Tomáš Jelínek
Charles University
Czech Republic
https://orcid.org/0000–0002–8521–4715

Multi-word Lexical Units with Repetition of Lexemes in Czech and Identification of Their Variants

Abstract. A study of the variability of multi-word lexical units with repetition of lexemes such as *Bůh dal, Bůh vzal* 'the Lord has given, the Lord has taken away' based on a large corpus is presented. Four case studies illustrate the extent of variability of such expressions. A database of multi-word expressions is described, with a special attention to expressions with repetition and their variability. The automatic identification of multi-word expressions with repetition in Czech texts is also explained.

Key words: *multi-word expressions, repetition of lexemes, variability, automatic identification of MWE in text*

1. Introduction

The variability of multi-word lexical units with repetition of lexemes is an interesting case of reproducibility of multi-word expressions which are recognized not only on the basis of a particular word or group of words but on the basis of their syntactic structure and the fact that a lemma is repeated in such a structure as well. For example, for the Czech saying *Bůh dal, Bůh vzal* 'the Lord has given, the Lord has taken away', we found many cases where the word *Bůh* 'God' is replaced by another word occurring twice in the expression, such as *stát dal*, *stát vzal* 'the state has given, the state has taken away'. Such variants are easily recognized as referring to the original multi-word expression by native speakers. In Czech, we have found about 90 such multi-word lexical units so far, with higher or lower variability. Some of these multi-word expressions are simple binomials such as *den za dnem* 'day after day', *kniha knih* 'the book of (all) the books' Čermák (2007: 414–

429), some are more complex. The phenomenon of repetition in language has been studied extensively, see for example Fisher (1995) or Frédéric (1985).

In this paper, we describe the variability of these multi-word lexical units in four case studies and show how we address this variability in the database of multi-word lexical units LEMUR we have developed (see below). We also explain automatic identification of variants of these multi-word expressions in texts.

2. Data sources

In order to study the variability of multi-word lexical units in use, large corpora are needed. Multi-word expressions (MWE) occur far less frequently than most single-word lexical units, even in their invariant form; finding their variants in individual texts or in smaller corpora is very difficult. In order to investigate the variability of multi-word lexical units, we need really large corpora containing billions of tokens. The above mentioned saying *Bůh dal, Bůh vzal* 'the Lord has given, the Lord has taken away' appears, for example, only 4 times in three different variants of one single type, in the corpus SYN2015, developed by Křen et al. (2015), which contains one hundred million words. In the SYNv6 corpus, compiled by Křen et al. (2017), which contains 4.8 billion tokens, we find 80 occurrences of this multi-word expression with significantly richer variability. 80 occurrences are perhaps not enough to base a research on, but at least it gives us a better picture of how this MWE is used and what the extent of its variability is.

The research presented in this article is therefore based on the SYNv6 corpus of written contemporary Czech. This corpus consists of 4.8 billion text positions, i.e. 4 billion words not including punctuation. The corpus is not balanced, the majority of texts are journalistic texts because of easier accessibility of this type of text in electronic form. It does not contain texts originating from the Internet (due to the higher proportion of non-standard language in the texts of this provenance).

3. Case studies

In this section, we present four Czech multi-word lexical units: *Bůh dal, Bůh vzal* 'the Lord has given, the Lord has taken away'; *čistému vše čisté* 'to the pure, everything is pure'; *podobný/podobat se jako vejce vejci* 'as alike as two peas in a pod, lit. similar as an egg to an egg'; *já na bráchu, brácha na mě* 'you scratch my back, I'll scratch yours, lit. me to brother, brother to me'. The first three case studies will show how diverse the lexical variability found

in the corpus for this type of MWEs is, the last case is interesting because of its structure and the possibility of replacing all autosemantic words in the multi-word lexical unit.

