisions on the structure of a treebank:

1. Structure of a tree must not obstruct marking any MWE.
2. Representation of a MWE has to enable identification of the same MWE automatically in the text.

How does one represent a MWE in a treebank? As the tree structure is non-linear, the best representation is a set of nodes that make up a particular MWE. This set has to be unambiguous, i.e. two different MWEs should not be represented by the same set of nodes.[7] On the other hand, slight variations in form of the same MWE should lead to the same representation so that the various forms of the MWE can be matched against each other. The set of nodes itself for a particular MWE highly depends on the treebank grammar and it is generally not guaranteed that every peculiar MWE can be mapped to a tree structure. For example, the MWE may contain a word that has been elided and does not have a corresponding node in the tree structure. In other cases, deep syntactic structure may contain a complex node spanning several surface words, some of which belong to a MWE and some of which do not; one would have to be able to mark only a part of a complex node in order to delimit the MWE properly.

The second principle leads to this aim: each and every instance of the particular MWE should be described by absolutely identical structure in data. In that case, it would be easy to find other instances of the same MWE automatically (using the same treebank or formalism). Following subsections illustrate that this is not as natural as it might seem.

4.1 Morphology

MWEs are hard to recognize automatically in an unprocessed text. Lemmatisation (or at least stemming) is the minimum requirement—even in English, not speaking of highly inflected languages such as Czech.

Consider the two instances of the German idiom *auf die lange Bank schieben* ("put off") in Examples (2) and (3) and Figure 3. The first one is in infinitive, the second one is passive. However, their lemmatised strings are identical, which makes it possible to recognize them as instances of the same MWE.

(2) *sentence:* Die EU dürfe die Entscheidung nicht **auf die lange Bank**
lemmatised: Der EU dürfen der Entscheidung nicht **auf der lang Bank**
lit.: The EU could the decision not **on the long bench**
schieben ...
schieben ...
shift ...
trans.: The EU could not put the decision off ...

[7] i.e. The structure of a subtree plus the words (lemmas) themselves.

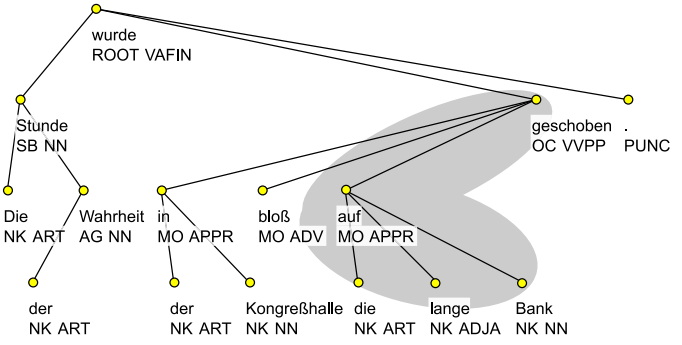

Fig. 3. An example from the CoNLL 2009 German treebank: *Die Stunde der Wahrheit wurde in der Kongreßhalle bloß auf die lange Bank geschoben. The moment of truth in the congress hall was just put off (lit. "shifted on the long bench").* The idiom has been passivised in this sentence but it still can be identified using lemmas.

(3) *sentence:* Die Stunde der Wahrheit wurde in der Kongreßhalle bloß **auf**
lemmat.: Der Stunde der Wahrheit werden in der Kongreßhalle bloß **auf**
lit.: The moment of truth was in the congress-hall just **on**
die lange Bank geschoben.
der lang Bank schieben.
the long bench shifted.
trans.: The moment of truth in the congress hall was just put off.

One might think that lemmatisation is a solved problem and that an annotated corpus is unlikely to lack it. As a matter of fact, out of the 23 treebanks from the CoNLL 2006 and 2007 shared tasks, lemmatisation was missing from significant number of them (Bulgarian, Chinese, Danish, German, Japanese and Swedish 2006 and English and Chinese 2007).

4.2 Word Form Alternations

There are many changes of word forms other than inflection mentioned in previous Section 4.1. Although these changes are more significant than morphological alternations, they still do not necessarily change the meaning of the MWE.

