PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 112 APRIL 2019 83–101

# Design of a Multiword Expressions Database

Pavel Vondřička

Charles University, Faculty of Arts, Institute of the Czech National Corpus, Praha, Czechia

**Abstract**

The paper proposes design of a generic database for multiword expressions (MWE), based on the requirements for implementation of the lexicon of Czech MWEs. The lexicon is aimed at different goals concerning lexicography, teaching Czech as a foreign language, and theoretical issues of MWEs as entities standing between lexicon and grammar, as well as for NLP tasks such as tagging and parsing, identification and search of MWEs, or word sense and semantic disambiguation. The database is designed to account for flexibility in morphology and word order, syntactic and lexical variants and even creatively used fragments. Current state of implementation is presented together with some emerging issues, problems and solutions.

## 1. Introduction

Multiword expressions (MWEs) have been long in the focus of the theoretical linguistics as well as NLP, especially during recent years. Breaking the seemingly clear borderline between lexicon and grammar, they obstruct each successful NLP task based on this traditional dichotomy.[1] As much as their actual definition differs, in its widest meaning reaching from proper names or fixed idiomatic expressions with non-compositional meaning to light verb constructions or seemingly free, but actually statistically idiosyncratic collocations,[2] so differ also the applications and implementations dealing with them. A particular language and the more or less limited scope of view also determine the complexity of their description and identification.

[1] As presented already by Sag et al. (2002).

[2] Sag et al. (2002) speak generally about *institutionalized phrases*, but also about simple *statistical affinity* – the phenomenon seems to be broader.

 Corresponding author: pavel.vondricka@ff.cuni.cz
Cite as: Pavel Vondřička. Design of a Multiword Expressions Database. The Prague Bulletin of Mathematical Linguistics No. 112, 2019, pp. 83–101. doi: 10.2478/pralin-2019-0003.

The PARSEME survey on MWE resources (Losnegaard et al., 2016) shows that many MWE lists and lexicons are still limited to contiguous sequences of words or lemmas, and that has also been the case of some approaches to Czech MWEs.[3] However, the limitations of this method become quickly apparent especially in Slavic languages with relatively free word order. Treatment of non-contiguous MWEs has therefore been lately addressed more intensively, although it has been called a "challenge" by Savary (2008).

In the current project *Between Lexicon and Grammar*[4] we aim at a complex description of Czech MWEs, targeting both various NLP applications and human users at the same time. The resulting database must therefore cover many diverse, incompatible or even contradictory requirements. Our goal is also to cover not only the explicit and exact use of established MWEs, but also as many as possible of their variants and unusual modifications appearing in various texts, where the creativity of language users[5] seems rather unlimited.

This goal does not only require some treatment of the non-contiguity, but also treatment of variable word order as well as variable lexical members. Components of MWEs may be more or less freely inflected or modified, omitted or even replaced by some rather unusual lexeme, while the core meaning of the MWE is still kept. Restrictions on the morphology and its various irregularities within MWEs have already been addressed by many projects,[6] but treatment of lexical variability seems to be a rather rare case. It has been addressed (at least to some degree) for example by Villavicencio et al. (2004) or Grégoire (2010) in the project DuELME, which obviously also takes into account difference in occurrence (frequency) of the particular variable components. The approach of Al-Haj et al. (2013) offers a very simple and elegant solution both to the lexical variability and the variability of word order at the same time. Nevertheless, our project still aims at even greater flexibility in the identification of MWEs or even their fragments.

## 2. Flexibility of MWEs and their core

As shown by Hnátková et al. (2017) and Jelínek et al. (2018), even seemingly fixed MWEs may appear in different variants or modifications. Some of their components may be optional and many components may vary or be modified. Often it is enough to keep just a core or fragment of the original MWE in order to recall the original meaning and construct a metaphor or other form of word play, where the rest of the

[3]See e.g. Pala et al. (2008).

[4]For more details see the articles by Hnátková et al. (2017), Hnátková et al. (2018) and Jelínek et al. (2018).

[5]Sometimes obviously caused also by their lack of knowledge of the established form or its correct language use.

[6]See Savary (2008); Oflazer et al. (2004); Al-Haj et al. (2013); Czerepowicka and Savary (2018), etc.

MWE may be twisted or replaced with some other construction in any way.[7] It is thus important to record, how every component of the MWE varies in the common use and how it may vary potentially.[8] In addition, it is useful to identify the minimal core components (fragments) that are able to identify the original meaning even if the rest of the components is missing or replaced. We have also found examples of multiple minimal cores which can convey the meaning; any of them may be used by the language user.[9]

All these possibilities of MWEs, which go far beyond the possibilities of simple words, raise the question of the identity of a MWE. As long as the meaning of the original MWE contributes to the message of the text, it should be considered its part and detected in the process of parsing. That should probably apply also in cases, where the identification of the MWE and its original meaning is not a necessary condition for the recipient to understand the message of the text; the use of a modified MWE may as well be just a part of the art form (e.g. to express irony, humour) or the authors desire to make an impression of wittiness and creativity, and not necessarily part of the message itself.

On the other hand, there is often a possibility of overlap with the literal use of the same combination of words. Such a risk varies a lot among different MWEs and depends mostly on the amount of their anomalies. The identification and disambiguation of MWEs thus remains a very difficult task in many cases.

