Multiword Expressions in Czech: Typology and Lexicon

Milena Hnátková, Tomáš Jelínek, Marie Kopřivová, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová and Pavel Vondřička

Faculty of Arts, Charles University, Prague

SlaviCorp 2018, September 24–26, 2018
Outline of the talk

Two main parts:

- **typology** of multiword expressions in Czech (V. Petkevič)
- presentation of the *lexical database* (P. Vondřička)
Need of a MWE lexicon for NLP tools (and their users)

MWEs play a significant role in any language, often standing in contrast to its standard grammatical properties.

Objectives:
- Description, typology and classification of MWEs
- Improvement of POS tagging, parsing, word sense disambiguation, semantic tagging
- Identification and search of MWEs in their standard form, but also of their fragments and variants – morphological, syntactic and lexical
Correct structure: \([\text{kyselinou } \text{Attr}[\text{šírovou}]]\)

‘sulphuric acid’
### Motivation

**MWE tagging in a corpus**

<table>
<thead>
<tr>
<th>Hits: 49,887,368</th>
<th>l.p.m. 0: 10,845.92 (related to the whole &quot;syn_v5&quot;)</th>
<th>ARF 0: 30,247,135.4</th>
<th>Result is shuffled</th>
</tr>
</thead>
</table>

#### Example:

- **Hospodářské noviny**
  - "(NYSE) se na tuto burzu ve skutečnosti odhrazuje méně než 35 procent." - Rychlost obchodů

- **Deníky Bohemia**
  - "V těchto dnech mají plně ruce práce, důležitá úkoly." - Technických služeb

- **Blesk**
  - "V závěru svého života." - Úspěchy

- **ForMen**
  - "Máme na způsobu jmenováný člen štýrského klubu." - Přístupové pravidla

- **Právo**
  - "V lekárnách nepochoď." - Jídelní lístky

- **24 hodina**
  - "Začíná se v poslední době." - Bude li štěstí na této události?

- **Marketing & Media**
  - "Máme na způsobu jmenování členů štýrského klubu." - Přístupové pravidla

- **Mladá fronta DÍNE**
  - "V budoucnu býchom se chutě podobat." - Historie

- **Deníky Bohemia**
  - "Ani jedno dobré slavo dobře slavo." - Čtenářům

- **Mladá fronta DÍNE**
  - "Čerstvě se tento vědění." - Konference

- **Deníky Bohemia**
  - "Novoroční VÝŠLAP na Zelenou horu si nenechalo." - Zábava
Identification of fragments and variants

Base MWE:

$vzít \ \text{rozum \ do \ hrsti}$
‘keep one’s wits about one’
lit. ‘take reason into the cupped hand’

Fragment:

$rozum \ \text{do \ hrsti}$
lit. ‘reason into the cupped hand’

Variant:

$Sebral \ \text{jsem \ ještě \ rozum \ do \ hrsti \ a \ utekl.}$
‘I still kept my wits about one.’
lit. ‘I gathered still reason into the cupped hand and fled.’
MWE lexical database

The desired content of the lexical database:

- 7000 MWEs in Czech with
- a detailed description of every MWE based on
- MWE classification/typology
- Focus on variability

Sources:

- *Slovník české frazeologie a idiomatiky* (Dictionary of Czech collocations and idioms) by F. Čermák et al.
- *FRANTALEX* – a list of idioms and collocations
- *Czech National Corpus*
- *Vallex* – a valency dictionary of Czech
Classification/Typology of MWEs

- We adopt the 3D classification used in the PARSEME\(^1\) project:
  - syntactic structure
  - fixedness/flexibility
  - idiomaticity

- Style/register typology
- Usage/global type typology

\(^1\)see https://typo.uni-konstanz.de/parseme/; Baldwin, T., Kim, S.N.: Multiword expressions. 2010
Style/Register

Standard:  *hořet zvědavostí*
‘be consumed with curiosity’
lit. ‘burn with curiosity’

Colloquial:  *brát něco hákem*
‘go at sth hammer and tongs’
lit. ‘take sth with a hook’

Dialect:  *bejt jako boží vědro*
‘be totally drunk’
lit. ‘be like a God’s bucket’

Expressive:  *dřina jako prase*
‘hard work’
lit. ‘toil like a pig’

Slang

Other
Usage/Global type

Proverb:  *(mít) co na srdci, to na jazyku*
‘wear one’s heart on one’s sleeve’
lit. ‘(have) what on heart, it on tongue’

Weather lore:  Únor bílý, pole sílí.
‘White February, fields are strengthened.’

Comparison/simile:  *držet se koho jako klíště*
‘cling to sth like a barnacle’

Citation:  *Nevědomost hříchu nečiní.*
‘Ignorance is bliss.’

