DEFINING COLLOCATION FOR SLOVENIAN LEXICAL RESOURCES

Iztok KOSEM
Faculty of Arts, University of Ljubljana; Jožef Stefan Institute

Simon KREK
Jožef Stefan Institute

Polona GANTAR
Faculty of Arts, University of Ljubljana

Kosem, I., Krek, S., Gantar, P. (2020): Defining collocation for Slovenian lexical resources. Slovenščina 2.0, 8(2): 1–27.

DOI: https://doi.org/10.4312/slo2.0.2020.2.1-27

In this paper, we define the notion of collocation for the purpose of its use in machine-readable language resources, which will be used in the creation of electronic dictionaries and language applications for Slovene. Based on theoretical and lexicographically-driven studies we define collocation as a lexical phenomenon, defined by three key aspects: statistical, syntactic, and semantic. We take lexicographic relevance as a point of departure for defining collocations within the typology of word combinations, as well as for distinguishing them from free combinations. Free combinations are (frequent) syntactically valid word combinations without lexicographic value and consequently there is no need for the description of their meaning, or syntactic role. Next, we distinguish collocations from all multiword lexical units (compounds, phraseological units and lexico-grammatical units) using the lexicographic view that multiword lexical units, whose meaning is not a sum of its parts, require a description of their meaning whereas collocations do not. In the final part, we return to the three aspects of collocation and their role in automatic extraction of collocational information from corpora. Semantic criterion or dictionary relevance of extracted collocations has particularly exposed the problem of semantically broad collocates such as certain types of adverbs, adjectives and verbs, and word which feature in different syntactic roles (e.g.

pronouns and adjuncts). We discuss a particular issue of collocations related to proper names and the decisions about their inclusion into the dictionary based on the evaluation of lexicographers.

Keywords: collocation, multiword lexical unit, word combination, Slovene, lexicography, dictionary database

1 INTRODUCTION

The inclusion of collocations in machine-readable language resources, which are used in the creation of electronic dictionaries and language applications, requires a detailed, yet general enough, definition of the notion of collocation. It is important that such a definition can be applied in the development of language technologies as well as in language description, in our case in the compilation of Dictionary of Modern Slovene (Gorjanc et al., 2017). Majority of studies that describe collocation as a lexically relevant phenomenon mention three key aspects: (i) statistical, which defines collocation as a statistically significant combination of two or more words, (ii) syntactic, which expects certain syntactic relations between words, and (iii) semantic, which presupposes that a collocation has a specific communication role. The latter aspect has made collocations since their "beginnings" (Firth, 1957; Altenberg, 1991; Sinclair, 1991) a lexical phenomenon that is lexicographically relevant and especially important for non-native speakers of a language (Palmer, 1933).

Considering these established notions of collocations, our paper has two aims. Firstly, we want to identify characteristics that define collocations as lexically relevant units. By this we mean that collocations are observed as an important part of lexis and worth including into language resources, intended for the creation of dictionaries, language tools and further computer processing (Klemenc et al., 2017). Secondly, we want to define collocations within all types of word combinations, especially in terms of their syntactic and semantic characteristics, which is important when considering their "place" in the dictionary database as well as their description aimed at human users.

The paper is structured as follows. First, the basic notions that describe collocation as a lexically relevant phenomenon are presented. Considering that collocation is a combination of at least two words, it means that we need to

consider its relation to all types of word combinations, taking into account the specifics of lexicographic workflow and automatic data extraction from corpora. In Section 3, we describe a typology developed in the compilation of Slovene Lexical Database (Gantar, 2015), which distinguishes between different types of lexicographically relevant multiword units. Next, we present parameters for automatic extraction of collocation candidates from the corpus, and discuss problematic points discovered during the evaluation. Automatically extracted collocation candidates that were deemed as bad or not relevant are divided into four groups according to their nature: problems in corpus annotation, problems related to statistical criteria, problems related to syntactic criteria, and problems related to semantic criteria (or dictionary relevance). We conclude the paper by discussing steps for improving automatic extraction of collocations from corpora, and offering some solutions for the presentation of collocations as dictionary units.

2 COLLOCATION AS A LEXICAL PHENOMENON

In the study of collocations, the approaches differ depending on how general or narrow the definition of collocation intends to be, and on the purpose of the definition, for example when including collocations in a dictionary. Although different approaches according to their purpose (different types of dictionaries, language learning, natural language processing etc.), focus on different characteristics of collocations, their definitions of collocation revolve around three criteria: statistical, syntactic and semantic.

2.1 Statistical criterion

One of the key characteristics when defining collocation is its statistical value, which must be higher than random, or as Atkins and Rundell (2008, p. 302) state, collocation is "a recurrent combination of words, where one specific lexical item (the 'node') has observable tendency to occur with another (the collocate) with a frequency higher than chance". A great body of research exists on measuring collocation strength or collocativity (e.g., Berry-Rogghe, 1973; Church and Hanks, 1990; Church et al., 1991; Biber, 1993; Manning and Schütze, 1999; Evert, 2004; Gries, 2013). There are different statistical methods, i.e. association measures, used. Association measures are regularly being compared, and

new ones proposed. Two good overviews of association measures are Wiechmann (2008) who compares 47 different association measures, and Pecina (2009) who conducts a comparison of more than 80 measures for collocation extraction. The general observations of the majority of such overview studies are aptly summarized by Evert (2009), namely that "different association measures will produce entirely different rankings of the collocates" (ibid., p. 1218) and "there is no ideal association measure for all purposes" (ibid., p. 1236).

As will be shown in the next sections, testing of automatic extraction of collocations for dictionary-making purposes has shown that the statistical criterion needs to be combined with semantic and syntactic characteristics of collocations. This is evidenced by findings such as that statistically relevant collocations are usually syntactically more flexible (Gantar et al., 2019) and that collocations containing semantically very general collocates, which are often also very frequent, are semantically less informative and consequently lexicographically less relevant.

2.2 Syntactic criterion

As evident from various definitions (Moon, 1998; Hausmann, 1989; Kilgarriff et al., 2004; Seretan, 2010; Baldwin and Kim, 2010; Fellbaum, 2015), collocations are also defined by syntactic relations in which they occur, as well as their internal syntactic relationships. It is worth noting that all word combinations are not possible or syntactically correct and all (frequent) syntactically correct word combinations are not collocations (see also Section 3.1 on the distinction between collocations and free word combinations). Therefore, when considering syntactic criteria in defining collocation one must also consider the number of elements and their lexical value (semantic or grammatical word classes[1] versus functional and modificational word classes), and relatedly also the order of elements in the collocation. Namely, the syntactic nature of word combinations allows for element insertion (e.g. **organizirati mizo* 'to organize a table' → *organizirati okroglo mizo* 'to organize a round table') and adaptation to the context with opening valency positions (*tekmovalni del* 'competition part' → *tekmovalni del programa* 'competition part of the programme').

