Slovenščina 2.0

Kolokacije v leksikografiji: obstoječe rešitve in izzivi za prihodnost

Collocations in Lexicography: existing solutions and future challenges

Let. 8 (2020), št. 2

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Slovenščina 2.0
Letnik/Volume 8, Številka/Issue 2, 2020

ISSN: 2335-2736

Založila/Published by
Znanstvena založba Filozofske fakultete Univerze v Ljubljani

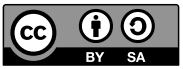
Izdal/Issued by
Center za jezikovne vire in tehnologije Univerze v Ljubljani

Za založbo/For the publisher
Roman Kuhar, dekan Filozofske fakultete

Publikacija je brezplačna./Publication is free of charge.

Publikacija je dostopna na/Avaliable at: dostopna na: https://revije.ff.uni-lj.si/slovenscina2/index

Revija izhaja s podporo Javne agencije za raziskovalno dejavnost Republike Slovenije./
This journal is published with the support of the Slovenian Research Agency (ARRS).



Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=24561667
ISBN 978-961-06-0360-3 (pdf)

KAZALO

SLOVENŠČINA 2.0: COLLOCATIONS IN LEXICOGRAPHY: EXISTING SOLUTIONS AND FUTURE CHALLENGES

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

Polona GANTAR
Faculty of Arts, University of Ljubljana

Kosem, I., Gantar, P. (2020): Slovenščina 2.0: Collocations in Lexicography: existing solutions and future challenges. Slovenščina 2.0, 8(2): i–vi.

DOI: https://doi.org/10.4312/slo2.0.2020.2.i-vi

Collocations have become an increasingly popular topic of lexicographic research and resources in recent years, something that has been also facilitated by the rapid progress in the field of electronic lexicography. There are ongoing debates about what a collocation actually is, what is its relation to other multiword expressions, how much collocational data should be included in the dictionaries and how it should be presented, and how collocational information should be encoded to make it useful for different purposes. This has prompted us to organize a workshop centred around the topic of collocations. The workshop was collocated with the eLex 2019 conference in Sintra, Portugal. 14 different presentations were given at the workshop, offering an insight into the work on collocation at different institutions around the world. The presentations sparked interesting and thought-provoking discussions, and it was clear that a publication was needed to present the state-of-the-art on collocation in more detail. This led to the preparation of this special issue of the journal Slovenščina 2.0, which contains seven contributions based on the workshop presentations. The contributions cover a wide range of topics related to collocations, in six different languages, giving this special issue a truly international focus and relevance.

The first two papers deal with the definition of collocation, but from two different perspectives. **Iztok Kosem**, **Simon Krek** and **Polona Gantar** provide

a definition of collocation, and the classification of collocation in the typology of word combinations. Motivated by the use of collocational data for lexicographic purposes, they present the main criteria that define collocation on the one hand, and describe the main features that distinguish them from other word combinations on the other. Another, but equally important perspective to defining collocation is offered by **Toma Tasovac**, **Ana Salgado** and **Rute Costa** who focus on the modelling and encoding of polylexical units, including collocations, with TEI Lex-0, using the Dictionary of the Portuguese Academy of Sciences as a case study. Given that the existing TEI Guidelines do not address the encoding of polylexical units in sufficient detail, this paper is a very important and much needed contribution to the fields of lexicography and digital humanities.

The next three papers cover three different aspects of collocations in the lexicographic workflow. **Maria Khokhlova** and **Vladimir Benko** present a study on Russian data in which the role of corpus size in the identification of collocations is examined. In addition to determining the minimum size of a corpus for collocational research, they analyse and compare the suitability of four different association measures for extracting collocations from corpora of different sizes. **Lana Hudeček** and **Milica Mihaljević** present the treatment of collocations in the Croatian Web Dictionary called *Mrežnik*, showing detailed examples of the collocational block, with supporting questions and phrases, for different types of headwords. Their paper also addresses methodological questions such as how to define collocation for such a project, and how to address the issues related to the unrepresentative nature of corpus data. **Sanni Nimb**, **Nicolai Hartvig Sørensen** and **Henrik Lorentzen** look at the dictionary post-publication stage, in particular at the role of collocational changes in the detection of new meanings, which can then be translated into the updates of the Danish monolingual dictionary. They present the results of a corpus study in which automatic extraction methods using bigrams were combined with manual annotations.

The paper by **Ene Vainik, Maria Tuulik** and **Kristina Koppel** brings the psycholinguistic perspective by comparing word associations with collocations in the Estonian language, with special emphasis on the role of different parts of speech. They indicate the potential applications of word associations

in lexicography, e.g. in writing definitions, and in language learning. The final paper of the issue by **Eva Pori, Jaka Čibej, Iztok Kosem** and **Špela Arhar Holdt** offers insights into the user evaluation of an automatically compiled Collocations Dictionary of Modern Slovene. Considering that automatic extraction methods are becoming more and more common in modern lexicography, it is useful to learn how different types of users, in this case, teachers, translators, proofreaders, and lexicographers, have reacted to the use of a dictionary containing rich, but sometimes problematic, collocational data.

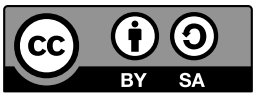
SLOVENŠČINA 2.0: KOLOKACIJE V LEKSIKOGRAFIJI: OBSTOJEČE REŠITVE IN IZZIVI ZA PRIHODNOST

Kolokacije so v zadnjih letih postale vse bolj priljubljena tema leksikografskih raziskav in z njimi povezanih virov, k čemur je pripomogel tudi hiter razvoj področja elektronske leksikografije. Številne diskusije potekajo o tem, kaj sploh je kolokacija, kako jo opredeliti do drugih večbesednih izrazov, koliko kolokacijskih podatkov vključiti v slovar, kako naj bodo predstavljeni uporabnikom ter kako kodirati kolokacijske podatke, da bodo uporabni za različne namene. Vse to nas je spodbudilo, da smo v okviru konference eLex 2019, ki je potekala v Sintri na Portugalskem, organizirali delavnico na temo kolokacij. Na delavnici je bilo predstavljenih 14 prispevkov, ki so ponudili vpogled v delo s kolokacijami na različnih ustanovah po svetu in sprožili vrsto zanimivih in stimulativnih razprav. Prav te razprave so spodbudile tudi potrebo po podrobnejšem opisu aktualnega stanja na področju kolokacijskih raziskav v samostojni publikaciji. Rezultat teh prizadevanj je pričujoča tematska številka revije Slovenščina 2.0 s sedmimi prispevki, ki izhajajo iz predstavitev na delavnici. Prispevki naslavljajo širok nabor tem v šestih različnih jezikih, zaradi česar je tematska številka res mednarodna, tako v zastopanosti kot relevantnosti obravnavanih tem.

Prva dva prispevka se lotevata opredelitve kolokacije z dveh različnih perspektiv. **Iztok Kosem, Simon Krek** in **Polona Gantar** opredelijo kolokacijo in njeno umestitev v tipologiji besednih kombinacij. Glavno vodilo pri tem je uporaba kolokacijskih podatkov za leksikografske namene, na podlagi katerega predstavijo tri glavne kriterije pri opredelitvi kolokacije in tudi glavne lastnosti, ki ločijo kolokacije od drugih besednih kombinacij. Drugačno, a enako pomembno perspektivo pri opredelitvi kolokacije predstavijo **Toma Tasovac, Ana Salgado** in **Rute Costa** s prispevkom o modeliranju in kodiranju večbesednih leksikalnih enot, vključno s kolokacijami, v formatu TEI Lex-0, pri čemer kot testni primer vzamejo Slovar Portugalske akademije znanosti. Glede na to da v obstoječih smernicah TEI kodiranje večbesednih leksikalnih enot ni dovolj poglobljeno predstavljeno, gre za zelo pomemben in dragocen prispevek tako za leksikografijo kot tudi digitalno humanistiko.

Sledijo trije prispevki, ki predstavljajo tri različne stopnje v postopku izdelave slovarskih virov. **Maria Khokhlova** in **Vladimir Benko** predstavita študijo na podlagi ruščine, v kateri preučujeta vlogo velikosti korpusa pri luščenju kolokacij. Določiti skušata minimalno velikost korpusa, ki je še ustrezna za kolokacijske raziskave, analizirata in primerjata pa tudi ustreznost štirih različnih statističnih mer pri luščenju kolokacij iz korpusov različnih velikosti. **Lana Hudeček** in **Milica Mihaljević** predstavita obravnavo kolokacij v Hrvaškem spletnem slovarju Mrežnik, ki vključuje prikaz različnih vprašanj in fraz za posamezne tipe kolokacij pri iztočnicah različnih besednih vrst. Avtorici se dotakneta tudi metodoloških vprašanj, kot je na primer opredelitev kolokacije za namene splošnega izhodiščno digitalno zasnovanega slovarja in reševanje problemov, povezanih s slabo reprezentativnostjo korpusnih podatkov. **Sanni Nimb, Nicolai Hartvig Sørensen** in **Henrik Lorentzen** raziskujejo možnosti uporabe kolokacijskih podatkov pri posodabljanju obstoječega danskega enojezičnega slovarja, zlasti vlogo sprememb v rabi kolokacij pri prepoznavi novih pomenov z namenom ugotoviti uporabnost postopka pri pripravi slovarskih posodobitev. V prispevku predstavijo rezultate korpusne raziskave, v kateri so uporabili kombinacijo avtomatskega luščenja bigramov in njihove ročne anotacije s strani leksikografov.

Prispevek **Ene Vainik, Marie Tuulik** in **Kristine Koppel** s primerjavo besednih asociacij in kolokacij v estonščini s poudarkom na vlogi besednih vrst prinaša tematski številki psiholingvistično perspektivo. Avtorice med drugim ponudijo razmisleke o izrabi rezultatov študije na področju leksikografije, npr. pri pisanju pomenskih definicij in pri poučevanju tujih jezikov. Tematsko številko sklene prispevek **Eve Pori, Jake Čibeja, Iztoka Kosma** in **Špele Arhar Holdt** o uporabniški evalvaciji avtomatsko izdelanega Kolokacijskega slovarja sodobne slovenščine. Metode avtomatskega luščenja podatkov so v sodobni leksikografiji vse pogosteje uporabljane, zato je koristno opazovati in analizirati odzive različnih tipov uporabnikov, v tem primeru učiteljev, prevajalcev, lektorjev in leksikografov pri uporabi slovarja, ki vsebuje sicer številne, a včasih problematične kolokacijske podatke.



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).

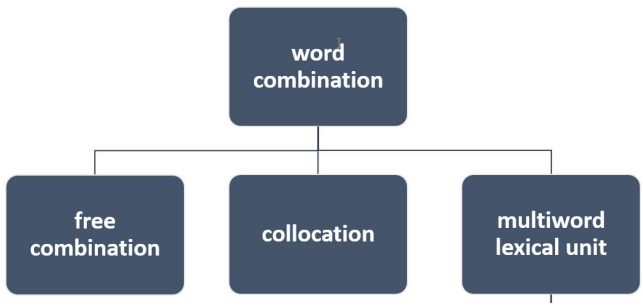


Figure 1: Collocations in word combination typology.

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]

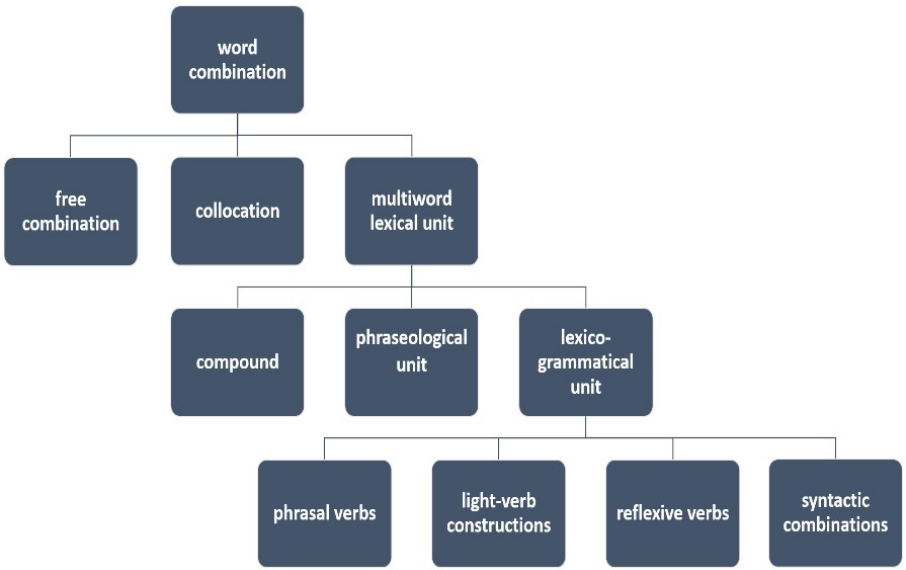


Figure 2: Divison of multiword lexical units.

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).

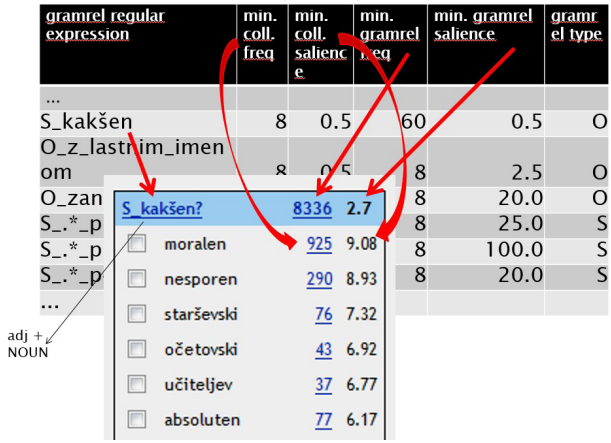


Figure 3: Parameter settings for different grammatical relations and their connections (red arrows) with a table of the syntactic structure adjective + NOUN, illustrated with the results for the noun *avtoriteta* ('authority') in the Word Sketch function.

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.

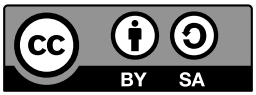
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

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



ENCODING POLYLEXICAL UNITS WITH TEI LEX-0: A CASE STUDY

Toma TASOVAC
Belgrade Center for Digital Humanities, Belgrade, Serbia

Ana SALGADO
NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal,
Academia das Ciências de Lisboa, Lisbon, Portugal

Rute COSTA
NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal

Tasovac, T., Salgado, A., Costa, R. (2020): Encoding polylexical units with TEI Lex-0: A case study. Slovenščina 2.0, 8(2): 28–57.

