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)

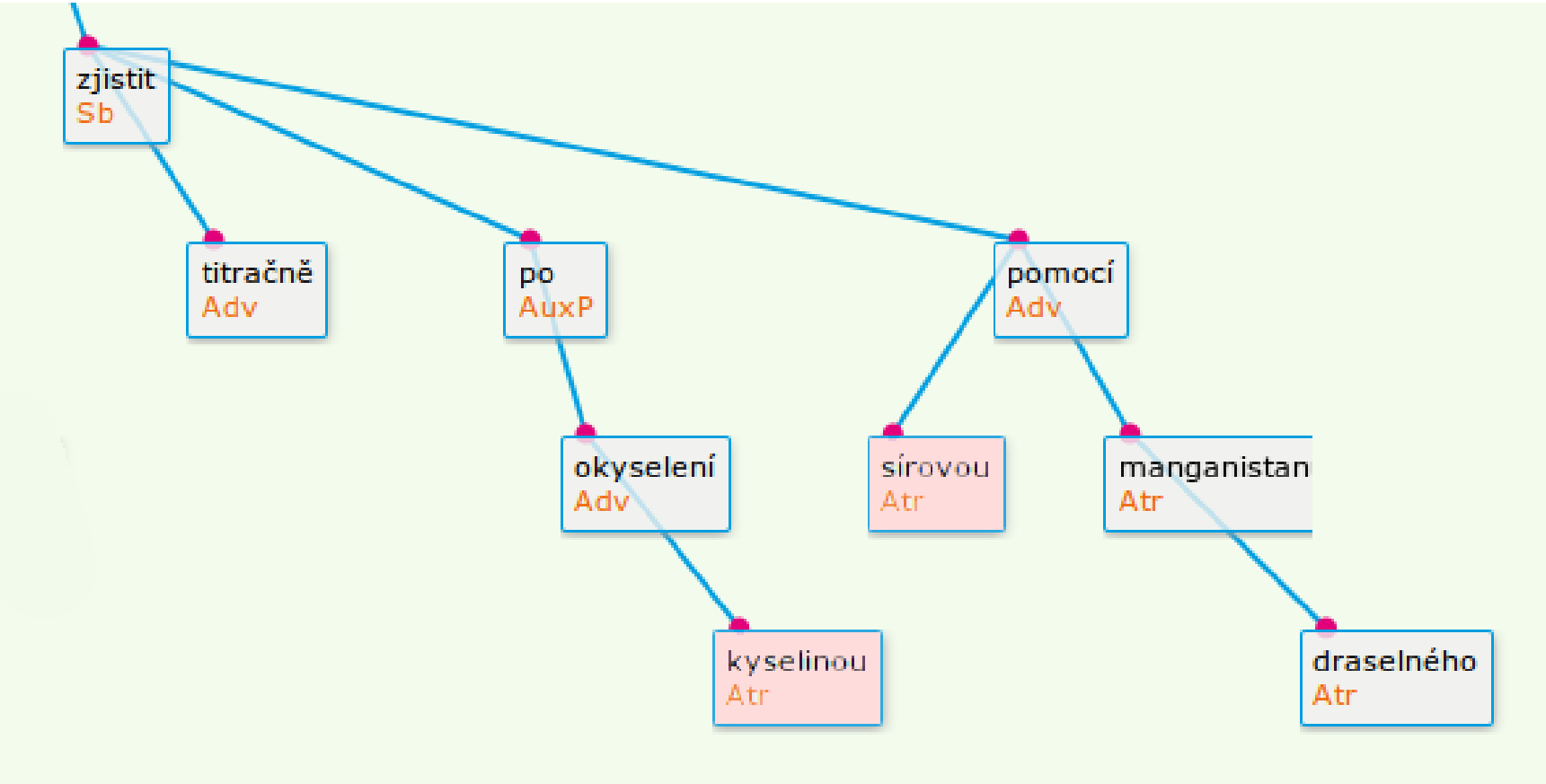
Need of a MWE lexicon for NLP tools (and their users)