3.1. Bůh dal, Bůh vzal

The expression *Bůh dal, Bůh vzal* (*jméno Boží budiž požehnáno*) 'The Lord has given, the Lord has taken away (the name of the Lord be blessed)' is of biblical origin, coming from the book of Job (Job, 1:21). It expresses reconciliation with some loss. A typical use of this phrase is illustrated by example (1) from the corpus:

(1) *„Zřejmě nakoupím nové vybavení a včelstva," nevzdává se postižený včelař Luboš Machatý. Ztrátu chovu bere optimisticky: „Je to příroda.* ***Bůh dal, Bůh vzal****."*

"Obviously, I will have to buy new equipment and honeycombs," says the affected beekeeper Luboš Machatý. He takes the loss of beehives optimistically: "That is nature. God has given, God has taken away."

There are 80 occurrences of this MWE in the SYNv6 corpus. This saying is very variable, only in 40% (27 occurrences) we find it in its original form with the noun *Bůh* 'God'. The reproducibility of this MWE is based on the use of the verbs *dát* 'to give' and *vzít* 'to take away' (usually in past tense) and on a structure with the repetition of the noun in the nominative case. The structure of the variable MWE can be expressed by the formula:

N_nom.$_a$ *dát*_past.part., N_nom.$_b$ *vzít*_past.part (lemma$_a$ = lemma$_b$).

The verbs *dát* 'give' and *vzít* 'take away' are usually in past tense, we find only two occurrences with a different tense (*stát dal, stát vezme* 'the state gave, the state will take away'). The past tense (expressed by the past participle of the verb) agrees in gender and number with the subject; in most cases (70 out of 80) the subject is masculine animate, the remaining cases are feminine (*voda dala, voda vzala* 'water gave, water has taken away'), or plural (*politici dali, politici vzali* 'the politicians gave...'). Exceptionally, in four cases, we find an inversed order of the verbs (*banka vzala, banka dala* 'the bank took away, the bank gave').

The greatest variability of this MWE is related to the pair of nouns in the nominative case with the same lemma. In 58 cases (72% of all occurrences), these nouns are lexical variations of the word God: *Bůh* 'God', *Pánbůh* 'the Lord God', *pán* 'Lord', *Hospodin* 'Lord', *Alláh* 'Allah'; in 13 cases, it is institutions: *stát* 'state', *televize* 'television', *Apple*, *ČEZ* (a Czech power company);

in 5 cases influential people: *politici* 'politicians', *bolševik* 'Bolshevik', *gosudar* 'tsar', *sudí* 'jury', *Jandák* (a Czech politician); the remaining 4 cases are impersonal forces: *příroda* 'nature', *voda* 'water', *kostka* 'cube', *náhoda* 'coincidence'. In one single case the saying occurs without the lemma repetition (i.e. the reproducibility of MWE is based only on the structure, verbs and the meaning of the words): *car dal, sověty vzaly* 'the tsar gave, the Soviets took away'.

In 7 cases, the MWE is followed by the phrase *jméno Boží budiž požehnáno* 'the name of God be blessed' in several variants. Otherwise, no syntactic variability has been found; for example, we have not found any occurrence of an attribute modifying the noun.

3.2. Čistému vše čisté

The saying *čistému vše čisté* 'to the pure everything is pure' also comes from the Bible (Titus 1:15); it means that honest men also consider others to be honest, which may be naïve. In the SYNv6 corpus, there are 137 occurrences of this MWE. It is used as in example (2):

(2) *Yoko Ono a John Lennon se k sexu stavěli velice otevřeně, přesně podle rčení „**čistému vše čisté**".*

Yoko Ono and John Lennon have been very open about sex, precisely by the saying "to the pure, everything is pure".