DOI: https://doi.org/10.4312/slo2.0.2020.2.28-57

The modelling and encoding of polylexical units, i.e. recurrent sequences of lexemes that are perceived as independent lexical units, is a topic that has not been covered adequately and in sufficient depth by the Guidelines of the Text Encoding Initiative (TEI), a de facto standard for the digital representation of textual resources in the scholarly research community. In this paper, we use the Dictionary of the Portuguese Academy of Sciences as a case study for presenting our ongoing work on encoding polylexical units using TEI Lex-0, an initiative aimed at simplifying and streamlining the encoding of lexical data with TEI in order to improve interoperability. We introduce the notion of *macro- and microstructural relevance* to differentiate between polylexicals that serve as headwords for their own independent dictionary entries and those which appear inside entries for different headwords. We develop the notion of *lexicographic transparency* to distinguish between those units which are not accompanied by an explicit definition and those that are: the former are encoded as `<form>`–like constructs, whereas the latter becomes `<entry>`–like constructs, which can have further constraints imposed on them (sense numbers, domain labels, grammatical labels etc.). We codify the use of attributes on `<gram>` to encode different kinds of labels for polylexicals (implicit, explicit and normalised),

concluding that the interoperability of lexical resources would be significantly improved if dictionary encoders would have access to an expressive but relatively simple typology of polylexical units.

Keywords: TEI, Lexicography, Language Resources, Polylexical Units, Interoperability

1 INTRODUCTION

A polylexical unit can be defined as a stable and recurrent sequence of lexemes that are perceived as an independent lexical unit by the speakers of a language. In the specialized literature, different authors with different theoretical backgrounds (Gantar et al., 2018; Fellbaum, 2016; Baldwin and Kim, 2010; Calzolari et al., 2002; Sag et al., 2001; Moon, 1998; Cowie, 1994, 1998; Mel'čuk, 1984–1999, 1998; among others) have referred to these morphosyntactic sequences as multiword expressions, collocations, phrasemes, phraseologies, idiomatic expressions, lexical combinations, and so forth. Each of these designations is often defined inside a particular theoretical linguistic framework.

At the same time, scholars have long recognised that polylexical units are essential components of lexical resources (Svensén, 2009; Atkins and Rundell, 2008; Fontenelle, 1997; Hausmann, 1979; Mel'čuk et al., 1984–1999; Zgusta, 1971). When including a polylexical item in a dictionary, lexicographers decide on the degree of its lexical independence based on several criteria from different fields of knowledge, including statistics, semantics, morphosyntax, pragmatics and/or, broadly speaking, culture. This kind of lexicographic judgement, enacted through a particular editorial policy *and* influenced by the conventions of a given lexicographic tradition, necessarily leads to multiple ways of capturing, classifying and presenting lexicographic knowledge about polylexical units. The lack of a more general agreement within the lexicographic community makes the process of encoding dictionaries particularly challenging: how can we identify, describe and consistently represent this type of linguistic phenomena in lexical resources if we do not agree on what they are and/or what to call them?

Unlike corpus linguists who try to describe linguistic evidence as it appears in recorded instances of genuine language use, or practising lexicographers who

try to systematise their knowledge about words and their meaning by laying it out in dictionary articles, dictionary encoders work on formally representing the concrete lexicographic content of existing dictionaries. This is an important distinction to be kept in mind in the context of what we are trying to achieve in this paper. When, in the rest of this paper, we discuss polylexical units, we will do so from the point of view of lexicographic data modelling, i.e. the process of explicitly marking up the structural hierarchies and the scope of particular textual components appearing in existing dictionary entries in order to convert them to electronic format as part of lexicographic digitisation workflow (Tasovac and Petrović, 2015). In other words, our starting point will be polylexical units as a stable and recurrent sequence of lexemes that are *perceived as independent lexical units by the lexicographers of a given dictionary*. Our focus will be on how these linguistic phenomena appear on a printed dictionary page and at which level of the dictionary microstructure. Our main goal will be to explore how these phenomena can be formally described using the recommendations of the Text Encoding Initiative (TEI),[1] in general, and TEI Lex-0,[2] in particular.

The encoding of polylexical units in dictionaries is a topic that has not been covered adequately and in sufficient depth by the TEI, a *de facto* standard for the digital representation of textual resources in the scholarly research community. We will discuss the challenges and propose some solutions to this problem. We will also argue that a typology of polylexical units for dictionaries encoding – especially given both the limited resources which are usually available for this kind of work *and* data interoperability as a worthy goal to pursue – need to be relatively general so that it can be used and applied by dictionary encoders in a straight-forward fashion.

The terminology we use in this paper aims to be supra-theoretical, and consequently, as neutral as possible, hence our preference for "polylexical units". We recognize, nonetheless, that the term "multiword expression" (MWE) is already widely used, including in the LMF standard, ISO 24613-1:2019. In this paper, we will, therefore, proceed as follows: when we refer to the linguistic structure of a lexical unit composed of two or more lexemes, we will use the term polylexical unit. In our discussion of TEI Lex-0, we will allow "MWE" as

1 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

2 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

an attribute value in order to provide better alignment with LMF and because the TEI Lex-0 community has already used this term.

This article is organised as follows: in Section 2, the lexicographic treatment of polylexical units is explored based on the Dictionary of the Portuguese Academy of Sciences (DLPC) as a case study. A TEI Lex-0 representation of polylexical units in DLPC is discussed in Section 3; and, finally, in Section 4, we offer some concluding remarks and some recommendations about the future work needed in this area.

2 LEXICOGRAPHIC TREATMENT OF POLYLEXICAL UNITS

Dictionaries by design describe systematised knowledge about words and their meanings through typographic conventions that are imbued with meaning and affected by a long tradition: the use of bold typefaces to signal the lemma or headword in a dictionary article; the use of abbreviations (especially in print dictionaries) for grammatical features or usage labels (Salgado et al., 2019a); the numbering of senses and the use of different typefaces for different elements in the hierarchy (definitions, examples, etc.). Experienced dictionary users can become quite proficient at understanding and navigating the structure of the dictionary by interpreting the dictionary's typographic features and the way these features may differ from one dictionary to another. Still, that kind of understanding, based on both knowledge and experience, is not something which can always be easily formalised.

Two main challenges are affecting the modelling of polylexical units in dictionaries, both of them related to the typographical constraints of the print-based, general-language dictionaries:

1. In most general-language dictionaries, polylexical units do *not* appear as headwords, i.e., independent lexical units in the dictionary macrostructure, but rather as sub-units within entries that have a monolexical headword; and

2. Polylexical units in dictionaries are not always *explicitly* labelled as such: they may be typographically singled out, using a particular typeface, but they are not always accompanied by the label which identifies the given unit as a "collocation", "idiom" or a "proverb".

The position of polylexical units in the dictionary and the benefits of lemmatisation have been discussed before (see Jónsson (2009) and Lorentzen (1996), for instance) but for our purposes, it is essential to note that when we suggest particular encodings of the Dictionary of the Portuguese Academy of Sciences, we will be following the structure and the conventions of that very dictionary. That means that we will not be trying to flatten the hierarchy or to encode all polylexical units using the same set of tags. We will be encoding them as they appear within the structure imposed by the dictionary itself.

As for the lack of explicit labels for particular types of polylexical units, we will, in the subsequent sections, explain the extent to which the types can be deduced from the entry structure. We will, in the process, also consult the Introduction to the Dictionary, which to some degree explains the structure from the point of view of the dictionary editors.

2.1 DLPC as a case study

The *Dicionário da Língua Portuguesa Contemporânea* (DLPC) is a monolingual Portuguese dictionary published by Academia das Ciências de Lisboa (2001). As such it is representative of the Academy tradition in European lexicography: large-scale and long-term dictionary projects, initiated and compiled by official national bodies established to record, maintain and promote authoritative accounts of language use (see Considine, 2014). It contains around 70,000 entries and was published in 2001 in two volumes, totalling 3880 pages. The PDF version of the printed dictionary was later converted into XML using a customised version of the P5 schema of the Text Encoding Initiative (TEI), while a custom-built dictionary writing system using TEI as a data model in the backend, was developed to serve as an editing environment for the new and improved online edition of the dictionary (Simões et al., 2016). Besides, the DLPC is currently being converted to the TEI Lex-0 format for data interoperability purposes (Salgado et al., 2019b).

We selected DLPC as a case study in our ongoing work on developing guidelines for encoding polylexicals in TEI Lex-0 for two reasons: (1) as a monolingual scholarly dictionary of the Portuguese language, DLPC covers a wide range of polylexical units from collocations to strongly lexicalised

expressions; and (2) because scholarly dictionaries, with their "pursuit of completeness concerning the entries relevant to subject matters" (see Kinable, 2015) typify detailed lexicographic information and elaborate microstructure, which can more often than not pose challenges in terms of consistent data modelling.

Given the lack of detail given to the encoding of polylexical units in the TEI Guidelines, the authors thought it was essential to take a single but complex dictionary as a starting point for our exploration of the topic in this paper. It goes without saying that further comparative work will be needed to validate and improve our recommendations. But it also goes without saying that the proposed mechanisms for marking up polylexical units in DLPC at different levels of the dictionary microstructure will generally be applicable to other dictionaries as well. While dictionaries may differ in terms of their "typographic view", i.e. page layout, column and line breaks, and their "editorial view", i.e. the sequential arrangement of individual tokens along with the use of specific font styles, punctuation and special symbols (the so-called "editorial" view), they are more easily comparable in terms of their "lexical view", i.e. the underlying structure and the types of information units contained in them.[3] While our focus on DLPC here is, above all, a matter of practicality, we will be using it as a springboard for illustrating broader encoding challenges.

Structurally speaking, we should distinguish two main types of polylexical items:

1. polylexical units which serve as headwords for their own independent dictionary entries;

2. polylexical units which appear inside entries for different headwords.

We will refer to the first category as the macrostructurally relevant polylexical units and the second as the microstructurally relevant polylexical units. The notion of relevance here is local – it refers only to the structure of the given dictionary.

3 On the difference between different "views"of the dictionary, see Section 9.5 "Typographic and Lexical Information in Dictionary Data" in the TEI Guidelines, https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DIMV.

2.1.1 Macrostructurally relevant polylexical units

In Salgado et al. (2019b), we identified four different types of headwords in DLPC: monolexical units, polylexical units, affixes and abbreviations. Polylexical headwords can be of two different types:

i) compounds ("palavras compostas" which are graphically realized as "palavras hifenizadas" ["hyphenated words"] (DLPC, 2001, p. XIV) (e.g. **decreto-lei** [decree-law], **franco-canadiano** [French Canadian], **pré-cristão** [pre-Christian); and

ii) Latin phrases ("locuções latinas") (e.g. **fiat lux** [let there be light]).

In the context of this particular dictionary and, more generally speaking in the Portuguese orthographic tradition, hyphenation is treated as a mark of lexicalisation and non-compositional meaning, which leads to lexicographic treatment at an entry-level. For instance, **lugar-comum** [commonplace] does not merely connote a common type of place [**lugar comum**]: the meaning of the hyphenated unit – an ordinary thing, a platitude or a cliché – cannot be obtained from its constituent parts. As such, it is considered, from the point of view of the lexicographer, headword material.[4] Latin phrases, which are used in the Portuguese language, are included in the DLPC macrostructure as entries of their own because they cannot be easily ascribed to particular Portuguese headwords.

2.1.2 Microstructurally relevant polylexical units

Microstructurally relevant polylexical units in DLPC fall into two distinct categories:

i) *lexicographically transparent* polylexical units, i.e., units which are *not* accompanied by an explicit definition; and

ii) *lexicographically non-transparent* polylexical units, i.e., units which *are* accompanied by an explicit definition.

4 The hyphen as a marker of semantic opaqueness, however, is, to a certain extent, a projection of lexicographic idealism. Many polylexicals which are are traditionally hyphenated in Portuguese dictionaries are written without the hyphen in common usage.

2.1.2.1 Lexicographically transparent polylexical units

Lexicographically transparent polylexical combinations in DLPC do not come with an explicit definition in addition to the general one already given for the sense of the headword under which they appear. The lexicographic assumption is that the user will be able to deduce their meaning from their individual components and their syntactic structure. These kinds of polylexical units serve as additional illustrations for the given sense. Still, they differ from typical full-sentence examples in that they stress the collocational aspects of the given headword: they function as lexicographical pointers to the user for how the given word is meaningful — and typically — used in combination with other words. The closeness of these polylexical combinations to actual examples in DLPC is signalled by their proximity next to each other in the dictionary entry, and by their common typographic features: both are set in italic typeface and grouped together inside a particular sense.

descalçar [dɨʃkałsár]. *v.* (Do lat. * *discalceāre*). **1.** Tirar aquilo que se tem calçado; despir os pés ou as mãos; tirar o calçado. ≠ CALÇAR, ENFIAR, PÔR. *Descalçou-se mal chegou a casa. Descalçou um chinelo e atirou-mo. Descalçou a criança e deitou-a.* *+ as botas, as luvas, as meias; + os sapatos.* **2.** Tirar aquilo que serve de apoio para que fique bem assente no chão; tirar o calço. *Descalçou a mesa e esta ficou a balançar.* **3.** Tirar as pedras que cobrem o pavimento. ≃ DESCALCETAR, DESEMPEDRAR. ≠ CALCETAR, EMPEDRAR. **4.** *Fam.* Tirar os recursos; deixar sem soluções, sem possibilidade de resolver alguma coisa; deixar descalço. ≃ DESAMPARAR, DESARMAR, DESPREVENIR. **descalçar a bota,** resolver uma dificuldade.

Figure 1: Descalçar [to remove] – DLCP (2001).

The monolexical lemma **descalçar** [to remove], as shown in Figure 1, has four numbered senses. The first sense consists of a definition “tirar aquilo que se tem calçado; despir os pés ou as mãos; tirar o calçado” [take off one’s shoes; undress one’s feet or hands], followed by three antonyms “calçar, enfiar, pôr” [to put on; to slip on] and three full-sentence examples. In addition, DLPC

lists two sets of typical collocates of the headword separated by a semicolon: + *as botas, as luvas, as meias* and + *os sapatos*. The plus sign is used as a label representing the headword, but the headword is stated only once in a given set: in other words, + *as botas, as luvas, as meias* is directly equivalent to *descalçar as botas, as luvas, as meias* [to remove one's shoes, one's boots, one's gloves], but indirectly equivalent to: *descalçar as botas, descalçar as luvas* and *descalçar as meias*. This is an example of lexicographic shorthand, typical of print dictionaries. In the given case, the user is expected to be able to decipher that the verb *descalçar*, in the given sense (removing something one is wearing), is typically used with objects such as shoes, boots or gloves.