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

Parsing



Correct structure: [kyselinou $_{Attr}$[sírovou]]
'sulphuric acid'

MWE tagging in a corpus

Hits: **49,887,368** | i.p.m. ⓘ: **10,845.92** (related to the whole "syn_v5") | ARF ⓘ: **30,247,135.4** | Result is shuffled 1 / 3,325,825

Line selection: simple | Attributes:

Hospodářské noviny	(NYSE) se na této burze ve	skutečnosti/ve_skutečnosti	odehraje méně než 35 procent .</s><s>Rychlost obchodů
Deníky Bohemia	čtvereční .</s><s>V těchto dnech mají plné ruce	práce/mít_plné_ruce_práce	ženy ze střediska údržby veřejné zeleně Technických služeb
Deníky Bohemia	potenciál .</s><s>* Zlato zMS jste získal na	podzim/na_podzim	a letošní MČR je na programu už za
Blesk	neštěstí došlo kvůli trhlině v tepelném štítu levého	křídla/levé_křídlo	, do níž se dostal horký vzduch vznikající
Reflex	Brychtovou - ze skla sochy , jaké nemají	obdoby/nemít_obdoby	.</s><s>- Připadá mi , že jste své
ForMen	se pěstuje Arabica Bourbon a Tipica .</s><s>Prvně	jmenovaná/prvně_jmenovaný	se pyšní nasládlou chutí , Tipica má mírně
Právo	, v lékárnách nepochodí .</s><s>" Máme na	skladě/mít_na_skladě	deset balení .</s><s>Češi ale o ně zájem
24 hodin	, že na dávky v hmotné nouzi nemá	nárok/mít_nárok	člověk , který nepracuje , ani není v
Marketing & Media	ohřívat) , nebo mrazené zboží .</s><s>I	když/i_když	na takové případy jsme už vybavení a nezaskočí
Mladá fronta DNES	chceme ukazovat ty sporty , které nejsou k	vidění/být_k_vidění	jinde .</s><s>V budoucnu bychom se chtěli podobat
Deníky Bohemia	1:4 (0:3) Zásadě se v poslední	době/v_poslední_době	herně daří a svěřenci trenéra Libora Lubase to
Mladá fronta DNES	koloniálu nás znají , tam se cítím v	pohodě/v_pohodě	, nikdy mě tam nikdo z ničeho neobvinil
Mladá fronta DNES	předává v dobrém stavu .</s><s>Ani jedno dobré	slovo/dobré_slovo	ovšem nemá pro piráty silnic , kteří zásobují
Mladá fronta DNES	na nákupy za hranice , o tom vědí	své/vědět_o_tom_svoje	.</s><s>Ceny jsou v německých obchodech často srovnatelné
Deníky Bohemia	deníku</s><s>NOVOROČNÍ VÝŠLAP na Zelenou horu si nenechalo	ujít/nenechat_si_ujít	na osm stovek lidí .</s><s>Užili si výlet

1 / 3,325,825

Identification of fragments and variants

Base MWE:

vzít rozum do hrsti

'keep one's wits about one'

lit. 'take reason into the cupped hand'

Fragment:

rozum do hrsti

lit. 'reason into the cupped hand'

Variant:

Sebral 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'

Formula

Term

Syntactic type

Noun phrase: *žabí muž*
'frogman'

Adjectival phrase: *všeho schopný*
'capable of anything (bad)'

Verb phrase: *stát za starou belu*
'not worth a damn'

Verbonominal (light verb) constructions

Adverbial phrase

...

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:

uhodily [*třeskuté*] *mrazy*

'[bitter] frosts struck'

We mark only exceptions:

běžet jako o ***závod***

'go like the wind'

lit. 'run like for a race'

Internal modifiability

By default, content words may be modified:

uhodily [*třeskuté*] *mrazy*
'[bitter] frosts struck'

We mark only exceptions:

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.DAT wished'
- (impossible) nominalization (of verbal MWEs)
- (impossible) adjectivization

Morphological restrictions

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

Number:

Kostky jsou vrženy.PL
'The die is cast.'

