Synonymy Between Theory and Practice: The Corpus-based Approach to Determining Synonymy in Lexicographic Descriptions in Croatian

Goranka Blagus Bartolec^[0000-0002-3577-7026]

Institute of Croatian Language and Linguistics, 10 000 Zagreb, Croatia
gblagus@ihjj.hr

Abstract. The aim of this research is to determine the degree of synonymy among words based on a corpus search in hrWaC 2.0 (Croatian Web Corpus, Version 2). First, we analyze the lexicographic determination of absolute synonymy at the level of the dictionary definition (where the independent meaning of the words is extracted from usage context). The major part of the research focuses on the usage of synonyms based on corpus data. The analysis includes two pairs of synonymous nouns in Croatian: *krevet* and *postelja* ‘bed’ and *bijes* and *gnjev* ‘rage’. The ten most common multiword units (MWUs) of these synonyms were selected as attested in the corpus according to raw frequency. According to the obtained results, we determined how many synonyms are distanced from each other when compared to other words, whereby these words obtain the attributes of partial synonyms instead of absolute synonyms. Despite the fact that the context permits the use of both synonyms, it was confirmed that only one is significantly prevalent in some MWUs. The corpus-based approach imposes a different manner of defining absolute synonyms in dictionaries – apart from describing the basic semantic features of the word, the definition must consider the contextual representation and the collocational perspective of synonymous words. Corpus results can improve synonymous relationships among words in lexicographic descriptions.

Keywords: Corpus-based Approach to Synonymy, Croatian, Lexicographic Synonymy, Multiword Units

1 Synonymy as a lexicographic feature

Determining synonymous relationships in dictionary descriptions is a fundamental task of every lexicographer, one which contributes to both the semantic features of words and connecting words that have a different signifier, but the same signified. At the level of the dictionary definition many words can be described with the same definition as absolute synonyms, which presupposes that they should be interchangeable in all meanings and usage contexts [1: 268], [2: 107]. Three contemporary dictionaries of Croatian [3], [4], [5] show many synonyms, i.e. words that are equal or similarly defined. At the cognitive level [1: 270], if we only consider certain semantic values, we can agree that some Croatian words behave as absolute synonyms (eg. *bedro* and *natkoljenica* ‘thigh’). When words from the dictionary are put into a real context

and co-occur with other words in different multiword units (MWUs), there is a gap between absolute synonyms and their behaviour in practice. Many synonyms, even if the context allows, cannot be replaced by a synonymous mate with some collocates. At the level of contextual representation, it is more precise to discuss partially synonymous relationships between words [1: 285], [2: 107], [6: 60–61] because, although they have the same meaning, they are not absolutely substitutable in all contexts. Corpus tools [7], [8] enable exhaustive searches that provide statistical and other data about the usage potential of synonyms in the same context. The following analysis focuses on this aspect.

2 The corpus-based approach to synonymy on the collocational level

According to [9: 122], corpus searches allow synonyms to be observed from two perspectives: (i) the co-occurrence approach, which implies the collocation potential of individual words at the MWU level, (ii) the substitution approach, which determines the degree of synonymy depending on the substitutability of a word in a specific context. These two approaches actually include frequency, collocability, and preferences by style and type of text [10: 20]. Two synonym pairs were selected for analysis: the concrete (countable) nouns *krevet* and *postelja* ‘bed’ and the abstract (uncountable) nouns *bijes* i *gnjev* ‘rage’. The analysis includes the 10 most common MWUs selected by raw frequency. As open class words, the most frequent verbs, adjectives, and nouns were chosen as collocation candidates, while conjunctions, prepositions, and pronouns, as closed class words, are not taken into account because they are not considered semantic constituents of MWUs.

2.1 The synonymous pair *krevet/postelja* ‘bed’

The synonyms *krevet* and *postelja* usually collocate with adjectival collocates on the left. The ratio of the lemma *krevet* to the lemma *postelja* in hrWaC is 85,994 to 9,179, which indicates that the lemma *krevet* appears about 9.4 times more often than the lemma *postelja*. This result clearly shows that the noun *krevet* prevails in everyday language use in many different contexts in Croatian. From the aspect of substitutability, the noun *postelja* is not used in the same context as the noun *krevet*. A search for stable MWUs with adjective collocation candidates was carried out with CQL using regular expressions [tag="A.*"] [lemma="krevet/postelja"]. We chose the ten most frequent adjectives that classify the noun by type or purpose. Adjectives that describe the quality of the noun or possessive adjectives (eg. *velik* ‘large’, *nov* ‘new’, *isti* ‘same’), *tuđ* ‘foreign’, *vlastit* ‘one’s own’, *cijeli* ‘whole’, *topao* ‘warm’, *prazan* ‘empty’, *mek* ‘soft’, *zajednički* ‘shared’, *udoban* ‘comfortable’) were excluded from the analysis, as they are components of free combinations, not of MWUs [11, 12].