This type of polylexical unit is classified as "co-ocorrente privilegiado" [privileged co-occurrent] in the Introduction to DLPC.[5] The sets separated by the semi-colon are described as "semantically and syntactically related blocks".[6] It appears, however, that this rule is not always followed consistently because the two sets we described above are semantically and syntactically indistinguishable: the difference in the gender of the collocate (*as botas* vs. *os sapatos*) is of no relevance to the construction of this particular type of polylexical unit.

2.1.2.2 Lexicographically non-transparent polylexical units

In DLPC, the treatment of lexicographically non-transparent polylexical units follows a minimal entry-like structure in which the polylexical unit itself is set in boldface (similar to a lemma) and accompanied by a definition (or a pointer to a definition under a different entry). These units can themselves be divided into two further categories, based on the position they take up in the entry microstructure:

1. those that are attached to particular senses; and
2. those that appear at the end of the entry, following the description of individual senses.

5 Privileged co-occurrent is a dependency relationship ("uma relação de dependência") which occurs between full words ("palavras plenas") such as nouns, adjectives, verbs and adverbs and other words in the construction of sentences ("na construção das frases") (DLPC, 2001, p. XXI).

6 "os co-ocorrentes são apresentados em blocos semântica e sintaticamente afins, separados por ponto e vírgula; dentro de cada bloco aparecem separados por vírgula." (DLPC, p. 2001, XXI).

Take, for instance, the following example (Figure 2):

bombeiro[1]**, a** [bõbɐ́jru, -ɐ]. *s.* (De *bomba*[2] + suf. *-eiro*). **1.** Pessoa que faz parte de um corpo organizado de combate a incêndios; o que trabalha com bombas de incêndio. *Sapadores bombeiros.* «*Bombeiro, o pai, apagava incêndios nas matas, trazia as chamas na vista, televisão nem pensar, entretinha-se para adormecer com o* Almanaque Cearense.» (M. O. BRAGA, *Lua*, p. 95). **bombeiro voluntário,** o que pertence a uma corporação com a obrigatoriedade de acudir a incêndios, acidentes, unicamente por filantropia. **corpo**[+] **de bombeiros. 2.** *Bras.* Pessoa que faz ou conserta bombas, canos. **bombeiro hidráulico,** *Bras.,* canalizador. **3.** *Bras. Fam.* Criança que, durante a noite, tem incontinência urinária.

Figure 2: Bombeiro [firefighter] – DLCP (2001).

The monolexical item **bombeiro** [firefighter], as shown in Figure 2, is a headword for an entry which has three distinct, numbered senses. The first sense has a definition written in regular typeface. Two unnumbered examples follow the definition in italic typeface; and of the two examples, the latter is a citation: it is surrounded by quotation marks and followed by a bibliographic reference inside brackets. Following the definition and the examples, the first sense of **bombeiro** has two polylexical items attached to it: **bombeiro voluntário** [volunteer firefighter] and **corpo de bombeiros** [fire brigade]. Both of these polylexical items appear in boldface, just like the lemma, but only the first of the two has a definition in regular typeface ("o que pertence a uma corporação com a obrigatoriedade de acudir a incêndios, acidentes, unicamente por filantropia") appearing after a comma, which is used as a field separator. The second polylexical item has no definition, but its other distinguishing feature is the superscript plus sign which appears after the word "corpo". In DLPC, this superscript label is used by convention to indicate that the given polylexical unit is defined under a different headword: **corpo**[+], in this case, can be thought of as a cross-reference: it tells the reader to look up the entry **corpo** in order to find the definition for **corpo de bombeiros.**

The Introduction to DLPC calls this type of polylexical units "combinatórias fixas" [fixed combinations].[7] They are attached to particular senses of the headword, and defined only once, the first time they appear in the dictionary. That is why **bombeiro voluntário** is defined under **bombeiro** and cross-referenced from **voluntário**, whereas **corpo de bombeiros** is defined under **corpo**, but cross-referenced from **bombeiro**.

Polylexical units that appear *outside* the sense structure are organised the same way as the "fixed combinations" described above: they have lemma-like headwords and can contain definitions, domain labels, etc. The difficulty, from the modelling point of view, is that DLPC does not use a delimiter or a label to separate the last sense in a given entry from the polylexical units that are not attached to a particular sense. That means that for all intents and purposes, a polylexical unit appearing at the end of an entry in DLPC is typographically indistinguishable from a polylexical entry appearing in the last sense of the given entry.

The Introduction to DLPC describes two types of polylexical units which appear outside the sense structure:

1. "locuções" [phrases]; and
2. "expressões idiomáticas ou fraseológicas" [idiomatic or phraseological expressions].

The two types of polylexical units appear in bold on the dictionary page, the only difference being in their labelling: "phrases" are labelled as such, whereas "idiomatic expressions" are not. Neither of the two terms is explicitly defined in the Introduction to the dictionary.

The entry for **dali**, a contraction of "de" (of, from) and "ali" (there), as shown in Figure 3, has two numbered senses. The definitions of the two senses, each of each describes one possible function of the compound preposition (indicating a point of origin of a movement; or indicating the origin of a person,

7 "Fixed combinations" are defined as "combinações de palavras cristalizadas ou em vias de cristalização, que funcionam frequentemente como verdadeiros compostos não hifenizados" [combinations of words crystallised or in the process of crystallisation, which often function as authentic non-hyphenated compounds] (DLPC, p. XXI) e.g. "pedra preciosa" [gemstone] or "sala de jantar" [dining room].

dali [dɐlí]. *contr.* Contr. da prep. *de* com o adv. *ali*. **1.** Indica ponto de partida de um movimento: daquele lugar, daquele ponto. *Fui dali para o médico, sem passar por casa.* **2.** Indica pessoa, entidade ou situação acabadas de referir e que constituem a fonte ou origem de alguma coisa. *Eu bem insisti junto dos serviços, mas parece que dali não virá a solução do problema. O meu avô está tão mal que os médicos já não esperavam nada dali.* **dali a nada,** *loc. adv.,* muito pouco tempo depois. *Dali a nada estava ele a chamar-me.* **dali a pouco (tempo),** *loc. adv.,* pouco tempo depois. *Dali a pouco chegou o meu irmão.* **dali em diante,** *loc. adv.,* a partir de então, desde esse momento. *Devido a um esforço maior tirou boa nota: dali em diante foi sempre aluno aplicado.* **dali para a frente,** *loc. adv.,* o m. que *dali em diante. Dali para a frente nunca mais senti dificuldades.* **dali por diante,** *loc. adv.,* o m. que *dali em diante. Dali por diante resolveu-me sempre qualquer problema.*

Figure 3: Dali [from there] – DLCP (2001).

thing or situation). From the typographic layout of the entry alone, it would be impossible to judge whether the five polylexical units **dali a nada**, **dali a pouco (tempo)**, **dali em diante**, **dali para a frente** and **dali por diante** are meant to be attached to the second sense or whether they appear outside the sense structure. Each of the polylexical units is explicitly labelled as *loc. adv.* [adverbial phrase].

The dictionary itself defines **locução** in its grammatical sense as a group of words that work, semantically and syntactically as a whole, equivalent to a single word.[8] The same sense also includes several different types of expressions: adjectival, adverbial, conjunctive, prepositional and verbal.

8 "Grupo de palavras que funcionam, semântica e sintacticamente como um todo, que equivalem a um só vocábulo. Rey and Chantreau (1993) underline the difference between lexical and grammatical phrases: "Locution [...] est exactement 'manière de dire', manière de former le discours, d'organiser les éléments disponibles de la langue pour produire une forme fonctionnelle. C'est pourquoi on peut parler de 'locutions adverbiales' ou 'prépositives', alors que ces mots grammaticaux complexes ne seraient jamais appelés des 'expression' (p. VI).

dura [dúɾɐ]. *s. f.* (Deriv. regres. de *durar*). **1.** Qualidade do que resiste ao tempo e ao uso. ≃ DURABILIDADE. **2.** Espaço de tempo entre o princípio e o fim de uma coisa. ≃ DURAÇÃO. **ser de pouca dura,** durar pouco tempo; passar depressa. *Foi amor de pouca dura.* **ser sol de pouca dura,** ser algo que, por ser bom ou agradável, dura pouco tempo. *Aquele bom-humor do chefe foi sol de pouca dura, pois começou logo a resmungar.*

Figure 4: Dura [durability; duration] – DLCP (2001).

The entry for **dura** [duration], on the other hand, as shown in Figure 4, has two numbered senses followed by two polylexical units: **ser de pouca dura** [to be short-lived] and **ser sol de pouca dura** [lit. to be a sun that does not last, i.e., to be a nine days' wonder]) without explicit labelling of the type of units that they are.

In DLPC proper, **expressão idiomática** has the domain label Linguistics and is defined as an expression that is peculiar to the language, usually because its meaning is not literal.[9] The **expressão fraseológica** [phraseological expression] is not defined in the dictionary.

3 REPRESENTING POLYLEXICAL UNITS IN TEI LEX-0

TEI is a *de facto* standard for the digital encoding of all types of written texts, ranging from standard books to poems, visiting other less straightforward documents, e.g., tables, mathematical formulae, cookery recipes or even music notation. It also defines how specific humanities resources, including morphologically annotated monolingual and parallel corpora, should be encoded. Chapter 9 of the TEI Guidelines[10] focuses specifically on the encoding of dictionaries and other types of lexical resources.

TEI Lex-0[11] (Romary and Tasovac, 2018) is a newer, stricter subset of TEI, which was launched in 2016 by the DARIAH Working Group on Lexical

9 "*Ling*. a que é peculiar a uma língua, geralmente devido ao facto de o seu significado não ser literal."

10 https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

11 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

Resources.[12] The goal of TEI Lex-0 is to establish a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources. TEI Lex-0 should not be thought of as a replacement of the Dictionary Chapter in the TEI Guidelines but rather as a "format that existing TEI dictionaries can be unequivocally transformed to in order to be queried, visualised, or mined uniformly".[13] In the context of the ELEXIS project,[14] TEI Lex-0 has been adopted, together with OntoLex, as one of the baseline formats for the ingestion of existing dictionaries into the ELEXIS infrastructure (McCrae et al., 2019). While TEI Lex-0 is being developed, some of its best-practice recommendations are also changing the recommendations of TEI Guidelines themselves.

3.1 Polylexical units in TEI Guidelines

The Dictionary Chapter of the TEI Guidelines is very sparse when it comes to recommendations for encoding polylexical units. The only mention of the adjective "multi-word" appears in the definition of the element `<term>`: "contains a single-word, multi-word, or symbolic designation which is regarded as a technical term" but this is not relevant for the encoding of polylexical units in general-purpose dictionaries.

TEI includes an element `<colloc>` (collocate), which is defined as containing "any sequence of words that co-occur with the headword with significant frequency" but, in a different example, "colloc" is used as an attribute value for the element `<usg>` (usage). It is precisely this type of ambiguity that TEI Lex-0 is trying to resolve.

The TEI Guidelines recommend the use of `<re>` (related entry) to encode "related entries for direct derivatives or inflected forms of the entry word, or for compound words, phrases, collocations, and idioms containing the entry word" with barely any useful examples, or discussion of how to encode different types of polylexical units. TEI Lex-0, on the other hand, does not include `<re>`. In TEI Lex-0, `<entry>` was made recursive in order to account

12 https://www.dariah.eu/activities/working-groups/lexical-resources/

13 https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.1

14 https://elex.is/

for nestable entry-like structures without the need to resort to <re>, a differently named element whose content model would be indistinguishable from <entry> itself. Eventually, the new content model of <entry>, which allows nesting, was adopted by TEI itself.

3.2 Encoding macrostructurally relevant polylexical units

In terms of modelling, polylexical units as headwords do not present any particular challenges for TEI Lex-0. Because they function as lemmas in dictionary entries, they need to be encoded with the required @type attribute on <form>. DLPC does not label them explicitly as polylexical, which is why previously in Salgado et al. (2019b), the authors recommended that this information be encoded as a @type attribute on <entry>. At the time, the goal was to differentiate entries based on their headwords as monolexical, polylexical, affixes and abbreviations. Nevertheless, for lexicographic work with digital lexical resources, it is crucial not only to be able to extract all polylexical units but also to have the possibility to individualize them. That is why we need to go one step further and develop a mechanism for encoding different types of polylexical units.

decreto-lei [dɨkrɛtulɐ̃j]. *s. m. Dir.* Acto normativo proveniente do Governo da República. *Actualmente, os decretos-leis são publicados na primeira série-A do Diário da República.* Pl. decretos-leis.

Figure 5: Decreto-lei [decree-law] – DLCP (2001).

```
<entry xml:lang="pt" xml:id="decreto-lei" type="polylexicalUnit">
   <form type="lemma">
      <orth>decreto-lei</orth>
      <pron>dɨkrɛtul'ej</pron>
   </form>
   <gramGrp>
      <gram type="mwe" value="composto"/>
      <gram type="pos" norm="NOUN">s.</gram>
      <gram type="gen">m.</gram>
   </gramGrp>
   <!--etc. -->
</entry>
```

In Figure 5, the only addition to the encoding suggested in Salgado et al. (2019b) is the inclusion of `<gram type="mwe" value="composto"/>` to mark up the particular kind of polylexicality, even though this type of entry-level polylexicals is not explicitly labelled as such. For a detailed explanation of how one can encode different types of polylexical units, regardless of whether the given dictionary uses explicit labels for them or not, see Section 3.4 in this paper.

The situation with Latin expressions is slightly different because they are explicitly labelled in DLPC as such. See Figure 6:

fiat lux *loc. lat.* Exprime o desejo de que se torne clara alguma coisa importante.

Figure 6: *Fiat lux* – DLCP (2001).