## 3. MWE entry: identity

The question of identity of a MWE opens the question of the identity of one single entry in the database. We consider the meaning as the main criterion for the distinction of single entries. The expression *jít přes čáru* 'cross the line' can have three different meanings, depending on the meanings of the noun *čára*: 1) the literal meaning commonly appearing in sports, where lines often demarcate a play field; 2) the colloquial meaning of borderline between countries, referring commonly to the phenomenon of illegal emigration to the Western Europe during the communist rule in

[7] This has been illustrated by the example of the biblical quote 'it is easier for a camel to go through the eye of a needle than for a rich man to enter into the kingdom of God' encountered in various forms and allusions in the data of the Czech National Corpus, such as: '*a camel would rather enter into a kingdom of heaven than* I would pass a thread *through the eye of the needle*', 'the bypass should be threaded through the area like *the camel through the eye of a needle*', '*it is easier to go through the eye of a needle than* to get access to EU funds' or 'Klaus forces two elephants to be pulled *through the eye of a needle*'. (Hnátková et al., 2017)

[8] Of course, the potential variability is actually unlimited, but observed variants may be presented according to real occurrence in corpus data.

[9] E.g. the expression *hodit flintu do žita* (lit. 'throw the rifle into the rye (field)', meaning 'throw in the towel', 'give up'), can either appear in forms such as 'throw [something] into the rye' (e.g. 'throw the camera into the rye', i.e. 'abandon the career of a photographer') or 'throw the rifle [somewhere]'. (Hnátková et al., 2017)

Czechoslovakia; 3) the most abstract meaning of any negative phenomenon exceeding some acceptable threshold. The literal meaning should not be covered by the database, while the other two meanings should probably have two independent entries.

However, the form of the MWE remains an important factor as well. In combination with the high variability of many MWEs, it may become difficult in many cases to decide whether some expression is still a variant of another expression with the same or very close meaning, or whether it should be considered an independent MWE. The pragmatic factor can therefore motivate the distinction of two or more partly identical MWEs with the same meaning[10] if one single (merged) description of the variability would be too complex. This concerns especially complicated dependencies among several components of a MWE which must or must not appear together. Some basic examples motivated by syntactic alternations are presented in Section 10 and those are still kept within a single entry, but more complex cases can be encountered.

Another question concerns possible derivations of MWEs – e.g. passivization or nominalization of a verbal MWE. We do not want to create separate entries for such derivations, unless their meaning, usage or behaviour differs significantly. Instead, the possibilities or restrictions imposed on the common types of derivation should be defined for every MWE entry, where such a possibility can be expected by the grammar.

## 4. MWE entry: structure

### 4.1. Requirements

A MWE consists of two or more components, understood – by definition and the name itself – as *words*. These "words" may be more or less fixed: some components may be realized by one particular word (lexeme) only, some by a choice of several different words (lexemes); some can be formed just by any lexeme or even a whole phrase of a particular type, just like any standard valency element.[11] Some components may be freely inflected or modified, while others are subject to various restrictions. The MWE therefore needs a definition by means of its components and their various possible realizations. It is important for us to be able to describe the features

[10]The meaning of MWEs is also often considerably more complex than meaning of simple words. The meaning of each individual component (word) still contributes to the meaning of the whole expression even for relatively fixed expressions, and the variations may always modify it to some degree.

[11]Like single word units, MWEs take valency elements as well. Some MWEs are just verbal phrases, where some of the valency elements are filled by fixed expressions while others remain open. Some MWEs can also take valency elements which none of the components would require or allow by itself. This has already been presented by Hnátková et al. (2017) on the example of the expression *dát na srozuměnou, že…* 'to make clear that…' (lit. 'to give on understanding that…'), where the valency slot for *that*-clause is not specified by any of the component parts.

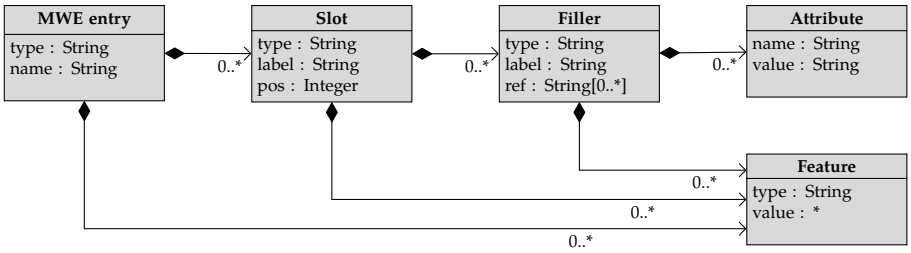


*Figure 1. MWE entry structure (basic model)*

and restrictions of the MWE, its components and their respective possible realizations, independently.

Since we want to describe the entries and their components both for the purpose of NLP parsing and for human users, there will be a need to record very different types of features. Different features also concern different levels of description. While valency, usage type or syntactic type are features of the whole MWE, internal modifiability concerns rather its components, and lexical idiomaticity is a feature specific to the particular lexeme used as a component. Some features may apply to several levels: style or register usually depend on the use of a particular lexeme, if there is a choice of several varaints. However, it may also be the feature of the whole MWE: the expression *mít něco na háku* ('couldn't care less', lit. 'to have st. on the hook') is rather colloquial despite of the fact that it consists solely of standard words – the metaphorical use is not standard anymore.