Table 1. Frequency distribution of the lemmas *krevet* and *postelja* and their first ten adjective collocation candidates in hrWaC. Common MWUs are in bold.

Query A. *, <i>krevet</i>	total: 12,039	Query A. *, <i>postelja</i>	total: 3,111
bračni krevet 'double bed'	1,596	bolesnička postelja 'sick bed'	492
bolnički krevet 'hospital bed'	950	samrtna postelja 'deathbed'	377
bolesnički krevet 'sick bed'	692	bračna postelja 'double bed'	273
vodeni krevet 'water bed'	223	bolnička postelja 'hospital bed'	204
pomoćni krevet 'additional bed'	165	samrtnička postelja 'deathbed'	101
slobodan krevet 'free bed'	154	hotelska postelja 'hotel bed'	98
hotelski krevet 'hotel bed'	153	turistička postelja 'tourist bed'	63
dječji krevet 'cot'	133	slobodna postelja 'free bed'	37
dodatni krevet 'extra bed'	102	privatna postelja 'private bed'	17
francuski krevet 'french bed'	84	registrirana postelja 'registered bed'	15

The search for the lemmas *krevet/postelja* showed the following: the lemma *krevet* appears with adjectives almost four times more often than lemma *postelja* [Table 1]. The ratio of matching and non-matching of adjective collocates for both lemmas is 50%. Common adjectives with lemmas *krevet* and *postelja* are *double* (lit. 'marriage'), *hospital*, *sick*, *free*, and *hotel*, and no others match. The substitutability criterion for the nouns *krevet* and *postelja* is not applicable with the adjectives *watery* and *French*. The corpus search showed that only *vodeni krevet* 'water bed', lit. 'watery bed' and *francuski krevet* 'French bed' appear as MWUs, and there are no matches for the MWUs *vodena postelja* i *francuska postelja*. It should be noted here that we specifically searched for the MWU *krevet/postelja na kat* 'bunk bed', lit. 'multi-storey bed'. The MWU *krevet na kat* appears 553 times, and the MWU *postelja na kat* appears 7 times. The MWU *postelja na kat*, according to the corpus results, appears in regionally marked texts (which offer tourist accommodation in the Croatian region of Dalmatia), and it is not affirmed in general language use. Including both the co-occurrence approach and substitution approach, the noun pair *krevet* and *postelja* are only partial synonyms, as the noun *krevet* prevails in usage and shares only 50% of the most frequent adjective collocates with the noun *postelja*. The noun *postelja* is limited to some contexts that reflect stylistic use (*samrtnička/samrtna postelja* 'death-bed') or regional/local language use (*slobodna postelja* 'free bed', *postelja na kat* 'bunk bed'), but the noun *krevet* prevails in neutral or ordinary language use (*slobodan krevet* 'free bed', *krevet na kat* 'bunk bed').

2.2 The synonymous pair *bijes/gnjev* 'rage'

The three Croatian dictionaries reviewed describe the abstract nouns *bijes* and *gnjev* 'rage' as absolute synonyms with the meaning of 'a very strong/great state of displeasure and anger'. Compared to the synonymous pair *krevet/postelja*, the synonyms *bijes* and *gnjev* have almost double the raw frequency, but since they are abstract meanings, they do not have strong collocation potential for MWUs. The corpus search was not limited to regular expressions, rather the first ten collocation candidates on the left (-1) and on the right (1) were chosen from the lemmas *bijes* and *gnjev*. Closed class words have been removed from the list, and adjectives, nouns, and verbs are taken into consideration.

Table 2. Frequency distribution of the lemmas *bijes* and *gnjev* and their first ten collocation candidates in hrWaC. Common MWUs are in bold.