Lemmas in their usual sense cannot provide for unification of the alternations mentioned below, since the alternations differ morphologically and a considerable number of lemmatisers would assign a different lemma to each of them. In order to capture the relation between morphologically different expressions for a semantically identical concept, we need a sort of *generalized lemma*, common for all word form alternations.[8]

[8] Functional Generative Description (FGD, [12]), the theory behind PDT, introduces such a generalized concept called the "tectogrammatical lemma". The deep (tectogrammatical) layer of PDT 2.0 assigns a `t-lemma` attribute to nodes but it fails to merge some of the alternations mentioned here.

An alternative approach would be to annotate each MWE with its exact lemma, and create links between "variants" in the lexicon. The drawback here would be the large amount of lemma variants (some of them created productively on a regular basis) all written in the lexicon. The additional complexity could however bring also some additional information, i.e. in case of lemmas whose relation can be described by lexical functions [18]. Some variants of lemmas cannot, however, be distinguished by a lexical function, e.g. variants of diminutives in Czech. Some of the (spelling) variants are even unified on the level of morphology, while some other are not, and we unify them only in the MWE lemmas. Thus we have decided to employ the simple and uniform approach of using the same MWEs for all lemma variants. We can list and further analyse and classify all the variant realisations of all MWEs at a later point. We view the application of a lexical function in this respect as a form of a modification of a MWE, very much like any other modification, with similar restrictions: Some words in some MWEs can be modified, while other words or even whole MWEs cannot. Thus the approaches can be complimentary in our view.

Gender Inflection. The first alternation type we want to mention is present in many languages, including English, French or German. Since gender inflection of nouns is not productive in any of these languages, the alternate forms are assigned separate lemmas. Examples include pairs like "waiter"/"waitress", "actor"/"actress", "écrivain" or "homme de lettres"/"femme de lettres" etc. Examples of such pairs used in Czech MWEs are quoted below. We believe that the core meaning of the MWE remains the same across genders and it should be differentiated by a flag, not by a separate MWE in the lexicon.

We indicate the approximate occurrence ratio in PDT 2.0 in parentheses.

(4) mistr / mistryně světa (ratio 76:1)
master / she-master of-the-world
world champion

(5) státní zástupce / zástupkyně (ratio 2:1)
public prosecutor / prosecutrix
prosecuting attorney

(6) poštovní doručovatel / doručovatelka (ratio 1:2!)
postman / -woman
postman / postwoman

We propose that in each of the pairs, both variants should map to the same generalized lemma. One may wonder whether the actual string representing the generalized concept in (6) should match the masculine form (as is the usual default), or the feminine form (because in this particular case it seems to be more common in Czech data), or somehow embrace both (e.g. *poštovní_doručovatel(ka)*). However, these are only technical subtleties that are not significant from the perspective of the general concept-oriented approach.

Abbreviations. Writing systems of most languages have a means of abbreviating words and long multiword named entities. Examples of abbreviated and unabbreviated forms

referring to the same concept are given in (7) and (8). Again, we propose that the corpus annotation assign the same generalized lemma to both members of each such pair.

(7) Václav Havel / V. Havel

(8) země bývalého Sovětského svazu / země bývalého SSSR
states of-former Soviet Union / states of-former USSR

Aspect. In quite a few languages (the Slavic family being an example) aspect alternation is lexicalised (or at least not fully productive), which means that perfective and imperfective verbs get different surface lemmas. The following Czech examples (9) and (10) illustrate aspectual variations of MWEs.[9]

(9) zaujímat stanovisko / zaujmout stanovisko
take a stand *imperfective* / *perfective*

(10) pohlavně zneužívat / pohlavně zneužít
sexually abuse *imperfective* / *perfective*

Diminutives. Unless diminutive formation is fully productive in a language, the diminutive typically gets a (surface) lemma different from the base word. Yet the core meaning of a MWE is usually preserved in a "diminutivized" variant such as in the following Czech example (11):

(11) rodinný dům / domek
family house / small-house

Others. For the sake of completeness we bring up some other related pairs, although it is arguable whether it is necessary to unify them all. They have very close meanings and one has to consider them carefully. The variants in (12) are lexical meronyms but their encompassing MWEs are almost synonymous (furthermore, the second one is rarely used). The second expression (13) has the same meaning, only the first form is fixed phrase and the second is less formal. The pair in (14) has exactly the same properties in English. And the last one (15) illustrates an ellipsis[10] of a part of a word; the two expressions are totally synonymous in the context of telecommunications.