### 4.2. Entry structure

For full flexibility, we define the entry pattern by means of *slots* and *fillers*, common terms used for this type of description in computational linguistics.[12] The entry itself consists of *slots* and *features* referring to the MWE as a whole. Slots represent the single components of the MWE (pattern), which is the syntagmatic dimension of the MWE. Slots consist of *fillers* and the slot-specific *features*. Fillers represent the paradigmatic dimension of the components: the possible variants which may be used to realize the particular component. The primary role of fillers is to represent actual (terminal) tokens to be matched in the data.

In the process of annotation of the data of the Czech National Corpus, the experimental identification of MWEs has been applied as the last step, on top of the tokenized, lemmatized, tagged and desambiguated texts. The current parser FRANTA

[12]The basic principles follow (in a simplified form) the proposal for a structured lexical description as described by Vondřička (2014).

| | |
|---|---|
| `lemma="ryba"` | noun *ryba* 'fish' |
| `tag="NNF[SP]1"` | in nominative singular or plural |
| `lemma="stát"` | verb *stát* 'stand / cost / happen' |
| `tag="V"` | in any form |
| `tag="AA"` | any common adjective |
| | in any form |

*Table 1. Example definitions of positional attributes for different types of fillers*

uses a combination of the positional attributes *lemma* and morphological *tag* to match token patterns in the data. The patterns have been defined in a special MWE list called FRANTALEX. This list has been used as the primary source of initial data for the MWE database. Therefore we define the fillers by means of a combination of positional attributes (*lemma* and *tag*) and their values that must be matched in the text in order to identify the MWE. However, the fillers may actually declare just any arbitrary positional attributes used to identify the matching tokens. Other restrictions, such as possible word order, modifications or transformations, can be defined by means of additional features. Figure 1 shows the scheme of the whole entry structure.

The attributes to be matched may also be underspecified: the *tag* value may contain just a prefix referring to the part-of-speech or a regular expression to match a custom choice of acceptable morphological forms. Specification of the *lemma* may be completely avoided in cases when just any lexeme of some particular part of speech or morphological category may fill the position, but its presence is still necessary (or typical) for the identification of the MWE (see Table 1 for examples). Of course, the filler may provide its own additional *features* as well.

For strictly fixed expressions, a slot will mostly contain only one possible filler defining the particular type of token to be matched. More flexible expressions may contain a list of several synonymous or otherwise alternative fillers. Since the fillers may also have their own features, it is possible to document their actual relative usage (e.g. by terms of corpus frequency) or further individual effects on the other slots or on the MWE as a whole. Such slots can be classified as *fixed* or "closed". In case of relatively *open* slots, the fillers may be underspecified as mentioned above. They may also represent only the most typical representatives of a relatively open semantic class. That is relevant in cases where such a group of acceptable fillers cannot be fully defined formally in an explicit way. We call this third kind of slots *semi-open*. Of course, such incomplete description can currently only be of limited use for a NLP parser, but it will still remain a useful hint for human users of the lexicon.

If we want a slot or filler to represent a whole phrase of some type (e.g. in the case of valency elements), we cannot use a combination of positional attribute values to match one single token anymore. We need to use specific features to define the phrase type (restriction) instead. Such description probably cannot be directly used

by a simple low-level parser such as FRANTA, but it can be useful for human users and later also for possible higher-level parsers operating also at the syntactic level.

### 4.3. Classification of features

Features are generic pairs of *type* (name) and *value*.[13] For easier organization and systematization of various types of features, we use a hierarchical system of specification of the *type* by means of a path in an arbitrary hierarchy of features, using colon as the separator. At the top level, features are classified as morphological, syntactic, semantic, statistical, related to the form, purely user- or editor-oriented notes, etc. Further levels are divided as needed: as specific groups of features, by particular theory, source of data, etc. This also allows us to store multiple similar features from different sources (or for different purposes) at the same time.

In case we need to include multiple alternative values of some type of feature, custom subspecification may be used. This applies especially to user notes, examples from real texts or statistical values. For example, the basic type of features for absolute frequency `:stats:fq:abs` is expected to be extended by additional custom subspecification of the corpus (and possibly subcorpus) used to acquire the frequency value, e.g. `:stats:fq:abs:BNC:fiction`. This allows the database to be searchable by features both using underspecification of the type (by means of a path prefix) or its full (sub)specification as needed.[14]

## 5. Multi-level and multi-purpose utilization of the structure

The flexible design allows for multi-purpose utilization of the entry components. Features can easily be used (and classified) both for purely technical purposes of NLP processing tools and to store information aimed at human users of the database, such as definitions, examples or notes.

Information may also be provided at several levels of description, also in parallel if needed: surface restrictions on the form or occurrence of particular components (such as those presented in Section 10) may actually result from regular alternations on higher levels (e.g. syntax), but at the time of the initial import of patterns from the FRANTALEX list, this information is provided just in the form of simple surface rules, restricting the possible occurrence of particular tokens (forms) in the sentence, and must be later reinterpreted manually in order to obtain a more appropriate, higher-level linguistic description. On the other hand, new MWEs created manually in the

[13] For the purpose of effective implementation and searching, a single feature record may actually have several values of different type in the database: string value, numeric value, etc. Currently, we do not want to utilize this technical property of implementation in the data model, but keep the feature as a purely atomic property.