Set phrase:  *na tom nesejde*
‘it makes no difference’; lit. ‘on it doesn’t descend’
Syntactic type

Noun phrase: \( \text{žabí muž} \)
  ‘frogman’

Adjectival phrase: \( \text{všeho schopný} \)
  ‘capable of anything (bad)’

Verb phrase: \( \text{stát za starou belu} \)
  ‘not worth a damn’

Verbonominal (light verb) constructions

Adverbial phrase
...

MWE lexical database  Syntactic structure

Hnátková et al.  Multiword Expressions in Czech...
SlaviCorp 2018  11 / 35
Syntactic structure and Valency

In the lexicon, every MWE is represented as
- a dependency tree
- a phrase-structure tree

Valency is included in syntactic trees, sometimes it is idiosyncratic:

*dát na srozuměnou, že...*
give on understanding that
‘to make it clear that...’

Neither *dát* nor *srozuměná* has a *that*-clause in their valency.
Fixedness/flexibility

- Variants
- Fragments
- Word order
- Internal modifiability
- Transformations
- Morphological restrictions
Word order

Czech is a language with “free word order”, but some structures/words have a fixed word order:

- inside NPs and PPs the word order is fixed, e.g.
  - Preposition precedes a Noun
  - Adjective precedes a Noun
- clitics are placed on the 2nd position

Only deviations from the standard word order are marked in lexical entries:

Constituents cannot be shifted:
  *nedá se svítit*

  ‘nothing can be done’; lit. ‘it’s impossible to (make) light’

Noun precedes adjective:
  *skokan zelený*

  ‘edible frog’; lit. ‘jumper green’
Word order

Czech is a language with “free word order”, but some structures/words have a fixed word order:

- inside NPs and PPs the word order is fixed, e.g.
  - Preposition precedes a Noun
  - Adjective precedes a Noun
- clitics are placed on the 2nd position

Only deviations from the standard word order are marked in lexical entries:

**Constituents cannot be shifted:**

*nedá se svítit*

‘nothing can be done’; lit. ‘it’s impossible to (make) light’

**Noun precedes adjective:**

*skokan zelený*

‘edible frog’; lit. ‘jumper green’
Internal modifiability

By default, content words may be modified:

\textit{uhodily} [\textit{třeskuté}] \textit{mrazy}

\textquote{[bitter] frosts struck’}

We mark only exceptions:

\textit{běžet jako o závod}

\textquote{go like the wind’}
lit. ‘run like for a race’
Internal modifiability

By default, content words may be modified:

\textit{uhodily} [třeskuté] \textit{mrazy}

‘[bitter] frosts struck’

We mark only exceptions:

\textit{běžet jako o závod}

‘go like the wind’

lit. ‘run like for a race’
Transformations

- passivization / active form impossible
  
  **MWE cannot be passivized:**
  
  *jak si přejete*
  
  ‘as you wish’
  
  **MWE cannot have the active form:**
  
  *budiž ti přáno*
  
  ‘enjoy it!’; lit. ‘let be you.\text{DAT} wished’

- (impossible) nominalization (of verbal MWEs)

- (impossible) adjectivization
Morphological restrictions

Restriction in number, mood, case, non-standard form...:

Number:

\[\textit{Kostky jsou vrženy.PL}\]

‘The die is cast.’

Mood:

\[\textit{stůj.IMP co stůj.IMP}\]

‘at all costs’; lit. ‘\textit{cost.IMP what.ACC cost.IMP}’

Case

Non-standard form

...
Types of idiomaticity

**Lexical:** some lexical items occur in MWEs only

*mírnyx týrnyx*

‘for no reason’; from German ‘mir nichts dir nichts’

**Morphological:** some morphological forms occur in MWEs only

*chca nechca* instead of *chtě nechtě*

‘nolens volens’

**Syntactic:** MWE breaks some syntactic rule

*ber kde ber*

‘take wherever you can’; lit. *take.IMP where take.IMP*

**Semantic:** non-compositional meaning of MWE

*zamilovaný až po uši*

‘head over heels in love’; lit. ‘in love up to ears’
Type of idiomaticity II

**Pragmatic:** MWE is used in specific situations

*Smím prosit?*
‘May I have this dance?’; lit. ‘May I ask?’