(12) občanský zákoník / občanský zákon — *meronym*
civil code / civil law

(13) náčelník generálního štábu / šéf generálního štábu — *synonym*
chief of-general staff / head of-general staff

(14) cenová regulace / regulace cen — *synonymous, though different*
price$_{adj}$ control / regulation of-prices *syntactic structures*

(15) telekomunikační systém / komunikační systém — *ellipsis*
telecommunication system / communication system

[9] Aspect is an exception that *is* unified in the `t-lemma` attribute of PDT 2.0, except for a few omissions.

[10] If we substitute the Greek prefix "tele-" in *tele*communication in (15) with its translation *remote*, the fact that it is an ellipsis becomes obvious.

4.3 Word Order

Lemmatisation is not sufficient (not even the generalized one) when the word order comes into play. In languages with a free word order, the same MWE can surface in various permutations. For instance, consider the phraseme "sehrát roli" (to play a role) in (16) and (17). The two instances differ in word order. The first sentence is neutral with respect to topic-focus articulation (i.e. it keeps the default Subject-Verb-Object order), whereas the second sentence accents the Subject ("communistic interpretation of history") by placing it into the focus position (resulting in the Object-Verb-Subject order).

(16) *sentence:* Klubíčko vztahů, které **sehrály roli** v této kauze,
lit.: Entanglement of-relations, that played role in this case,
se pokoušíme rozmotat…
we-are-trying to-disentangle…
trans.: We are trying to disentangle the entanglement of relations that played a role in this (legal) case.

(17) *sentence:* Svou **roli sehrál** i komunistický výklad historie.
lit.: Its role played even communistic interpretation of-history.
trans.: Even the communistic interpretation of history played its role.

The word order differences are a good reason why MWE detection should be done on dependency trees (as opposed to simple bracketing). Instead of looking at sequences of adjacent tokens, one can query parent-child pairs that remain the same regardless of word order. See Figure 4 for the dependency trees of (16) and (17).

4.4 Discontinuity on Surface

Discontinuous MWEs pose a problem similar to the word-order issue. Even a very lexicalised phrase (such as a verbal phraseme) can be disconnected with other words breaking in. In a phrase-based bracketing one would have to capture a MWE with gaps. The results would differ across sentences (different positions and sizes of the gaps) and there seems to be no reasonable algorithm to recognize them automatically.

Examples (18) and (19) illustrate that continuity is not related to MWE boundaries. There seems to be the phrase "hrát na nervy" ($\sim$ to fray one's nerves) twice – but only the first one (the one with gaps) is a real phraseme; the words in (19) came together just by coincidence.

(18) *sentence:* **Na nervy** to muselo **hrát** i našemu olympijskému vítězi.
lit.: **On nerves** it must-have **play** also to-our Olympic winner.
trans.: It must have been making nervous even our Olympic winner.

(19) *sentence:* Je to balzám **na nervy hrát** s Jenseny.
lit.: Is it balm **for nerves to-play** with Jensen's.
trans.: To play with Jensen's is like a balm for your nerves.

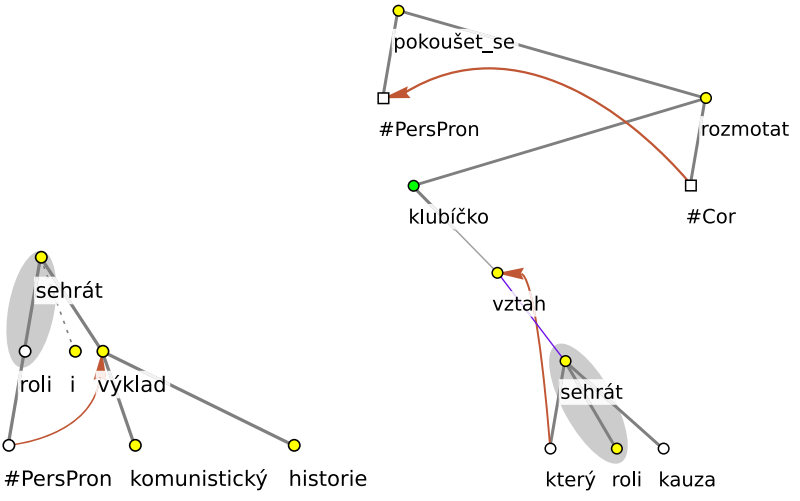

Fig. 4. Although the MWEs look diverse in the text (examples 16 and 17), they are identical and so are their subtrees. (It is not important whether a node is the left or the right son of its parent – the order of nodes represents the topic-focus articulation and does not affect the MWE.)