[14] The possibility of custom subspecification makes it possible to add additional information to the value of the feature and it thus makes the atomicity of the feature object rather ostensible.

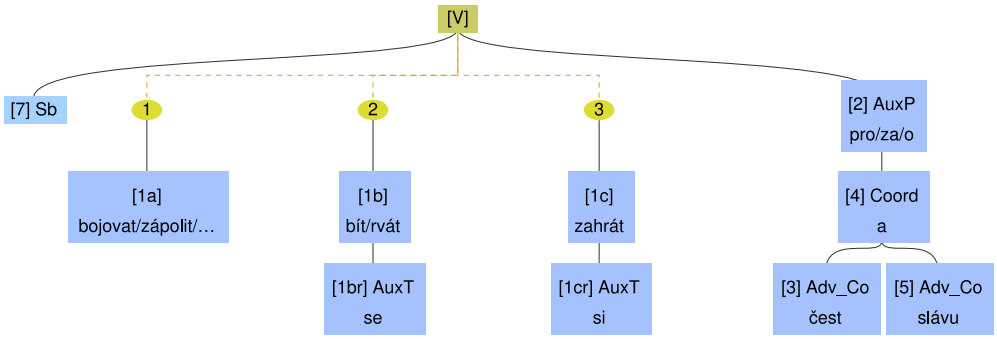


*Figure 2. Dependency structure for the expression* ***bojovat za čest a slávu***

database or imported from other sources – and already provided with the more abstract linguistic descriptions – must also allow for an automatic reinterpretation of the information back into the form of the (low-level) surface rules the parser is actually able to process and apply to the data lacking any higher-level annotation. This may actually also apply to such basic phenomena as grammatical agreement. Future advanced parsers may possibly utilize the higher-level descriptions (such as syntactic relations and alternations) more directly.

## 6. Representation of tree structures

In the database, it is desirable to capture tree structures such as dependency and constituency structure of an expression. Dependency relations between the components can easily be recorded in the form of slot features. One single feature is needed as reference to the parent slot and another one to identify the syntactic function of the component. Such relations can easily be projected into the resulting tree structure, as illustrated in the Figure 2 showing a combination of dependency tree with variant subtrees (explained later in Section 8).

However, constituency trees need non-terminal nodes and we need to be able to refer and assign features to them as well. Therefore, they should be represented by standalone objects in the database, equivalent to the slots. That is the reason why just grouping the components by means of features would not be a satisfactory solution.

The flat structure of slots and fillers does not allow for nesting. Previously, we have suggested to use recursive structures for lexical descriptions, where fillers can branch into further sequences of "subslots".[15] However, indexing and querying recursive data structures is still a very demanding task not well supported by the current database and search engines. Therefore the idea was abandoned. Instead, we decided

[15]See Vondřička (2014) for more details.

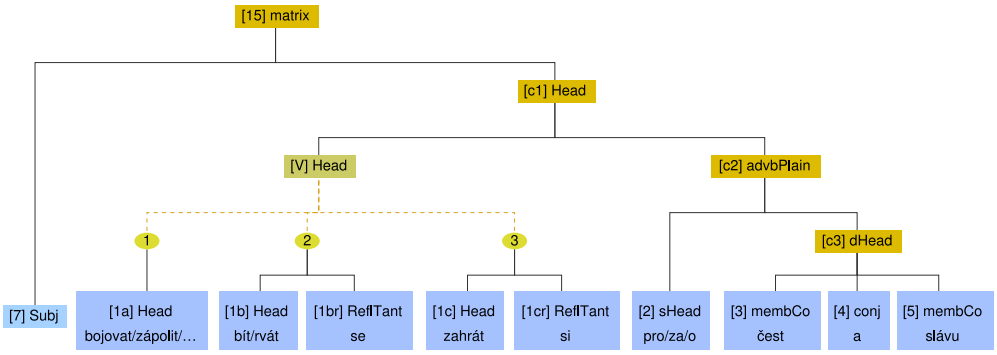


*Figure 3. Constituency structure for the expression* ***bojovat za čest a slávu***

to keep to the flat structure, but to allow fillers to refer to a sequence of other slots by means of their identifiers (labels). This brings the possibility to add non-terminal fillers (and their respective non-terminal slots). Again, the tree structure can easily be reconstructed by nesting the slots additionally, as seen in Figure 3.

Various advantages and disadvantages emerge from this design: indexing and searching for both terminal and non-terminal nodes is equally simple, but traversing relations between them in a single query is not supported by the search engine. That means that searching for MWEs by their structure – e.g. by syntactic (or other) relations – would be difficult to implement. Currently, we do not expect the need to search the database by tree structures, but in case this would be necessary, the structures can be reconstructed for all entries, encoded into some kind of searchable patterns and indexed separately by the same or a more appropriate engine.[16] Another advantage is the possibility to record several independent tree structures within a single entry, which corresponds well to the requirement of multifunctionality. A partial disadvantage is the potential need for treatment of possible partial trees, overlapping trees and orphan nodes.

## 7. The treacherous term “word”

The most problematic issue of “multi-word entities” is the fact that the term is based on the linguistically not well defined concept of “word”. Relying solely on the orthographic aspect of using space or punctuation marks as boundaries between “words” is very treacherous even in languages using Latin alphabet.[17] This can be well demonstrated by the English example of the triple acceptable spelling “airstream”,

[16] E.g. a graph database.

