Chapter 1

LEMUR: A lexicon of Czech multiword expressions

Hana Skoumalová[a], Marie Kopřivová[a], Vladimír Petkevič[a], Tomáš Jelínek[a], Alexandr Rosen[a], Pavel Vondřička[a] & Milena Hnátková[a]
[a]Charles University

This chapter describes a lexicon of Czech multiword expressions, designed to be useful both for human readers and for natural language processing tasks. Its entries use a rich typology of multiword expressions, based on their syntactic aspects, idiomaticity and flexibility, with a focus on the specific features of Czech multiword expressions with their significant variability, and a classification according to a traditional approach. The content and structure of the entries facilitate the use of the lexicon in natural language processing. The chapter also describes how the lexicon is implemented and used in parsing and for annotating multiword expressions in a corpus. The corpus and the lexicon are linked, so each entry in the lexicon includes examples from the corpus, and each annotated multiword expression in the corpus is linked with a corresponding lexical entry.

1 Introduction

In every language, multiword expressions (henceforth MWEs) represent a substantial part of the vocabulary, both in common and in specialist use. A lexicographical resource describing MWEs is therefore an obvious need. Such descriptions can be part of a standard lexicon or included in a dedicated lexicon of MWEs.

On the path from lexicon to grammar, many MWEs stay at least halfway between the two, some much closer to grammar than single-word lexemes. This

Hana Skoumalová, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondřička & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631

is even more pronounced in a language such as Czech, with its free word order and rich morphology, including intricate morphosyntactic agreement patterns. Considering the flexibility of many MWEs (not only in Czech), allowing for insertions, omissions, permutations, morphosyntactic transformations, the use of synonyms, and other manifestations of variability, a satisfactory solution calls for a highly elaborate scheme for the specification of lexical entries. As an answer to the need for a lexical resource up to the task we introduce LEMUR, a LExicon of MUltiword expRessions of Czech.

The chapter is structured as follows. §2 relates LEMUR to some existing common lexical resources, referring to its sources of inspiration and providing a concise summary of research on MWEs, with a specific emphasis on Czech. The extensive §3 introduces the components of a lexical entry together with the multidimensional taxonomy of MWEs. Next, §4 presents an overview of how the lexical entries are encoded and how the whole lexicon is implemented. In §5 we exemplify the current use cases of the lexicon: (i) as a resource for annotating MWEs in corpora and providing links between their occurrences in a corpus and the corresponding entries in the lexicon, and (ii) as an aid in tagging and parsing. The chapter concludes with a summary of achievements and pitfalls and some perspectives of the project (§6).

2 LEMUR related to other MWE lexicons and previous research

LEMUR was designed from the start as a richly structured database, with an interface suitable for use in lexicography, for teaching Czech as a foreign language, for studying theoretical issues of MWEs as entities between lexicon and grammar, and also for Natural Language Processing (henceforth NLP) tasks such as tagging, parsing and corpus annotation, including MWE identification and search, or word sense and semantic disambiguation.

2.1 LEMUR and other MWE lexicons

LEMUR is not the first lexicon of Czech MWEs. The standard reference lexicon of Czech phraseology (Čermák et al. 1983–2009) is an impressively large and detailed achievement, but its printed format and standard lexicographical approach favour the traditional manual look-up before other possible uses. Other resources focus either on an inventory of MWEs used for their identification in corpora, such as the FRANTALEX lexicon (Hnátková 2002, Kopřivová & Hnátková 2014),

or on extending a valency lexicon to include MWEs headed by verbs (Urešová 2009, Lopatková et al. 2014, Przepiórkowski et al. 2017). LEMUR differs from the above in its broad focus: to the best of our knowledge, its entries cover more types of MWEs and capture more properties of each MWE than any other resource (for a similarly rich resource for Bulgarian and Romanian, see Leseva et al. 2024 [this volume]). Moreover, it provides the option of bi-directional links between entries in the lexicon and occurrences of the multiword lexemes in a corpus.

In addition to MWEs listed in traditional phraseological dictionaries, i.e. proverbs, similes and sayings (Burger et al. 2007), the lexicon includes compound function words (mainly prepositions and conjunctions), scientific terms (Kováříková & Kovářík 2019), and typical collocates (for example *vydatná strava* 'nutrient food'). However, it does not include frequent co-occurrences of function words such as *ale i* 'but also', *že se* 'that REFL' etc.

LEMUR builds on FRANTALEX, which was based on Čermák et al. (1983–2009) and extended by additional MWEs, found in corpora, and variants of already known MWEs. The MWE typology used in the lexicon is a modification of the multi-dimensional taxonomy used in lexical templates within the PARSEME project,[1,2] and inspired by Baldwin & Kim (2010). An important addition to the taxonomy is the notion of morphological idiomaticity (see §3.3.2). While compiling the lexicon, we also addressed theoretical issues related to the variability of MWEs (Pasquer et al. 2018). As a major theoretical contribution, we see the design of a scheme describing the variability, together with detailed descriptions of variability of each MWE.

Last but not least, the entries include the syntactic structure of each multiword lexeme as dependency and constituency trees. This view of MWEs is important also for the section of the entry where possible valency requirements of a part or the whole of the MWE may be specified.

2.2 MWE research mainly from the Czech perspective

Research on MWEs intensified in the late 20th century. Properties and usage of MWEs have been studied from various angles. Some of the studies deal mainly with terminology (Bozděchová 2007, Temmerman 2000), while non-compositional MWEs are studied within the disciplines of phraseology, paremiology (Čermák 2007), and also comparative studies (Popovičová 2020: 12–16). With the de-

[1]https://typo.uni-konstanz.de/parseme/

[2]https://www.lexical-resource-semantics.de/wiki/index.php/Parseme_MWE_Template:_English (visited Nov 13, 2023)

velopment of language corpora, new possibilities for research on MWEs and collocations emerged: for terms (Kováříková 2017), and for phrasemes (for example Colson 2017). A new concept of terminology (Klégr 2016) and new demands for the identification of MWEs within large-scale data appear. In recent years, NLP focusing on the description of MWEs has become one of the fastest growing areas, obviously relevant also for Czech (Lichte et al. 2019, Sheinfux et al. 2019).

The description of phraseology is essential for language teaching (Čechová 2011: 66–67), lexicography (Čermák et al. 1983–2009) and linguistic theory. In order to handle concrete data, the properties of phrasemes become important. However, individual researchers differ even here (for example Čechová 2011, Čermák et al. 1983–2009): for Čechová, a MWE is characterized by the fixedness of form, while Čermák allows for the possibility of variability in some MWEs. We see variability as a complex phenomenon: some MWE properties, such as variation, fixedness and repetition of a lexeme, need to be described in a way more consistent with real-text data (Jelínek et al. 2018). Thanks to the availability of large corpora, variation and fixedness can be observed in more detail to see where grammatical categories alternate and where new MWEs with new lexical components emerge: for example plural and singular alternate in *cesta do pekla/pekel*, (lit. 'way to hell/hells'), while a new MWE *dávat logiku*, (lit. 'to give logic'), is derived from its original version *dávat smysl*, (lit. 'to give sense'). Corpora also help to identify and annotate monocollocable components within MWEs, i.e. words with restricted usage in one or few combinations only, and to mark MWEs containing such components. Moreover, in our approach we also annotate fragments of MWEs since MWEs often occur in fragmentary forms.

In the LEMUR lexicon, we also try to reconcile the different approaches and introduce a taxonomy of MWEs that encompass different linguistic domains. This makes it possible to search for units according to criteria used by different approaches, with an emphasis on the traditional Czech MWE categorization that reflects the current educational needs. In order to classify MWEs, we use a combination of the classification presented by Čermák (2007) and Moon (2007), with the addition of some new categories. In determining the type of idiomaticity, we adopt the PARSEME taxonomy, supplemented with categories related mainly to morphology (Hnátková et al. 2017).

3 Typology of MWEs

The MWE typology in LEMUR is inspired by the PARSEME project and by Baldwin & Kim (2010), which categorizes MWEs according to their

- *idiomaticity*: lexical, syntactic, semantic, pragmatic and statistical;
- *syntactic category*;
- *fixedness/flexibility* of "lexicalized phrases".

This typology was adopted and primarily extended with respect to: (i) specific properties of Czech, especially morphological idiomaticity, and (ii) the fact that the lexicon should be useful both for human users and software applications. For instance, in (1) the form *nosa* 'nose' is a nonstandard genitive form of the noun *nos* appearing exclusively in this MWE (cf. also §3.3.2). This fact is marked on *nosa* in the lexical entry.