1 The expression grammatical collocation can also be found in literature (cf. Benson et al., 1986).

As a result, automatic exctraction of lexically relevant collocations from the corpus warranted a careful description of syntactic structures (see Section 4 for more).

2.3 Semantic criterion

The semantic criterion is the most important criterion for distinguishing collocations from multiword lexical units and is at the same time the most difficult to specify. While statistical and syntactic criteria are more generally accepted, the body of research on collocations uses one of the two basic approaches when considering their lexical characteristics. The first approach sees collocations as a separate type of phraseological units which is partly or completely (semantically and syntactically) fixed and has become established through regular contextual use. This definition includes especially so-called "phraseological" or "strong" collocations which are limited in lexical choice of its components (Halliday, 1966; Cowie, 1981; Sinclair, 1991), and are a relevant part of mental lexicon.

An example of a phraseological collocation, as put forward by Halliday, is the expression *strong tea*. While the same meaning could be conveyed by the roughly equivalent *powerful tea*, this expression is considered excessive and awkward by native English speakers. On the other hand, there are approaches that define collocations more broadly, i.e. as word combinations that are not limited or exclusive but rather allow longer (open) lists of collocates (e.g. *herbal/camomile/pepermint/sage tea*). Atkins and Rundell (2008, p. 167) define collocations as "… salient phrases in corpus citations [that] yet seem to have no idiomatic meaning" and "… a significantly frequent grouping of words whose meaning is quite transparent" (ibid., p. 223).

In general it can thus be said that collocations found in general dictionaries are not treated as lexical units that require an explanation of their meaning.[2] The inclusion of collocations in dictionaries is due to the fact that they typically disambiguate meanings of polysemous words (e.g. *king crown; Czech crown; dental crown*) or are due to their widespread use typical of natural language

2 This is not always true of collocation dictionaries, especially if they are targeted at non-native speakers. Those dictionaries often include word combinations (e.g. compounds) that require explanations.

use (*pitch black*, *thick fog*; but not **thick black*). Their use is sometimes not only language-specific but also culture-specific (*take a walk*). We have thus selected the semantic criterion, or more specifically the lexicographer's decision about the semantic transparency of word combination and consequently its inclusion among lexical units, as the point of departure of our typology of multiword lexical units. In our typology, presented in the following sections, collocations are excluded from the narrower phraseological framework, which is especially important for their role in the dictionary database.

3 COLLOCATIONS IN RELATION TO OTHER WORD COMBINATIONS

The fact that the collocation is always a combination of at least two (usually lexical) words requires that we define their relationship towards other frequent word combinations (free combinations) that represent certain syntactic combinations, but usually do not feature in dictionaries. At the same time, collocations need to be defined in terms of their relationship towards different kind of word combinations that behave like lexical units (i.e. multiword lexical units), and thus require a semantic description, or occupy some pragmatic and communication role (see Figure 1).

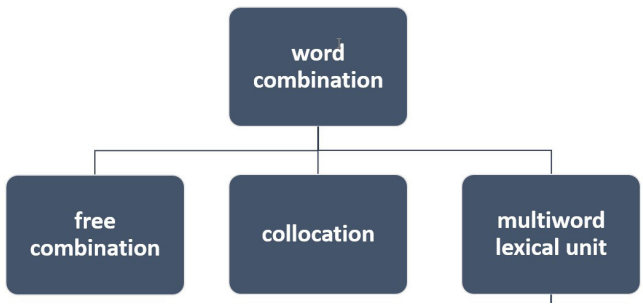

Figure 1: Collocations in word combination typology.

3.1 Collocations and free combinations

In our dictionary-driven typology collocations are distinguished from so-called "free" word combinations mainly on the basis of their lexicographic relevance. For example, certain word combinations, which can be very frequent but do not disambiguate meanings and contain delexicalised words, are

consequently semantically less informative. For example, free combinations such as *in pri tem* ('and then'), *nisem vedel* ('I didn't know'), *ta način* ('this way') etc. are not considered as lexical units. Considering all three aforementioned criteria, we can say that free combinations are, similar to collocations, often frequent word combinations, but differ from collocations in the fact that they do not have any lexicographic value.

It should be noted that syntactic combinations that exhibit characteristics of free combinations can become lexicographically relevant units if they take on certain connective, modificational or discourse roles in the text. For example, combinations such as *glede tega* ('about this') or *zaradi tega* ('because of this') have a role of text connectors, whereas the combination *samo malo* ('only a little' or 'just a moment') in certain contexts has a special discourse or pragmatic role and can be considered as a phraselogical unit.

3.2 Collocations and multiword lexical units

In defining collocations in relation to multiword lexical units (MLU),[3] i.e. different multiword units that belong to lexicon and in a dictionary, our main criterion is that MLUs need to exhibit some degree of idiomatic meaning or behaviour.[4] From the perspective of being considered for dictionary inclusion and description, they need to fulfil the criterion that their "meaning is more than the sum of the parts" (Atkins and Rundell, 2008, p. 167). This semantic criterion is, of course, relative and exclusively lexicographic. The judgement of a lexicographer whether a certain word combination requires its own semantic description or not depends on the type of dictionary and its target user(s) (human or computer).

To be able to distinguish collocations from MLUs and determine their role in the dictionary database, we divided MLUs into three groups (Figure 2).

3 Multiword expression and multiword lexical unit can be viewed as synonymous terms, however we decided for multiword lexical unit in order to stress the difference between units, which suggest a semantically independent whole, whereas expressions (and combinations) do not.

4 In this, we partially follow the definition of multiword expressions by Atkins and Rundell (2008), but it should be noted that under multiword expressions they also list transparent collocations which they define as "phrases ... [that] seem to have no idiomatic meaning" (ibid., p. 167).