DLPC labels the headword as *loc. lat.,* which stands for "locução latina" [Latin phrase]. This abbreviated label uses the same italic typeface in the same position as the label *s. m.* (substantivo masculino [masculine noun]), which we saw in the above example for **decreto-lei**. From DLPC's internal logic, one could argue that the label *loc. lat.* functions as a grammatical label. And yet, the two-partite structure of *loc. lat.* is internally different from that of *s. m.* While both part-of-speech and gender are grammatical categories, one can not say the same of *loc. lat.,* which combines grammatical and etymological information. Therefore, we recommend that this label be modelled as two different components: an `mwe` label for *loc. lat.,* which adequately represents the label of the source, and an `etym` element to explicitly mark up the language of origin.

```
<entry xml:lang="pt" xml:id="fiat_lux" type="polylexicalUnit">
   <form type="lemma">
      <orth>fiat lux</orth>
   </form>
   <gramGrp>
      <gram type="mwe" value="locução_latina">loc. lat.</gram>
   </gramGrp>
   <etym type="borrowing"><lang value="la"/></etym>
   <!--etc. -->
</entry>
```

The use of both grammatical and etymological tags is advantageous because it makes the same phrase findable in two different search contexts.

3.3 Encoding microstructurally relevant polylexical units

Microstructurally relevant polylexical units will be encoded differently in TEI Lex-0 depending on whether they are lexicographically transparent or not. Only the non-transparent ones will require full markup within an `<entry>` construct.

3.3.1 Encoding lexicographically transparent polylexical units

Following from our discussion in Section 2.2.2.1, the TEI Lex-0 encoding of lexicographically transparent polylexical units in DLPC should meet the following requirements:

1. each set of polylexical units should be grouped together to represent the microstructure of the entry adequately;
2. each polylexical unit should be identifiable as such for easy retrieval;
3. the explicit label "+" should be used only where it occurs in the dictionary text, but the implicit positioning of the headword in the given polylexical unit should be marked up as well.

Because lexicographically transparent polylexical units are *not* structured as mini-entries but are instead presented to the reader as a sequence of forms, we recommend to encode them as `<form>` elements:

```
<sense xml:id="descalçar.1">
   <!--etc.-->
   <form type="collocations">
      <form type="collocation">
         <orth>
            <ref type="oRef"><lbl>+</lbl></ref>
            <seg>as botas</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
      <pc>,</pc>
```

```
      <form type="collocation">
         <orth>
            <ref type="oRef"/>
            <seg>as luvas</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
      <pc>,</pc>
      <form type="collocation">
         <orth>
            <ref type="oRef"/>
            <seg>as meias</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
   </form>
   <pc>;</pc>
   <form type="collocations">
      <form type="collocation">
         <orth>
            <ref type="oRef"><lbl>+</lbl></ref>
            <seg>os sapatos</seg>
         </orth>
         <gramGrp>
            <gram type="mwe" value="co-ocorrente_privilegiado"/>
         </gramGrp>
      </form>
   </form>
   <pc>.</pc>
</sense>
```

The `<ref>` element of the type `oRef` (orthographic reference) is used to encode the position of the headword in the polylexical unit. Optionally, this element can contain a `<lbl>+</lbl>` to reflect the explicit headword substitution label.

3.3.2 Encoding lexicographically non-transparent polylexical units

A sense-related non-transparent polylexical unit can be encoded in TEI Lex-0 within an <entry> construct.[15] The type of the polylexical unit is indicated by the <gram> element, which is discussed in greater detail in the following section of this paper.

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="bombeiro">
   <form type="lemma">
      <orth>bombeiro</orth>
   </form>
   <!--etc. -->
   <sense xml:id="bombeiro.1">
      <!--etc. -->
      <entry xml:id="bombeiro_voluntario" xml:lang="pt" type="relatedEntry">
         <form type="lemma">
            <orth>bombeiro voluntário</orth>
         </form>
         <gramGrp>
            <gram type=”mwe” value=”combinatória_fixa”/>
         </gramGrp>
         <pc>,</pc>
         <sense xml:id=”bombeiro_voluntario.1”>
          <def>o que pertence a uma corporação com a obrigatoriedade de acudir
a incêndios, acidentes, unicamente por filantropia</def>
            <pc>.</pc>
         </sense>
      </entry>
      <entry xml:id="corpo_de_bombeiros" xml:lang="pt" type="relatedEntry">
         <form type="lemma">
            <orth>
               <ref type="entry"><seg>corpo</seg><lbl>+</lbl></ref>
               <seg>de bombeiros</seg>
            </orth>
         </form>
         <pc>.</pc>
      </entry>
   </sense>
   <!--etc.-->
</entry>
```

15 TEI and TEI Lex-0 diverge somewhat on how they allow this, but the end result is the same: in TEI Lex-0, the content model of <sense> allows elements from the class model.sensePart as its children, and <entry> is a member of this class; whereas in TEI <sense> has a broader content model which allows members of the class model. entryPart as its children.

Because sense-related polylexical units are modelled as nested entries, they can include domain labels as well. For instance (Figure 7):

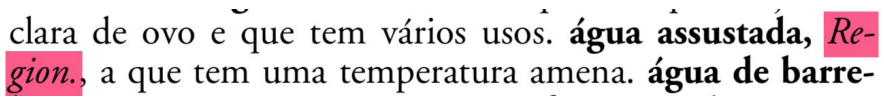

clara de ovo e que tem vários usos. **água assustada,** *Region.*, a que tem uma temperatura amena. **água de barre-**

Figure 7: Água assustada [mild water] – DLCP (2001).

```
<sense xml:id=”agua.4”>
    <!--etc.-->
    <entry xml:id="agua_assustada" xml:lang="pt" type="relatedEntry">
        <form type="lemma">
            <orth>água assustada</orth>
        </form>
        <pc>.</pc>
        <usg type="geographic" norm="regionalism">Region.</usg>
        <pc>,</pc>
        <sense xml:id=”agua_assustada.1”>
            <def>a que tem uma temperatura amena.</def>
        </sense>
    </entry>
    <!--etc.-->
</sense>
```

Sense-related polylexical units can themselves be polysemous. For instance (Figure 8):

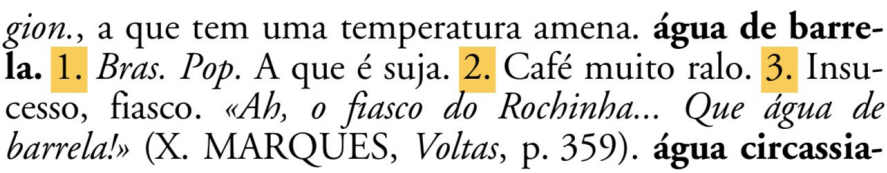

gion., a que tem uma temperatura amena. **água de barre-la.** 1. *Bras. Pop.* A que é suja. 2. Café muito ralo. 3. Insucesso, fiasco. *«Ah, o fiasco do Rochinha... Que água de barrela!»* (X. MARQUES, *Voltas*, p. 359). **água circassia-**

Figure 8: Água de barrela [dirty water; weak coffee; fiasco] – DLCP (2001).

```
<sense xml:id="agua.4">
    <!--etc.-->
    <entry xml:id="agua_de_barrela" xml:lang="pt" type="relatedEntry">
        <form type="lemma">
            <orth>água de barrela</orth>
        </form>
        <sense xml:id=”agua_de_barrela.1” n=”1”>
```

```
            <lbl>1.</lbl>
            <usg type=”geographic”>Bras.</usg>
            <usg type=”socioCultural”>Pop.</usg>
            <def>A que é suja.</def>
        </sense>
        <sense xml:id=”agua_de_barrela.2” n=”2”>
            <lbl>2.</lbl>
            <def>Café muito raro.</def>
        </sense>
        <sense xml:id=”agua_de_barrela.3” n=”3”>
            <lbl>3.</lbl>
            <def>Insucesso, fiasco.</def>
            <cit type=”example”>
                <quote>Ah, o fiasco do Rochinha... Que água de barrela!</quote>
                <bibl>
                    <author>X. MARQUES</author>
                    <pc>,</pc>
                    <title>Voltas</title>
                    <pc>,</pc>
                    <citedRange>p. 359</citedRange>
                </bibl>
            </cit>
        </sense>
    </entry>
    <!--etc.-->
</sense>
```

Entry-related polylexicals have the same structure as the sense-related ones, only they appear as children of the main entry:

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="dali">
   <form type="lemma">
      <orth>dali</orth>
   </form>
   <!--etc.-->
   <entry xml:id="dali_a_nada" xml:lang="pt" type="relatedEntry">
      <form type="lemma">
         <orth>dali a nada</orth>
      </form>
      <gramGrp>
         <gram type="mwe" value="locução_adverbial">loc. adv.</gram>
      </gramGrp>
      <pc>,</pc>
```

```
      <sense xml:id="dali_a_nada.1">
         <def>muito pouco tempo depois</def>
         <pc>.</pc>
         <cit type=”example”>
            <quote>Dali a nada estava ele a chatear-me.</quote>
         </cit>
      </sense>
   </entry>
   <!--etc.-->
</entry>
```

The same type of encoding applies to idiomatic expressions:

```
<entry type="monolexicalUnit" xml:lang="pt" xml:id="dura">
   <form type="lemma">
      <orth>dura</orth>
   </form>
   <!--etc.-->
   <entry xml:id="ser_de_pouca_dura" xml:lang="pt" type="relatedEntry">
      <form type="lemma">
         <orth>ser de pouca dura</orth>
         <gramGrp>
            <gram type=”mwe” value=”expressão_idiomática”/>
         </gramGrp>
      </form>
      <pc>,</pc>
      <sense xml:id=”ser_de_pouca_dura.1”>
         <def>durar pouco tempo; passar depressa</def>
         <pc>.</pc>
         <cit type=”example”>
            <quote>Foi amor de pouca dura.</quote>
         </cit>
      </sense>
   </entry>
   <!--etc.-->
</entry>
```

3.4 Encoding types of polylexical units

We saw above that some polylexical units in DLPC are explicitly labelled as such (for instance *loc. lat.* or *loc. adv.*, but some are not – for instance, hyphenated compounds as headwords, or idiomatic expressions. TEI Lex-0 should

provide a consistent but flexible mechanism for labelling types of polylexical units in dictionaries regardless of whether these labels exist explicitly in the dictionary source or not. We propose to encode this information using the existing TEI `gramGrp/gram` mechanism, in order to have the maximum flexibility to cover these three distinct types of labels:

1. *implicit labels*, i.e., those labels whose value can only be deduced from its typographical properties or its position in the entry structure, but are not present on the dictionary page (for instance, compounds as headwords in DLPC);
2. *explicit labels*, i.e. labels which appear on the dictionary page (for instance, *loc. adv.* in DLPC);
3. *normalised labels*, i.e. normalised versions of either implicit or explicit labels, which can be used to improve the interoperability of the labels.

The consistent labelling of polylexical units in a dictionary can be achieved by adopting the following principles:

1. Any polylexical unit should be identified by the presence of a generic element-attribute combination: `<gram type="mwe"/>`. Without any further classification, `<gram type="mwe"/>` does not tell us anything about the specific type of the polylexical unit.
2. Explicit labels should be encoded as text nodes of `gram`: `<gram type="mwe">loc. adv.</gram>`.
3. Implicit labels should be placed in the `@value` attribute.
4. Normalised values should be placed in the `@norm` attribute.

In addition to being encoded as text nodes, explicit labels should, for the sake of consistency with implicit labels, also use the `@value` attribute. This is to avoid situations in which some labels are encoded as text and some as attributes. The consistent use of the `@value` attribute for both explicit and implicit labels will make it easier to retrieve all labels of a specific type regardless of how they are labelled in the text of the dictionary. Also, it is important to emphasize that the `@value` and `@norm` attributes should be kept conceptually distinct: the former should be used as a locally non-ambiguous identifier of both

the explicit and implicit labels in a given dictionary; the latter, on the other hand, should be optionally used as a placeholder for a dictionary-independent classification of the local label.

```
<gramGrp>
    <!--NOT RECOMMENDED: explicitly labelled MWE as text node only -->
    <gram type="mwe">loc. adv.</gram>
    <!--RECOMMENDED: explicitly labelled MWE as text node + attribute -->
    <gram type="mwe" value="locução_adverbial">loc. adv.</gram>
    <!--implicitly labelled MWE-->
    <gram type="mwe" value="co-ocorrente_privilegiado"/>
    <!--more work needed: normalizing values-->
    <gram type=”mwe” value=”locução_adverbial” norm=”???”>loc. adv.</gram>
    <gram type=”mwe” value=”co-ocorrente_privilegiado” norm=”???”/>
</gramGrp>
```

A typology of labels for polylexical units that would work across multiple dictionaries and languages would be needed if we were to suggest possible values for the `@norm` attribute. Neither TEI nor TEI Lex-0 currently refers to any such typology. However, such a typology would be very helpful for any work on aligning multiple dictionaries, studying them in parallel or pooling various lexical resources together. For instance, in DLPC, the Latin phrase **habeas corpus** is a headword labelled as *loc. lat.* [Latin phrase] but the same polylexical unit in the *Grande Dicionário Houaiss da Língua Portuguesa* (Houaiss, 2015) is labelled as *loc. subst.* [locução substantiva; noun phrase] and “[lat.]”, which is an explicit label for Latin etymology. A typology of polylexical units would make it possible to normalize both explicit and implicit labels across different dictionaries.

4 CONCLUDING REMARKS

Our recommendations for encoding polylexical units using TEI Lex-0 show that TEI Lex-0 is fully capable of consistently marking up polylexical units as constituent parts of the dictionary macro- and microstructure, regardless of whether they appear as headwords in independent entries, or in nested entry-like structures inside entries for monolexical units. The use of nested `<entry>` elements to encode polylexical units inside dictionary entries is a robust mechanism which can take care of all kinds of lexicographic constraints

imposed on the description of polylexical units (polysemy, domain labels, grammatical labels etc.), whereas the combination of `<gram>` element and attributes `@type`, `@value` and `@norm` can be used consistently to encode explicit, implicit and normalised versions of the labels.

In this paper, we focused on the formal representation of polylexical units as they appear on the page of a single dictionary because we wanted to document the process of translating lexicographic and typographic conventions from linear text strings to hierarchical, tree-like structures using the vocabulary and syntactic constraints of TEI Lex-0. While further comparative work will be needed to validate our recommendations on a larger sample, the process we described in this paper and the markup solutions we proposed are sufficiently abstract to serve as a basis for marking up the lexical view of polylexical items in various dictionaries, even though we can expect to see more pronounced differences in their editorial and typographic views. When it comes to designing and applying TEI Lex-0 markup to dictionary entries, the question of whether a dictionary is a paper dictionary, a retrodigitised one or a born-digital resource is of little consequence: what matters is that one can consistently identify, represent and validate all the microstructural elements in a given dictionary entry using a standardised vocabulary.