Mood:

stůj.IMP *co stůj*.IMP
'at all costs'; lit. '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)

drát/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

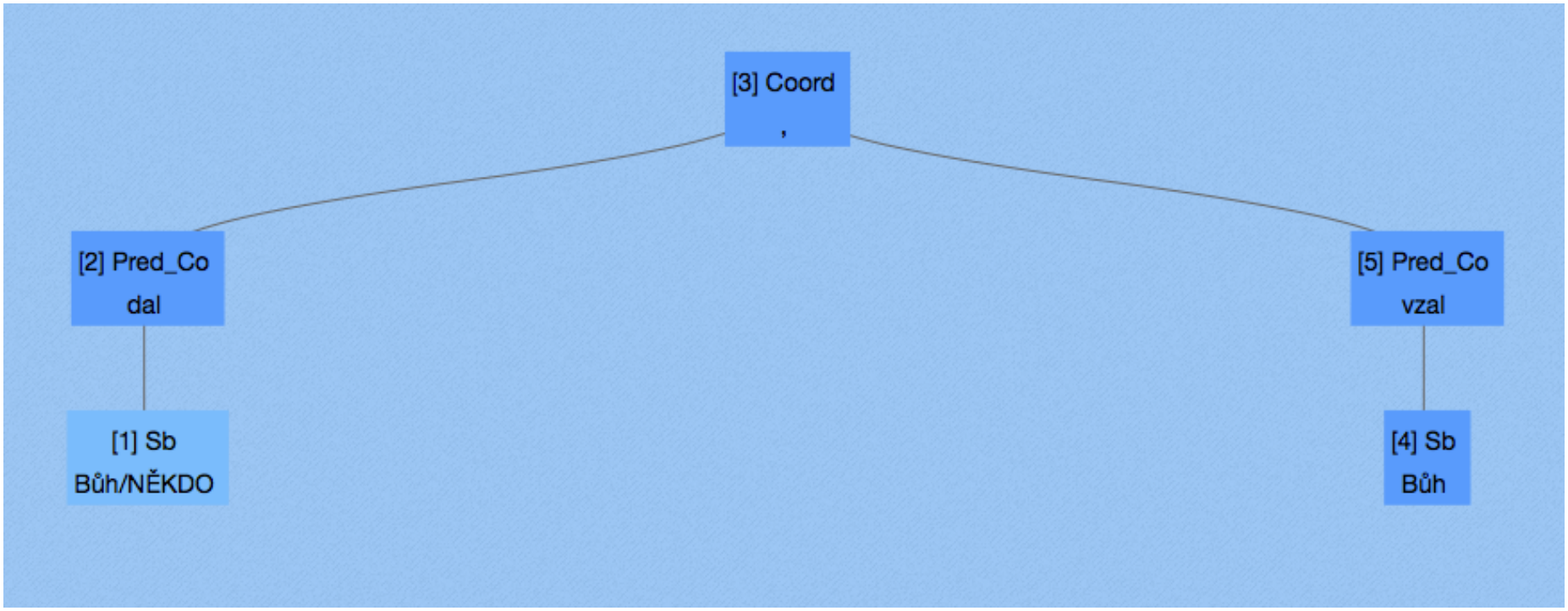
- variable components consisting of several tokens
- dependency relations
- constituency structure

Solution:

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

Dependency structure

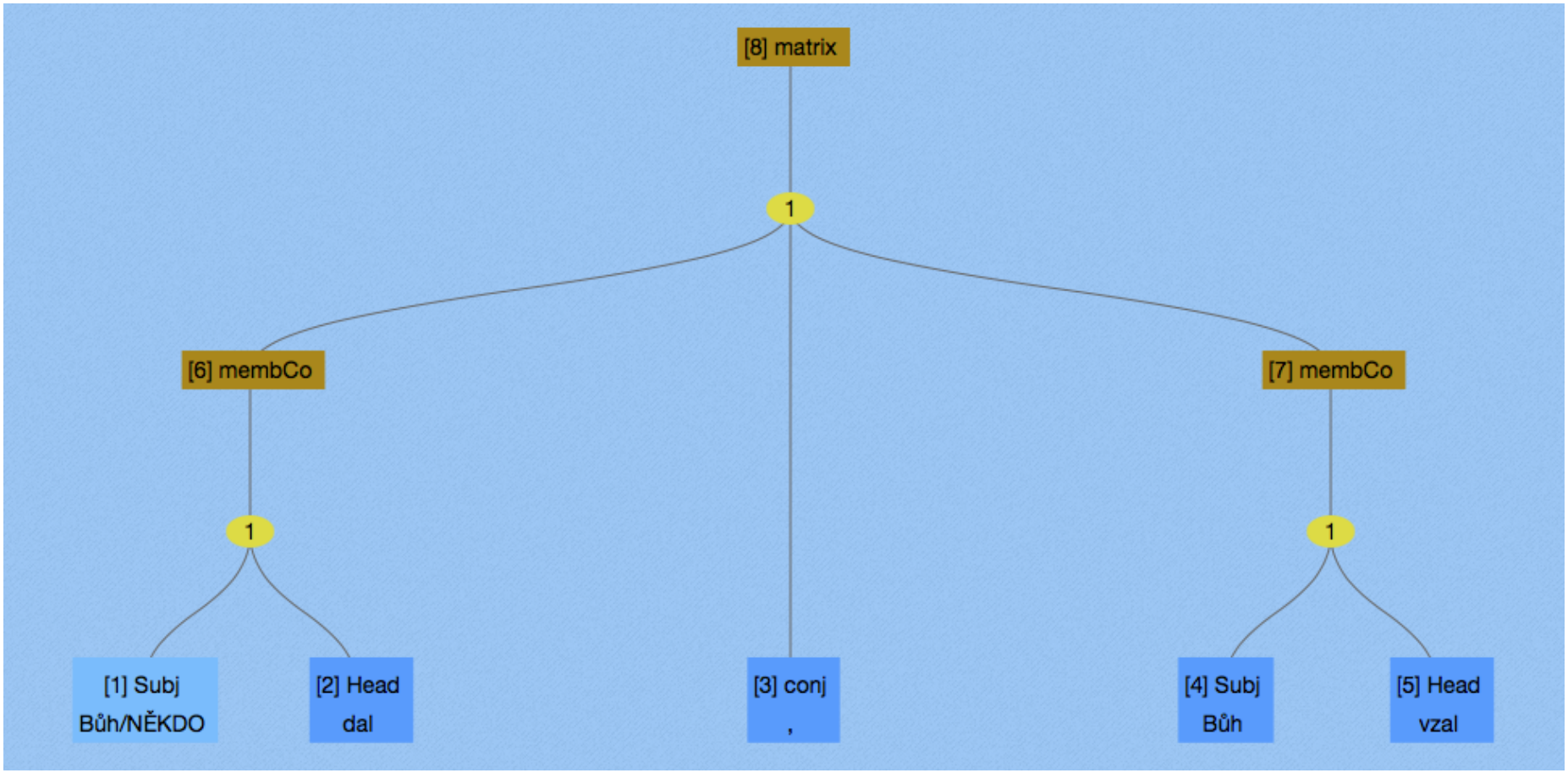
Bůh dal, Bůh vzal
(lit. God gave, God took)