Phraseological units and compounds require semantic description. The third group consists of different types of lexico-grammatical units such as light-verb constructions that represent typical syntactic combinations in known syntactic and semantic roles. These units are not a standard part of dictionaries, but when they are included, they come with certain lexico-grammatical information.[5]

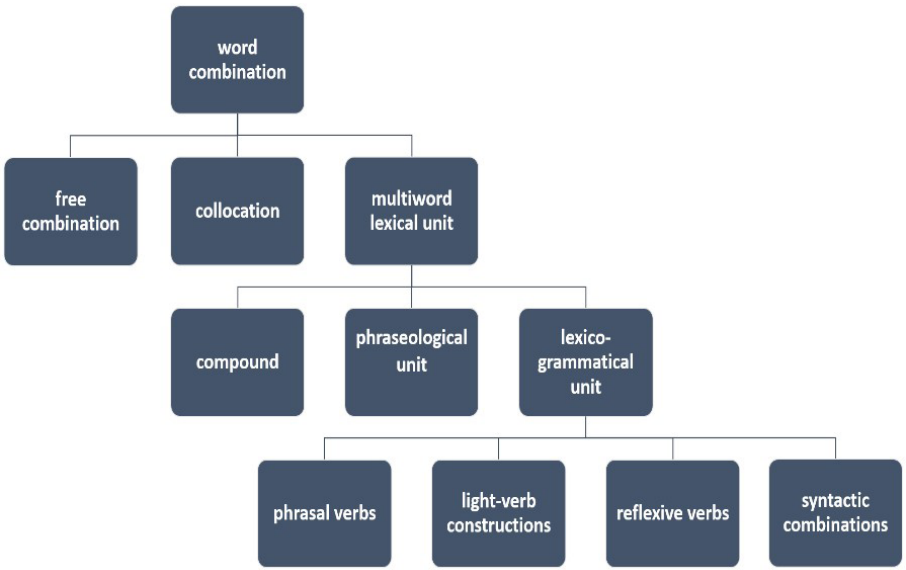

Figure 2: Divison of multiword lexical units.

3.2.1 Compounds

Compounds are a type of multiword lexical units that require a description in the dictionary, given that their meaning cannot be deduced from the meaning of each component. In other words, their meaning is more than a sum of their parts. The main characteristic that distinguishes compounds from phraseological units in our typology is that they as a whole do not have a metaphorical or expressive meaning; for example *topla greda* ('greenhouse' or 'greenhouse effect'): 1. A glass building in which plants are grown, 2. A process of the

5 C.f. phrase *more than* in the Macmillan online dictionary: https://www.macmillandictionary.com/dictionary/british/more-than

earth's surface warming up due to warmer atmosphere. Compounds typically carry a specific terminological or technical content, phenomenon or object; they normally have a concrete referent. The level of terminology varies, and sometimes it is difficult to determine their semantic independence that separates them from collocations; for example *trebušna votlina* ('visceral cavity'), *jedilna žlica* ('soupspoon'), *zeleni čaj* ('green tea'), *osnovna šola* ('elementary school') etc. The decision on whether these are terminological compounds or collocations is solely lexicographic, and is normally a part of dictionary's style guide. When including them into the dictionary database these compounds can feature as collocations connected with the meaning of one of their component elements, e.g. *šola* ('school' meaning institution): *osnovna šola* ('primary school', *srednja šola* ('secondary school'), *visoka šola* ('college') etc., and at the same time as terminological units that require a definition: *osnovna šola* ('primary school') as "an official institution offering certain education". In addition, compounds usually cannot be directly translated into another language, e.g. a direct translation of *dnevna soba* would be 'day room' rather than the actual translation 'living room'. Similarly, a certain compound in one language is not a compound or a multiword unit in another, e.g. *stara mama* in Slovene means *grandmother* in English. In fact, we are aware that languages such as German, Dutch and Norwegian are known for the high productivity of compounds, without space delimitation, however in such cases the formal criteron of single-word vs. multiword structure already acts as the main criterion of distinguishing collocations from compounds.

Also, compounds of terminological and semi-terminological nature are multiword lexical units that are of metaphorical origin, but their role is primarily denotative and not expressive, e.g. *črna luknja* ('black hole') as a space phenomenon. Such compounds can have a metaphorical meaning (among other meanings) which is consequently categorised in our typology under phraseological units.

3.2.2 Phraseological units

Phraseological units are also multiword lexical units with their own meaning. However, unlike compounds, phraseological units have a metaphorical meaning (also called figurative or connotative meaning). From the communication

perspective, this means that when using them, one wants to say something in a more noticeable or expressive manner, differently. Also, in language there is normally a more neutral term with a similar meaning, e.g. *to make a mountain out of a molehill* and *exaggerate*. We are therefore talking about phraseology (idiomatics) in its narrowest sense. It is worth pointing out that even within phraseological units we can find different types in terms of their structure and meaning, for example compound-like phraseological units (*začarani krog*, 'catch-22'), sentence phraseological units or proverbs and sayings (*čas je denar*, 'time is money', *počasi se daleč pride*, 'haste makes waste'), expressions with pragmatic and evaluative role (*za vraga,* 'damn', *kapo dol,* 'hats off'), and expressions in different adverbial (*ena na ena*, 'one on one', *bolj ali manj*, 'more or less') or communicative roles (*dober večer*, 'good evening', *vesel božič,* 'Merry Christmas').

3.2.3 Lexico-grammatical units

Another group of word combinations that needs to be distinguished from collocations (and free combinations) are lexico-grammatical units, i.e. frequent multiword units that also contain grammatical and function words. Unlike collocations, the role of lexico-grammatical units in the text is that of sentence or text organisation, which makes them relevant for dictionaries and thus differentiates them from frequent free word combinations. Another characteristic of lexico-grammatical units is that they show statistically significant co-occurrence in certain syntactic relations and are accompanied by predictable syntactic roles in their context.