(1) podle *_nosa_ poznáš kosa
by nose.SG.GEN recognize.2SG.PRS blackbird.SG.ACC
'someone's character can be recognized by her/his deeds'

Moreover, we also extended PARSEME's approach in the following respects. In our approach, lexical idiomaticity (§3.3.1) encompasses not only MWE components that are not part of the conventional lexicon of Czech, such as MWEs consisting of foreign loans, for example (LAT) *mutatis mutandis*, but also (possibly almost) monocollocable words, for example *překot* in *o překot* 'headlong', negative only forms, for example *neličená radost* 'genuine pleasure', macaronic structures, that is, structures combining Czech and foreign words, for example *by voko* 'by guesswork' and other. Syntactic idiomaticity is not restricted only to MWEs whose syntax is not derived from that of their components, since we also annotate their deviations from standard syntax, such as anacolutha (cf. 31), attraction (cf. 32), idiosyncratic valency (cf. 33), aposiopeses (cf. 34), ellipses (cf. 35), zeugmas and others. For capturing semantic idiomaticity, we use a 4-grade annotation scale where a MWE can be: (i) always non-compositional, i.e. not explicitly derivable from its components, for example *nebrat si servítky* (lit. 'not to take napkins'), 'not to mince one's words'), (ii) rarely compositional, for example *kukaččí vejce* 'cuckoo's egg', (iii) often compositional, i.e. often non-idiomatic, for example *vlčí doupě* 'wolf's den', and (iv) always non-idiomatic, literal, for example *přísný pohled* 'stern look'. As for syntactic structure (§3.1), MWEs are identified by their syntactic category (determined by MWE's head) and assigned a dependency and a constituency tree. Moreover, the (im)possibility of MWEs' syntactic transformations (passivization, nominalization, adjectivization) is also annotated. As to MWE fixedness and flexibility, MWEs are specified for the (im)possibility

of their lexical, morphological/morphosyntactic and syntactic variation, including internal modification of their components and/or word order fixedness or freeness.

Generally, each lexicon entry contains descriptions of two types of MWE properties: some concern the MWE as a whole, others are related to its components (words). In addition to the three characteristics adopted from the PARSEME project, each entry in the lexicon is described by its lemma and superlemma, definition, examples of usage, style marker, usage type, i.e. a collocation type classified according to the traditional Czech phraseological taxonomy, and a basic classification of adverbial MWEs. The detailed description of these features follows.

- Lemma is a string (= sequence of concatenated word forms) constituting a MWE in its prototypical form that identifies the MWE in the lexicon and in an annotated corpus, for example *mateří kašička* 'royal jelly'. Via its lemma, the MWE can be searched both in the lexicon and in the corpus.

  In the case of synonymous MWEs, we have to decide whether they will be included under one lemma or whether the lexicon will contain two lemmas. If they differ only in meaning, but all other properties are the same, for example, *černá díra* 'black hole' (a scientific term vs. collocation meaning that money disappears somewhere with no visible benefit), the dictionary will contain only one lemma with two definitions and two sets of examples. However, if there is variability in the lexical setting of one of the lemmas, or constraints on syntactic transformations, word order changes, etc., we will introduce two lemmas, as in (2).

  (2) jít přes čáru
  go over line
  'to cross the border illegally/to cross the line'

  In the meaning 'to cross the border (illegally)', we can use synonyms of the verb *jít* 'go' that differ in aspect or prefix: *jít/přejít/přecházet/chodit přes čáru*. In the meaning 'to behave in an unacceptable way', only the imperfective aspect of the verb *jít* is possible, but the verb *být* 'be' can also be used here to describe the state when someone has crossed an imaginary boundary of decency (*jít/být přes čáru*).

- Superlemma is a representative of a list of lemmas that have at least one word in common and are semantically related. These are, for example, converse MWEs such as (3a) and (3b), or two related, but different MWEs such

as (4a) and (4b). Note that (4b) is not a standard nominalization of (4a) (unlike (4b), (4a) is a comparison/simile containing the conjunction *jako* 'like'),[3] but both share the same superlemma.

(3) a. dát ultimátum
give ultimatum
'to give an ultimatum'

b. dostat ultimátum
get ultimatum
'to get an ultimatum'

(4) a. hrát si jako kočka s myší
play REFL like cat with mouse
'to play like a cat with a mouse'

b. hra kočky s myší
play of cat with mouse
'a cat and mouse game'

A superlemma is always an existing lemma, or a fragment thereof and must consist of at least two words. For the examples given, the superlemmas are *dát ultimátum* 'give an ultimatum' and *hra kočky s myší* 'play of a cat with a mouse', respectively. The superlemma is actually only a label indicating a list of semantically linked lemmas, and we select the shortest lemma or fragment from the list as the superlemma.

- Definition is an informal gloss of the MWE's meaning. Most glosses are adopted from Čermák et al. (1983–2009).
- Examples are from corpora of contemporary Czech, representing real usage.
- Stylistic marker classifies both the MWE and its components (words) from the viewpoint of style. The following values are distinguished:
  - standard: used commonly in written texts: *být upoután na lůžko* 'to be confined to bed';

[3]The standard nominalization would be *hraní si na kočku a myš* where *hraní* 'playing' is a paradigmatically derived deverbal noun.

- colloquial: used mainly in spoken communication and understandable in every part of the country: *Co jí to **vlezlo do hlavy**?* (lit. 'What crept into her head?') 'Where did she get that idea?';
- dialect: the whole MWE or one of its components is part of a particular dialect spoken only in part of the country. Such a MWE is included in the lexicon since it occurs in corpus texts (in fiction or regional newspapers); dialect phraseology as such is not included in the lexicon. For instance, in (5) *náčeňovou hadrou* '(with a) dish cloth' is a dialectal expression;
- slang: *naprat to pod klacek* (lit. 'to nail it under the stick') 'to hoof the ball into the net';
- other: literary expressions, mainly from the Bible or classical (Greek, Roman) literature, and other sayings: *překročit Rubikon* 'cross the Rubicon'.

(5) lepší než náčeňovou hadrou přes papulu
better than dish.INS cloth.INS over gob
'better than a poke in the eye with the sharp stick'

In addition to the above categories, every word and every MWE can be marked as having an expressive meaning. The words *rachot*, *bengál*, *varvas*, *bordel* denote 'rumble' in different styles and all are expressive. On the other hand, ***vylít** někomu **boty*** (lit. 'pour out one's shoes') 'throw someone out on their ear', consists of non-expressive terms, but the entire MWE is expressive.

Generally, the style values are mutually exclusive except for the expressive value that can be assigned to a word or to a MWE together with some other value.

- Usage type is based on a classification common in the Czech linguistic literature (Čermák 2016) and the lexeme-specific data from Čermák et al. (1983–2009). The following values are distinguished:
  - proverb: *Chybovat je lidské.* 'To err is human.';
  - weather lore: a traditional saying used to predict or interpret weather patterns, or to suggest what people should do on certain dates (6);

- comparison/simile: a collocation typically formed by a verbal or adjectival phrase containing an expression to which something is compared (7);
- citation: part of another text presented verbatim and taken over from literature, film etc.: *Knihy mají své osudy.* (lit. 'Books have own fates.') 'Books have their own destiny.';
- foreign collocation: a collocation taken over unchanged from a foreign language (typically Latin, Greek, English, German, French): (FRE) *raison d'être*, (ENG) *by the way*;
- scientific/professional term: *diferenciální rovnice* 'differential equation';
- multiword function words: used mostly as prepositions or conjunctions: *bez ohledu na* (lit. 'without regard to') 'regardless of';
- (non-specific) verbal MWE: a semantically non-compositional MWE including a verb form as its governor (8);
- non-verbal MWE: a semantically non-compositional MWE not including a verb: *něžné pohlaví* (lit. 'gentle sex') 'the fair sex';
- quasiphraseme: collocation composed of an abstract noun and one of the very limited set of phase verbs (inchoative, durative, terminative): *věnovat pozornost* (lit. 'donate attention') 'pay attention'; it is usually difficult to find single-verb equivalents for these MWEs;
- sentential phraseme: a phraseme differing from a proverb, a weather lore or a citation: *a co ty?* (lit. 'and what you?') 'and what about you?';
- open phraseme/set phrase: a MWE requiring a continuation, typically routine formulation introducing a text or conversation (Coulmas 1981, Aijmer 1996), which is typically further expanded: *jen si představte…* 'just imagine…';
- (usual) collocation: a collocation based on semantic/selectional restrictions only: *úhlavní nepřítel* (lit. 'principal enemy') 'arch-enemy';

(6) Na svatého Jiří vylézají hadi a
On saint.GEN George.GEN creep out snakes.NOM and
štíři.
scorpions.NOM.
'On Saint George's Day the serpents and scorpions creep out.'

