Marie Kopřivová
Charles University
Czech Republic
https://orcid.org/0000–0001–7390–0753

Variability of Czech Verbal Phrasemes: Case Study of *dát* ('to give')[1]

Abstract. The paper is concerned with the topic of variability of Czech verbal idioms and its representation in a database of a multi-word expressions. In terms of material, it is based on SYN2015, a representative corpus of contemporary written Czech, which is equally divided into fiction, non-fiction, and newspapers and magazines. This corpus features an automatic annotation of multi-word units. The verb *dát* 'to give' serves as a case study, being one of the most frequent verbal components of Czech verbal idioms, right after the verbs *být* 'to be' and *mít* 'to have'.

Key words: *verbal idiom, corpus, multi-word expression, variability*

1. Introduction

The verb *dát* ('to give') is the third most frequently used verb in Czech phrasemes[2]. It is a ditransitive verb with arguments in the dative and the accusative. It retains this valency in most phrasemes. Its imperfective aspect counterpart is the verb *dávat*, which ranks seventh, with the interesting property that 40% of its occurrences are in phrasemes.

The semantically opposite action is denoted by verbs *dostat* and *dostávat* ('to get'), which rank 12th and 240th, respectively. As phraseme components, they are less frequent than the verb to give (see Table 1). This frequency

[1] This study was written within the project Between Lexicon and Grammar, supported by the Grant Agency of the Czech Republic, reg. No 16-07473S.

[2] The term phraseme is used here especially with regard to the formal criteria for its definition (see Čermák, 2007, 83), which includes the term quasi-phraseme to (see Čermák, 2007, 104), which are verbo-nominal structures with an abstract noun.

disproportion is one of the signals that hint at the semantic bleaching of this verb, which is also manifested in collocations (cf. also the valency lexicon by Lopatková et al.).

Table 1 shows the most frequent verbs used in phrasemes in the SYN2015 corpus. It is possible to see here that the verbs *dát* and *dávat* (to give) have the highest proportion of occurrences within phrasemes compared to the other verbs.

For Czech corpora, automatic annotation of phrasemes can be performed using the FRANTA tool (Kopřivová – Hnátková, 2014), which is based on the Dictionary of Czech Phraseology and Idioms (Čermák et al., 2009). This annotation has so far been applied only to written corpora. Each collocation[3] component has a collocation lemma and a collocation tag. This annotation should be improved, supplemented and clarified by the newly-developed database LEMUR (see Hnátková et al., Jelínek, 2019).

Table 1. The most frequent verbs occurring within phrasemes in the SYN2015 corpus

Rank	Verb		Occurrences			Different idioms
			Total	Phrasemes		
1	být	be	4,044,082	122,363	3%	3444
2	mít	have	734,066	94,341	13%	1417
3	dát	give	112,636	25,792	23%	529
4	říci	tell	180,576	15,555	9%	170
5	jít	go	160,655	14,059	9%	391
6	dělat	do	75,480	13,720	18%	339
7	dávat	give	29,075	11,983	41%	161
8	stát	stand	182,776	10,115	6%	239
9	vzít	take	44,928	9,060	20%	224
10	vědět	know	151,747	9,009	6%	176
11	nechat	let	54,787	8,311	15%	161
12	dostat	get	82,741	8,270	10%	307
240	dostávat	get	12,052	352	3%	27

Source: own research.

[3] This annotation is also applied to other types of collocations, such as terms.

2. Data sources

For the analysis of the verb *dát* (to give) in corpus texts, we used two corpora of written Czech. The first corpus, SYN2015, is a representative corpus of contemporary written Czech, with balanced proportions of fiction, nonfiction, and newspapers & magazines (one third each). It contains 100 million word forms, lemmatization and POS and MWE tagging. Circa 4% of the word forms are marked as components of an MWE (mostly phrasemes). It was selected for analysis because it includes different types of texts in and it is balanced.

However, the range of attested phrasemes depends on the size of the corpus. This can be seen when automatic tagging of MWEs is applied to a large corpus, such as SYN_v6, which counts some 4 billion word forms. It is only in this larger corpus that some little-used phrasemes and proverbs appear, albeit with occurrence rates in the single digits. This unbalanced corpus of written language, with a predominance of journalistic texts, was the starting point for an analysis complementing the types and variants of verbal phrases with verbal component *dát* (to give).

3. The verb *dát* (to give)

The Dictionary of Czech Phraseology and Idioms catalogues 466 different phrasemes under the verb *dát* (to give). In the SYN2015 corpus, 529 different phrasemes are annotated, and in the SYN_v6 corpus, 712 different phrasemes. These numbers also include cases where the verb is associated with a reflexive pronoun (*se* or *si*). In the SYN2015 corpus, the numbers are separated: *dát* alone 349 different phrasemes, *dát* + *si* 106 different phrasemes, *dát* + *se* 74 different phrasemes.