Constituency structure

Bůh dal, Bůh vzal

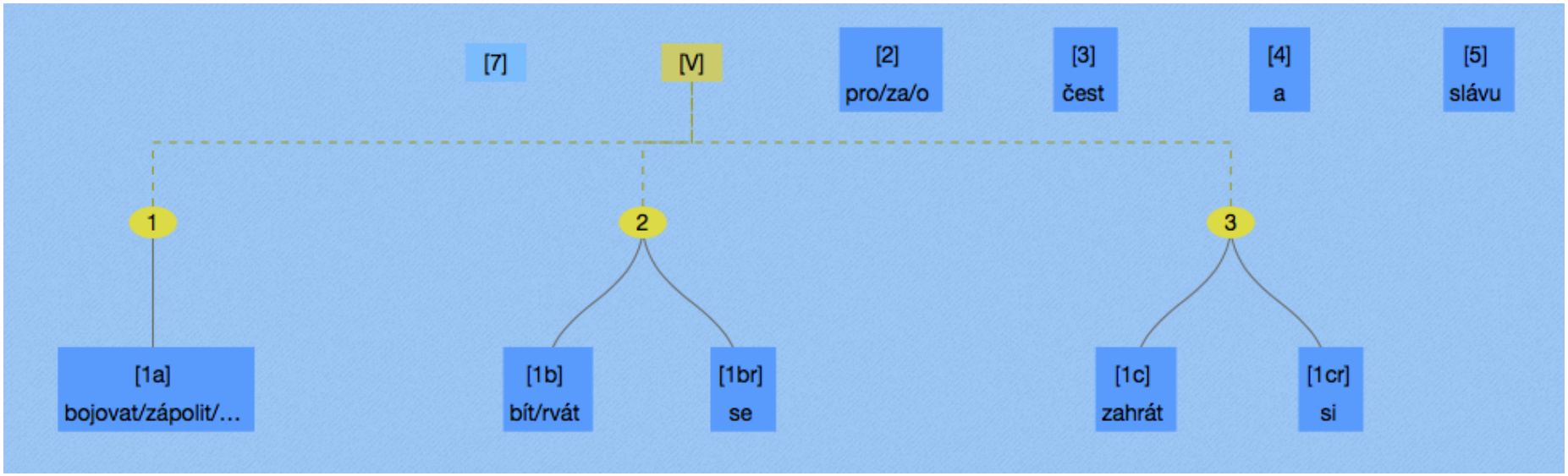
(lit. God gave, God took)