The structure of the variable MWE can be expressed by the formula:

$$\text{A\_nom.}_a \textit{ vše}\text{\_nom. A\_dat.}_b \ (\text{lemma}_a = \text{lemma}_b).$$

There is significantly less variability than for the previous saying, 118 occurrences (86% of all occurrences) are in the invariant form *čistému vše čisté*; there are 5 occurrences of a variant *hebkému vše hebké* 'to the soft, everything is soft' which is a name of a musical performance cited by the journals; we find 3 occurrences of *mrtvému vše mrtvé* 'to the dead everything is dead'; each of the other 11 variants appears only once: *živý* 'alive'; *hloupý* 'stupid', *chytrý* 'smart', *dobrý* 'good', *velký* 'big', *nízký* 'low', *tajný* 'secret'; *sladký* 'sweet', *česnekový* 'garlicky', *bílý* 'white', *ruský* 'Russian'. Seven times the saying occurs in the variant with the verb *být* 'to be' (only with the adjective *čistý* 'pure'), as in example (3):

(3) *Některé zvěsti, které kolují o Hrabalovi, věru nepatří do čítanek (ale **čistému je vše čisté**).*

Some rumors about Hrabal do not belong to textbooks (but to the pure, everything is pure).

3.3. Podobat se / podobný jako vejce vejci

The third of our case studies presents a fixed comparison: *podobat se / podobný jako vejce vejci* 'as alike as two peas in a pod', lit. 'resemble/similar as an egg to an egg'. The meaning of this comparison is that two objects or persons are very similar, almost indistinguishable. The reflexive only verb *podobat se* 'resemble' or adjective *podobný* 'alike' can occur in the vicinity of the words *jako vejce vejci* 'as an egg nom. an egg dat.' (which have a fixed order), not necessarily just in an adjacent position (in the corpus, we were searching for cases where *podobat se* or *podobný* occur at a distance of maximum ten tokens to either side, regardless of sentence boundaries). Typical use is shown in example (4). 2180 occurrences of this MWE and its variants appear in the SYNv6 corpus.

(4) *Webové stránky mnohých fakult jsou si* ***podobné jak vejce vejci****, přitom studium na nich je naprosto odlišné.*

The web sites of many faculties are similar as an egg to an egg, while studying at them is totally different.

The structure of the variable MWE can be expressed by the formula:

podobný/podobat se... jako/jak N_nom.$_a$ N_dat.$_b$ (lemma$_a$ = lemma$_b$).

Whereas the lexical variability in this comparison is very small compared to the above-described expressions, the frequency in the corpus is, on the contrary, far higher. Out of the total number of 2180 occurrences, 2166 (99.4%) are examples of the invariant form (*podobat se / podobný ... jako vejce vejci*), with a uniform division between the verb *podobat se* 'to resemble' (1129) and the adjective *podobný* 'alike' (1046). There is a strong preference for the longer form of the conjunction *jako* 'as' (2050), compared with the shorter form *jak* 'as' (116).

Apart from the noun *vejce* 'egg', a number of other nouns repeated twice appear in this comparison, but with a very low frequency. There are altogether only 14 such occurrences in the SYNv6 corpus. The most common is *podobný ... jako oko oku* 'alike as an eye to the eye' (3 occurrences); then there are objects that appear in large numbers: *tráva* 'grass', *kapka* 'drop', *hvězda* 'star'; animals: *pták* 'bird', *vrána* 'crow', *vlk* 'wolf', *blecha* 'flea'; people: *dvojče* 'twin', *Číňan* 'Chinese' (probably because the Asians are difficult to distinguish for the Czechs), *Hitler* (used in a context about a court battle between two companies using a similar design resembling Hitler) and *telenovela* 'soap-opera' (used in a context pointing to small differences among TV shows).

For the basic form *... jako vejce vejci*, 6 examples of a noun attribute modification can be found in the corpus, in five occurrences the same adjective is repeated, twice *dračí vejce* 'dragon egg', once *pštrosí vejce* 'ostrich egg', *kukaččí vejce* 'cuckoo egg', *zlaté vejce* 'golden egg'.

In the case of (5), a single such occurrence in the corpus, the adjectives differ in a contextually dependent use of the original fixed comparison where the complete similarity of compared objects is modified by unequal wealth (*chudé vejce* 'poor egg' vs. *bohatší vejce* 'richer egg').