Query <i>bijes</i> 'rage'	total: 22,189	Query <i>gnjev</i> 'rage'	total: 6,004
izazvati bijes 'to cause rage'	620	Božji gnjev 'God's rage'	295
izljev bijesa 'outburst of rage'	567	izazvati gnjev 'to cause rage'	172
navući bijes 'to fill with rage'	368	pravedan gnjev 'just rage'	137
napad bijesa 'attack of rage'	339	pravednički gnjev 'righteous rage'	122
ispad bijesa 'outburst of rage'	298	navući gnjev 'to fill with rage'	102
nalet bijesa 'flurry of rage'	229	gnjev javnosti 'rage of the public'	71
iskaliti bijes 'to wreak rage on sb.'	194	velik gnjev 'big rage'	71
bijes javnosti 'rage of the public'	182	opravdan gnjev 'justified anger'	63
napadaj bijesa 'attack of rage'	180	izljev gnjeva 'outburst of rage'	59
provala bijesa 'outbreak of rage'	124	pun gnjeva 'full of rage'	59

The substitutability of the synonyms *bijes* and *gnjev* in the same collocation context according to the obtained results [Table 2] is 40% (*izazvati* 'to cause', *navući* 'to fill with', *javnost* 'public', *izljev* 'outburst',). The lemma *bijes* often co-occurs with verbs (*cause*, *wreak*, *fill with*) and nouns (*outburst*, *flurry*, *attack*, *outbreak*) which specifies a process or action. The noun *gnjev*, in addition to verbs and nouns with meanings of action, usually co-occurs with adjectives or nouns that denote the holder of rage (*God's*, *righteous*, *the public*) which indicate stylistic (ethical or religious) contexts of use. The noun *bijes* has been expanded in ordinary, non-stylistic use.

3 Conclusion: Absolute synonymy on the collocational level?

The corpus-based approach indicates stylistic and regional limitations in the use of some words that dictionaries define as absolute synonyms. However, the corpus does not provide an answer to the question as to why synonyms are not substitutable in all contexts. Is contextual non-substitutability simply a matter of the communications habits, or is synonymous substitution truly impossible in some contexts? In order to answer this question, in addition to the semantic potential of synonymous pairs, it would be necessary to explore the semantic potential of the most common collocates and to determine why only one synonymous mate co-occurs with some collocates while the other does not have the same potential. The lexicographer should not take corpus results for granted, which implies that care should be taken in defining absolute synonyms. The corpus suggests that (i) absolute synonymy, which is based on the cognitive principle of determining "certain semantic properties in common" [1: 270], should be distinguished at the level of dictionary definitions so that several words can have the same definition in dictionaries, and (ii) partial synonymy depends on both the communication context and the collocates with which synonymous words most often co-occur. Given that stable collocations should confirm the definitions of words in the dictionary, the results of these corpus searches evidently show that absolute synonymy is a very limited lexical phenomenon [2:107], and that it is difficult to associate two dictionary entries that match in both their definitions and in their frequent MWUs.

Acknowledgments

This paper is written within the research project Croatian Web Dictionary – MREŽNIK (IP-2016-06-2141), financed by the Croatian Science Foundation. The research presented in this paper is mostly based on the lexicographic sources of the Institute for Croatian Language and Linguistics.

References

1. Cruse, A.: Lexical semantics. Oxford University Press, Oxford (1986).
2. Edmonds, Ph., Hirst, G.: Near-synonymy and lexical choice. *Computational Linguistics* 28(2), 105–144 (2002)
3. Birtić, M. et al.: Školski rječnik hrvatskoga jezika. Institut za hrvatski jezik i jezikoslovlje – Školska knjiga, Zagreb (2012).
4. Hrvatski jezični portal, <http://hjp.znanje.hr/>.
5. Rječnik hrvatskoga jezika. Šonje, J. (eds.), Leksikografski zavod „Miroslav Krleža“ – Školska knjiga, Zagreb (2000).
6. Lyons, J.: *Linguistic Semantics: An Introduction*. Cambridge University Press, Cambridge (1995).
7. Ljubešić, N., Erjavec, T. hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In: Habernal, I., Matousek, V. (eds.) *Text, Speech and Dialogue 2011*, pp. 395–402. Springer, Heidelberg (2011).
8. Ljubešić, N., Klubička, F. {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: Bildhauer, F., Schäfer, R. (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29–35. Association for Computational Linguistics, Gothenburg: (2014).
9. Gries, S., Otani, N.: Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34, 121–150. (2010).
10. Kjellmer, G.: Synonymy and corpus work: On *almost* and *nearly*. *International Computer Archive of Modern and Medieval English* 23, 19–28 (2003).
11. Apresyan, Y.: *Leksicheskaya semantika. Sinonimicheskie sredstva yazyka*. Nauka, Moscow (1974).
12. Mel’čuk, I.: Collocations and Lexical Functions. In: Cowie, A. P. (eds.) *Phraseology: Theory, Analysis and Applications*, 23–53. Clarendon Press. Oxford (1998).