(7) líný jako veš
lazy like louse
'lazy as a bear'

(8) na tom nesejde
on it descend.NEG.3SG.PRS
'it makes no difference'

- Adverbial MWEs are classified by the four basic semantic categories (place, time, manner, circumstance):
  - adverbials of place: *na pokraji* 'on the brink of';
  - adverbials of time: *dnem i nocí* 'day and night';
  - adverbials of manner: *po vzoru* 'on the pattern of';
  - adverbials of circumstance: *u příležitosti* 'on the occasion of'.

3.1 Syntactic structure

As another feature inspired by the PARSEME project, each entry is characterized by its syntactic type, i.e. a syntactic category it constitutes in the sentence: NP, AdjP, VP (distinguishing content verb and categorial/light verb phrases), AdvP, PP, or compound preposition/conjunction/interjection, clause, compound sentence. Moreover, for every MWE, the lexicon specifies its dependency and constituency structure, both represented as syntactic trees. Another possible way to capture the syntactic structure of the MWE is a catena (Osenova & Simov 2024 [this volume]). Dependency trees (including syntactic functions) are produced by a parser. In the past, it was TurboParser (Martins et al. 2013), but today it is a parser from the NeuroNLP2 tools (Ma et al. 2018). The parses are manually checked, and then constituency trees are derived from dependency trees using a rule-based conversion system.

Whenever a MWE requires some of its parts to be a lexically unspecified constituent, the syntactic head (verb or adjective) is provided with information on its valency (Rosen & Skoumalová 2018). If necessary, entries may specify the valency of the whole MWE. This is the case, for example, for some constructions consisting of a verb and a nominal or prepositional object: they may take a complement which is required neither by the verb nor by the object. Thus, the MWE in (9) can be complemented, for example, by a *that*-clause, while such a clause can complement neither the verb *dát* 'give' nor the noun *srozuměnou*.

(9) dát na srozuměnou
give on understanding.SG.ACC
'to let know'

Thus, each MWE is described by its syntactic structure, syntactic type and – if syntactically non-standard – also by the kind of its syntactic idiomaticity (see §3.3.3).

3.2 Variability/flexibility

Variability is understood in several different meanings (Hnátková et al. 2017):

- lexical variability: some positions in a MWE can be occupied by synonyms;
- morphological variability: a MWE can possibly occur in various morphological forms;
- word order variability: specific/anomalous free or fixed word order within (parts of) a MWE;
- syntactic transformations: passivization, nominalization, adjectivization, etc.;
- insertion of modifying elements in between the standard MWE template/pattern, i.e. syntactic *modifiability* of MWE components;
- omission of words resulting in *fragments* of standard MWEs.

Unless specified otherwise, we assume that MWEs behave in the same way as regular constructions and contain morphologically standard forms. Hence, we only indicate violations of default properties and rules of grammar. For instance, one of the general properties in Czech is its free word order, thus only specific word order configurations in MWEs are indicated in their lexicon entries.

It is important to account for variability on various levels of linguistic description since one of the objectives of the lexicon is to make it possible to identify not only MWEs in their standard, canonical forms (expressed, for example, in their lemmas) but also their modifications of various kinds. It is often the case that language users modify standard MWEs in a creative way. The lexicon entries cover the kinds of variability listed in §3.2.1, §3.2.2, and §3.2.3 below.

3.2.1 Lexical variability

Lexical variability can be indicated in each lexical position, where appropriate, ranging from a specific word to a choice of several variants (for example synonyms) to a completely free choice determined only by an appropriate word class. A special case of this type of variability is a MWE where a certain lexeme is repeated, while this lexeme can be chosen from several variants (Jelínek 2020). For instance, in the Biblical saying (10) the lexeme *Bůh* 'God' is repeated in the second clause, which has the opposite meaning. This MWE can be seen as a template where both positions occupied by *Bůh* 'God' are in fact containers that might be filled with (almost) arbitrary, but identical nouns (for example *Život dal, život vzal* 'Life gave, life has taken away'; *Bolševik dal, bolševik vzal* 'Bolsheviks gave, Bolsheviks has taken away').

(10) Bůh dal, Bůh vzal.
God gave, God took.
'The Lord gave, and the Lord has taken away.' (the Book of Job 1,21)

Single-word lexical synonyms within a MWE can sometimes have the form of a multiword microstructure, for example a non-reflexive verb can be expressed by its reflexive synonym consisting of a verb and its reflexive particle (*se/si*) as a free morpheme, which need not occupy an adjacent position. This results in different syntactic structures of MWE's synonymous variants and may complicate a successful identification of such a MWE in texts.

3.2.2 Morphological variability

Due to the rich morphological system of Czech, MWEs can occur in various morphological forms, for example verbs can differ in aspect (perfective, imperfective, biaspectual), nouns can appear in various cases or numbers, adjectives or adverbs can occur in the comparative or superlative degree, etc. The morphological richness is illustrated by examples (11), (12) and (13). The following MWE represented by the same lexical entry can appear in two variants, reflected in the entry: in the nominative plural *houby* 'mushrooms' or genitive plural *hub*:

(11) přibývat jako houby/hub po dešti
multiply like mushroom.PL.NOM/GEN after rain.
'to spring up like mushrooms'

For instance, the MWE *nebrat konce* 'to be no end to [something]' can appear in two variants: the noun *konec* 'end' is typically in the genitive of negation

(*konce*), but it can also, rarely, be in the accusative case (*konec*), satisfying the object valency requirement of the transitive verb *brát* 'take':

(12) pořád to *nebere* *konce/konec*
always it take.NEG.3SG.PRS end.SG.GEN/ACC
'there is no end to it'

Paradoxically, fossilized constructions with the obsolete genitive of negation are typical examples of morphological variability.

Verbs in Czech, as in other Slavic languages, express aspect lexically. For instance, a single lexical entry can include both the perfective and imperfective variant of a verb:

(13) *koupit/kupovat* něco *za babku*
buy.PFV/IPFV something for old woman
'to buy something dirt cheap'

Since aspect is a lexical rather than morphological category, aspectual variability is treated as lexical variability. For instance, there are MWEs permitting only one aspectual variant of a verbal lexeme: in (8) the perfective verb form *nesejde* 'descend' cannot be replaced by its imperfective counterpart *neschází*.

3.2.3 Word order variability

Although free word order is a typical trait of Czech, constructions with a fixed word order do exist: the position of prepositions within prepositional phrases, the position of prepositional phrases within noun phrases or the position of clitics within clauses or sentences. Free word order applies to clausal constituents. In the entries, only anomalies concerning both free and fixed word order are captured. For instance, in a MWE consisting of a verb and its syntactic object (14) the verb *dělat* and its object noun *aféru* can appear in either order – this regular syntactic fact is not recorded in the lexicon entry.

(14) *dělat* z něčeho *aféru*
make from something affair.ACC
'to make a big deal about something'

Word order variations can also be due to standard grammar rules (concerning, for example, the position of clitics) or topic-focus articulation. The MWE in (15a) appears in sentence (15b) where the verb and the reflexive particle, components

of the inherently reflexive verb *rovnat se* 'match', occur in the reversed order, separated by the verb form *nemohla* 'could not'. Again, this standard grammatical word order is not indicated in the entry.

(15) a. nemoci se rovnat
can.NEG.INF REFL match
'to be no match for somebody'

b. Co se týče rozpočtů, *se* nepálská studia
What REFL concern.3SG.PRS budgets, REFL Nepali studies
nemohla indickým *rovnat.*
can.NEG.PST Indian.DAT match.INF
'In terms of budgets, Nepali studies could not match Indian studies.'

On the other hand, in the anomalous syntactic structure in (16) the noun *slova* 'word', an attribute in the genitive case, precedes its syntactic head *smyslu* 'sense'; this kind of reversed word order is very rare and appears – as well as other word order anomalies – primarily in MWEs, duly indicated in their lexical entries.

(16) v nějakém slova smyslu
in some.LOC word.GEN sense.LOC
'in some sense of the word'

3.2.4 Syntactic transformations

MWEs related by the same or similar meaning can appear in various syntactic structures that are derived by syntactic transformations from a basic variant. We account for transformations of the following three types, marking only the structures and patterns that are idiosyncratic with respect to the standard grammar of Czech:

- Passivization/depassivization. The following features can be specified in lexical entries:
  - MWE cannot be passivized: a flag specified for MWEs headed by a transitive verb that cannot be passivized in this particular MWE (as an exception to the general rule stating that every transitive verb can be passivized). For instance, the verb *spatřit* 'see' can be passivized in general, but cannot be passivized in (17).
  - MWE cannot occur in the active form, for example the MWE in (18) exists only with the passive form *přáno* 'wished'.