Lexico-grammatical units include phrasal verbs and light-verb constructions, reflexive verbs, and syntactic combinations. Phrasal verbs include a verb and a preposition, often followed by a predictable valency position, e.g. *priti do* [*sprememb, dogovora, napredka* ...] 'result in [a change, an agreement, progress]'. Examples of light-verb constructions, which are formed by a verb that carries "less meaning in such constructions than in many other contexts" (Atkins and Rundell, 2008, p. 175) and a noun, include *biti v dvomih* 'to be in doubt', *imeti mnenje* 'to have an opinion'. Reflexive verbs contain a combination of a verb and a reflexive clitic; in many cases, a reflexive clitic is always found with the verb (e.g. *zdeti se* 'to appear'; in other cases, the reflexive and

non-reflexive use of a verb have different meanings (e.g. *ločiti se* 'to have a divorce' vs. *ločiti* 'to split'). Syntactic combinations overlap with free combinations without any specific syntactic role, and also with pragmatic phraseological units (*to je to*, 'this is it'). They can have different roles in a sentence, for example they can be (a) adverbials (*na prostem,* 'in the open', *pred leti,* 'years ago', *zadnje čase,* 'recently', *kar nekaj* 'quite a few'), (b) discourse markers (*po besedah*, 'as stated by', *v bistvu*, 'actually') and c) text connectors (*glede na,* 'according to', *medtem ko* 'while', *po eni strani – po drugi strani,* 'on the one hand – on the other hand').

4 COLLOCATION AS A DICTIONARY UNIT

So far, we defined collocation as a lexical phenomenon, i.e. as a string of words which (a) is statistically relevant, (b) has a predefined syntactic structure and (c) needs to be semantically transparent and meaningful. We also juxtaposed collocations with other word combinations, from free combinations on the one hand to multiword lexical units with their own meaning on the other. We now need to also consider the criterion of dictionary relevance. In this section, we present statistical, syntactic in semantic criteria when extracting collocations from a corpus with the aim of including them into digital dictionary database for Slovene. Furthermore, we outline the parameters for selection of those extracted collocation candidates that are suitable for inclusion in the Collocations Dictionary of Modern Slovene (Gorjanc et al., 2017).

4.1 Automatic extraction of collocation candidates

Automatic extraction of collocations from a corpus was conducted with the aim of creating a large digital dictionary database, with several satellite dictionary databases (Klemenc et al., 2017), including the database of collocations dictionary. The extraction was done in two stages, with each stage consisting of several extraction-evaluation iterations (Krek et al., 2016). The methodological decision was that automatically extracted data will be used for the Collocations Dictionary of Modern Slovene and immediately presented to the users, followed by regular updates of entries after lexicographic analysis (Kosem et al., 2018).

4.1.1 Statistical parameters

In the first stage of automatic extraction, collocation candidates were extracted from the Gigafida reference corpus for Slovene (Logar et al., 2012), using a sample of 2,500 lemmas from the Slovene Lexical Database (Gantar et al., 2016). We used grammatical relations[6] in the Sketch Engine tool (Kilgarriff et al., 2004), using the Sketch Grammar for Slovene, written especially with automatic extraction in mind (Krek, 2016). Moreover, good examples for each collocation were extracted using the GDEX tool and the configuration for Slovene (Kosem et al., 2011). The second iteration of the extraction was conducted on 35,989 lemmas[7] and contained over seven million collocations and slightly less than 35 million corpus examples (Krek et al., 2016). Both iterations of data extraction used the same lists of grammatical relations per word class, with lemmas divided into different frequency groups. Each frequency group per word class used different settings for the following parameters: minimum frequency of a collocate, minimum frequency of a grammatical relation, minimum salience (logDice value) of a collocate, minimum salience (logDice value) of a grammatical relation (Figure 3). All groups of lemmas shared the same limit of extracted collocates per grammatical relation and examples per collocation. More on the procedure of how exact parameter values were set can be found in Gantar et al. (2016).

One additional step used in the second iteration was the inclusion of collocations with higher raw frequency. This was done because we found that logDice sometimes gives low ranking to highly frequent and relevant collocations, which meant that the exported data, while focussing on statistically more relevant collocations, could include an insufficient number of collocations for highly frequent and polysemous words to represent all the senses. Consequently, we performed and merged two extractions (using the same maximum limit of collocations per grammatical relation), one with collocations ranked by logDice, and the second one with collocates ranked

6 Grammatical relations or gramrels are used in a narrow sense of the Sketch Engine terminology in this paper; they represent the definitions of syntactic structures in the sketch grammar.

7 The initial list contained 50,000 lemmas, but was reduced to 35,989 after removing the noise in the lemma list, excluding proper names and lemmas with frequency under 400 occurrences in the corpus (deemed to contain very little useful collocational data).

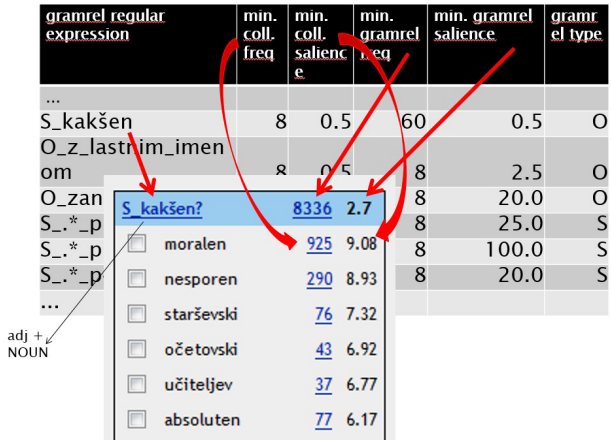

Figure 3: Parameter settings for different grammatical relations and their connections (red arrows) with a table of the syntactic structure adjective + NOUN, illustrated with the results for the noun *avtoriteta* ('authority') in the Word Sketch function.

by raw frequency. Expectedly, there was often a significant overlap between the two lists.

4.1.2 Syntactic structures

The first stage of automatic extraction of collocations used grammatical relations, defined in the sketch grammar file in the Sketch Engine tool. The grammatical relations included syntactic structures that were identified during lexicographic analysis. Initially, 528 syntactic structures were used (Krek et al., 2016), with noun and verb structures being the most common, but syntactic structures with prepositions (and nouns in different cases) are also prevalent (Table 1), as is also the case in collocations dictionaries for other languages.

Table 1: *Common collocation structures in collocations dictionary database*

	Most common collocation structures (Collocationas dictionary database)	Number of structures in the Collocationas dictionary database
1	NOUN + $NOUN_{GENITIVE}$	1,783
2	VERB + $NOUN_{ACCUSATIVE}$	1,672
3	ADJ + NOUN	1,609
4	VERB + $NOUN_{GENITIVE}$	1,598
5	VERB + PREP + $NOUN_{INSTRUMENTAL}$	1,193

It is noteworthy that in the word sketch, collocates under grammatical relations are listed as individual words and in lemma form.[8] Thus, in a morphologically rich language like Slovene, collocate and the headword often need to be put in the correct form to adequately reflect their use in a particular grammatical relation. This can be because of gender and/or number agreement of the headword and the collocate *(rdeč -> rdeč**a** jagod**a**; jesenski -> jesensk**o** list**je**)*, or because the headword or the collocate need to be in a certain case (i.e. *olupiti jabolk**o***$_{\text{accusative}}$; *črv v jabolk**u***$_{\text{locative}}$). Moreover, additional elements (e.g. prepositions, conjunctions) were missing in relations with more than two elements, however in such cases the third element was always found in the same form. We solved this issue by automatically postprocessing the extracted data where each element of the grammatical relation (headword, collocate, preposition) was automatically attributed with their role in the collocation (using different tags) and written in the correct form (e.g. correct gender, case, number).