For the purposes of the analysis, phrasemes were divided into nine formal groups. These groups were then analyzed in the SYN_v6 corpus. We only describe two groups in more detail.

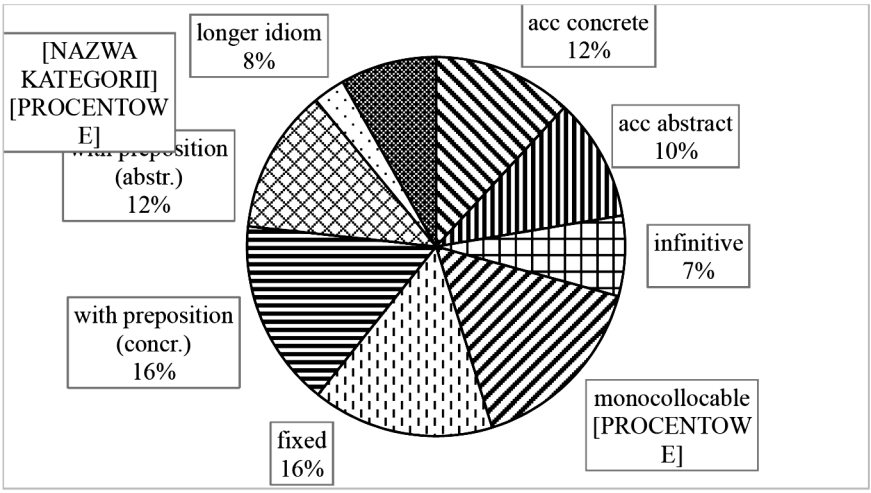
3.1. Dividing the phrasemes into groups

1. verb + acc (abstract noun)
   Example: *dát přednost (to give priority)* *to prefer*
2. verb + acc (concrete noun)
   Example: *dát pusu (to give mouth – kiss)* *to kiss*

3. verb + infinitive
   Example: *dát vědět (to give to know)* *to alert*
4. verb + preposition + abstract noun
   Example: *dát za pravdu (to give for the truth)* *to agree*
5. verb + preposition + concrete noun
   Example: *dát k ledu (to give to the ice)* *put on ice*
6. verb + monocollocable word
   Example: *dát bacha* *watch it/be careful*
7. verb in negation
   Example: *nedat dopustit (do not give allow)* *to believe/to protect someone*
8. fixed idiom
   Example: *to dá rozum (it gives sense)* *it is common sense*
9. longer idiom
   Example: *dát ruku do ohně (put hand in fire)* *to be sure*

The distribution of these groups in the representative corpus SYN2015 is shown in Figure 1. The proportions are derived from the frequencies of all occurrences, not different types of phrasemes. The largest groups are fixed idioms, idioms with a monocollocable component and phrasemes with a preposition. For further detailed analysis, we chose the combination of the verb with the accusative.

Figure 1. Distribution of the *dát* phraseme groups in SYN2015



Source: own research.

3.2 Analysis of *dát* + acc

Using the SYN_v6 corpus, we searched for instances of the accusative following immediately after the verb and which were *not* annotated as phrasemes. This was the case of 61% of all occurrences. For a detailed analysis, the 115 most common nouns were selected, with their frequency ranging from 25,225 occurrences to 400 occurrences.

3.2.1. verb + acc (abstract noun)

A larger part of the analyzed sample consists of combinations of verb and abstract noun. These are classical verbo-nominal combinations (light verb constructions) which were missing in the dictionary, e.g.: *dát hlas* ('vote'), *možnost* ('possibility'), *odpověď* ('answer'), *příkaz* ('command'), *svolení* ('permission'), *podmínku* ('condition'). They are in competition with equivalent verbal expressions and their frequency increases at their expense.

The most frequent collocations are those that do not have a one-word equivalent, e.g.: *dát šanci* ('chance'), *prostor* ('space'), *průchod* ('passage'), *příležitost* (opportunity), *podnět* (impulse).

Their frequency is on the rise, in some cases perhaps under the influence of English[4].

Some of these expressions are found in the dictionary with another verb, e.g.: *dát prostor – vytvořit* (space), *dát důvěru* (*vyjádřit důvěru*), *dát maximum* (*vložit maximum*). Together with the increasing frequency, there is a change of meaning, shifting to a more general plane, which confirms the process of grammaticalization of this verb. Individual expressions get simplified: more common vocabulary is used, which is less demanding and therefore corresponds to an informal style.

3.2.2. verb + acc (concrete noun)

With concrete nouns, the verb conserved its original meaning in many cases (e.g. to give a present, money, put your feet on the table).