(17) spatřit světlo světa
see light world.GEN
'to come into the world'

(18) nebylo mu přáno
be.NEG.3SG.N.PST he.DAT wish.PASSP
'he was out of luck'

- Nominalization. We assume that a verb in a MWE can be nominalized. If this is not the case, such a verb is flagged appropriately. For instance, in (19) the reflexive verb *hodit se* 'be suitable' cannot be nominalized and this negative fact is recorded in the MWE.

(19) hodit se jako pěst na oko
be suitable REFL like fist on eye
'to be completely out of place'

- Adjectivization. Similarly as with nominalization, it is assumed that generally a verb in a verbal MWE can be adjectivized. If not, such a MWE is marked appropriately. In (20), the impersonal neuter verbal participle *došlo* 'it got to' cannot be adjectivized and this fact is duly recorded as this MWE's property.

(20) *došlo* *na* má *slova*
get.3SG.N.PST on my words
'my words came true'

3.2.5 Insertion

Normally, content words within MWEs can be syntactically modified; typically, adjectives modify nouns, adverbs modify verbs, adjectives or adverbs, etc. Such regular syntactic structures are not reflected in the annotation of MWEs. For instance, the MWE in (21a) can appear in a text as in (21b).

(21) a. *nechávat si* něco *pro sebe*
keep REFL.DAT something.ACC for oneself
'to keep something to oneself'

b. Navrátil *si* krutou informaci *nechával* dlouho jen
Navrátil REFL.DAT cruel.ACC information.ACC keep.PST long.ADV only
pro sebe.
for himself.
'Navrátil kept the harsh information only to himself for a long time.'

In (21b), the modifying adverbs *dlouho* 'for a long time' and *jen* 'only' are inserted in between the components of the standard MWE.

However, there are MWEs whose components cannot be modified, i.e., no insertions in between their components are allowed: this fact is specified in MWE entries where appropriate. For example, in the MWE *něco k snědku* 'something to eat' the monocollocable noun form *snědku* cannot be modified. There are words, however, such as *příslovečný* 'proverbial' or *doslova* 'literally', which can modify almost any MWE of the appropriate syntactic category. Indeed, lexical entries do not specify the availability of such insertions.

3.2.6 Omission/Fragments

MWEs can sometimes appear in their reduced forms – as *fragments*, with the same meaning as the entire MWEs. Our ambition is to recognize MWEs not only in their full, canonical form but also in their partial, fragmentary form. For instance, the lexicon contains the following entry in its standard form:

(22) *hoří* někomu *koudel* *u zadku*
burn.3SG.PRS somebody.DAT oakum.NOM at backside
'somebody is in a tight corner'

Such an entry contains information on possible fragments (represented by identifiers of individual words and of (sub)structures) as central, nuclear parts of the MWE. This approach enables the user to identify even fragmentary MWEs in a text. Thus we can find fragments of standard MWEs such as (23), where only the fragment *hoří koudel* 'burns oakum' remains while the sequence *u zadku* 'at backside' is missing.

(23) Měl asi pocit, že mu *hoří* *koudel*
have.3SG.PST probably feeling, that he.DAT burn.3SG.PRS oakum
kvůli Karlovi.
because of Karel.
'He had a feeling that he is in a tight corner because of Karel.'

In some MWEs there may be two representative fragments that allow for identifying such MWEs in texts. For instance, the standard MWE (24) contains two fragments that might identify the original full-fledged standard MWE: (i) *mazat [někomu] med* 'spread [someone] with honey', (ii) *med kolem huby* 'honey around the gob'. Both fragments are marked in the entries of such MWEs.

(24) *mazat* někomu *med kolem huby*
spread somebody.DAT honey around gob
'soft-soap someone/butter someone up'

In this way, we also capture various modifications leading to reduced versions of standard MWEs, reflecting the authors' creativity.

3.3 Idiomaticity

We stick to the definition of idiomaticity proposed by Baldwin & Kim (2010), adopted also in the PARSEME project:

> In the context of MWEs, idiomaticity refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels. A given MWE is often idiomatic at multiple levels... (Baldwin & Kim 2010: 4)

In particular, we distinguish between lexical, morphological, syntactic, semantic, pragmatic and statistical idiomaticity. The types of idiomaticity used in the PARSEME project were extended by morphological idiomaticity to capture Czech word forms which do not exist outside the specific MWEs. For instance, in the MWE (25) the adjective *pitomá* 'stupid' is a non-inflected feminine form, but in the MWE it is used as an expressive form that morphologically does not agree with the masculine noun form *kluk* 'boy':

(25) kluk pitomá
boy.NOUN.M stupid.ADJ.F
'stupid boy'

Below, the types of idiomaticity are described in detail.

3.3.1 Lexical idiomaticity

Lexical idiomaticity concerns MWEs containing lexically idiomatic word forms or lexemes. The following kinds of lexical idiomaticity are distinguished:

- Monocollocable word forms (26). The word *zadost* 'satisfaction' can exist in this MWE only. Such monocollocable words are often components of terms such as *kysličník osmičelý* 'osmium tetroxide'.

  (26) učinit zadost
  make satisfaction
  'do justice'

- Almost monocollocable word forms, i.e. forms associated with a very limited set of collocates: *zorný úhel* 'angle of vision / point of view'.
- Negative only word forms (27).

  (27) nedílná součást
  undivided part
  'integral part'

- Foreign loans: for example (28), a collocation loaned from German, phonetically and orthographically modified.

  (28) mírnyx týrnyx
  mir nichts dir nichts (GER)
  lit. 'nothing to me, nothing to you'
  'casually / as if it was nothing'

- Macaronic structures: for example, the following collocation consisting of the Latin preposition *per* and the Czech noun *huba* assigned the Latin morph *-m* (29).

  (29) per *huba-m
  via.LAT gob.CZE.F-LAT.F.SG.ACC
  'orally / by word of mouth'

- Other, such as verbatim translations: *potřást hlavou* 'shake one's head' instead of *zavrtět hlavou* 'turn one's head', or adaptations of foreign loans: *mandatorní výdaje* 'mandatory expenses' instead of *závazné/povinné výdaje*.

In the lexicon entry, every lexically idiomatic word form in a MWE is marked. Moreover, a single idiomatic form can be marked with multiple kinds of lexical idiomaticity at the same time. In the lexicon, we also plan to mark each MWE as containing/not containing a lexically idiomatic word form.

3.3.2 Morphological idiomaticity

Morphological idiomaticity concerns a morphologically non-standard morphological form existing only within a MWE. For instance, in the MWE *chca nechca* 'nolens volens', the forms *chca*, *nechca* are non-standard, the standard forms being *chtě nechtě* with the same meaning.

Similarly to lexical idiomaticity, every morphologically idiomatic word form in a MWE is indicated. We also plan to mark each MWE as containing/not containing a morphologically idiomatic word form.

Forms used in MWEs are sometimes licensed by rhyme, as in (30), where *sloupích* 'columns' is a non-standard variant of the standard form *sloupech*.

(30) jména hloupých na všech *sloupích
names stupid.PL.GEN on all.PL.LOC columns.PL.LOC (intended)
'names of the stupid are on all columns'

3.3.3 Syntactic idiomaticity

Syntactic idiomaticity accounts for the following kinds of syntactic anomalies always concerning the entire MWE. They are marked on the MWE where appropriate.

- Anacoluthon: as in the modified New Testament saying (31).

  (31) Kdo po tobě kamenem, ty po něm chlebem.
  who at you stone.INS, you at him bread.INS.
  lit. 'Whoever throws a stone at you, offer him bread.'
  'Do not repay anyone evil for evil.'

- Attraction: as in (32), where the imperative form *padni* 'fall' is repeated in the subordinate clause *komu padni* 'to-whom fall'. The entire construction follows the *imperative wh-word imperative* template, which is realized by several different phrasemes.

  (32) Padni komu padni.
  Fall.IMP who.DAT fall.IMP.
  'Come what may.'

- Idiosyncratic valency: for instance, a noun in the obsolete genitive of negation as object of a negated transitive verb, the standard form being in the accusative case (33).

(33) nemám námitek
have.NEG.1SG.PRS objections.GEN
'I have no objections'

- Aposiopesis: unfinished sentence, as in (34).

(34) Já bych tě nejradši ...
I would you.ACC most preferably
'As for you, I wish I could...'

- Ellipsis:[4]

(35) *Nevím, co* [mám dělat] *dřív.*
Know.NEG.1SG.PRS what [have.1SG.PRS do.INF] sooner.
'I do not know what to do first.'