4.1.3 Semantic criteria

There were no specific semantic criteria set for the automatic extraction of collocations. We could say that the selection of grammatical relations already indirectly determined some semantics, as only lexical word classes (with the exception of prepositions and conjunctions in trinary grammatical relations, i.e. relations containing two lexical words and one function word) were used as collocation components. Also, the verb *biti* ('be') was excluded as a collocate in nearly all grammatical relation containing verbs. Other than that, no other criteria were used, as we wanted to induce semantic criteria (and potentially other statistical and syntactic criteria) from the evaluation with the users.

4.2 Evaluation

Evaluation of the automatically extracted collocation data comprised of three separate studies. The first one was conducted with dictionary users (students, translators etc.) on the initiallly extracted data for 2,500 lemmas (Krek et al., 2016), which were available online as the Database of the Collocations

8 It has to be mentioned that the COLLOC directive in the Sketch Engine enables the extraction of collocations as bigrams/trigrams and in particular word forms, but this directive was introduced after the extraction has already been performed.

Dictionary. The focus was more on the interface features (layout of information, clarity etc.), but included also questions on the presentation of collocations and on the benefits and shortcomings of automatically extracted data.

The second study was done with lexicographers (and linguists) on the 35,989 lemmas dataset, using the Pybossa platform. Lexicographers inspected 17,576 collocations in 143 different grammatical relations for 333 different lemmas (Pori and Kosem, 2018), with at least three lexicographers "voting" on each collocation. They were presented with the information of the grammatical relation, collocation and one example, and were given various options. The optional answers were grouped into Yes, No and I don't know, however Yes and No options had suboptions, e.g. Yes had the suboption that the collocation is OK but the form displayed is not, for example when the collocation should have been in plural. The first findings of the study, with focus on grammatical relations containing adverbs, were presented in Pori and Kosem (2018).

The third study by Pori et al. (2020) combined the approaches of both previous studies by focussing on the user perceptions of automatically extracted collocational data for 35,989 lemmas, as presented in the Collocations Dictionary of Modern Slovene. One important aspect of the study is the fact that lexicographers represent one of the user groups, and their perceptions of the value and problems of automatically extracted data can be directly compared with other types of users.

The findings of all three studies, which point to problems of automatic collocation identification and extraction and are relevant for this paper, can be divided into four interconnected topics:

- shortcomings related to corpus data,
- shortcomings related to syntactic criteria,
- shortcomings related to statistical criteria,
- shortcomings related to dictionary relevance.

4.2.1 Shortcomings related to corpus data

Many errors that occur during automatic extraction of collocation stem from problems in corpus annotation, i.e. lemmatisation (e.g. **piliti alkohol -> piti*

alkohol) and part-of-speech tagging (e.g. mixing between adjectives and adverbs (**težek do alkohola* 'difficult to alcohol' -> *težje do alkohola* 'more difficult to get alcohol') or between adjectives and nouns (**premagati poljski* 'beat Polish' – *premagati poljsko* 'beat Poland') that share forms. The first stage of automatic extraction was conducted on the Gigafida corpus, which was automatically tagged using the JOS tagset, with the accuracy of tagging reaching 97.88% at lemma level, and 91.34% at the level of all morphosyntactic tags (Grčar et al., 2012). Quite problematic for syntactic criteria were also errors in annotation of cases when the forms were the same, e.g. nominative and accusative of inanimate nouns, or genitive singular and nominative plural of feminine nouns.

Collocation identification was also influenced by certain linguistic decisions related to corpus annotation. For example, in hyphenated forms such as *sladko-kisla omaka* ('sweet-sour sauce'), each part of the hyphenated combination was annotated separately; thus, only collocations such as *sladka omaka* ('sweet sauce') and *kisla omaka* ('sour sauce') were extracted. Similarly, nominalised adjectives such as *zaposleni* ('the employed') were annotated as adjectives and thus not found in grammatical relations containing nouns.

4.2.2 Shortcomings related to syntactic criteria

The problems of corpus annotation also affected syntactic criteria, or better said, the quality of collocational output at different grammatical relations. The sketch grammar is tagset-based, which means that grammatical relations must be defined via tags rather than e.g. syntactic relation identified by parsers. Aforementioned problems of incorrect case annotation therefore resulted in wrong grammatical relation attribution, e.g. **botrovati alkohol* ('causes alcohol'; verb + $\text{noun}_{\text{accusative}}$) rather than *alkohol botruje* ('alcohol causes'; $\text{noun}_{\text{nominative}}$ + verb). Similarly, adjectives could be incorrectly identified as attributive even when used only predicatively, e.g. **priložena miška* ('included mouse') instead of *miška je priložena* ('mouse is included') or **kriv hormon* ('responsible hormones') instead of *hormoni so krivi* (hormones are responsible (for)). Such combinations, while syntactically correct, do not form meaningful collocations, which means that the expected syntactic relation had to be more narrowly defined on the syntactic/tree level.

There were also cases when one grammatical relation was a limited version of another one, often resulting in duplication of collocations. For example, the collocation *vulkanskega izvora* ('of volcanic origin') was extracted in the grammatical relation adjective$_{genitive}$ + noun$_{genitive}$; however, the genitive form was also included in the grammatical relation adjective + noun (agreement in all possible cases) as the collocation *vulkanski izvor* ('volcanic origin'). Yet, such collocations have different syntactic roles, as an attributive or subject/object respectively. Thus, it is important to define grammatical relations more narrowly in such cases.

