Towards a database of Czech multi-word expressions

Milena Hnátková and Alexandr Rosen
Institute of Theoretical and Computational Linguistics, Faculty of Arts
Charles University in Prague, Czech Republic

This contribution aims to present: (i) a lexical resource and tool used to identify and categorize multi-word expressions in a tagged, lemmatized and parsed corpus of Czech and (ii) a follow-up plan to design and build a MWE database, based on a taxonomy of MWEs compatible with the PARSEME template. The MWE entries should allow for linking with other resources, including other lexica, annotated corpora, grammars and NLP tools.

1 Linear annotation based on a MWE dictionary

In a tagged, lemmatized and parsed 100M representative corpus of Czech,[1,2] about 3.8M words are identified as part of a MWE. The discovery of MWEs, performed by a tool called *FRANTA*, is rule-based, using a modified and extended list of about 39.7 thousand MWE entries [3, 8, 9] (henceforth *FrantaLex*), extracted from a published dictionary of MWEs [11]. Each MWE in the corpus is assigned one of the six MWE types, adopted from [11]:[3] non-verbal (722K/8.2K), verbal (523K/19.0K), clausal (139K/5.2K), similes (18K/5.8K), proverbs (4K/1.4K) and compound conjunctions (64K/35).

The MWE types above are subject to variability in: (i) word order due to information structure or syntactic context; (ii) inflection due to syntactic, pronominal or referential agreement, syntactic government or different register; (iii) lexical replacement (by a synonym, pronoun, aspectual verbal variant), insertion or deletion; (iv) syntactic alternations (passivization, negation); (v) internal modification; (vi) updating (coinage of a new MWE based on an existing one). The encoding of variability potential has a technical solution in the *FrantaLex* format,[4] but a principled test-based taxonomy and encoding is previewed for the MWE database project.

FRANTA is applied to morphologically tagged, lemmatized and disambiguated texts, using a pattern matching algorithm, targetting lemmas and tags. Syntactic structure and functions are not used in the process.[5] All parts of an identified MWE receive *collocational lemma* (col_lemma), i.e. a string denoting the specific MWE, and *collocational type* (coll_type).[6]

[1] SYN1015fr, an in-house release of SYN2015, a part of the Czech National Corpus (CNC); a much larger MWE-annotated SYN release 4 (about 2.5G) is due to be released soon, see http://wiki.korpus.cz/doku.php/en:cnk:uvod.

[2] A part of the work reported here has been supported by the Grant Agency of the Czech Republic within the project *Between Lexicon and Grammar*, grant no. 16-07473S.

[3] The numbers stand for frequency of the given type in the corpus, followed by the number of corresponding entries in *FrantaLex*. *FrantaLex* lists the individual MWE types separately. Non-verbal MWEs can be non-inflected (multi-word particles, adverbials, frozen prepositional expressions, together about 2.6K entries) or inflected (nominal expressions, about 5.6 entries).

[4] See [3] for details. The format allows to specify forms, lemmas, tags, insertions and word order variations.

[5] Tagging and lemmatization was done by a hybrid tagger [2, 5], parsing by a tool combining several parsers [6].

[6] At present, words belonging to two instances of the same MWE within a single sentence cannot be distinguished.

Example (1) shows a sentence including two MWEs: a clausal type (coll_type = SZ) *jak se zdá* 'as it seems' and a verbal type (coll_type = VZ) *zahodit flintu do žita* 'to give up'. The keywords, used as a MWE handle to count frequencies, are specified as SH and VH, respectively.

(1) *Jak se zdá, nezahodil ještě flintu do žita*
as REFL seems [he] didn't throw yet rifle into rye
'He doesn't seem to have given up yet.'

The annotation is illustrated in (2). Following the type is the abbreviated coll_lemma, the word's syntactic function and a link to its head in a dependency parse. Although the MWE annotation is compatible with syntactic structure and functions, the currently used concordancer [10] cannot handle non-linear structures: syntactic trees and MWEs can only be searched or displayed as individual words.

(2) $^{1}Jak_{\text{SZ:jsz,Adv:3}}$ $^{2}se_{\text{SZ:jsz,AuxT:3}}$ $^{3}zdá_{\text{SH:jsz,Pred\_Pa:4}}$, $^{4}nezahodil_{\text{VZ:zfdž,Pred}}$ $^{5}ještě_{\text{Adv:4}}$
$^{6}flintu_{\text{VZ:zfdž,Obj:4}}$ $^{7}do_{\text{VZ:zfdž,AuxP:4}}$ $^{8}žita_{\text{VH:zfdž,Atr:7}}$

Several other CNC corpora are annotated in their in-house releases by *FRANTA*, including a 1.5M spoken corpus. The annotation of idioms including body parts in this corpus is evaluated with precision and recall around 80%, depending on the specific body part [7]. For another partial evaluation on a spoken corpus see [4].

2 The plans

FrantaLex includes more typological information about each MWE than the corpus annotation shows. The table below shows examples for each distinction.