- Idiosyncratic word order: for instance, an adjective exceptionally (with respect to the grammatical system of Czech) follows its nominal syntactic governor: *mše svatá* (lit. 'mass holy').
- Other: ungrammatical/non-standard syntactic structures, contaminations, zeugmas, etc., such as (36), where the verb form *nevidím* 'I do not see' immediately follows a preposition *od* 'from' and *do* 'to', respectively, thus forming an ungrammatical structure:

(36) od nevidím do nevidím
from see.NEG.1SG.PRS to see.NEG.1SG.PRS
lit. 'from I can't see till I can't see' | 'all the time / without interruption'

3.3.4 Semantic idiomaticity

Semantic idiomaticity concerns a MWE's semantic (non-)compositionality, i.e. (non-)metaphoricity, viewed as the relative frequency of how often the MWE also appears in its compositional/literal meaning (as to the degree of compositionality of nominal MWEs, cf. also Schulte im Walde (2024 [this volume])). We use the following scale:

[4]The brackets in example (35) are used for marking the ellipsis.

- MWE is always non-compositional, i.e. always idiomatic – the situation described by the MWE can never happen in the real world:

  (37) mít ocelové nervy
  ‘to have nerves of steel’

- MWE is rarely compositional, i.e. it is often idiomatic:

  (38) *strouhat* někomu *mrkvičku*
  grate somebody.DAT carrot
  ‘to express Schadenfreude’

- MWE is often compositional, i.e. rarely idiomatic:

  (39) hrát si na schovávanou
  play REFL at hide and seek
  ‘to play hide and seek’

- MWE is always compositional, i.e. non-idiomatic, literal:

  (40) dlouhodobá investice
  ‘long-term investment’

3.3.5 Pragmatic idiomaticity

A MWE is pragmatically idiomatic if it is used in specific situations. For instance, a standard invitation to a dance sounds as in (41).

(41) Smím prosit?
May.1SG ask?
‘May I have the pleasure (of this dance)?’

3.3.6 Statistical idiomaticity

Usual, frequent, semantically non-idiomatic collocations reflecting selectional restrictions in usage fall within this category. Some of their components have a very limited collocability potential. The components of such MWEs can hardly be replaced by synonyms, for example *vydatný déšť* ‘heavy rain’, or similarly in (42) where the adjective *dezolátní* is unlikely to be replaced by a synonym. In addition to usual collocations, we regard as statistically idiomatic also terms such as *bezkontextová gramatika* ‘context-free grammar’, and multiword function words (multiword prepositions and conjunctions).

(42) být v dezolátním stavu
be in desolate state
'be in a state of neglect'

4 Design of the database

4.1 Basic data model

For full flexibility required by the potential variability of the expressions (see §3.2), we define the entry pattern by means of *slots* and *fillers*.

The entry unit consists of *slots* and *features* referring to the MWE as a whole. Slots represent the 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 a particular component (slot). The primary role of fillers is to represent actual (terminal) tokens to be matched in the data. They are defined by means of a combination of token attributes and their values that must be matched in the text data in order to identify the MWE as a whole (for slots and fillers see examples in §4.3.2). Other possible restrictions, such as those concerning word order, modifications or transformations, can be defined by means of additional features. Figure 1 shows the scheme of the entry structure.[5]

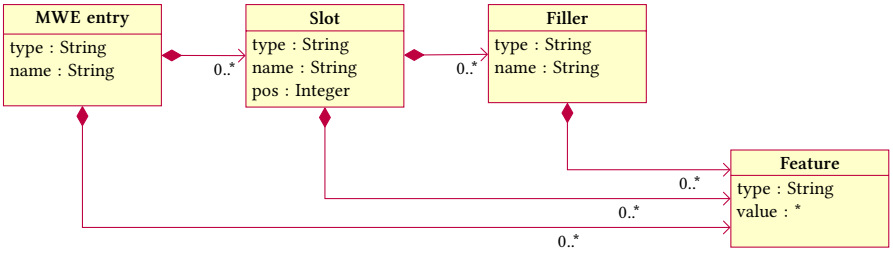

Figure 1: MWE entry structure (basic model)

All the container objects (entries, slots and fillers) have an arbitrary *name* and a type. Types are defined as a path in a hierarchy of categories defined in a separate metadatabase. This helps to achieve a better organization and systematization of object types. For example, the atomic features may be easily classified by

[5]The structure follows (in a simplified form) basic principles of the proposal for a structured lexical description presented first in Vondřička (2014) and has been described in full detail in Vondřička (2019). In the current version of the database, the structure has been further simplified mainly by replacing filler attributes and references by dedicated features.

the linguistic layers they belong to (form, morphology, syntax, semantics, pragmatics, statistical properties, etc.). At a different level of classification, they can be grouped, for example by a particular purpose, linguistic theory or relative to a particular corpus. This also allows us to store multiple similar features from different sources or for different purposes at the same time. Features can easily be used (and classified) both for purely technical purposes of NLP processing tools and for storing information aimed at human users of the database, such as definitions, examples or notes.

In case we need to include multiple alternative values of some type of feature, additional custom subspecification may be used. This applies especially to user notes, examples from real texts or statistical values. For example, the basic type of feature 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, for example `:stats:fq:abs:BNC:fiction`. This allows the database to be searchable by features using underspecification of the type (by means of a path prefix) or its full (sub)specification as needed.

As described above, the fillers are expected to match more or less specific tokens in the text data. In our case, the data is already morphologically analyzed and disambiguated. This allows for underspecification of token attributes to be matched by using incomplete matching patterns or even regular expressions. We can, for example, match some lemma generally, independently of its particular morphological form (which is especially useful for verbs), or we can restrict the form more finely to some specific morphological category (number, case, person, etc.). It is also possible to match just some particular part of speech, such as adjective, demonstrative pronoun, etc. In case of specific valency requirements, it would be even more practical to match just whole syntactic phrases of a particular type instead of listing all their possible morphological realizations. While this is also possible in principle, unfortunately, we do not have a syntactic parser for Czech reliable enough to build upon. Therefore we need to identify MWEs solely by their realization in form of tokens on the surface level.

4.2 More advanced variability and structures

The basic model as described above makes it possible to deal with variability only at the level of single tokens, since one slot only corresponds to a single token possibly realized by various (single token) fillers. However, variability often concerns multiple tokens: prepositional phrases, periphrastic word forms, etc.

Instead of implementing a recursive database structure, we decided to keep it flat for practical reasons.[6] Instead, we implement recursion on top of the flat structure: we allow fillers to refer to other slots or their sequence. This effectively creates non-terminal fillers (and potentially slots) within the structure and allows us to build a kind of tree structure. In this way, we can define components grouping alternative multi-token variants.

Since both slots and fillers can also be typed, we can easily differentiate terminal and non-terminal slots (and fillers) of different types. This allows us to define additional virtual structural relations among the terminal tokens such as constituency structures for potential syntactic analysis. A side effect of this "broken" virtual recursion is thus the possibility to define multiple alternative (full or partial) tree structures of the core terminal slots, with all the obvious advantages and disadvantages.[7]

More complex dependencies have also been already registered, for example, several optional components which may either occur exclusively, but not all at the same time, or which must actually either appear all together, or not at all. Another type is represented by example (10) (cf. §3.2.1), showing a variable component used repeatedly. Some of these could be (in theory) easily marked at the syntactic level, but as explained above, we can currently only rely on the surface form (with morphological analysis at its best) and therefore we need more primitive methods to group, relate and classify some slots using additional dedicated supporting features to give proper hints to the parser.

As mentioned in §3.2.6, the creativity (or lack of knowledge) of language users may eventually go far beyond the bounds of any common variability and the MWE may be modified or reduced up to the point where it is just barely recognizable as the original MWE, so that we call it a *fragment*. For this purpose, we add another special feature for each more complex entry: the minimal list(s) of the necessary components which must necessarily occur in the text in order to make an association with the original MWE possible at all.

[6] Indexing, querying and processing recursive data structures is still a demanding task, not very well and efficiently supported by the current database and search engines.

[7] Among the advantages: multi-purpose or multi-theory use and multi-dimensionality of the core database; disadvantages: additional complex requirements on consistency and validity management, need for interpretation and filtering of the basic data on higher application levels. Querying the structure of the MWEs would also be rather difficult to implement, but this functionality is currently not needed.

4.3 Implementation

The database has been implemented as a part of a more generic database of corpus annotation units, sharing a common infrastructure and principles. ElasticSearch[8] is used as back-end 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 latest front-end user interface is designed using ReactJS.[9] It uses the API and metadata about all defined object types (a kind of *configuration* also managed by the API) to create a customized and highly configurable user interface on the fly.