As we could see in the penultimate section of this paper, the interoperability of encoded lexical resources would be significantly improved if dictionary encoders would have access to a typology of polylexical units that was both expressive and straightforward enough to apply when modelling lexical data. It would be safe to say that very detailed typologies, like the one proposed by Bergenholtz (2013), which includes twenty different types of MWEs, would be challenging to implement in practice. That is why more work on the classification of polylexical items *specifically for encoding purposes* will be necessary. One could argue that there is "no hope of finding a single classification or taxonomy of polylexical units that can be used for all purposes" (Sailer, 2018, p. vi), but a comparative study of multiple dictionaries in different languages would bring us one step closer to proposing, discussing and eventually agreeing on a sensible typology that could be used in the context of TEI Lex-0 as a set of attribute values for normalizing local lexicographic classifications. We hope to pursue this line of work in the future.

Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure), and by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

REFERENCES

Dictionaries

Dicionário da Língua Portuguesa Contemporânea. (2001). João Malaca Casteleiro (Eds.), 2 vols. Lisboa: Academia das Ciências de Lisboa and Editorial Verbo.

Dictionnaire des Expressions et Locutions. (1993). Alain Rey and Sophie Chantreau (Eds.). Col. Les Usuels. Paris: Éd. Dictionnaires Le Robert.

Grande Dicionário Houaiss da Língua Portuguesa. (2015). Instituto António Houaiss Bloco Gráfico, Lda. Lisboa: Círculo de Leitores.

Websites

DARIAH WG = *Lexical Resources and the H2020-funded European Lexicographic Infrastructure (ELEXIS)*. Retrieved from https://github.com/DARIAHERIC/lexicalresources/tree/master/Schemas/TEILex0 (23. 2. 2020)

TEI Consortium (Ed.) = *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (2019). Version 3.5.0. [Last updated on 29th January 2019, revision 3c0c64ec4.] TEI Consortium. Retrieved from http://www.tei-c.org/Guidelines/P5/ (23. 2. 2020)

Other

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

Baldwin, T., & Kim, S. (2010): Multiword Expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed., pp. 267–292). Boca Raton, USA, CRC Press.

Bergenholtz, H., & Gouws, R. (2013). A Lexicographical Perspective on the Classification of Multiword Combinations. *International Journal of Lexicography*, *27*(1), 1–24. doi: 10.1093/ijl/ect031

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2002)* (pp. 1934–1940). Spain: Las Palmas, Canary Islands.

Considine, J. (2014). Academy Dictionaries 1600-1800. Cambridge, New York: Cambridge University Press.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The Encyclopedia of Language and Linguistics* (pp. 3168-3171). Oxford, UK: Pergamon.

Cowie, A. P. (Ed.). (1998). *Theory, Analysis, and Applications*. Oxford: OUP.

Fellbaum, C. (2016). Treatment of Multi-Word Units. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 411–424). Oxford: Oxford University Press.

Fontenelle, T. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Tübingen: Niemeyer.

Gantar, P., Colman, L., Parra Escartín, C., & Martínez Alonso, H. (2018). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography, 32*(2), 138–162. doi: 10.1093/ijl/ecy012

Hausmann, F. J. (1979). Un Dictionnaire des Collocations Est-Il Possible? *Travaux de Linguistique et de Littérature, 17*(1), 187–195.

ISO 24613-1 (2019). Language Resource Management — Lexical Markup Framework (LMF) — Part 1: Core Model. Genève: Organisation Internationale de Normalisation.

Jónsson, J. H. (2009). Lemmatisation of Multiword Lexical Units: Motivation and Benefits. In H. Bergenholtz, S. Nielsen & S. Tarp (Eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (pp. 165–194). Bern: Peter Lang AG.

Kinable, D. (2015). Reflections on the Concept of a Scholarly Dictionary. *Kernerman Dictionary News*, *23*, 11–2.

Lorentzen, H. (1996). Lemmatization of Multi-word Lexical Units: In Which Entry? In M. Gellerstram et al. (Eds.), *Proceedings of the 7th EURALEX*

International Congress on Lexicography: Part I (pp. 415–421). Goteborg, Sweden: Goteborg University Department of Swedish.

McCrae, J. P., Tiberius, C., Khan, F., Kernerman, A., Declerck, T., Krek, S., Monachini, M., & Ahmadi, S. (2019). The ELEXIS interface for interoperable lexical resources. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference* (pp. 417–433). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_37.pdf

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., & Polguère, A. (1984–1999). Dictionnaire Explicatif et Combinatoire du Français Contemporain. *Recherches lexico-sémantiques, IV*. Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A. P. Cowie (Ed.), *Phraseology, Theory, Analysis, and Applications* (pp. 23–54). Oxford: Oxford University Press.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Romary, L., & Tasovac, T. (2018). TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *Proceedings of the 8th Conference of Japanese Association for Digital Humanities* (pp. 274–275). Retrieved from https://tei2018.dhii.asia/AbstractsBook_TEI_0907.pdf

Sailer, M., & Markantonatou, S. (2018). *Multiword expressions: Insights from a multilingual perspective (Phraseology and Multiword Expressions): Vol. 1*. Berlin: Language Science Press. doi: 10.5281/zenodo.1182583

Salgado, A., Costa, R., Tasovac, T., & Simões, A. (2019a). Improving the Consistency of Usage Labelling in Dictionaries with TEI Lex-0. *Lexicography: Journal of ASIALEX 6*(2), 133–156. doi: 10.1007/s40607-019-00061-x

Salgado, A., Costa, R., & Tasovac, T. (2019b). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the*

eLex 2019 Conference, 1–3 October, 2019, Sintra, Portugal (pp. 417–433). Brno: Lexical Computing CZ, s.r.o. Retrieved from https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_23.pdf

Simões, A., Almeida, J. J., & Salgado, A. (2016). Building a Dictionary using XML Technology. In *Open Access Series in Informatics (OASIcs). 5th Symposium on Languages, Applications and Technologies (SLATE'16): Vol. 51* (pp. 14:1–14:8). Germany, Dagstuhl: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary Making*. Cambridge: Cambridge University Press.

Tasovac, T., & Petrović, S. (2015). Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. 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. 384–396). Ljubljana/Brighton: Institute for Applied Slovene Studies and Lexical Computing Ltd. Retrieved from https://elex.link/elex2015/proceedings/eLex_2015_25_Tasovac+Petrovic.pdf

Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia; The Hague/Paris: Mouton.

KODIRANJE VEČBESEDNIH LEKSIKALNIH ENOT S TEI LEX-O: ŠTUDIJA PRIMERA

Modeliranje in kodiranje večbesednih leksikalnih enot oz. pogostih nizov leksemov, ki jih obravnavamo kot samostojne leksikalne enote, je tematika, ki v smernicah Text Encoding Initiative (TEI) ni ustrezno in dovolj poglobljeno predstavljena, čeprav je TEI v raziskovalni skupnosti de facto standard pri delu z elektronskimi besedili. V prispevku na primeru Slovarja Portugalske akademije znanosti predstavimo nekatere rešitve pri kodiranju večbesednih leksikalnih enot v formatu TEI Lex-0, iniciative, katere namen je poenostaviti in racionalizirati kodiranje leksikalnih podatkov s TEI in posledično izboljšati interoperabilnost. Vpeljemo pojem makro- in mikrostrukturne relevantnosti z namenom razločevati med večbesednimi leksikalnimi enotami, ki so samostojne slovarske iztočnice, in tistimi, ki se nahajajo v geslih enobesednih iztočnic. Vpeljemo tudi pojem leksikografske transparentnosti za razlikovanje med enotami, ki nimajo razlage, in tistimi, ki jo imajo; prve so kodirane v okviru elementa <form>, slednje pa v okviru elementa <entry> in lahko vsebujejo nadaljnje omejitve (številke pomenov, področne oznake, slovnične oznake ipd.). V elementu <gram> vpeljemo uporabo atributov za kodiranje različnih tipov oznak za večbesedne leksikalne enote (implicitne, eksplicitne in normirane). Prispevek zaključimo s sklepom, da bi se interoperabilnost leksikalnih virov močno izboljšala, če bi avtorji slovarskih shem imeli dostop do bogate, a relativno enostavne tipologije večbesednih leksikalnih enot.

Ključne besede: TEI, leksikografija, jezikovni viri, večbesedne leksikalne enote, interoperabilnost

SIZE OF CORPORA AND COLLOCATIONS: THE CASE OF RUSSIAN

Maria KHOKHLOVA
St Petersburg State University

Vladimir BENKO
Slovak Academy of Sciences

Khokhlova, M., Benko, V. (2020): Size of corpora and collocations: the case of Russian. Slovenščina 2.0, 8(2): 58–77

DOI: https://doi.org/10.4312/slo2.0.2020.2.58-77

With the arrival of information technologies to linguistics, compiling a large corpus of data, and of web texts in particular, has now become a mere technical matter. These new opportunities have revived the question of corpus volume that can be formulated in the following way: are larger corpora better for linguistic research or, more precisely, do lexicographers need to analyze bigger amounts of collocations? The paper deals with experiments on collocation identification in low-frequency lexis using corpora of different volumes (1 million, 10 million, 100 million and 1.2 billion words). We have selected low-frequency adjectives, nouns and verbs in the Russian Frequency Dictionary and tested the following hypotheses: 1) collocations in low-frequency lexis are better represented by larger corpora; 2) frequent collocations presented in dictionaries have low occurrences in small corpora; 3) statistical measures for collocation extraction behave differently in corpora of different volumes. The results prove the fact that corpora of under 100 M are not representative enough to study collocations, especially those with nouns and verbs. MI and Dice tend to extract less reliable collocations as the corpus volume extends, whereas t-score and Fisher's exact test demonstrate better results for larger corpora.

Keywords: collocations, Russian corpora, corpus size, corpus linguistics, statistical measures

1 INTRODUCTION

Over the past 10 years, corpora have dramatically increased in size, giving lexicographers much more data than ever before. At the same time, however, this has brought up the question whether we really need those amounts of texts or we can be satisfied with less. The issue is not that simple: corpora, on the one hand, are expected to attest such units by generating a sufficient number of examples; on the other hand, lexicographers and language users should not be overloaded with large bulks of examples.

The size of corpora is also relevant when applied to the task of describing collocability. Is there any correlation between the size of the corpus and the extracted collocations? Can we find more collocations in larger corpora?

We would like to answer the following question: What would be the benefit of using larger corpora? In our study, we analyze the behaviour of Russian collocations using corpora of different volumes. The aim of the paper is threefold. First, to conduct a case study of low-frequency lexemes and analyze their collocations. Secondly, to investigate a number of frequent collocations presented in several dictionaries. Thirdly, to apply statistical measures to collocation extraction from corpora and to interpret possible interrelation between the results and volume.

2 BACKGROUND

The issue of data volume is of importance. For a long time, the amount of data was objectively limited by technical capacities. The Brown corpus comprised 1 million words, the British National Corpus (BNC) amounted to 100 million words, the Russian National Corpus (RNC) has more than 600 million words. The volumes of newly compiled Giga-word corpora can exceed dozens of billions of words.

Linguists understand volume as a concept in different ways. Earlier, a compilation of frequency dictionaries was associated with the question of what amount of data would suffice to describe most frequent lexical units in a language. This question is also relevant in the context of sample reliability or in the context of (foreign) language learning, i.e. what is the minimal amount of lexical units – and, hence, the minimal corpus volume – that students should memorize to learn a language.

Speaking about corpora as samples from larger populations we can mention that the Russian frequency dictionary by Steinfeld (1963) required a 400-thousand-word sample, whereas dictionaries compiled by Zasorina (1977) and Lenngren (1993) are based on a 1 million-word sample; the new dictionary by Lyashevkaya and Sharoff (2009) features a sample of approximately 100 million words. It should be noted that Piotrowski et al. (1977) showed that 1600–1700 most frequent words can be reliably described using a sample of 400 thousand words.

Different works discuss the question of how large a corpus should be. This question is especially crucial in the studies of rare words and word combinations. Sinclair (2005) rightly points out that the occurrences of two or more words are far less frequent than ones of a single word. There are not too many works dealing with the ideal volume of texts required to search collocations. Brysbaert and New (2009) discuss the sufficient corpus volume depending on word frequency distinguishing between high- and low-frequency lexis. Piperski (2015) performs a case study of the same words in two corpora of different sizes, namely the main subcorpus from RNC (230 million words) and ruTenTen (14.5 billion words). The author claims that corpora cannot provide evidence for non-existence of collocations but they can be used to prove their existence. And in this case, even a single example in a corpus is enough.

Finding suitable collocation candidates is quite popular in linguistic research and statistical association measures are widely used for this task. They have their practical application to collocation selection and identification adopted in corpus tools. The dependency between the behaviour of association measures and corpus size was the main focus of a number of research studies. Daudaravičius (2008, p. 650) mentions that “the values of MI grow together with the size of a corpus, while the Dice score is not sensitive to the corpus size and score values are always between 0 and 1”. Rychly (2008) proposes logDice as the measure that is not affected by the size of the corpus and takes into account only frequency of a node and of a collocate. It can be used for collocation extraction from large corpora and is successfully implemented in Sketch Engine (Kilgarriff et al., 2014). Also relevant is the study by Evert et al. (2017) who evaluated not only association measures but also various corpora, co-occurrence contexts and frequency thresholds applied to automatic

collocation extraction and thus tuning statistical methods. The results show that sufficiently large Web corpora (exceeding 10 billion words) perform similarly or even better than the carefully sampled BNC.

Taking these findings into account, a new question is to be considered: how do corpora of different sizes represent multi-word expressions or collocations? In our paper, we analyze quantitative properties of collocations that were found in corpora of different sizes and present some findings on low-frequency collocations.

3 METHODOLOGY

Our previous experiments showed that high-frequency nouns (Khokhlova, 2017) and their ranking positions in both 1-billion-token and 14 billion-token subsets produced the same results, but this was different for low-frequency nouns. For low-frequency data, three corpora did not show much coincidence with ranking shown in the Russian frequency dictionary by Lyashevskaya and Sharoff (2009). Hence, this issue requires a more detailed investigation.