**Statistical:** “selection restriction”, terms, compound prepositions...

*silný čaj*
‘strong tea’

*kyselina sírová*
‘sulphuric acid’
Conclusion

A complex typology of Czech MWEs was presented with PARSEME typology being enhanced primarily with:

- global/usage type
- style/variety
- syntactic trees including valency
- fragments/variants and core
- morphological idiomaticity

A MWE lexical database was designed

Future plans:

- Enlarging the database
- Refining the typology (if needed)
- Database integration with parsing
Initial requirements for the lexical database

- main challenge: variability
- MWE – a sequence of tokens/types; different levels of specificity:
  - particular lexeme; one or more possible morph. forms
  - choice of lexemes
  - any word in particular form (valency?)
  - a phrase of some type (obligatory or typical valency)
- features – both for the MWE and its components:
  - *morphological idiomaticity* is a feature of a component
  - *syntactic type* is a feature of the whole MWE
  - *style* may depend on a particular component or be the feature of the MWE (e.g. “mít něco na háku”)
- purpose: both *lexicographical* (human users) and *technical* (machine parsing)
Structure of a lexical database entry

- **slots** – the individual components of the MWE (syntagmatic dimensions)
- **fillers** – possible types, which may appear as variants of a component (paradigmatic dimension)
- **features** – custom `name=value` pairs assignable both to the MWE entry, or to a particular slot or a filler

- a MWE entry is a sequence of slots
- a slot is an enumeration of possible fillers:
  - **fixed** – the component may only be realized by one of the listed types
  - **open** – the component may be any lexeme (in a particular form)
  - **semi-open** – the component is usually realized by one of the listed lexemes, but other synonyms or semantically related expressions may sometimes appear
Structure of a lexical database entry

- **slots** – the individual components of the MWE (syntagmatic dimensions)
- **fillers** – possible types, which may appear as variants of a component (paradigmatic dimension)
- **features** – custom `name=value` pairs assignable both to the MWE entry, or to a particular slot or a filler

- a MWE entry is a sequence of slots
- a slot is an enumeration of possible fillers:
  - **fixed** – the component may only be realized by one of the listed types
  - **open** – the component may be any lexeme (in a particular form)
  - **semi-open** – the component is usually realized by one of the listed lexemes, but other synonyms or semantically related expressions may sometimes appear
Representation of different component types

- single token: slots with one or more possible fillers; a filler defines particular type by means of positional attributes (lemma, tag); tag as prefix, regular expressions used to match acceptable forms, e.g.
  - fixed case, singular or plural: lemma="ryba", tag="NNF[SP]1"
  - any verbal form: lemma="stát", tag="V"
  - any common adjective: tag="AA"


  => fillers represent exact types to be matched by the parser

- phrases (valency): open slots without fillers, features define type and morphological restrictions on the contents


  => slots may also represent more abstract units for lexicographical description (human users) or advanced parsing
Variable components (slots with several fillers)

dráť/vtírat se na mysl
(lit. force/intrude into (one’s) mind)
Various verbs: drát se, vtírat se
Advanced description: tree structures

- variable components consisting of several tokens
- dependency relations
- constituency structure

Solution:
- dependency: relations between slots (features)
- constituents: “non-terminal” slots; fillers may refer to a sequence of other slots
Advanced description: tree structures

- variable components consisting of several tokens
- dependency relations
- constituency structure

Solution:

- dependency: relations between slots (features)
- constituents: “non-terminal” slots; fillers may refer to a sequence of other slots
Dependency structure

Bůh dal, Bůh vzal
(lit. God gave, God took)
Constituency structure

Bůh dal, Bůh vzal
(lit. God gave, God took)
Constituency structure – additional “non-terminal” slots
Multi-token variants

bojovat pro/za/o čest a slávu
(lit. fight for honor and glory)
Various verbs: bojovat, zápolit; bít se, rvát se; zahrát si
...additional “non-terminal” slot grouping the variants
...variability within the constituent structure
...variability within the dependency structure?
Challenges: repeating the same (variable) lemma

Bůh/X dal, Bůh/X vzal
(lit. God/X gave, God/X took)

The value of the lemma attribute of the filler in the 4\textsuperscript{th} slot is a reference to the lemma used in the 1\textsuperscript{st} slot.
Challenges: dependencies among “optional” components

Two optional components, but at least one must be present:

  míť  NĚCO  pro  SVOU  vlastní  potřebu
  míť  NĚCO  pro  SVOU  potřebu
  míť  NĚCO  pro  vlastní  potřebu
  * míť  NĚCO  pro  potřebu
(lit.  to have  ST.  for  ONE’S  own  need/use)

Exclusive syntactic alternations:

  naložit  NĚKOMU  NĚCO  na  bedra
  naložit  NĚCO  na  NĚČÍ  bedra
  * naložit  NĚKOMU  NĚCO  na  NĚČÍ  bedra
  * naložit  NĚCO  na  bedra
(lit.  to load  SO.  ST.  on  loins/shoulders
or  to load  ST.  on  SO.’S  loins/shoulders)
Technical implementation

Implemented on a prototype of a more generic experimental database of annotation units.

- Elasticsearch used as backend (Java; Lucene based search engine)
- generic abstraction/model with a REST API (Python)
- highly configurable web interface (Angular, bootstrap, d3.js)