4.3.1 Populating the database with entries

To populate the LEMUR database with entries, we use an automatic conversion from the FRANTALEX lexicon, which contains lemmas and tags describing the syntactic type of the lemmas. Lemmas in FRANTALEX are divided into individual variants, for example, *jít přes čáru* 'to cross the line' and *být přes čáru* 'to be beyond the line', whereas in LEMUR we group these variants under one lemma. Also, tags for syntactic types are converted into lemma descriptions. Syntactic structures are generated using a dependency parser. After a manual check they are automatically converted to constituent tree structures. The rest of the information about each lemma has to be added manually.

The FRANTALEX lexicon consists of about 49,000 lemmas, of which about half have been transferred. LEMUR contains about 16,000 lemmas, but these include grouped lemmas from the original lexicon. A test corpus, which corresponds to the SYN2020 corpus (Jelínek et al. 2021), is annotated with more than 1.3 million collocations from the FRANTALEX lexicon and more than 722,000 collocations from the LEMUR database.

4.3.2 User interface

The user interface shows all important information about a lexical entry (its components and features) in a form suitable for human readers. Figure 2 shows the lemma *mazat někomu med kolem huby* (lit. 'spread honey around someone's gob') 'butter somebody up'. Individual words, fillers, fill the numbered slots, where some positions can be occupied by several synonyms, variant fillers. For instance,

[8] https://www.elastic.co
[9] https://reactjs.org

slot [1] is filled with the variant verb fillers *mazat* and *namazat* ‘smear’, slot [5] is filled with the variant noun fillers *huby* ‘gob’, *pusy* ‘mouth’ and *ústa* ‘mouth’. Below the lemma, definitions and examples from the corpus are given, as shown in Figure 3. This is followed by an option to search for examples in the corpus.

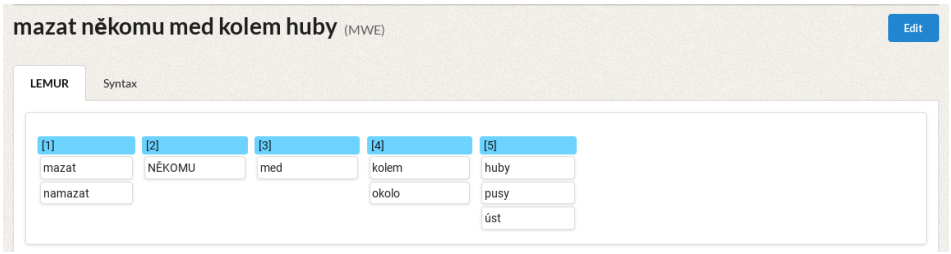

Figure 2: MWE lemma in user interface

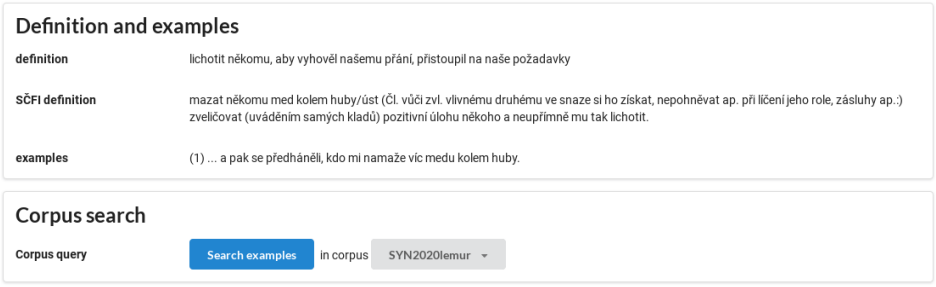

Figure 3: Definitions, examples and search

The interface also dynamically generates charts representing syntactic structures of the MWE from its flat list of slots and their fillers and links (relations) between them. In Figures 4 and 5 we show the dependency and constituent structures of the phrase *bojovat pro čest a slávu* ‘fight for honour and glory’ with all the lexical variants in place of the verb as well as the preposition.

In these charts, blue nodes represent terminal slots of the MWE indicating also their actual possible fillers (for example the variable choice of prepositions *pro*, *za* and *o* in the slot [2]). The dark yellow slots represent the (non-terminal) phrase nodes in the constituent structure. The light yellow non-terminal slot [V] represents the verb, which may be realized by three different types of verbs: (1) simple non-reflexive verbs (*bojovat* ‘fight’, *zápolit* ‘compete, wrestle’, etc.), (2) reflexive verbs using the accusative reflexive pronoun *se* (*bít se* ‘struggle, wrestle’, *rvát se* ‘brawl’) and (3) a reflexive verb using the dative reflexive pronoun

si (*zahrát si* ‘play, act the part of’). Since the latter two types consist of two tokens, a simple list of fillers within a singular terminal slot would not be sufficient. Therefore, the verb slot is defined as a non-terminal *variant-slot*, which branches both charts into three alternative sub-trees numbered by the respective fillers *1*, *2* and *3* (shown as small elliptical yellow nodes).[10]

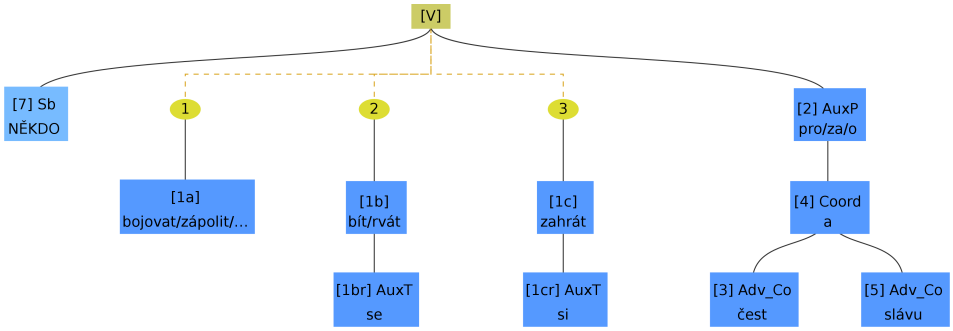

Figure 4: Dependency structure with multiple choices

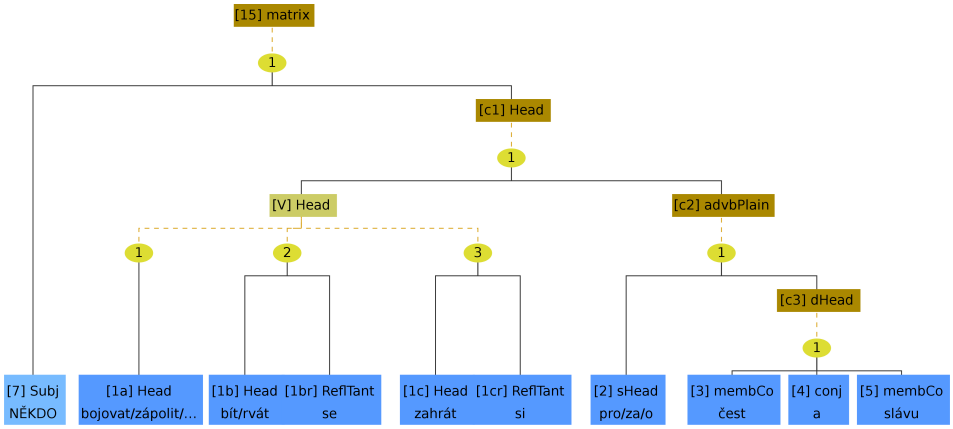

Figure 5: Constituent structure with multiple choices

In addition to the browsing mode, the database also allows editing of individual entries. In the editing mode, there is more information available that is not

[10] In Figure 5 (constituent structure), the fillers of all the non-terminal slots are shown as elliptical yellow nodes purely for consistency reasons, despite the fact that there is otherwise always just one filler in each of the slots and therefore no other case of branching. Terminal (single token) slots do not have their fillers branched out externally in order to keep the tree as compact as possible. For other caveats concerning the visualizations see Vondřička (2019).

normally displayed, but can be queried when searching the lexicon. For example, various grammatical constraints, such as the occurrence of the verbal MWE only in the active voice, or only in the singular, or only in the 3rd person, etc., are expressed by constraints on the morphological tag or on the "verbtag", a positional attribute which describes the properties of the entire verb form (whether simple or compound) such as person, voice or tense (see Jelínek et al. 2021). In Figure 6 we see the collocation *vyjít najevo* 'come to light', where we use a verbtag expressed by a regular expression to set the constraint that the verb can occur only in the 3rd person, in the active voice, or in the infinitive.

slot 1		**slot 2**	
lemma:	**vyjít**	lemma:	**najevo**
tag:	**V**	tag:	**D**
verbtag:	**V.A3.. | VFA---**		

Figure 6: Morphological constraints on a MWE member

5 Practical use of the LEMUR lexicon

The lexicon can be used as a standard phraseological lexicon, but it can also be used in corpus annotation and it also has potential applications in NLP.