In this study, we use a collection of Russian corpus data developed within the framework of the Aranea Project (Benko, 2014). We randomly sampled the largest Araneum Russicum Maximum corpus to obtain three smaller subcorpora of total 1 million words (1 M hereafter), 10 million words (10 M hereafter), and 100 million words (100 M hereafter) respectively. The sampling procedure was document-based and worked on sets of 1,000 documents. Out of each set, the first 1,000-n documents were obtained, and the 1,000-n ones were deleted. This approach allowed to preserve all document metadata in the sampled corpus. Although the procedure is not strictly random, it proved to be sufficient for large corpora without extra sophisticated randomization required.

The aim of our experiments was to test the following hypotheses:

1. Low-frequency lexis and its collocations are better represented in large corpora (exceeding 100 million words);
2. Frequent collocations presented in dictionaries have low occurrences in small corpora;
3. Certain statistical measures perform better on small corpora, whereas others require larger corpora.

It can be somewhat problematic to find data about low-frequency lexis or at least to understand what kind of collocations belong to the low-frequency group. Authors of the Macmillan English Dictionary for Advanced Learners (2002) make a clear distinction between high-frequency core vocabulary and less common words using different fonts and the star symbol.

Russian dictionaries, on the other hand, do not provide such information. Thus, frequency dictionaries are the only ones that can provide quantitative data for individual words (but not collocations). The dictionary by Lyashevskaya and Sharoff (2009) provides data for 20,000 lemmata. In the first part of our experiment, we selected lexical items from the end of the list that can produce collocations. Those were ranked between position 19,687 to 20,004 and had the same frequency, i.e. 2.6 instances per million (ipm). Nouns and adjectives were the most representative groups, but verbs and adverbs were also analyzed.

When developing a gold standard for Russian collocability (Khokhlova, 2018a), we produced a list of collocations presented in different Russian dictionaries and introduced a notion of dictionary index, i.e. the number of dictionaries that include a given collocation. The higher the dictionary index, the more frequent and widely used the collocation is. Less frequent collocations have lower dictionary index scores. In the first experiment of our study, we evaluate corpora with those collocations that have minimal dictionary index score.

Along with studying the behavior of low-frequency lexemes and their collocations, we conducted a case study of frequent collocations from the gold standard, i.e. the ones that showed the highest dictionary index scores. For this task we selected 20 collocations which were described in four different Russian dictionaries (explanatory and specialized ones, for example, for language learners).

In the last phase of our experiment, we extracted *adjective+noun* collocations (based on the morphosyntactic annotation by TreeTagger (Schmid, 1994) from each of the above mentioned subcorpora using four association measures (t-score, MI, Dice coefficient and Fisher's exact test) (Evert, 2004; Pecina, 2009) and compared top 500 candidates. These measures were chosen as they are based on different statistical principles and have demonstrated efficiency in prior experiments (Khokhlova, 2018b). Having applied the

frequency threshold (at least 3), we extracted bigrams[1] from three subcorpora. Here are some examples: *Rossiyskaya Federatsiya*[2] 'Russian Federation', *elektronnaya pochta* 'e-mail', *vannaya komnata* 'bathroom', *rabochiy stol* 'work table', *evropeyskaya strana* 'European country' etc. Collocations that were used for evaluation are largely based on the gold standard and insufficient; therefore, we had to rely on linguistic assessment as well.

Then, we analysed the top 500 candidates. Altogether, we extracted the following number of bigrams:

- 1 M: 9,862;
- 10 M: 51,745;
- 100 M: 368,055.

There were no dictionaries of Russian collocations that would be large enough in volume and, thus, information on collocational restrictions (that can be used for data evaluation) had to be obtained from other types of dictionaries and resources.

4 RESULTS

4.1 Results for low-frequency collocations

For our case study we selected 25 adjectives, 8 nouns, 10 verbs and 8 adverbs and thus investigated the following lexical items: adjectives *bezotkaznyy* 'failproof, unfailing', *daveshniy* 'recent', *kinetisheskiy* 'kinetic', *neprerekayemyy* 'incontestable', *priglushennyy* 'muted', *slovarnyy* 'lexicographic', *neproglyadnyy* 'impenetrable', *okkupatsionnyy* 'occupational', *opryatnyy* 'neat', *pogrebal'nyy* 'funeral', *rassuditel'nyy* 'sober', *tyagovyy* 'tractive', *bezdumnyy* 'thoughtless', *vitoy* 'twisted', *neproshenyy* 'undesired', *nerazlichimyy* 'indiscernible', *bessrochnyy* 'perpetual', *mezhlichnostnyy* 'interpersonal', *orkestrovyy* 'orchestric', *zazhitochnyy* 'prosperous', *neprelozhnyy* 'inviolable', *obsharpannyy* 'shabby', *smertonosnyj* 'pestilent', *kishechnyj* 'intestinal', *tseleustremlennyy* 'purposeful'; nouns *inkvizitsiya* 'inquisition', *rassloyeniye*

1 The term "bigram" denotes combinations of two adjacent words.

2 Henceforth, the examples originally written in Cyrillic are given in Latin transliteration.

'stratification', *eroziya* 'erosion', *podlodka* 'submarine', *pischevareniye* 'digestion', *sedmitsa* 'week', *ontologiya* 'ontology", *kholuy* 'toady'; verbs *vydelyvat'* 'to curry', *zavyvat'* 'to wail', *pronzat'* 'to pierce, to impale', *teshit'* 'to amuse, to please', *vlepit'* 'to slap', *pokolebat'* 'to shake', *zayedat'* 'to eat', *poloskat'* 'to rinse, to gargle', *ostudit'* 'to cool', *privivat'* 'to implant, to instil'.

We scrutinized and evaluated the concordance output against the gold standard.

Table 1 represents the results of the analysis for collocations with low-frequency adjectives. The first column lists the lemmata, other columns give the number[3] of concordance lines in total (in the 1 M, 10 M and 100 M corpora) and with appropriate nouns (marked as collocations) for the 1 M, 10 M and 100 M corpora respectively. We considered as appropriate those lexical combinations that are recurrent in the written language. Thus, out of 20 concordance lines of output, all 20 may turn out to contain interesting word form collocates.

Table 1: *Results for low-frequency adjectives*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
bessrochnyy	2	2	24	23	249	248
bezdumnyy	0	0	18	8	51	32
bezotkaznyy	2	1	15	13	132	120
daveshniy	0	0	0	0	21	14
kineticheskiy	10	10	25	23	180	178
kishechnyy	11	11	101	95	210	208
mezhlichnostnyy	0	0	34	34	148	148
neprelozhnyy	0	0	9	9	82	78
neprerekayemyy	0	0	5	4	34	34
neproglyadnyy	0	0	5	5	33	32
neproshenyy	2	2	7	7	26	20
nerazlichimyy	0	0	1	0	41	11
obsharpannyy	0	0	1	1	35	35
okkupatsionnyy	0	0	7	7	92	88

3 Here and in the following tables we mean instances (i.e. absolute frequencies) in columns with numbers.

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
opryatnyy	0	0	12	6	130	88
orkestrovyy	0	0	4	4	69	69
pogrebal'nyy	2	2	17	17	149	149
priglushennyy	1	1	7	7	239	187
rassuditel'nyy	0	0	6	1	84	31
slovarnyy	1	1	47	47	447	441
smertonosnyj	3	3	18	18	114	104
tseleustremlennyy	4	2	48	23	221	133
tyagovyy	0	0	2	2	205	203
vitoy	3	3	14	14	156	147
zazhitochnyy	3	3	18	18	133	116

One can observe that despite the same low-frequencies found in the dictionary by Lyashevskaya and Sharoff (2009), lexical items show a significantly different behaviour, i.e. their frequencies vary as well as the number of collocates. The analysis suggests that a 1 M corpus is evidently not enough to produce a sufficient number of examples illustrating low-frequency collocations. More than 50% of adjectives were missing in the given sample. In the 1 M corpus only two lexical items (*kineticheskiy* 'kinetic' and *kishechnyy* 'intestinal') produced 10 and 11 collocations respectively (ranging from 1 up to 3 instances) that can be accounted for their narrow semantic meaning and hence restricted collocability (e.g. *kishechnaya infektsiya* 'enteric infection', *kishechnaya muskulatura* 'intestinal muscles', *kineticheskaya energiya* 'kinetic energy'). More extensive corpora would likely yield larger numbers of relevant examples.

More than a half of concordance lines in the 10 M and 100 M corpora can be seen as a source of collocations without any filtration (e.g. *priglushennyy, slovarnyy, neproglyadnyy* etc). This fact can suggest that in case of low-frequency lexis the increase of texts does not necessarily result in overflow with data and false examples.

Among irrelevant candidates one can find also other instances, i.e. errors in lemmatization (e.g. *vitoj* 'twisted' in *dolche vitoj* 'dolce vita' was lemmatized

as *vitoj* 'twisted' instead of Latin *vita* 'vita'), erroneous part-of-speech tagging (e.g. adjectives instead of adjectival nouns), mistakes and typos.

The findings of the case study for a number of adjectives are reported next.

Priglushennyy 'muted': in the 1 M corpus we found only one rare occurrence *priglushennoye urchaniye* 'muted growl'. The 10 M and 100 M corpora contained collocates representing one lexical group of colour, e.g. *tsvet* 'colour', *gamma* 'colour scheme', *ottenok* 'tint', *pigment* 'pigment', *terrakotovyy* 'terracota' and *zelenyy* 'green'. There were also examples with *golos* 'voice', *shum* 'noise', *zvon* 'toll of the bell'.

Orkestrovyy 'orchestric': only two collocations occured in the 10 M corpus, namely *orkestrovaya jama* 'orchestra pit' and *orkestrovaya partitura* 'orchestra score'. The 100 M corpus gave a wide range of collocates with the sememe 'music', e.g. *aranzhirovka* 'arrangement', *partiya* 'play', *rakovina* 'shell', *syuita* 'suite'.

The evidence suggests that the results obtained for the 1 M corpus include collocates that belong to lexical periphery – not the frequent ones. This is somewhat unexpected, hence the most frequent collocates tend to be found only in larger corpora.

Table 2 shows the results for low-frequency nouns.

Table 2: *Results for low-frequency nouns*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
eroziya	4	2	109	75	484	421
inkvizitsiya	2	1	29	14	134	64
kholuy	0	0	0	0	11	5
ontologiya	2	0	35	20	65	36
pischevareniye	6	6	126	108	1,044	725
podlodka	1	1	18	11	117	51
rassloyeniye	2	2	29	22	239	211
sedmitsa	4	4	11	8	109	100

Rassloyeniye 'stratification': there are only two occurrences in the 1 M corpus, a term *rassloyeniye vina* 'wine stratification' and *sotsial'noye*

rassloyeniye 'social differentiation'. The former has a highly specific and narrow meaning while the latter can be called a collocation. In the 10 M corpus one can find other meaningful examples, e.g. *rassloyeniye strany* 'stratification of country' or *obschestva* 'of society', *rassloyeniye nogtey* 'nail splitting' or *komponentov* 'segregation of components'.

Podlodka 'submarine': the most frequent collocate turns to be *atomnyy* 'atomic' that can be found both in the 1 M and 10 M corpora. The 10 M corpus also contains two verbal collocates, e.g. *zatonut'* 'to founder' and *topit'* 'to sink'. The 100 M corpus gives more examples, e.g. *prishvartovat'* 'to moor', *unichtozhit' 'to destroy'*, *stoyat'* 'to stay, *idti 'to go'*, *chodit' 'to go'*.

Pischevareniye **'**digestion**'**: the given noun is the only one showing wide collocability, i.e., we find collocates among adjectives, nouns and verbs. Compared to other nouns it has the highest frequency.

Sedmitsa 'week': The 1 M corpus shows only adjective collocates, e.g. *Svetlyy* 'Easter' and *Strastnoy* 'Holy'. The 10 M corpus does not add any valuable collocations with adjectives, except for one occurrence of *syrnaya sedmitsa* 'shrovetide'. The 100 M corpus includes only one example of noun collocate *sedmitsa mytarya i fariseya* 'the week of the Publican and the Pharisee'.

Kholuy 'toady': among all the nouns, it proved to have the lowest frequency; no occurrence was found in the 1 M and 10 M corpora.

It is also true for nouns (as it was the case for adjectives) that although we see the same low-frequency according to the frequency dictionary (Lyashevkaya and Sharoff, 2009), the number of examples and hence collocations is different. The noun *pischevareniye*, for example, shows more than 1,000 occurrences.

We can see that small corpora produce even fewer collocates for nouns than for adjectives. There are virtually no collocations with verbs, whereas those with nouns and adjectives prevail.

Table 3 presents the results for low-frequency verbs and their collocations.

Table 3: *Results for low-frequency verbs*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100M	100M (collocations)
ostudit'	3	3	21	9	208	156
pokolebat'	2	2	10	9	68	46
poloskat'	1	1	22	21	170	123
privivat'	4	4	28	28	260	209
pronzat'	1	1	4	3	47	42
teshit'	0	0	9	6	76	63
vlepit'	0	0	2	0	16	8
vydelyvat'	0	0	3	3	41	37
zavyvat'	0	0	3	2	25	19
zayedat'	0	0	9	6	103	79

Despite the fact that the verbs selected for the experiment are polysemous and should therefore demonstrate wide collocational preferences, they tend to get the lowest number of collocations in smaller corpora, as opposed to nouns and adjectives. Both the 1 M and 10 M corpora do not yield a sufficient number of examples.

Although the frequency of the verbs is the same (2.6 ipm) in the dictionary (Lyashevkaya and Sharoff, 2009), it varies widely in corpora, e.g. from 0.16 up to 2.25 ipm.

Vydelyvat' 'to curry': only the 100 M corpus shows collocability of verbs with nouns.

Zavyvat' 'to wail': in the 10 M corpus there are two examples of a subject collocating with a verb, e.g. *v'yuga* 'snowstorm' and *veter* 'wind'.

The average percentage of the data filtering for nouns and verbs is higher than for adjectives, i.e. the output results show irrelevant occurrences, mistakes, typos, other noise or word usage without any collocates. Adjectives tend to be part of noun groups (not always, though), whereas nouns and verbs can be used more often as independent lexical units. Therefore, corpora exceeding 100 M are more efficient in representing collocability of low-frequency nouns and verbs.