(5) *Pořady České televize se většinou **podobají** těm na soukromých stanicích **jako chudé vejce bohatšímu vejci**.*
Czech TV shows are usually similar to those in private stations as poor eggs to richer eggs.

Exceptionally (also only once in the whole four-billion word corpus), we find a different variant of the fixed comparison *podobný jako vejce vejci* in which only the first part is used to recognize the MWE, the second part is changed and lemmas are not repeated (6):

(6) *Ten mladík tam je mi **podobný** asi **jako vejce hromádce kuřecích kostí**.*
The young man there is similar to me as an egg to a bunch of chicken bones.

3.4. Já na bráchu, brácha na mě

The last of our MWE case studies *já na bráchu, brácha na mě* 'you scratch my back, I'll scratch yours', lit. 'me to brother, brother to me' is an example of a complex MWE where its reproducibility is based primarily on lemma repetition and morphosyntactic structure, there are no autosemantic words that could not be changed in a variant. The saying is used as a negative assessment of cases where people (typically powerful) help each other, usually in an unethical or illegal way (cronyism), as in example (7). There are 521 occurrences of this MWE in the SYNv6 corpus (not including fragments).

(7) *Máme tady z toho dojem klientelismu – pověstného **já na bráchu, brácha na mě**. Je to vidět na firmách, které v tomto obvodu vítězí ve veřejných zakázkách.*
There is an impression of cronyism – the proverbial you scratch my back, I'll scratch yours. It manifests itself with the companies that are winning in public procurement in this area.

The structure of this MWE is expressed by the formula:

P/N_nom.$_a$ *na* N_acc.$_b$, N_nom.$_c$ *na* P/N_acc.$_d$
(lemma$_a$ = lemma$_d$ & lemma$_b$ = lemma$_c$).

In the invariant form, the personal pronoun forms *já*, *mě* 'I, me' is repeated at the first and last position, the noun *brácha* 'bro, brother' (a colloquial form of the word *brother*) is found following the first preposition *na* 'on, to' and also in front of the second preposition *na*. In some variants, both the noun *brácha* and the pronoun *já* can be replaced by a noun (denoting a person). What is especially interesting about these variants, is the fact that with such a substitution, only two prepositions *na* 'on, to' and the parallel morphosyntactic structure (a noun in nominative – preposition *na* – a noun in accusative – comma – a noun in nominative – preposition *na* – a noun in accusative) are taken over from the original multi-word expression, yet it is recognized in an appropriate context without difficulty. Between the two parallel nominal groups, there is usually a comma, but the conjunction *a* 'and' or a dash can be there as well.

In total, 521 of these multi-word lexical units (with the whole structure, see below) occur in the SYNv6 corpus out, of which 464 occurrences (89%) appear in the basic, invariant form, 53 variants occur (10%) with the noun *brácha* being replaced (E.g.: *já na soudruha, soudruh na mě* 'me to comrade, comrade to me') and 4 occurrences can be found with both positions replaced by another noun. In case only the noun *brácha* is replaced (53 occurrences), the replacing noun often denotes another family member such as *ségra* 'sis, sister', *žena* 'woman, wife', *táta* 'dad', *děda* 'grandpa', *švára* 'brother-in-law', *kmotr* 'godfather' (22 occurrences in total) or friends: *kamarád* 'friend' (masc.), *kamarádka* 'friend' (fem.), *kámoška* 'friend' (fem., colloquial); the other occurrences are mostly influential persons and institutions: *ministr* 'minister', *vláda* 'government', *polda*, 'cop, policeman', or proper names (of politicians, etc.) as in example (8). One variant of this MWE appears 12 times: *já na Háchu, Hácha na mě*. It is a title of a historical theater play about *Hácha*, a Czech politician, the title is a word game using paronymy between the nouns *brácha* and the name *Hácha*.