5.1 Annotation of MWEs in corpora and linking with the lexicon

Occurrences of MWEs in text corpora are identified and the corpus annotation is extended by the MWE lemma and type, assigned as new attributes of every token recognized as part of a MWE (in addition to the standard annotation of individual words in terms of POS tags and lemmas). Corpus users can then search for MWEs by their MWE lemma (if they know it) or they can combine various types of linguistic annotation in one search, such as a verb in imperative which is a part of a syntactically idiomatic MWE or any form of the noun *holub* 'pigeon' being part of a MWE. Using the Corpus Query Language in the KonText search environment (Machálek 2020) of the Czech National Corpus, the latter query would be specified as `[lemma="holub" & mwe_lemma!=""]` (`mwe_lemma` is not empty, i.e. the token with the lemma *holub* is part of an identified MWE). The user would thus find several MWEs in their context such as *pečení holubi lítají do huby* (lit. 'roasted pigeons fly into the mouth') 'expectation of profit without effort' or *točit se jako holub na báni* (lit. 'to turn around as a pigeon on a temple dome') 'to turn around constantly'.

Each MWE occurrence in the corpus is linked to the corresponding lexical entry in the lexicon, so that the corpus user can consult the lexicon directly, see Figure 7 (`mwe_lemmas` are shown in bold characters after slashes). In the opposite direction, it is possible to view occurrences of a given MWE in a corpus when browsing the MWE lexicon, as shown above in Figure 3.

Figure 7: Corpus concordance linked to MWE lexicon

5.2 Use of the lexicon in POS tagging and parsing

A morphological tagger may use a module that identifies some frequent MWEs in order to decide about the most likely tag using the knowledge of such MWEs rather than general linguistic rules or a stochastic model unaware of these phenomena (Hnátková & Petkevič 2017). Since MWEs are sometimes morphologically or syntactically irregular as in (36), their identification, including tagging with a special module, helps to increase the tagging accuracy of the whole corpus. For instance, in (43) the general morphosyntactic rules of Czech cannot fully disambiguate case and number of the noun *bratrství* ‘brotherhood’ (following the preposition *na* ‘on’, requiring accusative or locative, the noun *bratrství* can be interpreted as ACC.SG, LOC.SG or ACC.PL), whereas the morphologically fully disambiguated entry *připít na bratrství* helps to disambiguate this MWE within a sentence as indicated:

(43) připít na bratrství
drink on.ACC-VAL brotherhood.SG.ACC
'raise one's glass to brotherhood'

The proverb in (44) includes *štěstí* 'good luck', a highly syncretic noun form, whose interpretations are difficult to disambiguate without a MWE lexicon listing the disambiguated morphological categories. Even in this context, but without the knowledge of the proverb, the ambiguous form *štěstí* 'good luck' can mistakenly be parsed as dative, modified by *odvážnému* 'to the brave'.

(44) Odvážnému štěstí přeje.
brave.M.SG.DAT good luck.NOM favour.3SG.PRS
'Fortune favours the brave.'

5.3 Use of the lexicon in parsing

The lexicon of MWEs contains information about the syntactic structure of each MWE. In parsing, this information can be used to automatically correct the syntactic annotation of MWEs by comparing the annotation made by the parser with the annotation specified in the lexicon for each identified MWE. If they differ, the automatic annotation can be replaced with the annotation from the lexicon, unless this would result in an overall incorrect structure, such as a looped tree, in which case the correction is not performed.

As an example, consider a simple MWE type: a noun followed by an adjective. This is a typical structure of Czech terms, for example *anděl strážný* (lit. 'angel guardian') 'guardian angel', *kudlanka nábožná* (lit. 'mantis devout') 'praying mantis', *kyselina sírová* (lit. 'acid sulphuric') 'sulphuric acid', etc. However, apart from terms, the typical word order in Czech is adjective–noun. The parser cannot acquire sufficient "knowledge" of Czech terms from the limited training data, and even the use of methods based on word embeddings (creating a mathematical representation of words using extensive "raw" language data, see Mikolov et al. 2013) does not completely remove this handicap. When a term of the noun-adjective type is followed by another noun, or by an adjective and a noun, the parser decides in 58% of cases that the adjective pre-modifies the noun to its right, sometimes even when the adjective cannot agree with the following noun in number, gender or case, as in *představený kláštera Matky Boží řádu trapistů* 'the abbot of the monastery of the Mother of God of the Trappists', including the term *Matka Boží* 'Mother of God', where the parser identified the adjective *Boží* 'of God', 'divine' as a modifier of the following noun *řádu* 'order', instead of the

preceding noun *Matky* 'mother'. By providing the correct syntactic structure for *Matka Boží*, the lexicon could be used to rectify this error.

The parser used for syntactic annotation is based on neural networks (Ma et al. 2018) and trained on the data of the Prague Dependency Treebank (Hajič et al. 2018). Its overall results reach the state of the art but it struggles to correctly parse MWEs with unusual syntactic structures.

Experiments in correcting syntactic annotation were performed in the past (Jelínek 2019): syntactic structures of MWEs identified in corpora were checked against the syntactic structures exported from the MWE lexicon. When they differed, the (supposedly erroneous) structures were replaced by the structures from the MWE lexicon. The whole sentence was then checked to make sure an incorrect sentence structure did not result from this intervention. Manual analysis of the results showed that using the information from the lexicon, syntactic annotation was corrected in 88% of the identified syntactic annotation errors for MWEs, while in only 2% of the cases was an error introduced into the annotation by the intervention. Overall, however, the number of interventions was small (partly due to the relatively low number of MWEs in the lexicon at the time of the experiment) and the overall success rate of the syntactic annotation was almost unaffected by the experiment (less than 1 word per 100,000 was corrected in this way). However, there are now significantly more MWEs in the lexicon, so we consider applying the module for the automatic correction of MWE parses in the next syntactically annotated corpus due to be released in 2025. This will still mean a relatively small improvement in the overall success rate, but we expect that the syntactic annotation of MWEs will improve noticeably, especially since some structures in MWEs are really unusual and thus unmanageable for the parser. This has not been tested yet, however.

6 Conclusion

To answer the need for a lexicon of Czech MWEs, we designed and implemented a lexical database, coping with the variability and structure of multiword lexemes. To achieve that, lexical entries support descriptions from a number of angles. Thus, each entry specifies aspects such as the MWE's lemma, definition, examples, style, syntactic structure, idiomaticity and variability.

Following the taxonomy proposed by Baldwin & Kim (2010) and used in the PARSEME project, we use multiple types of both idiomaticity and variability, i.a. lexical, morphological or syntactic. While the types of idiomaticity describe the MWE's inherent properties, lexical specifications of variability describe the

MWE's behaviour in language use. This concerns the cross-linguistically common phenomena of internal modification (insertion) and the use of MWE fragments (omission).

In addition to its use in a standard way for lexical lookup, the lexicon can also be used as a resource for various NLP tools, such as taggers or parsers. Moreover, lexical entries can be linked with occurrences of the multiword lexemes in a corpus, supporting both lexical lookup and corpus search directly from the corpus or the lexicon, respectively. There are also plans for LEMUR to be linked with an emerging standard reference lexicon of Czech: the *Academic dictionary of Contemporary Czech* (Kochová & Opavská 2016a,b).[11]

Last but not least, the lexicon is being extended by adding new entries or by specifying additional features within existing entries. This (to a large extent manual) effort gradually alleviates the problem of insufficient coverage: the current number of tokens at the time of writing approaches 16,000, while the number of MWE occurrences identified and annotated using the FRANTALEX lexicon in the SYN corpus release 11 is about 49,000. However, we need a strategy for further expanding the lexicon. The lacunae that come up most often in real texts deserve to be filled first, thus helping to reach a better coverage with least efforts.

In the near future, we will add all FRANTALEX lemmas to LEMUR and use the LEMUR database to annotate a new experimental version of SYN corpus (Hnátková et al. 2014). This corpus will be accessible to interested lexicographers and linguists. The feedback they provide will be valuable for the further development of the lexicon.

The still insufficient coverage aside, we believe that LEMUR is built on solid foundations and hope that it turns out to be a useful resource for many purposes. Eventually, its design and structure may serve also other languages than Czech.