Having come to a preliminary conclusion that there is a need to further expand the volume of corpora, we also studied a number of syntactic relations[4] based on 100 M and 1.2 G corpora. We looked at the neighborhood of low-frequency nouns and analyzed the output by filtering out typos, errors in lemmatization etc. in order to count lemmata examples only. Table 4 represents the number of attributive and verbal collocations.

Table 4: *Number of different collocations for nouns*

	adjective + noun (100 M)		adjective + noun (1.2 G)		verb + noun, noun + verb (100 M)		verb + noun, noun + verb (1.2 G)	
	all forms	lemmata	all forms	lemmata	all forms	lemmata	all forms	lemmata
eroziya	77	31	1,328	78	106	46	2,919	79
inkvizitsiya	26	16	564	72	26	19	1,225	87
kholuy	6	6	13	10	0	0	9	3
ontologiya	20	16	246	43	9	4	298	22
pischevareniye	53	19	1,784	73	266	41	6,945	57
podlodka	32	18	582	62	30	18	964	81
rassloyenoye	72	30	743	66	64	33	1,230	82
sedmitsa	64	12	688	22	11	8	501	55

With the expansion of corpus volume, the number of collocations increases as well as the amount of noise or irrelevant cases. Additional data filtering is therefore needed. When the corpus volume increases by 10 times, the number of concordance lines per collocation also increases by at least 10 times (strictly speaking, on average, 18 times for the nouns under consideration).

To be more specific, preliminary results of our study have shown that higher absolute frequency of a particular lexical item does not always mean a larger number of syntactic relations for the lexical item (despite the greater number of collocates typical of each relation).

4.2 Results for frequent collocations from dictionaries

The dictionary index (Khokhlova, 2018a) designates the number of dictionaries which present the given collocation. Large values of the index imply

4 The analysis was made on the Russian word sketch grammar in Sketch Engine (Khokhlova, 2010; Kilgarriff et al., 2014).

that the collocation is reproduced quite often and thus should be learnt by heart (if we speak about the learners of Russian). Theoretically, the maximum is equal to the number of dictionaries, that is 6 for the *adjective + noun* model, but in practice the maximum number of dictionaries in which the collocation was fixed was 4. The gold standard comprises more than 15,000 collocations for the given model and only 61 examples were described in 4 dictionaries (so there is no example to be recorded in all 6 dictionaries). We randomly selected 20 frequent collocations from this list and analyzed them across the corpora. Table 5 presents the results sorted by the number of occurrences in the 100 M corpus.

Table 5: *Frequency distribution of selected collocations from the gold standard*

		1 M	10 M	100 M
yarkiy primer	'vivid example'	3	65	533
vysokiy rezul'tat	'high result'	1	43	532
bol'shoy uspekh	'big success'	6	50	357
grubaya oshibka	'great error'	1	8	125
vysokaya pribyl'	'high profit'	0	15	79
glubokaya blagodarnost'	'deep gratitude'	0	3	68
polnaya tishina	'complete silence'	1	11	62
polnaya pobeda	'complete victory'	1	12	55
bogatyy urozhay	'bountiful harvest'	0	9	50
glubokiy krizis	'deep crisis'	0	5	44
glubokoye udovletvoreniye	'deep satisfaction'	0	1	31
shirokiy razmakh	'wide scope'	0	0	24
ostraya bor'ba	'fierce struggle'	0	1	21
general'noye srazheniye	'decisive battle'	0	1	15
goryachaya lyubov'	'hot love'	0	4	14
zheleznaya distsiplina	'iron discipline'	1	3	10
gomericheskiy khokhot	'homeric laughter'	0	1	8
zhguchiy vopros	'burning question'	0	0	6
shirokoye sotrudnichetsvo	'wide cooperation'	0	0	2
zheleznyy kharakter	'strong character'	0	0	2

Even in the case of frequent collocations from the gold standard the 1 M corpus yields no results and hence cannot be used as a source of linguistic

evidence. The 10 M corpus also contains a small number of collocations. The collocation frequencies are significantly higher in the 100 M corpus and this can be accounted for by high frequencies of either the node or the collocate.

4.3 Results of automatic extraction

In the course of further experiments we used statistical measures to extract bigrams setting frequency cutoff threshold of f=3 and then the bigrams were evaluated bigrams against the dictionary data, and by native-speaker inspection. The analysis also revealed a large amount of morphological mistakes and errors in lemmatization. For example, *zloy dukhi* 'evil perfume' instead of *zloy dukh* 'evil spirit'; *pal'movom masle* 'palm oil' (the lemma for the adjective stands in the prepositional case) instead of *pal'movoye maslo*.

Table 6 presents the number of collocations extracted by each of the association measures from the 1 M, 10 M and 100 M subcorpora respectively.

Table 6: *Number of collocations per subcorpus*

	1 M	10 M	100 M
MI	229	97	54
t-score	484	492	495
Dice	301	186	114
Fisher	454	490	499

The analysis suggests that MI and Dice tend to extract fewer collocations from a larger corpora, retrieving examples with typos and mistakes. This can lead us to the hypothesis that vast collections of text data will have more non-collocations (for example, free phrases) and, thus, top lists will also contain such senseless word combinations (or even hapax legomena, if there is no frequency threshold). Dice coefficient also focuses predominantly on terms, proper names and set phrases, e.g. *nashatyrny spirt* 'liquid ammonia', *gadkiy utenok* 'ugly duckling'. Compared to other measures, Fisher's exact test extracted the largest number of collocations.

Table 7 shows numbers of shared bigrams found by each measure in different corpora.

Table 7: *Numbers of shared bigrams (by subsets)*

	1 M/10 M	10 M/100 M	1 M/100 M
MI	38	31	1
t-score	275	427	262
Dice	96	63	13
Fisher	241	424	233

When we compare lists extracted by different measures, we can see that MI and Dice do not tend to extract the same collocations in the corpora of different volumes. The percentage of the intersection declines with the increase of difference between corpus volumes, resulting in a smaller amount of bigrams. T-score and Fisher's exact test demonstrate contrasting behaviour, i.e. the highest number of the identical bigrams is extracted from the 10 M and 100 M corpora while the 1 M/10 M and 1 M/100 M pairs show almost the same number.

Table 8 demonstrates the number of the same bigrams found in the 1 M, 10 M and 100 M corpora, respectively. Here the results suggest that the measures can be again divided into two groups according to the behaviour, namely, the first group contains MI and Dice, whereas in the second are t-score and Fisher's exact test.

Table 8: *Number of the shared bigrams (breakdown by measures)*

	1 M				10 M				100 M			
	MI	t-score	Dice	Fisher	MI	t-score	Dice	Fisher	MI	t-score	Dice	Fisher
MI	500	8	350	32	500	0	347	0	500	0	366	0
t-score		500	80	385		500	46	393		500	4	396
Dice			500	134			500	71			500	8
Fisher				500				500				500

Tables 9 to 11 show the number of the identical bigrams that were found in the 1 M, 10 M, and 100 M corpora, respectively, by measures. The comparison was made between corpora of different sizes. Measures from the above mentioned two groups show lower numbers of identical bigrams with the increase of corpus size.

Table 9: *Number of identical bigrams (1 M vs 10 M by measures)*

	MI (1 M)	t-score (1 M)	Dice (1 M)	Fisher (1 M)
MI (10 M)	38	0	35	4
t-score (10 M)	6	275	43	222
Dice (10 M)	62	35	96	57
Fisher (10 M)	15	248	62	241

Table 10: *Number of identical bigrams (1 M vs 100 M by measures)*

	MI (1 M)	t-score (1 M)	Dice (1 M)	Fisher (1 M)
MI (100 M)	1	0	1	0
t-score (100 M)	2	262	33	211
Dice (100 M)	11	2	13	6
Fisher (100 M)	25	241	57	233

Table 11: *Number of identical bigrams (10 M vs 100 M by measures)*

	MI (10 M)	t-score (10 M)	Dice (10 M)	Fisher (10 M)
MI (100 M)	31	0	31	0
t-score (100 M)	0	427	38	370
Dice (100 M)	54	5	63	8
Fisher (100 M)	0	375	60	424

5 CONCLUSION AND FURTHER WORK

Though final conclusions might be too early to formulate, we can say that larger corpora do not always have an advantage, especially in situations when most frequent phenomena are studied. Depending on the mode of analysis, larger amounts of data may even turn into an obstacle, especially if the research has to observe time limits. Nevertheless, the results for low-frequency lexis prove the fact that corpora of less than 100 million words are not sufficient to represent collocations. In terms of our study, this can be partly accounted for by rich flectional nature of Russian morphology and a relatively free word order.

We should mention that frequent collocations which are described in several

dictionaries cannot be found in smaller corpora. The results suggest that in order to properly represent these collocations in dictionaries, one needs corpora exceeding 100 million words.

The results are largely based and depend on the quality of data, which raises again the question of how to prepare a corpus, especially to study low-frequency phenomena. The evidence obtained for infrequent lexis can differ for other text types or domains and, thus, metatextual annotation can be taken into account in further experiments.

From the perspective of various association measures used to identify collocations, we have shown that not all of them work well for larger corpora. Our observation can be summarized as follows:

- MI and Dice extract more terms, typos, hapax legomena, errors in lemmatization with the increase of volume, and thus perform better on smaller corpora;
- t-score and Fisher's exact test extract more good collocations from larger corpora.

We believe that the relationship between the corpus size, and the number and "quality" of extracted collocations is a fascinating topic to study; a similar research should be performed on different corpora and/or languages as well.

Acknowledgments

This work was supported by the grant of the Russian Science Foundation (Project No. 19-78-00091).

REFERENCES

Dictionaries, corpuses and digital resources

Lyashevskaya, O., & Sharoff, S. (2009). *The Frequency Dictionary of Modern Russian based on the Russian National Corpus data* [Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo Korpusa Russkogo Yazyka)]. Moscow: Azbukovnik.

Macmillan English Dictionary for Advanced Learners. (2002). Macmillan Education.

Steinfeld, E. (1963). *Frequency dictionary of the Contemporary Russian language* [Chastotnyy slovar' sovremennogo russkogo literaturnogo yazyka]. Tallin.

The British National Corpus, (Version 3) (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/ (1. 5. 2020)

The Russian National Corpus [Natsional'nyy korpus russkogo yazyka]. Retrieved from http://www.ruscorpora.ru (1. 5. 2020)

The Brown Corpus. Retrieved from http://korpus.uib.no/icame/manuals/brown/index.htm, https://www.sketchengine.eu/brown-corpus/ (1. 5. 2020)

Zasorina, L. (1977). *Frequency dictionary of the Russian language* [Chastotnyy slovar' russkogo yazyka]. Moscow: Russkiy yazyk.

Other

Benko, V. (2014). Aranea Yet Another Family of (Comparable) Web Corpora. *Text, Speech and Dialogue. Proceedings of the 17th International Conference, TSD 2014, 8–12 September, 2014, Brno, Czech Republic.* LNCS 8655 (pp. 257–264). Springer International Publishing Switzerland.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Daudaravičius, V. (2010). The influence of collocation segmentation and top 10 items to keyword assignment performance. *Computational Linguistics and Intelligent Text Processing*. *Proceedings* of the *11th International Conference, CICLing 2010, 21–27 March, 2010, Iasi, Romania* (pp. 648–660). Berlin: Springer.

Evert, S. (2004). The Statistics of Word Cooccurrences Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Available at http//purl.org/stefan.evert/PUB/Evert2004phd.pdf (20. 2. 2020)

Evert, S., Uhrig P., Bartsch S., & Proisl, T. (2017). E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference, 19–21 September, 2017, Leiden Netherlands* (pp. 531–549). Leiden: Lexical Computing.

Khokhlova, M. (2010). Building Russian Word Sketches as Models of Phrases. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress, 6–10 July, 2010, Leeuwarden* (pp. 364–371). Ljouwert: Fryske Akademy – Afûk.

Khokhlova, M. (2017). Big data and word frequency: Measuring the consistency of Russian corpora. *Quantitative Approaches to the Russian Language* (pp. 30–48). Routledge, Taylor & Francis.

Khokhlova, M. (2018a). Building a Gold Standard for a Russian Collocations Database. In J. Čibej et al. (Eds.), *Lexicography in Global Contexts. Proceedings of the XVIII EURALEX International Congress* (pp. 863–869). Ljubljana: Ljubljana University Press, Faculty of Arts.

Khokhlova, M. (2018b). Similarity between the Association Measures a Case Study of Noun Phrases. In *Proceedings of the 12th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018* (pp. 21–27). Brno Tribun EU.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7–36.

Pecina, P. (2009). *Lexical Association Measures. Collocation Extraction*. Prague Institute of Formal and Applied Linguistics.

Piotrowski, R. G., Bektaev, K. B., & Piotrowskaya, A. A. (1977). *Mathematical Linguistics* [Matematicheskaya lingvistika]. Moskva: Vysshaya shkola.

Piperski, A. (2015). To be or not to be: Corpora as Indicators of (Non-)Existence. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, *1*(14), 515–522.

Rychly, P. (2008). A lexicographer-friendly association score. *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008* (pp. 6–9). Brno: Masaryk University.

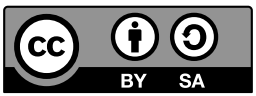
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Sinclair, J. (2005). Corpus and Text — Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1–16). Oxford: Oxbow Books. Retrieved from http://users.ox.ac.uk/~martinw/dlc/chapter1.htm (1. 5. 2020)

VELIKOST KORPUSOV IN OBSEG KOLOKACIJ NA PRIMERU RUŠČINE

Potem ko se je na področju jezikoslovja razmahnila uporaba informacijskih tehnologij, je izdelava obsežnih korpusov, sploh tistih s spletnimi besedili, postala zelo enostavna naloga. Nove priložnosti pa so zopet oživile vprašanja o velikosti korpusa: so večji korpusi boljši za jezikoslovne raziskave, natančneje, ali morajo leksikografi posledično analizirati večje količine kolokacij? Prispevek predstavi eksperimente, v katerih smo iskali kolokacije redkejših besed s pomočjo korpusov različnih velikosti (1 milijon besed, 10 milijonov besed, 100 milijonov besed in 1,2 milijardi besed). Izbrali smo redke pridevnike, samostalnike in glagole iz Ruskega frekvenčnega slovarja in preverili sledeče hipoteze: 1) kolokacije redkejše leksike so bolje zastopane v večjih korpusih; 2) pogoste kolokacije iz slovarjev se redko pojavljajo v manjših korpusih; 3) statistične mere za luščenje kolokacije dajejo različne rezultate pri korpusih različnih velikosti. Rezultati dokazujejo, da korpusi, manjši od 100 milijonov besed, niso dovolj reprezentativni za preučevanje kolokacij, sploh tistih, ki vsebujejo samostalnike in glagole. Statistični meri MI in Dice sta pri luščenju kolokacij manj zanesljivi, sploh pri večjih korpusih, po drugi strani pa t-score in Fisherjev natančni test kažeta boljše rezultate prav pri večjih korpusih.