The evaluation made it clear that certain grammatical relations contained much more noise, i.e. they contained many more bad collocation candidates. Whereas certain grammatical relations exhibited issues in general, at many different lemmas (e.g. noun + noun$_{genitive}$), others were problematic only at certain types of lemmas (e.g. inanimate nouns in the grammatical relation verb + noun$_{accusative}$). Furthermore, certain grammatical relations (e.g. verb + noun$_{genitive}$) contained such an overwhelming percentage of noise that they were excluded from the collocations dictionary altogether.[9]

A problem related to good/bad collocation identification at certain grammatical relations, especially those with errors in case annotation, is related to the fact that at first glance such collocations look good (e.g. *izolirati bakterije* 'isolate bacteria' in the relation verb + noun$_{genitive}$; when it is verb + noun$_{accusative}$ (in plural); only when considering both their form and the grammatical relation they are found in one can discard them as bad. This is of course more problematic when lay users, which perhaps pay less attention to accompanying grammatical information, are confronted with automatically extracted data.

4.2.3 Shortcomings related to statistical criteria

We have already mentioned problems linked to the selection of statistical method for collocation, which led to additional extraction of collocations ranked by raw frequency. Moreover, the parameters set for extraction had to be adjusted for different groups of lemmas according to their word class, grammatical relation, and corpus frequency. Despite these rather detailed

9 These grammatical relations may of course be added to the subsequent versions of the collocations dictionary.

criteria, problems were still observed on both ends of frequency ranking, i.e. at very frequent and very rare lemmas. For very frequent lemmas, the lists of extracted collocations were often too short, especially in the most common grammatical relations, resulting in non-coverage of certain (still salient) senses of the words. In fact, in such cases, the maximum number of collocations was often the only criterion that had to be used, as all the other were not even met (e.g. minimum collocation frequency). Similar problem with left out collocations was observed at very rare lemmas (i.e. rare as on the bottom end of our threshold of 400 hits in the corpus), but the reason was different; the problem occurred mainly because of collocation dispersion, i.e. there were many collocations in the grammatical relation belonging to the same semantic type (and representing the same sense), and while their joint frequency was very high, their individual frequency was below the minimum threshold and they were thus not extracted.

Additional issues that have come up during the evaluation were heavily linked to aforementioned errors in corpus annotation, and relatedly, errors in grammatical relation attribution. First and foremost, this includes collocation candidates that were always errors, and pushed down the ranking (and sometimes off the list of extracted data) other, good, collocations. However, there were also cases when syntactic problems were not absolute, i.e. the collocation was good but its statistics was misleading as the concordances included many incorrectly identified cases, in certain cases to the level where the number of good collocation examples was even below the minimum threshold of 4. For example, *čakati nastop* 'await a performance' is a good collocation in the verb + $\text{noun}_{\text{accusative}}$ structure, but examples contained many (incorrect) cases of *nastop čaka* 'a performance awaits'.

Collocation ranking is also interesting from the perspective of dictionary users. While one of the association measures seems the logical choice for collocation ordering in a dictionary as it reflects the nature of collocation, our initial research (Arhar Holdt, in press) has shown that this is not in line with the expectations of the users who clearly prefer (or expect?) frequency. Further evidence that this problem is not trivial is the practice of some dictionaries (e.g. see Hudeček and Mihajlević, 2020) that avoid any mention of statistics and list collocations by alphabet (only). In the case of our dictionary of

collocations, we used a solution where logDice ranking was used as the default one, and an option of switching to alphabetical ranking was made available to the users.

4.2.4 Shortcomings related to dictionary relevance

The evaluation of automatically extracted collocational data from the perspective of dictionary relevance was conducted manually and with the aim of identifying criteria for the selection of collocations for our database, and for the presentation in the dictionary interface. We focussed mainly on determining the informative value of collocations (strong vs. weak collocations), the informative value of the entire grammatical relation, and the predominant form of collocation in corpus examples.

Evaluation clearly identified different levels of collocability between collocation elements, which considerably determine the dictionary relevance of the collocation. As already discussed at the typology of word combinations, collocations can exhibit very strong internal link (e.g. *trda tema* 'pitch black', *debela denarnica* 'thick wallet'). On the other hand, there are headwords without any strong collocates, where "just about any word can (and does) combine with words like these [*house*, *buy* and *good*], as long as the combination makes sense."[10] While we did not exclude words like *house* and *buy* from our lemma list, collocations evaluated as weak often included semantically broad collocates such as certain types of adverbs (Pori and Kosem, 2018), e.g. *malo* 'little', *zelo* 'very', adjectives (e.g. proper adjectives like *slovenski* 'Slovenian', *angleški* 'English' etc. and temporal adjectives like *nov* 'new', *star* 'old', *nekdanji* 'recent', *bivši* 'former'), verbs (e.g. the verb *biti* '*be*' and modal verbs), and words which feature in different syntactic roles (e.g. pronouns, adjuncts, certain adverbs, e.g. *kar* 'quite', *nekaj* 'some', *samo* 'only', *okoli* 'about', *veliko* 'many').

While these weak collocations were not considered relevant for the inclusion in the dictionary, they were still kept in the database because they met statistical and syntactic criteria and might be relevant for some other resource. In fact, it is important to note that the record of all good (strong and weak)

10 M. Rundell: How the dictionary was created: http://www.macmillandictionaries.com/features/how-dictionaries-are-written/macmillan-collocations-dictionary/.

and bad collocation candidates should be kept in the database, and used for comparison in future automatic extractions, so that the duplication of work is avoided.

Interestingly, certain collocation candidates containing weak collocates often represent a part of units belonging to other word combinations in our typology. Such collocation candidates themselves are often semantically nonsensical and parts of other lexico-grammatical units, e.g. **formalen smisel* 'formal sense' is actually part of *v formalnem smislu* 'in a formal sense', or *zveza z gradnjo* 'relation to contruction' is actually part of *v zvezi z gradnjo* 'in relation to construction'. Continuous adding syntactic relations identified through (bad) collocations to our list enables the extraction of such units from the corpus, as well as avoiding identification of bad collocations.

A very specific issue in terms of dictionary relevance of collocation candidates were collocations related to proper names, i.e. collocations that are proper names themselves and often reflect some cultural or language (e.g. *Vesele Štajerke* 'Happy Styrians', which is the name of a band) and collocations with a collocate that is a proper name (e.g *prestolnica Lombardije* 'capital of Lombardy'). Such cases are not clear cut, which was also evident from the level of (dis)agreement among evaluators; while cases like *Vesele Štajerke* were seen as irrelevant for the collocations dictionary by all the evaluators,[11] *prestolnica Lombardije* showed less agreement as many believed the collocation was relevant as it was a representation of a highly salient and sense indicative combination *prestolnica* + country/region. In sum, while there are good arguments to include these types of collocations in dictionaries (see e.g. Hudeček and Mihaljević, 2020), we decided to treat such collocations separately as multiword named entities in the database.