(8) ***Já na Obamu, Obama na mě!*** *(...)* ***Já na bráchu, brácha na mě!*** *Jak se zdá, tohle heslo výborně zafungovalo i v případě netradičního spojenectví zpěvačky Beyoncé Knowles (27) a příštího amerického prezidenta Baracka Obamy (47).*

Me to Obama, Obama to me! (...) Me to brother, brother to me! This slogan seems to have worked well in the case of the unconventional alliance of singer Beyoncé Knowles (27) and the next US president Barack Obama (47).

In all four cases where both the pronoun *já* and the noun *brácha* were replaced, the replacing nouns were always the names of politicians as in example (9):

(9) ***Klaus na Majora, Major na Klause**. Český a britský premiér v sobě nalezli zalíbení.*

Klaus on Major, Major on Klaus. The Czech and British Prime Ministers have found mutual sympathy.

However, not all results matching the abovementioned formula are variants of the MWE *já na bráchu, brácha na mě*. It is necessary to carefully verify that there are no other reasons for using two parallel constructions with the preposition *na*: sometimes, this preposition is used as related to a verb, adjective or noun with valency *na* + accusative (frequent in Czech) as in example (10), where the use of the preposition *na* is motivated by the noun *žaloba* 'lawsuit' and the repetition of lemmas is due to the mutual relationship of both persons (*Zelníček* and *Vovsík*).

(10) *(...) do souboje, v němž se nyní **žalobou** ohání **Zelníček na Vovsíka** a **Vovsík na Zelníčka**.*

(...) into a duel in which Zelníček Vovsík and Vovsík Zelníček now threaten each other with a lawsuit.

Apart from the abovementioned variants, the saying *já na bráchu, brácha na mě* presents yet another type of variability: the use of a fragment, in particular the first part of the MWE *já na bráchu* 'me on brother' (often with the punctuation mark of ellipsis indicating that the MWE is not complete). The fragment sufficiently represents the whole expression as in example (11). Such occurrences can be found in the examined corpus 180 times. The saying is even frequently condensed into one noun *jánabráchismus*, used as a less formal synonym for *klientelismus* 'cronyism'; in the corpus we find 168 such occurrences.

(11) *Firma, kterou prosadil starosta, udělala s městem obchod a vydělala za rok bezmála dvacet milionů. Obvyklý obchod stylem „**já na bráchu...**".*

The company promoted by the mayor made a deal with the city and earned nearly twenty million in a year. The usual business style "me on brother...".

4. LEMUR, a database of Czech multi-word expressions

The case studies presented in the third part of this paper illustrate the variability of multi-word lexical units with the repetition of lemmas and, more generally, the variability of Czech multi-word expressions. So as to record this variability and to make it accessible to users for both their own study and the use of language processing tools, we have developed a database of multi-word lexical units LEMUR (**le**xicon of **mu**lti-word exp**r**essions). Currently it contains about 5000 MWEs: sayings, proverbs, weather lore, fixed comparisons, multi-word prepositions etc., hundreds of which were manually annotated in detail. The database contains a range of information about each of the multi-word lexical units, including their basic form, definition, syntactic structure, example of use, variability, idiomaticity, etc., as described in other articles, especially by Hnátková et al. (2017).

The database records the variability or fixedness of each MWE on the levels of word forms, word order, syntax, and lexicon. Generally, the database assumes that the modifications (like word order changes due to topic-focus articulation, passivizations or nominalizations) typical for the same Czech constructions are possible for any MWE, unless explicitly stated otherwise in the database.

4.1. Variation and fixedness in the LEMUR database

The question of word form variation or fixedness relates only to some MWEs in which there is no free choice of word forms with the same grammatical categories. For example, in the MWE *podle nosa poznáš kosa* 'someone's character can be recognized from her/his face', lit. 'after the nose you recognize the blackbird', the word form *nosa* is an unusual form of genitive singular of the noun *nos* 'nose', used almost exclusively in this expression. No other form of the noun *nos* can be used here (because of the rhyme). Otherwise, we assume a free variability of word forms (for example, in the expression *já na bráchu, brácha na mě* 'you scratch my back, I'll scratch yours', the accusative form *mě* 'me' of the pronoun *já* 'I' can be replaced with another equivalent form *mne* 'me', such examples were found in the corpus).