[17] This issue has been well discussed e.g. by Savary (2008).

"air stream" and "air-stream". Similar phenomena concern also Czech and become especially urgent when dealing with many colloquial MWEs, where there is no standard established for their spelling.[18]

A tendency to split composed words in Czech seems quite obvious lately, probably by the influence of English spelling which does not make this distinction. On the other hand, there has traditionally been an opposite tendency of merging more or less established prepositional phrases in the standard language: a continuum between already established adverbs (or prepositions) such as *včas* ('in time'), *dohromady* ('together') and less established combinations such as *na příklad/například* ('for example'), where both spellings are still in use, and *do ztracena/doztracena* ('(peter out) to nothing'), where the single-word spelling is still much less common, despite the fact that the noun alone can be extremely rarely encountered in other contexts. The situation becomes especially unstable in case of many colloquial exclamations, such as *pro Boha/proboha!* ('for God('s sake)!'), which usually only appear expressively in direct speech. The borderline between words and MWEs becomes quickly very unclear, and it proves much more as a problem from the practical perspective of low-level NLP parsers, rather than from the theoretical point of view of linguistics.

Such alternatives cannot be simply automatically merged in the source text either, since ambiguity may still exist as seen in the example (1). There is often a tendency to make a distinction between the adverbial, prepositional or particle meaning (such as 'for example') and the original literal meaning ('at example/exercise') by merging the words together as in the variant (1-a), in analogy to other already established adverbs of this type, and also to avoid confusion like in the example (1-b). However, this is not always the rule and the less common or lexicalized the combination is, the more unpredictable the spelling is. In such cases, individual factors such as education, age and conservatism play an important role.

(1) a. Podívejte se například na příklad číslo 7.
Look for example at example number 7.
'For example, look at the example/exercise number 7.'
b. Podívejte se na příklad na příklad číslo 7.

The problem also arises when dealing with standard multiword components of MWEs such as reflexive verbs or other analytic forms. The system of slots and fillers (representing always a single token only) can only deal with single-word alternations, but it cannot deal with alternation of non-reflexive verbs with reflexive verbs requiring an additional reflexive pronoun. One possible solution is to use non-terminal slots for such variants again.[19] In this case, the fillers do not represent the terminal to-

[18] Not to mention the fact, that the CNC project aims at retaining and mapping also non-standard phenomena in the language, including those classified by many people as "mistakes".

[19] Another solution is to define separators as fully-fledged components of the MWE, which may be omitted (Savary, 2008; Czerepowicka and Savary, 2018), or to define MWE components at the level of mor-

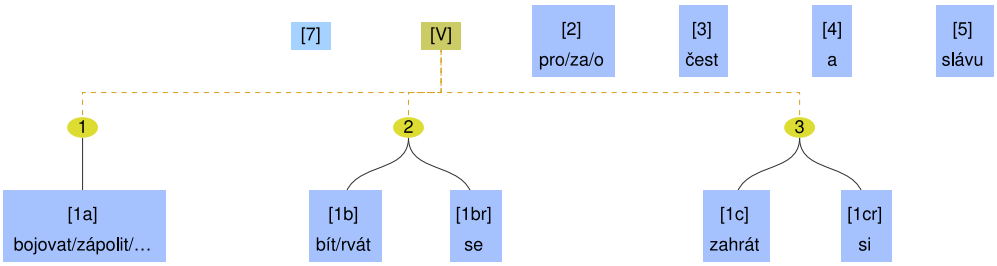


*Figure 4. Variants of the expression* ***bojovat za čest a slávu***

kens directly, but refer to other (terminal) slots or sequences of several slots that may alternate. Since the order of slots does not necessarily represent a real word order, even more complex alternations or dependencies can be defined by means of these non-terminal slots.

If we extend the possibility of reference beyond the limits of the entry itself, we can also describe MWEs containing (embedding) other MWEs, such as *držet/viset (jen) na čestné slovo* (lit. 'hold/hang (just) on a word of honour', meaning 'fixed/fastened in an unreliable, wonky way') embedding the MWE *čestné slovo* ('word of honour').

## 8. Alternatives requiring a different number of tokens in practice

The variability beyond the limits of simple words can be demonstrated on the example of the expression *bojovat za čest a slávu* ('fight for honor and glory'). Three different prepositions can be used in this expression (*za, pro* and *o*) and these can be listed as fillers of a single slot. Unfortunately, this does not apply to all the verbs which may also alternate here. The variant tree-structure of the expression can be seen in Figure 4.

Beside simple verbs, reflexive verbs can also appear in this expression, and these require a reflexive particle – i.e. an additional token. For this reason, they cannot fit into a simple list of alternative terminal fillers representing a single token anymore. In addition, both types of Czech reflexive verbs can occur here: those using reflexive pronoun in accusative case (e.g. *bít se*) and those using reflexive pronoun in dative case (e.g. *hrát si*).[20] For this purpose we have created a non-terminal "variant-slot" labeled `V`, representing all the various verbs, with three fillers (shown in the figure as

phemes (Al-Haj et al., 2013). However, this would be difficult to implement in our case, where the identification of MWEs is currently applied at the end of the liguistic analysis of data, which have previously been processed by a tokenizer and tagger unaware of the existence of MWEs.