Statistics is an essential part of collocation, and this goes beyond its constituent parts. A very important part of collocation not only at its identification but also in presentation to dictionary users is its predominant form. Two frequently problematized issues during evaluation was number for nouns and degree for adjectives. Semantic characteristics of several headwords either require or prefer non-singular form (plural or dual), e.g. **stresti bonbon* 'dispense

11 In general we consider encyclopaedic information as not relevant for the collocations dictionary.

bonbon' instead of *stresti bonbone* 'dispense bonbons', or *finančna težava* 'financial trouble' instead of *finančne težave* 'financial troubles'. Similarly, typicality of collocation can be limited to the adjective in a certain form e.g. superlative, as in **blizek sorodnik -> najbližji sorodniki* 'closest relatives'.[12] All these collocations, if presented in the 'basic form', do not reflect typical use or even appear strange, which means that future extractions should consider the predominant form. A similar approach is already used in the Sketch Engine word sketches in the form of longest-commonest match (Kilgarriff et al., 2015), however the feature still needs improving as it does not always provide a result or often offers a sequence which is longer than the collocation.[13]

5 CONCLUSIONS

Collocations are a highly relevant type of word combinations, and are defined by three types of criteria: statistical, syntactic and semantic. As shown in the paper, all three types are heavily interlinked, and each brings different decisions and problems. Equally important as these three types of criteria for any dictionary project is defining collocations in relation to other word combinations, i.e. free combinations and multiword lexical units; as we pointed out free combinations do not have any lexicographic value, whereas multiword lexical units do but they also require a description as their meaning is more than the sum of their parts. By knowing the typology in detail one can make better decisions as to which category the candidate word combination belongs.

Yet, as our evaluation of automatically extracted collocational data has shown, practical application of a theoretical framework brings new challenges, associated with the quality of corpus annotation, the purpose of the dictionary, and the expectations and needs of dictionary users. The challenges are mainly two-fold, with the common theme being the amount of collocations. Firstly, there is the need to separate the wheat from the chaff, i.e. bad collocation candidates from

12 We intentionally do not provide an English translation for the bad collocation candidate, as in English a collocation with *close* in its basic form and *relative* actually exists, whereas in Slovene the word form (and lemma) *blizek* is merely an artifical contruct of the basic form of this particular adjective (and is very rarely found in the corpus, and never with *sorodnik*).

13 This function in the Sketch Engine can be useful when identifying bad collocates or multiword units such as *v zvezi z gradnjo* 'in relation to construction' mentioned above.

the good ones, caused by problems in corpus annotation or problems stemming from the identification of collocation on the basis of part-of-speech tags. Secondly, there is the question of dictionary relevance, the decision of which cannot be left (only) to statistical measures for collocation identification but is rather mainly semantic, and driven by the target users of the dictionary.

What our experience has shown is that the collocation is defined by statistical, syntactic, and semantic criteria, however these criteria are not set in stone, and cannot be generalized across the language (i.e. they cannot be the same for different types of words). Constant evaluation and improvement of the criteria is required. The Slovenian language as a morphologically rich language is particularly problematic as far as the syntactic criteria are concerned. Our efforts to improve the quality of automatic collocation identification are currently directed mainly in this direction. Thus, we are testing the extraction of collocations from a parsed corpus, using 76 collocational structures that have been 'translated' from the definitions of grammatical relations for a part-of-speech tagged corpus. Initial results are promising and this approach seems to definitely solve a few existing problems (e.g. collocation form in terms of case and number as well as typicality, and the amount of bad candidates), but is likely to require some fine-tuning.

We are not neglecting the statistical and semantic aspects, though. On the statistical level, we are exploring the measures such as deltaP (Gries, 2013) to determine the symmetry of collocations, i.e. to establish which collocations are relevant only for one of its constituent parts. On the semantic level, we want to explore the characteristics of weak collocates and prepare stop lists, probably for different groups of lemmas. Most importantly, we are including all these activities in our efforts to compile a common digital database for Slovene where collocations, and all other word combinations, will be available to the research community and creators of language resources.

Acknowledgements

The authors acknowledge that the project *Collocation as a basis for language description: semantic and temporal perspectives* (J6-8255) was financially supported by the Slovenian Research Agency, and acknowledge the financial support from the Slovenian Research Agency (research core funding No.

P6-0411, *Language Resources and Technologies for Slovene*) and P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731015.

REFERENCES

Altenberg, B. (1991). Amplifier Collocations in Spoken English. In S. Johansson & A. B. Stenström (Eds.), *English Computer Corpora. Selected Papers and Research Guide* (pp. 127–147). Berlin/New York: Mouton de Gruyter.

Arhar Holdt, Š. (in press). Razvrstitev kolokacij v slovarskem vmesniku: uporabniške prioritete. In *Kolokacije kot temelj jezikovnega opisa: od statistike do semantike*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In *Handbook of Natural Language Processing* (2nd ed.). CRC Press, Taylor and Francis Group.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI Dictionary of English Word Combinations*. John Benjamins, Amsterdam.

Berry-Rogghe, G. L. (1973). The computation of collocations and their relevance in lexical studies. In *The computer and literal studies* (pp. 103–112). Edinburgh/New York: University Press.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing 8*(4), 243–257.

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, *6*(1), 22–29.

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting Online Resources to Build a Lexicon* (pp. 116–164). Erlbaum, Hillsdale, NJ.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. In A. P. Cowie (Ed.), Lexicography and its Pedagogical Applications [Thematic issue]. *Applied Linguistics 2*(3), 223–235.

Evert, S. (2004). The statistics of word cooccurrences: Word pairs and collocations. PhD Thesis, University of Stuttgart.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook: Vol. 2* (pp. 1212–1248). Berlin/New York: Mouton de Gruyter.

Fellbaum, C. (2015). Syntax and grammar of idioms and collocations In T. Kiss & A. Alexiadou (Eds.), *Syntax: Theory and analysis: Vol. 2* (pp. 776–802). Berlin/New York: Mouton de Gruyter.

Firth, J. R. (1957). Modes of Meaning. *Papers in Linguistics* 1934–51. London: Oxford University Press.