Similarly, the MWE database contains expressions for which the order of components cannot be changed, or no other word can be inserted between the components (e.g. the word order in the saying *čistému vše čisté* 'to the pure everything is pure' must be preserved, and no word can be inserted between the components, except for the verb *být* 'to be': *čistému **je** vše čisté*).

We also note whether there are any limitations of syntactic modification or syntactic variability, e.g. in case a verbal MWE cannot be passivized or nominalized or if it is not possible to modify a component of a MWE by another word. For example, in the MWE *Bůh dal, Bůh vzal* 'the Lord has given, the Lord has taken away', no syntactic variability is possible: the MWE cannot be passivized or nominalized, none of the components can be modified by other sentence members.

The lexical variability within the MWE is, on the contrary, explicitly defined. If a component is partially or completely lexically variable, a list of possible lemmas is recorded or other restrictions (if any) are noted in the database entry.

If a given MWE can be represented by a fragment, that is, a part of the multi-word lexical unit that identifies it and is used independently, as in example (11): *já na bráchu...*, these options are also described.

4.2. A database entry

In the database entry, which is used both for editing records and (at least so far) for viewing, a "slot" is defined for each MWE component. This slot represents a position in the MWE that can be filled with either a fixed or variable lemma. For example, the entry for the MWE *čistému vše čisté* 'to the pure everything is pure' is represented by four slots:

1. *čistému*: in the invariant form, the adjective *čistý* 'pure' in dative masculine singular; otherwise any adjective in the dative case.
2. *je*: optional verb *být* 'to be' in the third person of the present tense.
3. *vše*: nominative singular of the pronoun *všechno* 'all': forms *vše* or *všechno*.
4. *čisté*: adjective in nominative singular neuter, the same lemma as in the first slot.

For the whole MWE, the impossibility of word order changes, insertion of words or component modification is recorded.

5. Automatic identification of multi-word units with repetition of lemmas

In order to study multi-word lexical units in real use, it is not only necessary to have an extensive database of MWEs, but such expressions have to be identified in texts (in corpora), marked and linked to the database. For automatic annotation, we use (for now) a modified version of the system FRANTA, described by Kopřivová & Hnátková (2014), which identifies

MWEs based on word forms, lemmas and morphosyntactic tags. When any MWE is found, it is assigned a special MWE lemma. Via these lemmas, the MWEs in texts are linked to the LEMUR database.

In most cases, multi-word lexical units with lemma repetition are identified only in their basic, invariant form; moreover, some more frequent lexical variants of such MWEs are also identified. A more general approach concerning the identification of a morphosyntactic pattern and testing for the lemma identity is very costly in terms of computer resources (it is slower by at least two orders of magnitude), therefore new variants are identified solely in a few frequent cases comprising important lexical variability.

In the future, we are planning to develop new software that would directly use the export of information from the database, so it would not be necessary to link database entries with lemmas manually entered into the FRANTA system. The FRANTA system is not flexible enough and needs to be manually edited when adding new MWEs.

For more complex MWEs, it will probably be always necessary to concentrate only on documented lexical variants. A typical example of such complex MWE is *já na bráchu, brácha na mě* (see part 3.4), in which all autosemantic words can be replaced, an automatic query for any such variants would have to be based only on a morphosyntactic pattern, a frequent preposition and a test of lemma repetition. Moreover, the morphosyntactic pattern (N_nom. *na* N_acc., N_nom. *na* N_acc.) is not sufficiently distinctive, and may be motivated by the valency of another nearby word as in example (10), which does not exemplify the given MWE. In order to identify variants of this multi-word lexical unit with both lemmas replaced, it is therefore necessary to understand the text and thus to check it manually.