Abbreviations

ACC-VAL	valency (required) accusative
AdjP	adjective phrase
AdvP	adverbial phrase
GER	German
ENG	English
LAT	Latin
MWE	multiword expression
NP	noun phrase
NLP	natural language processing
PASSP	passive participle
PP	prepositional phrase
VP	verb phrase

[11] https://slovnikcestiny.cz/

Acknowledgements

Thanks to the reviewers as well as the editors for their valuable comments that helped to improve and enrich this paper.

The work on this paper was supported by the grant *Czech National Corpus* LM2023044 under Large Research, Development and Innovation Infrastructures program, and by the grant *Multiword Units for Digital Learning* TQ01000177 under TA ČR VS SIGMA - DC3 program.

References

Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity*. London: Routledge.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Fred J. Damerau & Nitin Indurkhya (eds.), *Handbook of natural language processing*, 2nd edn., 267–292. Boca Raton, FL: CRC Press, Taylor & Francis Group.

Bozděchová, Ivana. 2007. Teorie terminologie v historických a obsahových proměnách. In *Sborník příspěvků věnovaných profesorce PhDr. Marii Čechové, DrSc.* 65–74. Univerzita J. E. Purkyně, Ústí nad Labem.

Burger, Harald, Dmitri Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.). 2007. *Phraseology: An international handbook of contemporary research*. Berlin: Walter de Gruyter.

Čechová, Marie. 2011. *Čeština: Řeč a jazyk*. Praha: SPN.

Čermák, František. 2007. *Czech and general phraseology*. Prague: Karolinum.

Čermák, František. 2016. Frazeologie a idiomatika. In Petr Karlík, Marek Nekula & Jana Pleskalová (eds.), *Nový encyklopedický slovník češtiny*, 1st edn., 530–532. Praha: Nakladatelství Lidové noviny. https://www.czechency.org/slovnik/FRAZEOLOGIE%20A%20IDIOMATIKA.

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

Colson, Jean-Pierre. 2017. The IdiomSearch experiment: Extracting phraseology from a probabilistic network of constructions. In Ruslan Mitkov (ed.), *Computational and corpus-based phraseology*, 16–28. Cham: Springer.

Coulmas, Florian (ed.). 1981. *Conversational routine: Explorations in standardized communication situations and prepatterned speech*. The Hague: Mouton.

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 & Zdeněk Žabokrtský. 2018. *Prague Dependency Treebank 3.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2621.

Hnátková, Milena. 2002. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky. *Slovo a Slovesnost* 63(2). 117–126. http://sas.ujc.cas.cz/archiv.php?art=4064.

Hnátková, Milena, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová & Pavel Vondřička. 2017. Eye of a needle in a haystack. In Ruslan Mitkov (ed.), *Computational and corpus-based phraseology*, 160–175. Cham: Springer. DOI: 10.1007/978-3-319-69805-2_12.

Hnátková, Milena, Michal Křen, Pavel Procházka & Hana Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)*, 160–164. Reykjavík: ELRA. https://aclanthology.org/L14-1267/.

Hnátková, Milena & Vladimír Petkevič. 2017. Morphological disambiguation of multiword expressions and its impact on the disambiguation of their environment in a sentence. *Jazykovedný Časopis* 68(2). 145–155. DOI: 10.1515/jazcas-2017-0025.

Jelínek, Tomáš. 2019. Using a database of multiword expressions in dependency parsing. In Kamil Ekštein (ed.), *Text, speech, and dialogue: 22nd international conference*, 19–31. Cham: Springer. DOI: 10.1007/978-3-030-27947-9_2.

Jelínek, Tomáš. 2020. Multi-word lexical units with repetition of lexemes in Czech and identification of their variants. In Joanna Szerszunowicz & Martyna Awier (eds.), *Reproducible multiword expressions from a theoretical and empirical perspective*, 141–153. Białystok: University of Bialystok. http://hdl.handle.net/11320/11351.

Jelínek, Tomáš, Marie Kopřivová, Vladimír Petkevič & Hana Skoumalová. 2018. Variabilita českých frazémů v úzu. *Časopis pro moderní filologii* 100(2). 151–175. https://casopispromodernifilologii.ff.cuni.cz/magazin/2018-100-2-2/.

Jelínek, Tomáš, Jan Křivan, Vladimír Petkevič, Hana Skoumalová & Jana Šindlerová. 2021. Syn2020: A new corpus of Czech with an innovated annotation. In Kamil Ekštein, František Pártl & Miloslav Konopík (eds.), *Text, speech, and dialogue*, 48–59. Cham: Springer. DOI: 10.1007/978-3-030-83527-9_4.

Klégr, Aleš. 2016. Lexikální kolokace: Základní přehled o vývoji pojetí. *Časopis pro moderní filologii* 98(1). 95–103. http://hdl.handle.net/20.500.11956/96860.

Kochová, Pavla & Zdeňka Opavská. 2016a. Akademický slovník současné češtiny: Z přípravy Akademického slovníku současné češtiny. *Naše řeč* 99(2). 57–83. https://ujc.avcr.cz/miranda2/export/sitesavcr/ujc/zakladni-informace/pracovnici/files/KochovaOpavskaASSC.pdf.

Kochová, Pavla & Zdeňka Opavská (eds.). 2016b. *Kapitoly z koncepce: Akademického slovníku současné češtiny*. Praha: Ústav pro jazyk český AV ČR.

Kopřivová, Marie & Milena Hnátková. 2014. From dictionary to corpus. In Vida Jesenšek & Peter Grzybek (eds.), *Phraseology in dictionaries and corpora*, 155–168. Maribor: Filozofska fakulteta Maribor.

Kováříková, Dominika. 2017. *Kvantitativní charakteristiky termínů*. Praha: Nakladatelství Lidové noviny – Český národní korpus.

Kováříková, Dominika & Oleg Kovářík. 2019. Automatic identification of academic phrases for Czech. In Gloria Corpas Pastor & Ruslan Mitkov (eds.), *Computational and corpus-based phraseology (EUROPHRAS 2019)* (Lecture Notes in Computer Science 11755). Cham: Springer. DOI: 10.1007/978-3-030-30135-4_17.

Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.

Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.

Lopatková, Markéta, Václava Kettnerová, Eduard Bejček, Karolína Skwarska & Zdeněk Žabokrtský. 2014. *VALLEX 2.6.3: Valency lexicon of Czech verbs*. Prague: Karolinum Press.

Ma, Xuezhe, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig & Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics*, 1403–1414. Melbourne: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1130.

Machálek, Tomáš. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the twelfth Language Resources and Evaluation conference*, 7003–7008. Marseille: ELRA. https://aclanthology.org/2020.lrec-1.865.

Martins, André, Miguel Almeida & Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, 617–622. Sophia: Association for Computational Linguistics. https://aclanthology.org/P13-2109/.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*. DOI: 10.48550/arxiv.1301.3781.

Moon, Rosamund. 2007. Corpus linguistic approaches with English corpora. In Harald Burger, Dmitri Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.), *Phraseology: An international handbook of contemporary research*, 1045–1059. Berlin: Walter de Gruyter.

Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.

Pasquer, Caroline, Agata Savary, Jean-Yves Antoine & Carlos Ramisch. 2018. Towards a variability measure for multiword expressions. In Marilyn Walker, Heng Ji & Amanda Stent (eds.), *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, Volume 2 (Short Papers)*, 426–432. New Orleans, LA: Association for Computational Linguistics. DOI: 10.18653/v1/N18-2068.

Popovičová, Snežana. 2020. *Česká a srbská frazeologie: Na cestě ke dvojjazyčnému frazeologickému slovníku*. Praha: Karolinum.

Przepiórkowski, Adam, Jan Hajič, Elżbieta Hajnicz & Zdeňka Urešová. 2017. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography* 30(1). 1–38. DOI: 10.1093/ijl/ecv048.

Rosen, Alexandr & Hana Skoumalová. 2018. No way to have your say out of the frame: Specifying valency of multi-word expressions. *Prace Filologiczne* 2018(72). 301–320. https://www.ceeol.com/search/article-detail?id=732446.

Schulte im Walde, Sabine. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 269–308. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998645.

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 35–68. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579035.

Temmerman, Rita. 2000. *Toward new ways of terminology description: The sociocognitive approach*. Amsterdam, Philadelphia: John Benjamins.

Urešová, Zdeňka. 2009. Building the PDT-VALLEX valency lexicon. In *On-line proceedings of the fifth corpus linguistics conference*. University of Liverpool. https://ufal.mff.cuni.cz/~uresova/web.pdf/2012-CLC-Building%20the%20PDT-VALLEX.pdf.

Vondřička, Pavel. 2014. *Formalized contrastive lexical description: A framework for bilingual dictionaries*. München: LINCOM.

Vondřička, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101. https://ufal.mff.cuni.cz/pbml/112/art-vondricka.pdf.