[20]The question may be raised whether the meaning can really be considered identical in all these particular cases, but that does not change the situation in principle.

elliptical nodes with numbers 1, 2 and 3) for the three classes of verbs: the first one refers to the terminal slot listing simple verb fillers only, the second one refers to the sequence of slots listing the reflexive verb(s) requiring accusative and the reflexive pronoun in accusative itself, and the third one refers to the sequence of slots listing the verbs requiring dative and the reflexive pronoun in dative itself. All the other slots are terminal slots and remain orphans in this partial tree-view. Slots are labelled arbitrarily, by default by numbers in the order of addition.[21]

The variants branching into their own subtrees also make the visualization of syntactic trees more complicated. As shown in Figure 3, the alternating subtrees fit quite well into the constituency tree, at least as long as the variants correspond to the syntactic subtrees of the whole expression. However, their visualization within dependency trees is in principle impossible without adding a third dimension to the scheme: the non-terminal node breaks the principle of direct dependency between terminals, since it represents several terminals at the same time and their dependencies cannot point to all of them individually (see Figure 2). Therefore, the verbal dependencies need to point to the non-terminal node as their parent. The non-terminal node branches again into the three different verb groups it represents, but this is not a relation of dependency anymore. Their reflexive particles may then depend on the verbs directly again. In the scheme, we try to visualize the different type of relation again by means of different type of lines between the non-terminal node and the verbs.

## 9. Non-terminal slots

The database thus currently uses two types of non-terminal slots: slots for complex variants (multi-token alternations) and slots for non-terminal nodes within the constituency structure. Technically, only the fillers can actually represent terminals or non-terminals, but we want to avoid mixing terminal and non-terminal fillers in a single slot, so that the slots can also be clearly classified as the terminal and non-terminal nodes they are supposed to represent.

Slots representing valency elements – i.e. whole phrases of some type as described in Section 4.2 – represent actually a third type of non-terminal slots in the database, even though they do not refer to other components within the entry itself.

## 10. Internal dependencies

Several types of internal dependencies between the components of a MWE have already been encountered, which make the process of parsing and MWE identification more complex. One of them concerns MWEs using some lexeme repeatedly – these may also have variants or modifications concerning the repeated lexeme itself. The expression *Bůh dal, Bůh vzal* (lit. 'God gave – God took') can be used in various mod-

[21] In the figure, the slot labeled by No. 7 represents a generic subject of the verbal phrase.

ifications such as *život dal, život vzal* ('life gave, life took'), *čas dal, čas vzal* (time), *stát dal, stát vzal* (state/government), etc. All the modifications are based on a repeated lexeme which may vary itself. Therefore we need to define the lemma of the consequent slot (filler) as a reference to the lemma actually used in the first slot. For this purpose we currently define a special placeholder[22] with reference to another slot as the value of the filler's lemma. In case we encounter more complex dependencies of this type, we might need to find another appropriate solution.[23]

Another type of dependencies concerns optional components. These are often projections from a higher level, such as syntactic alternations. However, on the surface level of the parser they need to be specified as well. In the expression *mít NĚCO pro (svou) (vlastní) potřebu* ('to have ST. for (one's) (own/personal) use') both the possessive pronoun and the adjective are optional, but at least one of them must be present to specify the possessor of the 'use'. The expression *naložit NĚCO na NĚČÍ bedra* ('load ST. on SO.'s shoulders') alternates with the form *naložit NĚKOMU NĚCO na bedra*: on the surface level the addressee can either be expressed by an indirect object in dative, or as the possessor (attribute) of the 'shoulders'. If we want to define the MWE in one single entry, we must define both the indirect object and the possessive adjective as optional components, but indicate their mutual exclusivity in some way. The optimal solution to this kind of problems is still in discussion.

The variability of verbs in the example in Section 8 is another example of a low-level projection from a higher level, where the necessity of an additional token – the reflexive particle – would probably be easier to declare as a lexical or syntactic feature of the reflexive verbs.

## 11. Minimal fragments

It has also been mentioned in Section 2 that MWEs can also take part in some text in the form of creatively used fragments.[24] These fragments may go far beyond the common limits of variability or optionality of the MWEs components. It seems therefore useful to list the minimal combinations of components which have been proved to be sufficient to trigger the meaning of the MWE even if it has been heavily modi-

[22] We currently use the form `${target-slot-label}`

[23] The expression *hlava nehlava* (lit. 'head non-head', meaning 'without any regards'), where the repeated form is negated, has been discussed as an example of a prototype of a more general pattern applicable in theory to any other word (specifying closer the addressee of the (lack of) 'regards') as well. However, changing the base lexeme would also imply a change of the meaning, so that this construction in the generalized form belongs rather to the domain of grammar or some kind of 'multi-word word-formation', rather than to the lexicon directly.

[24] As demonstrated in detail by Hnátková et al. (2018) and Jelínek et al. (2018).

fied. For this purpose we use a separate feature containing the list of slot identifiers which are capable to represent a minimal core fragment of the MWE.[25]

## 12. Available sources of data

The primary source of MWEs for the database is its predecessor, FRANTALEX. Rather than a proper database, it is a list of about 36000 patterns (simple rules) for the parser FRANTA. This parser has been used to identify and tag MWEs in the corpora of contemporary Czech within the Czech National Corpus. These patterns have mostly been based on the descriptions included in the traditional Czech Phraseological Dictionary (Čermák et al., 1983–2009), but they have been extended by actual observations of the corpus data: common variants of the MWEs missed by the parser or incorrectly identified combinations of the same tokens having their original literal meaning (false positives).