In combination with the reflexive pronoun *si*, as in *dát si*, the meaning is "order/have some food or drink". The collocations *dát si panáka* (have an alcoholic drink) and *dát si jídlo* (have some food) are annotated. Other proto-

[4] The English influence can be seen for example on the verb *dát* which newly occurs in spoken Czech with sole accusative valency, meaning "to handle something". These cases were previously covered in the Czech language by the verb *udělat* (to do), e.g. *dát zkoušku* (pass the exam).

typical representatives included the following 9: *pivo* (beer), *káva//kafe* (coffee), *čaj* (tea), *cigareta* (cigarette), *sklenička/sklenka* (little glass), *oběd* (lunch), *jídlo* (food). The shift that is captured by the corpus is not only from eating and drinking to other enjoyments such as smoking, but also towards metonymic equivalents of beverages (*sklenička* – glass). The glass here acts as a synonym for *panák* – which can express the amount of a drink or type of a drink – alcoholic.

Six other nouns come from sports news and are closely related to the second most common annotated collocation, *dát gól* (to score a goal). One is a synonym (*branku*), the remaining ones are either other football "terms" (*dát míč, balón* – pass the ball, *dát hattrick*, *penalty*) or pertain to another type of sport (*dát koš* – to score in basketball).

The remaining cases are verbo-nominal collocations which are used instead of their one-word verbal equivalent: *dát inzerát – inzerovat* (give an ad).

Conclusion

This partial analysis of the collocations of the verb *dát* (to give) has shown the importance of new corpus data. As it has been ten years since the *Dictionary of Czech Phraseology and Idioms* was released, a number of common phrasemes or their variants are missing. Using the corpus can reveal variants when verbs are alternated and determine which verb should be entered as default in the dictionary. It is necessary for users to be able to find all the variants if they know a less frequent one. The analysis showed the need to add new phrasemes to the repertoire.

The analysis also demonstrates the ongoing grammaticalization of the verb *dát*. Also, of course, it confirms that phrasemes with this verb are often used. With respect to the needs of the new LEMUR database and in the interest of better annotation of collocations in corpora of Czech, it is necessary to supplement this analysis of the verb data by analyzing the verb *dávat* (to give, imperfective) and verbs with opposite meaning *dostat*, *dostávat*. Together with the further analysis of phrasemes, this will form the basis for a better theoretical description of Czech phraseology.

Literature

Čermák, Fantišek 2007. *Frazeologie a idiomatika česká a obecná*. Praha: Karolinum.

Čermák, František et al. 2009. *Slovník české frazeologie a idiomatiky*. Vol. 1–4. Praha: Leda.

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

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

Jelínek, Tomáš. 2021. Multi-word lexical units with repetition of lexemes in Czech and identification of their variants. (this volume).

Kettnerová, Václava. 2017. Syntaktická struktura komplexních predikátů v češtině. In: *Slovo a slovesnost*, 78, 3–24.

Sag, Ivan A.; Baldwin, Timothy; Bond, F.; Copestake, Ann; Flickinger, Dan. 2002. Multiword expressions: a pain in the neck for NLP. In: *Computational Linguistics and Intelligent Text Processing: Third International Conference, Cicling 2002*. Alexander Gelbukh (ed.). Berlin–Heidelberg: Springer. 1–15.

Corpora and tools

Křen, Michal; Cvrček, Václav; Čapka, Tomáš; Čermáková, Anna.; Hnátková, Milena; Chlumská, Lucie; Jelínek, Tomáš; Kováříková, Dominika; Petkevič, Tomáš; Procházka, Pavel; Skoumalová, Hana; Škrabal, Michal; Truneček, Petr; Vondřička, Pavel; Zasina, Adrian J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA. 2522–2528.

Křen, Michal; Cvrček, Václav; Čapka, Tomáš; Čermáková, Anna; Hnátková, Milena; Chlumská, Lucie; Jelínek, Tomáš; Kováříková, Dominika; Petkevič, Tomáš; Procházka, Pavel; Skoumalová, Hana; Škrabal, Michal; Truneček, Petr; Vondřička, Pavel; Zasina, Adrian J. *Corpus SYN, version 6 from 18.12.2017. Ústav Českého národního korpusu FF UK, Praha 2017*. http://www.korpus.cz.

Lopatková, Markéta; Žabokrtský, Zděnek.; Kettnerová, Václava. *VALLEX 2.5. Valency Lexicon of Czech Verbs*. http://ufal.mff.cuni.cz/vallex.

Wariantywność czeskich frazemów werbalnych: Studium przyopadku „*dát*" ('dać')

Streszczenie

Artykuł poświęcony jest wariantywności czeskich idiomów werbalnych i ich występowaniu w bazie połączeń wyrazowych. Materiał został zaczerpnięty z SYN2015, reprezentatywnego korpusu współczesnego pisanego języka czeskiego, w którym uwzględniono proporcjonalnie teksty fikcyjne, tekst niefikcyjne, prasę i czasopisma. Korpus umożliwia automatyczną anotację związków wyrazowych. Czasownik *dát* 'dać' posłużył jako przedmiot studium przypadku, ponieważ jest jednym z najczęstszych komponentów czasownikowych występujących w czekich idiomach, zajmując trzecie miejsce po czasownikach *být* 'być' i *mít* 'mieć'.