Constituency structure – additional “non-terminal” slots

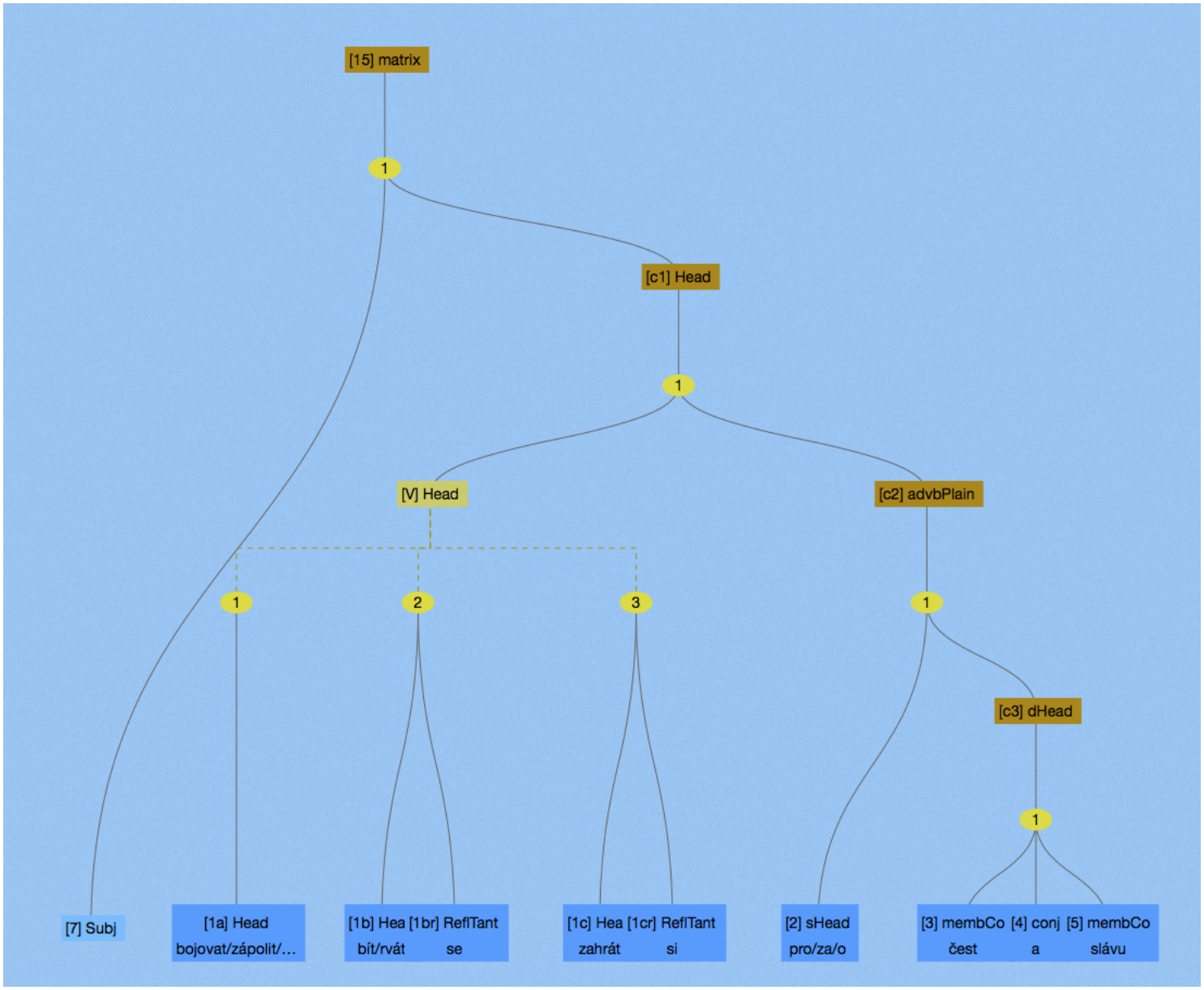
Multi-token variants

bojovat pro/za/o čest a slávu
(lit. fight for honor and glory)
Various verbs: bojovat, zápolit; bít se, rvát se; zahrát si