As mentioned before, FRANTA identifies MWEs as combinations of particular tokens identified by their lemma and morphological tag, with limited possibilities to define restrictions on gaps between them and the acceptable variability of the word order. Because of the simplicity of the parser, which operates at the surface level of a morphologically tagged text only, and the relatively free word order in Czech, many of the patterns actually identify different variants of one and the same MWE, and in some cases even variations with a different word order only. Such patterns must therefore be manually combined into single (but more complex) descriptions, before they can be imported into the new database as base for new MWE entries. As long as the FRANTA parser or a similar surface-level tool is used to identify the MWEs in the corpora, we must also be able to reverse the process and generate all the alternative rules from the merged complex entries in the new database with updated information.

Additional process is used to generate syntactic structures (dependency and constituency) for the existing patterns, both those imported from FRANTALEX and those created manually. They are generated by a syntactic tagger[26] trained on the data of the Prague Dependency Treebank[27] for each MWE, manually checked and added to its entry in the database. The constituency structures are then created by conversion from the manually corrected dependency structures. The dependency relations are added as features to the existing terminal slots, while constituency structures require adding new slots for non-terminal nodes.

[25]This solution is very similar to the more general solution presented by Al-Haj et al. (2013), which is also used to describe variable word order as well as optional and alternative use of different components and their mutual surface dependency.

[26]See Martins et al. (2013).

[27]See Hajič et al. (2018).

Other sources of MWEs, light-verb constructions (LVC) and named entities (NE) based on real data are also available from the development and annotation of the Prague Dependency Treebank and related projects.[28] These offer also a higher-level annotation, but they are to some degree limited to the texts of the PDT. Overlap with the primary source can also be expected and the possibilities of utilizing and merging the different sources will need closer inspection.

## 13. Practical issues

The desire for a multi-purpose resource uniting various different sources of data brings some unavoidable issues or pitfalls to be resolved in order to keep consistency of the data across the lexicon. The first one is the variability of the data sources based on different approaches and with different goals in mind. While the FRANTALEX database is a set of raw surface patterns based on actual observations of MWE variability in the annotated texts of the Czech National Corpus, the syntactic annotation offers higher-level abstractions of many of these observations and variations. However, the parsing and identification of MWEs will still need to be applied to syntactically unparsed raw data, and the need to project the higher-level abstractions from the database to the surface dependencies in the form of simple rules will remain necessary.[29]

The generic database structure also offers several possible solutions to many phenomena. Again, we can take the difference between a higher level classification and a surface description from FRANTALEX as an example: while a valency dictionary would define a valency slot as a phrase of some type, e.g. a prepositional phrase specifying the preposition and the case of the nominal phrase to be used (i.e. one single open non-terminal slot in the database), FRANTALEX will provide a pair of components (terminal slots): the fixed preposition and an open slot for a noun, possibly marked as 'open for modification'.

The pragmatic approach of the FRANTALEX database may thus be in conflict with the desire for theoretical purity (systematicity) and conceptual consistency. This will need to be resolved in order to make the database a unified resource.

Another issue is the dependency of the database on the current state of tokenization and morphological annotation of the data to be parsed for MWEs. The database must also try to account for possible common mistakes in morphological analysis or disambiguation. In the case of rule-based disambiguation, this may result in a circular dependency: the identification of MWEs depends on the morphological disambiguation, and the disambiguation may depend on the identification of MWEs. This problem concerns especially the issue with the variability of words alternatively split or merged by various language users. The morphological tagger currently cannot be

[28] E.g. Vallex (Lopatková et al., 2016; Kettnerová et al., 2012) or SemLex (Straňák, 2010).

[29] As already mentioned in Section 12.

expected to tag a combination of two words in the same way as one single composed word, especially if there is real danger of ambiguity: the single word *například* can (or must) be unambiguously tagged as a particle, but the phrase *na příklad* can safely[30] be analyzed as a combination of a preposition and a noun only, at least as long as a closer and unambiguous syntactic analysis of the whole sentence is not available.

The dependency of the fillers on a particular tagger or tagset can (to some degree) be reduced by defining multiple positional attributes (even virtual or planned future attributes) prefixed by some kind of "namespace" in a similar manner as the classification of feature types.

## 14. Implementation and user interface

The database has currently been implemented as a part of a more generic database of corpus annotation units, sharing a common infrastructure and principles. Elasticsearch is used as backend engine for searching and storing the entries in the form of JSON documents. A data model written in Python is used as an intermediate abstraction, providing a generic API.

The API also provides management of metadata about all object types stored in the database. The types of entries, slots and fillers can be classified in the same way as the types of the features. Each object type can also be provided with descriptions and definition of its contents, access restrictions and visualization hints, requirements on subspecifications, methods of editing and presentation for different users and purposes and so on. Definitions of the features may, for example, specify a particular type of value(s), so that basic input validation and appropriate searching criteria may be automatically applied by the interface.

Current frontend user interface is designed using the Angular.js and Bootstrap frameworks. It uses the API and the metadata to create customized and highly configurable user interface on the fly. We expect it to be able to present and visualize the data in different ways suitable for different types of users. Other user interfaces can also be created for more specialized purposes, using the generic API.