Similarly as in Section 4.3, we argue for the dependency structure as the basis for MWE detection. A dependent node of a MWE ("nervy" in this case) is connected to the governing node ("hrát"), no matter how far it is or in which direction. On the other hand, the parts of the would-be MWE in (19) are unrelated in the dependency tree, which blocks them from being considered as a MWE.

Dependency subtrees (with word order information stripped) provide sufficient means of representation for a vast majority of MWEs. They adhere to the second principle and assign the same representation to all instances of a MWE, regardless of word order and gaps. Unfortunately, there are still phenomena that cause problems.

4.5 Ellipsis

In (20) both "Ministry of the Interior" and "Ministry of Defense" should be recognized as MWEs. The problem is that there is only one word "ministry". The annotation mechanism would have to enable reusing one node in two different MWEs. Even if it did, a surface-oriented dependency tree (where there is a 1-1 mapping between nodes and tokens) would not provide enough information to detect the MWEs automatically (there would be no dependency link between "ministry" and "interior" or "defense", respectively).

(20) dvě klíčová ministerstva – vnitra a obrany
two key ministries – of-the-Interior and of-Defence

This example illustrates why we need a deep syntactic tree in which elided nodes are reconstructed. Figure 6 illustrates how the example is structured in the tectogrammatical

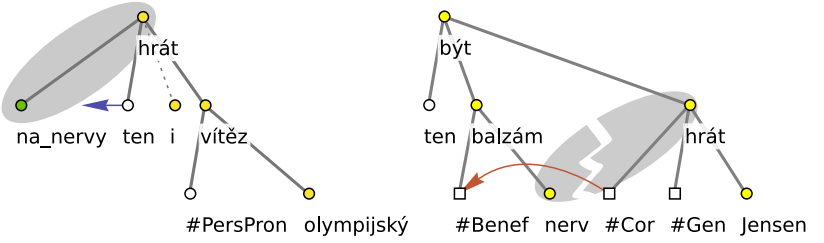

Fig. 5. Seemingly two occurrences of the phraseme "hrát na nervy" (∼ to fray one's nerves) in Examples (18 and 19). The dependency tree of (19) indicates that the idiomatic interpretation would be false. Such MWE across subtrees cannot be correct.

layer of the PDT 2.0. Thanks to the generated copies of "ministerstvo", the links to the required attributes are readily available and the MWEs can be detected easily.

Finally, there is an even worse problem with coordination: a coordination of two modifiers ascribed to an already modified noun, see (21):

(21) *coord.:* základní a náhradní vojenská služba
lit.: basic and substitute military service
trans.: military service and unarmed service

To be able to recognize both MWEs, we would like to see two complete (and disjunct) subtrees, one for "základní vojenská služba" and another for "náhradní vojenská služba". Node reconstruction in a deep syntactic tree could achieve that by generating copies of both the nodes "vojenská" and "služba". Unfortunately, this is not the case in PDT 2.0 where only the noun is copied, see Figure 7.

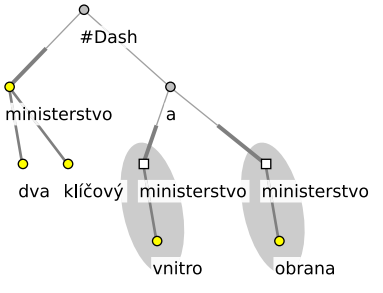

Fig. 6. A coordination with generated nodes (displayed as squares) enables annotation of words elided in the text (20)

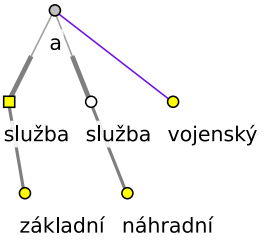

Fig. 7. The word "vojenský" (military) modifies the whole coordination in PDT 2.0, instead of modifying each coordinated node "služba" (service) with the same meaning

To summarize this section, we propose that elided nodes in coordination should be regenerated by copying *and* that their modifiers should be copied along, i.e. modification of the whole coordination should not be allowed.