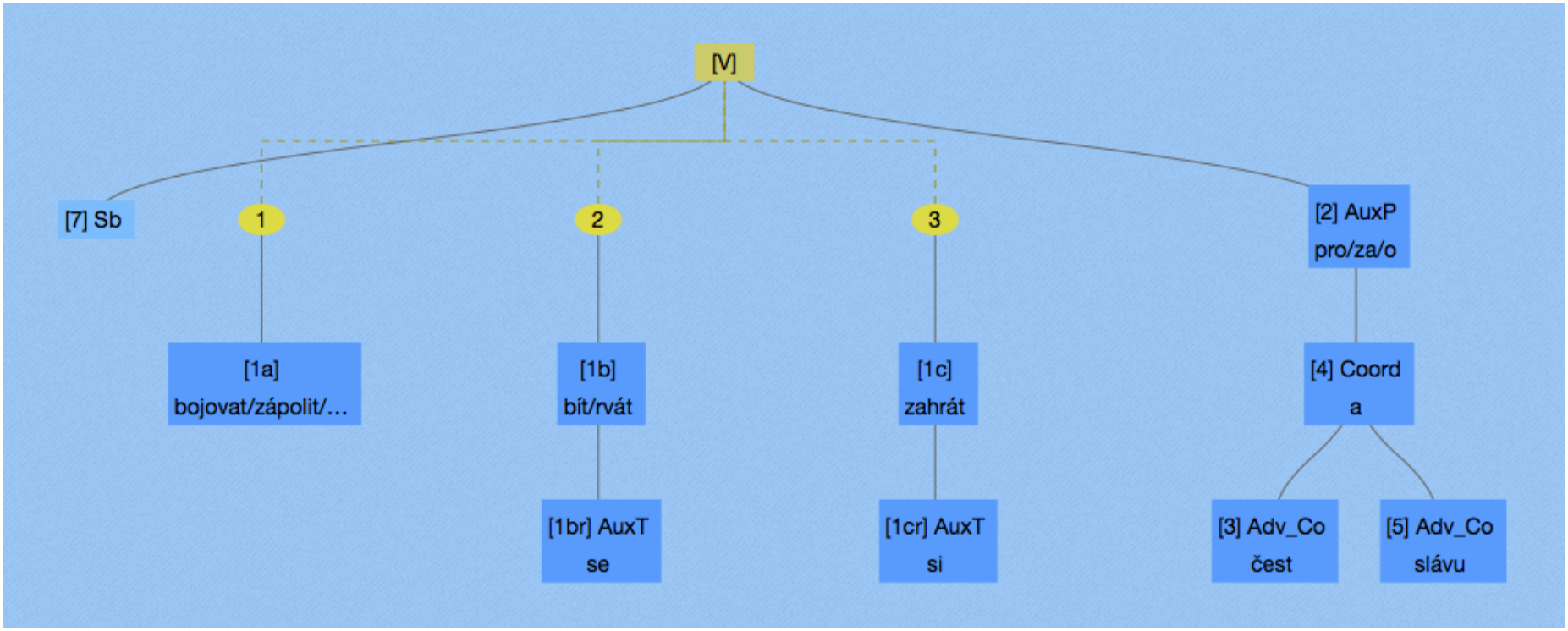


. . . additional “non-terminal” slot grouping the variants

...variability within the constituent structure

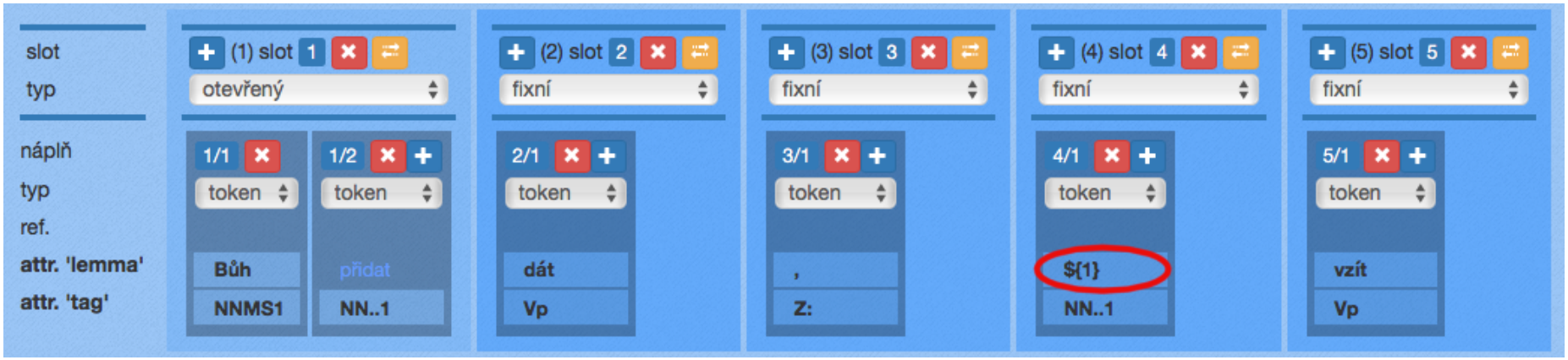


. . . variability within the dependency structure?



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 4th slot is a reference to the lemma used in the 1st slot.

Challenges: dependencies among "optional" components

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

	mít	NĚCO	pro	SVOU	vlastní	potřebu	
	mít	NĚCO	pro	SVOU		potřebu	
	mít	NĚCO	pro		vlastní	potřebu	
*	mít	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)