## 15. Conclusion

We have described the generic framework used to encode and manage the database of Czech MWEs and the principles of their encoding for various purposes. The main innovation is its open and flexible structure, aimed at multiple levels of description and multiple purposes, including linguistic description aimed at human users and the effort to cover also creative use and modifications of established MWEs in real lan-

[30] Underestimation of these distinctions leads frequently to wrong analysis in many taggers with excessive interpretative ambitions. Morphological taggers simply cannot be as smart as some linguists would like them to be.

guage use, both theoretically and formally. Several problems of the encoding strategy and their possible solutions have been discussed.

The classification of MWEs applied in this project and the actual contents of the database entry have been previously described in detail by Hnátková et al. (2017). We have not dealt here with the generation of the morphology of MWEs either, since in our project the MWEs are currently being identified in texts analyzed previously. A possible deeper integration of the MWE database into the process of disambiguation and parsing of Czech textual data remains an open question for further research.

## Acknowledgements

This paper is part of the project *Between Lexicon and Grammar* (2016–2018), supported by the Grant Agency of the Czech Republic (reg. no. 16-07473S), and the implementation of the *Czech National Corpus* project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

## Bibliography

Al-Haj, Hassan, Alon Itai, and Shuly Wintner. Lexical Representation of Multiword Expressions in Morphologically-complex Languages. *International Journal of Lexicography*, 27(2): 130–170, 12 2013. ISSN 0950-3846. doi: 10.1093/ijl/ect036. URL `https://doi.org/10.1093/ijl/ect036`.

Czerepowicka, Monika and Agata Savary. SEJF – A Grammatical Lexicon of Polish Multiword Expressions. In Vetulani, Zygmunt, Joseph Mariani, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 59–73, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93782-3.

Grégoire, Nicole. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1):23–39, Apr 2010. ISSN 1574-0218. doi: 10.1007/s10579-009-9094-z. URL `https://doi.org/10.1007/s10579-009-9094-z`.

Hajič, Jan, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Hajičová, Jiří Havelka, Petr Homola, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Petr Pajas, Jarmila Panevová, Lucie Poláková, Magdaléna Rysová, Petr Sgall, Johanka Spoustová, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. Prague Dependency Treebank 3.5, 2018. URL `http://hdl.handle.net/11234/1-2621`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, and Pavel Vondřička. Eye of a Needle in a Haystack. Multiword Expressions in Czech: Typology and Lexicon. In Mitkov, Ruslan, editor, *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017, London, UK, November 13–14,*

*2017, Proceedings*, volume Lecture Notes in Computer Science, vol. 10596, pages 160–175, Cham, 2017. Springer International Publishing. ISBN 978-3-319-69805-2. doi: 10.1007/978-3-319-69805-2_12. URL https://doi.org/10.1007/978-3-319-69805-2_12. ISBN: 978-3-319-69805-2.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, and Pavel Vondřička. Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. *Korpus – gramatika – axiologie*, (17/2018): 3–22, 2018. ISSN 1804-137X.

Jelínek, Tomáš, Marie Kopřivová, Vladimír Petkevič, and Hana Skoumalová. Variabilita českých frazémů v úzu. *Časopis pro moderní filologii (Journal for Modern Philology)*, 100(2): 151–175, 2018. ISSN 0008-7386 (Print), ISSN 2336-6591 (On-line).

Kettnerová, Václava, Markéta Lopatková, and Eduard Bejček. The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In Fjeld, Ruth and Julie Torjusen, editors, *Proceedings of the 15th EURALEX International Congress*, pages 434–443, Oslo, 2012. Department of Linguistics and Scandinavian Studies, University of Oslo.

Lopatková, Markéta, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. *Valenční slovník českých sloves VALLEX*. Karolinum, Praha, 2016. ISBN 978-80-246-3542-2.

Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. PARSEME Survey on MWE Resources. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

Martins, A., M. Almeida, and N. A. Smith. Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers. In *Annual Meeting of the Association for Computational Linguistics – ACL*, pages 617–622, August 2013.

Oflazer, Kemal, Özlem çetinoğlu, and Bilge Say. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 64–71, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613186.1613195.

Pala, Karel, Lukáš Svoboda, and Pavel Šmerk. Czech MWE Database. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, Alexander F., editor, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pages 1–15, London, 2002. Springer.

Savary, Agata. Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, (1-2):1–53, 2008. URL `https://hal.archives-ouvertes.fr/hal-01023019`.

Straňák, Pavel. *Annotation of Multiword Expressions in The Prague Dependency Treebank*. PhD thesis, Univerzita Karlova v Praze, Prague, Czech Republic, 2010.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. Lexical Encoding of MWEs. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 80–87, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613186.1613197`.

Vondřička, Pavel. *Formalized contrastive lexical description: a framework for bilingual dictionaries*. LINCOM GmbH, München, 2014. ISBN 978-3-86288-428-5.

Čermák, František et al. *Slovník české frazeologie a idiomatiky (SČFI)*, volume 1–4. Academia/Leda, Prague, 1983–2009.

**Address for correspondence:**
Pavel Vondřička
`pavel.vondricka@ff.cuni.cz`
Institute of the Czech National Corpus
Faculty of Arts, Charles University
nám. Jana Palacha 1/2
CZ-11638 Praha 1, Czech Republic