5 Conclusion

We have discussed a number of linguistic phenomena that affect representation and automatic detection of multiword expressions in corpora. Each phenomenon led to a type of additional information that is needed in the corpus in order to detect MWEs properly. Such information can be added either manually in annotated corpora, or by previous steps of automatic processing of the text.

The following features of a treebank have been identified as useful for appropriate and efficient representation of MWEs:

- Surface lemmatisation to overcome the impact of inflection.
- Generalized lemmatisation to unify surface lemmas referring to the same semantic concept.
- Dependency structure to abstract from word order variation and discontinuity.
- Restoring nodes for elided words. In case of coordinated modifiers, restoring can be achieved relatively easily by copying the modified node to each coordination member.

We have tested our proposals while annotating MWEs in PDT 2.0, using the deep syntax of its tectogrammatical layer. They proved to be helpful from the perspective of both the principles set in Section 4.

Our annotated data, a lexicon of MWEs in PDT 2.0, and the tools we have used are freely available at `http://ufal.mff.cuni.cz/lexemann/mwe/`.

Acknowledgement

The research has been supported by the grant P406/11/1499 of the Czech Science Foundation, the grant GAUK 4307/2009 of the Grant Agency of Charles University in Prague, and projects no. LC536, MSM 0021620838, and "Clarin", no. 7E08026 of the Ministry of Education of Czech Republic.

References

1. Mel'čuk, I.A., Polguère, A.: A formal lexicon in The Meaning-Text Theory (or how to do lexica with words). Computational Linguistics 13(3-4), 261–275 (1987)
2. Ren, Z., Lü, Y., Cao, J., Liu, Q., Huang, Y.: Improving statistical machine translation using domain bilingual multiword expressions. In: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 47–54. Association for Computational Linguistics, Singapore (2009)
3. Bejček, E., Straňák, P.: Annotation of multiword expressions in the Prague dependency treebank. Language Resources and Evaluation (44), 7–21 (2010)
4. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, p. 1. Springer, Heidelberg (2002)
5. Baldwin, T., Bannard, C., Tanaka, T., Widdows, D.: An empirical model of multiword expression decomposability. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions, pp. 89–96. Association for Computational Linguistics, Morristown (2003)

6. Pecina, P.: Lexical Association Measures: Collocation Extraction. Studies in Computational and Theoretical Linguistics, vol. 4. Institute of Formal and Applied Linguistics, Prague (2009)
7. Baldwin, T.: Multiword expressions, CSSE, University of Melbourne (2004)
8. Straňák, P.: Annotation of Multiword Expressions in The Prague Dependency Treebank. PhD thesis, Charles University in Prague (2010)
9. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z.: Annotation on the tectogrammatical level in the Prague Dependency Treebank. annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep. (2006)
10. Buchholz, S., Marsi, E.: CoNLL-X shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), pp. 149–164. Association for Computational Linguistics, New York City (2006)
11. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 915–932. Association for Computational Linguistics, Praha (2007)
12. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Academia/Reidel Publ. Comp., Praha (1986)
13. Kahane, S.: The Meaning-Text Theory. In: Dependency and Valency, Handbooks of Linguistics and Communication Sciences, vol. 25(1-2), p. 32. De Gruyter, Berlin (2003)
14. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: A corpus annotated with semantic roles. Computational Linguistics Journal 31(1) (2005)
15. Meyers, A.: Using treebank, dictionaries and GLARF to improve NomBank annotation. In: Proceedings of The Linguistic Annotation Workshop, LREC 2008, Morocco (2008)
16. Xue, N.: Labeling Chinese predicates with semantic roles. Computational Linguistics 34(2), 225–256 (2008)
17. Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M.: The SALSA corpus: a German corpus resource for lexical semantics. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Citeseer (2006)
18. Mel'čuk, I.: Lexical functions: A tool for the description of lexical relations in a lexicon. In: Wanner, L. (ed.) Lexical Functions in Lexicography and Natural Language Processing. SLCS, vol. 31, pp. 37–102. John Benjamins, Amsterdam (1996)