At present, we are unable to search automatically for new, not yet identified multi-word lexical units with lemma repetition in general (e.g. using association measures and testing for lemma repetition). The best way is to identify a new MWE using standard methods (association measures etc.), check it manually and, if there is a repetition of lemmas in the MWE, search for lexical variability preserving lemma repetition.

6. Conclusions

The variability of Czech multi-word lexical units with repetition of lemmas was presented. This type of MWE is a phenomenon deserving more detailed study: it is interesting to observe lexical variability related to MWE reproducibility, i.e. whether a multi-word expression is recognizable after

replacing the original repeated lemmas or to record what the extent of such variability is and how often such replacement occurs. MWEs of a similar character appear in many European languages, some expressions are used in many languages (***crows*** *will not pick out another* ***crow****'s eyes*; ***corbeaux*** *avec* ***corbeaux*** *ne se crèvent jamais les yeux*; ***vrána vráně*** *oči nevyklove*; ***во́рон во́рону*** *глаз не вы́клюет*), others are language specific (*a* ***friend*** *in need is a* ***friend*** *indeed*; *il faut* ***manger*** *pour* ***vivre****, et non pas* ***vivre*** *pour* ***manger***; ***дру́жба дру́жбой****, а* ***слу́жба слу́жбой***), so it is possible to examine the variability of similar multi-word expressions across languages.

We have shown that the variability of individual MWEs can be defined and recorded in a database that is both human-readable and computer-readable. Linking corpora in which MWEs are labeled and lexical databases containing a detailed description of such expressions will allow for a more precise and easier research of multi-word lexical units.

Acknowledgement

This paper is part of the project *Between Lexicon and Grammar* (2016–2018), supported by the Grant Agency of the Czech Republic, reg. no. 16-07473S. This project is a follow-up of the project entitled *The Grammar-Based Treebank of Czech* (2013–2015, cf. Skoumalová et al. 2014; Petkevič et al. 2015a, 2015b) and devoted to automatic parsing driven by a formal HPSG-like grammar of Czech.

Bibliography

Čermák, František. 2007. *Frazeologie a idiomatika česká a obecná*. Prague: Karolinum.

Fischer, Andreas. 1994. *Repetition*. Günter Narr Verlag, Tübingen.

Frédéric, Madeleine. 1985. *La répétition: Étude linguistique et rhétorique*. Tübingen: Max Niemeyer Verlag.

Hnátková, Milena et al. 2017. Eye of a Needle in a Haystack. Multiword Expressions in Czech: Typology and Lexicon. In: Mitkov, Ruslan (ed.). *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017*, London, UK, November 13–14, 2017, Proceedings. Springer: Berlin–Heidelberg, p. 160–175.

Kopřivová, Marie; Hnátková, Milena 2014. *From Dictionary to Corpus. Phraseology in Dictionaries and Corpora*. Maribor, Filozofska fakulteta Maribor, p. 155–168.

Křen, Michal et al. 2017. *Corpus SYN, version 6 from 12/18/2017*. Prague: Institute of Czech National Corpus. http://www.korpus.cz.

Křen, Michal et al. 2015. *Corpus SYN2015: a representative corpus of contemporary written Czech*. Institute of Czech National Corpus FF UK, Prague 2015. http://www.korpus.cz.

Wielowyrazowe jednostki leksykalne z powtórzonymi leksemami w języku czeskim i identyfikacja ich wariantów

Streszczenie

Niniejszy artykuł przedstawia wyniki badania zmienności jednostek wielowyrazowych z powtórzonymi leksemami, takich jak *Bůh dal, Bůh vzal* 'Bóg dał, Bóg wziął', opartego na dużym ilościowo korpusie. Cztery studia przypadków ilustrują zakres zmienności tego rodzaju połączeń wyrazowych. Opisano bazę danych związków wyrazowych, ze szczególnym uwzględnieniem ustabilizowanych konstrukcji z powtórzonymi komponentami i ich wariantów. Ponadto omówiono automatyczną identyfikację analizowanych frazeologizmów.