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete. Retrieved from http://www.ff.uni-lj.si/sites/default/files/Dokumenti/Knjige/e-books/leksikografski.pdf

Gantar, P., Colman, L., Parra Escartín, C., & Marínez Alonso, H. (2019). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography, 32*(2), 138–162.

Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography, 29*(2), 200–225.

Gorjanc, V., Gantar, P., Kosem, I., & Krek, S. (Eds.). (2017). *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts.

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec & J. Žganec Gros (Eds.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Gries, S. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*(1), 137–165.

Halliday, M. A. K. (1966). Lexis as a Linguistic Level. *Journal of Linguistics, 2*(1), 57–67.

Hausmann, F. J. (1989). Le dictionnaire de collocations. In F. J. Hausmann et al. (Eds.), *Wörterbücher: ein internationales Handbuch zur Lexikographie* (pp. 1010–1019). Berlin/New York: De Gruyter.

Hudeček, L., & Mihaljević, M. (2020). Collocations in Croatian Web Dictionary – Mrežnik. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 8*(1).

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Lorient: France.

Kilgarrif, A., Baisa, V., Rychlý, P., & Jakubíček, M. (2015). Longest–commonest Match. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (Eds.), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference* (pp. 397–404). Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Klemenc, B., Robnik Šikonja, M., Fürst, L., Bohak, C., & Krek, S. (2017). Technological design of a state-of-the-art digital dictionary. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (Eds.), *Dictionary of Modern Slovene: Problems and Solutions* (pp. 10–22). Ljubljana: Ljubljana University Press, Faculty of Arts.

Kosem, I., Husák, M., & McCarthy, D. (2011). GDEX for Slovene. In I. Kosem & K. Kosem (Eds.), *Electronic Lexicography in the 21st Century: New applications for new users. Proceedings of the eLex 2011 Conference, 10–12 November, 2011, Bled, Slovenia* (pp. 151–159). Ljubljana: Trojina, Institute for Applied Slovene Studies.

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts, 17–21 July, 2018, Ljubljana, Slovenia* (pp. 989–997). Ljubljana: Ljubljana University Press, Faculty of Arts. Retrieved from https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Krek, S. (2016). Leksikografska orodja za slovenščino: slovnica besednih skic. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (Eds.), *Slovar sodobne slovenščine: problemi in rešitve* (pp. 358–378). Ljubljana: Ljubljana University Press, Faculty of Arts.

Krek, S., Gantar, P., Kosem, I., Gorjanc, V., & Laskowski, C. (2016). Baza kolokacijskega slovarja slovenskega jezika. In T. Erjavec & D. Fišer (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities, September 29th–October 1st, 2016, Ljubljana, Slovenia* (pp. 101–105). Ljubljana: Academic Publishing Division of the Faculty of Arts.

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Massachusetts: The MIT Press, Chap. 5. Collocations.

Moon, R. (1998). *Fixed Expressions and Idioms, a Corpus-Based Approach*. Oxford: Oxford University Press.

Palmer, H. E. (1933). *Second Interim Report on English Collocations, Submitted to the Tenth Annual Conference of English Teachers under the Auspices of the Institute for Research in English Teaching*. Tokyo: Institute for Research in English Teaching.

Pecina, P. (2009). Lexical association measures and collocation extraction. *Language Resources and Evaluation, 44*(1–2), 137–158.

Pori, E., & Kosem, I. (2018). In the Search of Lexicographically Relevant Collocation: The Example of Grammatical Relations Containing Adverbs. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 6*(2), 154–185. doi: 10.4312/slo2.0.2018.2.154-185

Pori, E., Kosem, I., Čibej, J., & Arhar Holdt, Š. (2020). The attitude of dictionary users towards automatically extracted collocation data: a user study. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 8*(1).

Seretan, V. (2010). *Syntax-Based Collocation Extraction* (1st ed.). Berlin, Heidelberg: Springer-Verlag.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wiechmann, D. (2008). On the computation of collostruction strength. *Corpus Linguistics and Linguistic Theory 42*, 253–290.

OPREDELITEV KOLOKACIJ V LEKSIKALNIH VIRIH ZA SLOVENŠČINO

V prispevku definiramo pojem kolokacije za namene vključitve v strojno procesljive jezikovne vire, ki bodo služili izdelavi elektronskih jezikovnih priročnikov in različnih jezikovnih aplikacij za slovenščino. Na podlagi teoretičnih in slovarsko usmerjenih študij definiramo kolokacijo kot leksikalni jezikovni pojav, pri čemer izhajamo iz treh ključnih vidikov: statističnega, skladenjskega, in pomenskega. Kot izhodišče za opredelitev kolokacij znotraj vseh besednih kombinacij v jeziku in za ločevanje kolokacij od prostih besednih zvez štejemo njihovo slovarsko relevantnost. Proste besedne zveze v jeziku obstajajo kot (pogoste) skladenjsko ustrezne besedne kombinacije, ki pa nimajo slovarske vrednosti v smislu pomenskega opisa ali opisa njihove skladenjske ali gramatične vloge. Nadaljnja delitev temelji na slovarsko-semantičnem kriteriju, ki ločuje kolokacije od vseh drugih slovarsko relevantnih enot na podlagi leksikografske odločitve, da besedna zveza potrebuje opis pomena (t. i. večbesedne leksikalne enote). Pri naši opredelitvi kolokacije ne potrebujejo pomenskega opisa, kar jih v temelju ločuje od zvez z neidiomatičnim pomenom (stalne besedne zveze), različnih frazeoloških enot pa tudi od t. i. leksikalno-gramatičnih enot, ki imajo primarno besedilno povezovalne in druge skladenjske vloge. Pri opredeljevanju kolokacij kot slovarskih enot se znova vrnemo k trem ključnim kriterijem, ki jih podrobneje opišemo z vidika avtomatskega luščenja kolokacijskih podatkov iz korpusov. Slovarska relevantnost izluščenih kolokacij je izpostavila predvsem problem semantično odprtih kolokatorjev, kot so določeni tipi prislovov, pridevnikov in glagolov, in besed, ki se pojavljajo v različnih skladenjskih vlogah (e.g. zaimki in členki). Posebej opišemo problem lastnoimenskih kolokatorjev in odločitve pri vključevanju takih primerov v slovar na podlagi evalvacije med leksikografi.

Ključne besede: kolokacija, večbesedna leksikalna enota, besedna kombinacija, slovenščina, leksikografija, slovarska baza