Category	Verbal	Inflected	Continuous	Example
conjuction	no	no	yes	*a právě proto* 'and that's why'
adverbial	no	no	yes	*v neposlední řadě* 'last but not least'
	yes	no	yes	*od nevidim do nevidim* 'from dusk to dark'
particle	no	no	yes	*tak jako tak* 'anyway'
	yes	no	yes	*stůj co stůj* 'come what may'
preposition	no	no	yes	*v závislosti na* 'depending on'
proverb	no	no	yes	*komu čest, tomu čest* 'give them their due'
	yes	no	yes	*mrtvý prd ví* 'the dead don't care'
	yes	yes	yes	*jak si usteleš, tak si lehneš* 'you'll end up as you deserve'
	yes	yes	no	*tonoucí se stébla chytá* 'any port is good in a storm'
noun phrase	no	yes	yes	*očitý svědek* 'eye witness'
simile	no	yes	yes	*utahaný jako kotě* 'tired as a kitten'
	yes	yes	no	*podobat se jako vejce vejci* 'be like two peas in a pod'
clausal	yes	no	yes	*ted' babo rad'* 'so what now?'
	yes	yes	yes	*kde se vzal tu se vzal* 'out of nowhere'
	yes	yes	no	*div se hanbou nepropadl* 'he was overwhelmed with shame'
verbal	yes	yes	no	*mít máslo na hlavě* 'be guilty'
quasi-phrasemes	yes	yes	no	*klást otázky* 'ask questions'

The typology used in *FrantaLex* will be mapped on a taxonomy compatible with the PARSEME template, focusing on non-trivial cases of mapping and applying linguistic tests to arrive at the classification by syntactic category and structure, flexibility and idiomaticity.

The structure of an entry in the resulting MWE database should allow for its conceptual and formal integration with other lexical and text resources, but also with theoretical and NLP frameworks dealing with Czech. Candidates for extending the lexical database will be extracted primarily on the basis of two measures of MWEs' fixedness: obligatoriness and proximity as "P-collocations" [1].

We hope that a well-founded MWE typology with appropriately designed formal representation of MWE types will help to bridge the gap between lexicon and grammar.

References

[1] Václav Cvrček. *Kvantitativní analýza kontextu (Quantitative analysis of context)*. Nakladatelství Lidové noviny, Praha, 2014.

[2] Jan Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, 2004.

[3] Milena Hnátková. Značkování frazémů a idiomů v Českém národním korpusu s pomocí Slovníku české frazeologie a idiomatiky (Tagging of phrasemes and idioms in the Czech National Corpus with the help of the dictionary of Czech phraseology and idiomatics). *Slovo a slovesnost*, 63(2):117–126, 2002.

[4] Milena Hnátková and Marie Kopřivová. Identification of idioms in spoken corpora. In *Proceedings of the Seventh International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, pages 92–99, 2013.

[5] Milena Hnátková, Vladimír Petkevič, and Hana Skoumalová. Linguistic Annotation of Corpora in the Czech National Corpus. In *Trudy meždunarodnoj konferencii Korpusnaja lingvistika – 2011 (Proceedings of Corpus Linguistics – 2011)*, pages 15–20. St.-Petersburg State University, Institute of Linguistic Studies (RAS), Russian State Herzen Pedagogical University, 2011. ISBN 978-5-8465-0005-5.

[6] Tomáš Jelínek. Combining dependency parsers using error rates. In *Text, Speech and Dialogue – Proceedings of the 19th International Conference TSD 2016*. Springer, 2016. In print.

[7] Marie Kopřivová. Evaluating automatic idiom annotation in spoken corpora: the case of somatic idioms. In *SLOVKO 2015 – Natural Language Processing, Corpus Linguistics, Lexicography*, volume Katarína Gajdošová and Adriána Žáková, pages 72–76. Slovak National Corpus, Slovak Academy of Sciences, RAM-Verlag, 2015.

[8] Marie Kopřivová and Milena Hnátková. From dictionary to corpus. In Vida Jesenšek and Peter Grzybek, editors, *Phraseology in Dictionaries and Corpus*, volume 97 of *Zora*, pages 155–168, Maribor, 2014. Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Mariboru.

[9] Marie Kopřivová and Milena Hnátková. "Lemmatizing idioms" in Czech corpora. In *EUROPHRAS 2014: La phraséologie: ressources, descriptions et traitements informatiques, 10.–12. 9. 2014*, 2014.

[10] Pavel Rychlý. Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno, 2007.

[11] František Čermák, Jiří Hronek, Jaroslav Machač, Renata Blatná, Miloslav Churavý, Vlasta Červená, Jan Holub, Marie Kopřivová, Libuše Kroupová, Vladimír Mejstřík, Milan Šára, and Alena Trnková. *Slovník české frazeologie a idiomatiky*, volume 1–4. LEDA, Praha, 2nd edition, 2009.