Ključne besede: kolokacije, ruski korpusi, velikost korpusa, korpusno jezikoslovje, statistične mere



Slovenščina 2.0, 2020 (2)

COLLOCATIONS IN THE CROATIAN WEB DICTIONARY – *MREŽNIK*

Lana HUDEČEK, Milica MIHALJEVIĆ
Institute of Croatian Language and Linguistics

Hudeček, L., Mihaljević, M. (2020): Collocations in the Croatian Web Cictionary – Mrežnik. Slovenščina 2.0, 8(2): 78–111

DOI: https://doi.org/10.4312/slo2.0.2020.2.78-111

The *Croatian Web Dictionary – Mrežnik* project aims to create a free, monolingual, easily searchable, hypertext, born-digital, corpus-based dictionary of the Croatian standard language. Collocations play an important role in *Mrežnik*. At the outset of the *Mrežnik* project, the concept of collocations and their presentation was modelled after the *elexiko* project. However, this concept was modified during the project on the basis of corpus analysis. This paper will outline the presentation of collocations of headwords of different word classes. Some important issues connected with collocations in *Mrežnik* are collocation extraction methods, collocations as a means of differentiating meanings and extracting new meanings, the use of stylistic and terminological labels in collocations, and the relationship of collocations with normative and pragmatic notes, definitions, and subentries.

Keywords: collocations, Croatian, e-dictionary, *Mrežnik*, born-digital dictionary

1 INTRODUCTION

Collocations have received a great deal of attention in recent years. This is not surprising, as they can be considered "the building blocks of language and … fundamental units of language" (Sinclair, 2004, p. 213). They constitute a major challenge for linguists, lexicographers, native speakers, dictionary users, and language learners alike. The challenge for the linguist is how to define them and differentiate them from other multiword expressions.[1]

1 On collocations in Croatian (cf. Blagus Bartolec, 2014); on collocations in Croatian for non-native speakers (cf. Ordulj, 2018).

Collocations are also an important part of a dictionary entry. The *Oxford Collocations Dictionary* defines collocations as "the way words combine in a language to produce natural-sounding speech and writing" (McIntosh, 2018, p. V). A narrower view is that collocations are an unpredictable combination of lexical units, i.e. "a combination that cannot be produced based on the regular syntactic or semantic properties of the units involved" (Granger, 2012, p. 216). For lexicographic purposes, collocations can be defined as "a recurrent combination of words, where one specific lexical item ('the node') has an observable tendency to occur with another (the collocate), with a frequency far greater than chance" (Atkins and Rundell, 2008, p. 223).

Automated procedures for extracting collocations have been developed and are continually being improved. This means that lexicographers can very quickly obtain large quantities of collocational information, which facilitates dictionary compilation; however, this also poses difficulties for the lexicographer, as collocations entail a number of methodological problems and force the lexicographer to take certain decisions.

Collocations also present a challenge for dictionary users, who often "cannot be sure where to find collocations; a universal format, be it with regard to placement or typography, has yet to be realized" (Durkin, 2016, p. 37).[2] Yet, Durkin (ibid.) notes that born-digital dictionaries do not have space restrictions as print dictionaries, which means that collocations can be provided in the entries of both components of the collocaton.[3]

As linguists from the Institute of Croatian Language and Linguistics provide language advice daily, we know that many user questions are connected with multiword expressions. Although native speakers tend to use collocations more intuitively, some language advice also relates to collocations, e.g. in

2 "Cobuild6 and LDOCE5, for example, give collocations a separate status in the microstructure, listing (and, if necessary, explaining) them in a self-contained box (…). Thus, users can locate the data immediately without looking through the entire entry" (Durkin, 2016, p. 37). Cobild6 is the sixth edition of the Cobuild dictionary and LDOCE5 is the fifth edition of the *Longman Dictionary of Contemporary English*.

3 However, in born-digital dictionaries, there is still the risk of information death, wherein the user is overwhelmed by the abundance of information on the screen; thus, the choice of what to include and how to present it in the dictionary interface still remains.

the administrative style. This was the reason a special project on Croatian collocations was launched – the *Croatian Collocation Database* (CCD).[4] Special attention was paid to collocations in the *Mrežnik* project, which is the focus of this paper, for the same reason. Some entries in *Mrežnik* are linked to collocations in the *Croatian Collocation Database* (cf. Hudeček and Mihaljević, 2019a).

1.1 *Mrežnik*

Croatian is still one of a few national languages that does not have a freely available online corpus-based dictionary compiled according to the rules of contemporary e-lexicography or systematic research on e-lexicography; this was the reason for starting the *Croatian Web Dictionary – Mrežnik* project (cf. Hudeček and Mihaljević, 2017a; Hudeček and Mihaljević, 2017b; Hudeček, 2018). *Mrežnik* is a four-year project (1st March 2017 – 28th February 2021) financed by the Croatian Science Foundation. The result of the project will be a free, corpus-based, born-digital, monolingual, easily searchable, hypertext, normative online dictionary of the Croatian standard language. It will become the central meeting point of all language resources compiled at the Institute of Croatian Language and Linguistics, and will thus become a long-term project after the initial four-year period.

Mrežnik is a hypertext dictionary, as its entries and sub-entries are interconnected, as well as linked with entries in databases created within the framework of the *Mrežnik* project[5], as well as with databases being created by project collaborators or other Institute members within the framework of other

4 "The *CCD* is primarily based on traditional lexicographic and lexicological settings of multiword lexical units (…), so that the main plan is to put together in one database the most common Croatian multiword lexical units by defining their semantic types and context of use. The database will be a useful source to be included in other more advanced MWE sources (Croatian and international) for the development of tools that enable the extraction of MWEs on the basis of their semantic and lexical features (…)" http://ihjj.hr/kolokacije/english/about/.

5 The databases created in parallel with the creation of the dictionary are: a language advice database (http://jezicni-savjetnik.hr/), language advice for schoolchildren (http://hrvatski.hr/savjeti/), a conjunction database with a description of groups of conjunctions and their modifications, a database of explanations of the origins of idioms (http://hrvatski.hr/frazemi/), a database of ethnics and ktetics (http://hrvatski.hr/etnici-i-ktetici/).

projects (cf. Hudeček and Mihaljević, 2019a). In addition to the module for adult native speakers of Croatian, the dictionary includes a module for school-children and a module for non-native speakers of Croatian (cf. Mihaljević, 2018). *Mrežnik* is based on the *Croatian Web Repository Online Corpus*[6] and the *Croatian Web Corpus*.[7] As it is a corpus-based and not a corpus-driven dictionary, *Mrežnik* takes all other available print and web sources into account in addition to these two corpora. This means that, while the collocations are primarily based on Word Sketches[8] and the aforementioned corpora, other collocations can be added to the dictionary even if they are not attested in the corpora, but the compiler intuitively knows that they are commonly used in Croatian and can be found on the web. The reason for this approach is that there is currently no representative corpus of the Croatian language, and the aim is for the collocations to be representative of the Croatian (standard) language and not of the available corpora.

In order to present the approach to collocations in *Mrežnik*, the paper focuses on the problem of collocation extraction for *Mrežnik* and the compilation of the collocational blocks for different word classes. Furthermore, it also shows how collocations help differentiate between meanings of polysemous words, when and how pragmatic and normative notes explaining the usage of collocations are added, when stylistic and terminological labels are used, and how collocations help the lexicographer differentiate between the meanings of quasi-synonyms and recognize meanings not yet recorded in Croatian dictionaries.

2 MULTIWORD EXPRESSIONS IN *MREŽNIK*

Collocations are multiword expressions and in order to differentiate them from other multiword expressions, a brief overview of multiword expressions and the approach to them in *Mrežnik* will be provided. According to Atkins and Rundell (2008, p. 167), multiword expressions (MWE) "are a central part of the vocabulary of most languages, and need to be accounted for in the dictionary… All fixed and semi-fixed phrases are important, and worth recording during the analysis process of dictionary writing."

6 http://riznica.ihjj.hr/index.hr.html

7 http://nlp.ffzg.hr/resources/corpora/hrwac/

8 https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/

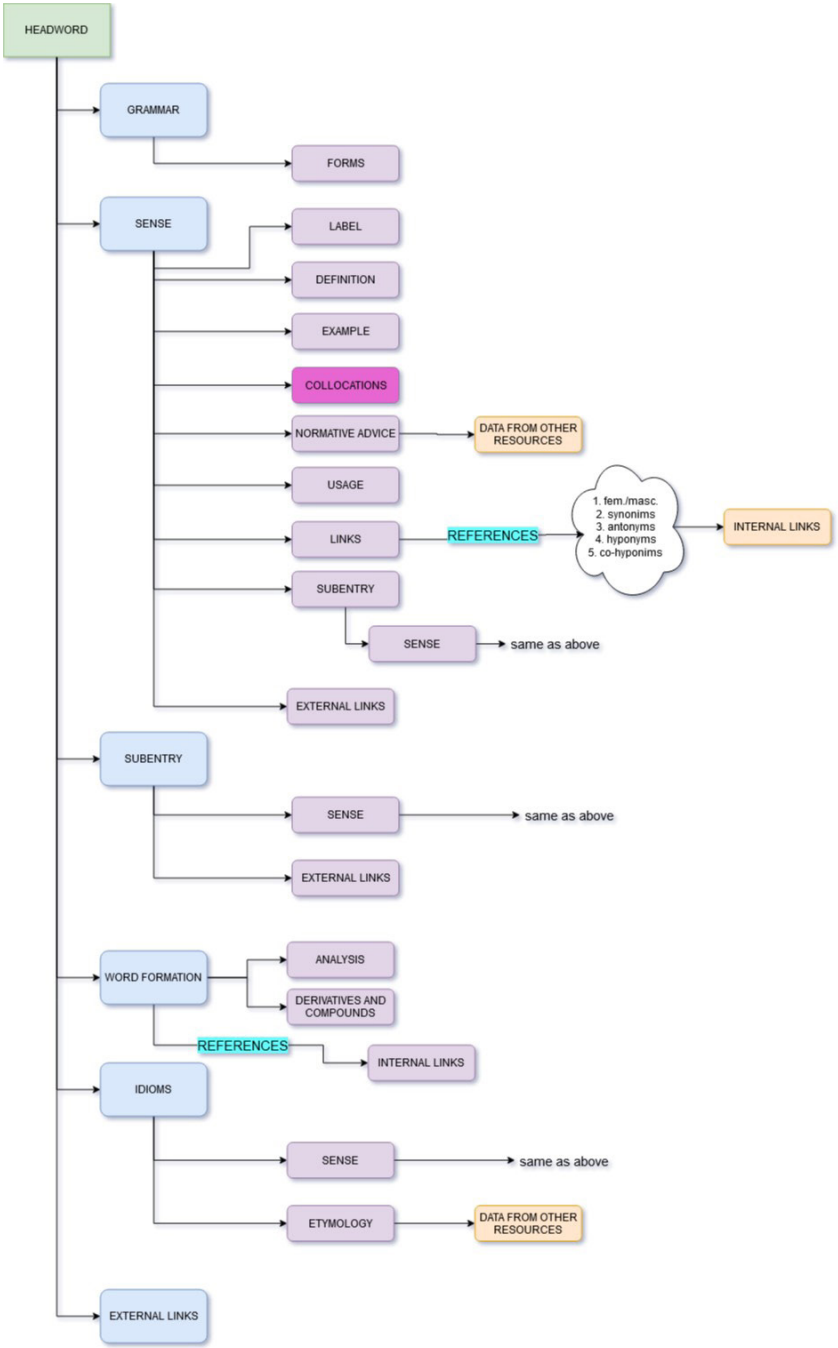


Figure 1: The microstructure of *Mrežnik*.

In the dictionary microstructure of *Mrežnik*, shown in Figure 1, multiword expressions can be presented in subentries (as headwords are always single words), the idiom section (which includes similes, catch phrases, quotations, and proverbs), and the collocational section. We briefly present the approach to multiword expressions used in the subentries and the idiom section before shifting the focus to collocations.

The subentries present terms or phrases the meaning of which cannot be derived from the sum of their constituent parts, e.g. *majčina dušica* (*majčina* = 'mother's', *dušica* = 'little soul', *majčina dušica* = 'thyme'), or when at least one word has a change in meaning, e.g. *morski pas* (*morski* = 'sea', *pas* = 'dog', *morski pas* = 'shark'). However, some frequent terms that can be derived from the sum of their constituent parts are also presented as subentries, especially if they can be linked to the *Struna* terminological database,[9] e.g. the subentries for the entry *broj* ('number') are the mathematical terms *prirodni broj* ('natural number'), *redni broj* ('ordinal number'), *glavni broj* ('cardinal number'). As the terms *redni broj* and *glavni broj* are also linguistic terms, they are also linked to *Hrvatska školska gramatika* ('Croatian School Grammar').[10] However, rare and lesser-known multiword terms are not always treated as subentries in *Mrežnik*. Some of the less frequent terms are provided in the collocational block, in which case they are not accompanied with a definition. The subentries also present some phrases, e.g. the entry *trokut* ('triangle') includes the subentries *ljubavni trokut* ('love triangle'), *ljubavni četverokut* ('love rectangle'), and *Bermudski trokut* ('Bermuda triangle').

The idioms section is compiled by a specially trained phraseologist. Idioms are linked to the database of explanations of the origins of idioms (*Frazemi. Hrvatski u školi*).[11] Some idioms are connected with the articles from the journal *Hrvatski jezik* that provide their etymology (section *Od A do Ž* 'from A to Z').[12]

9 http://struna.ihjj.hr

10 http://gramatika.hr

11 http://hrvatski.hr/frazemi/

12 https://hrcak.srce.hr/hrjezik

3 COLLOCATIONS IN *MREŽNIK*

Collocations are presented in the collocational block and in sentence examples. There are two basic criteria for choosing good sentence examples in *Mrežnik*: a) they contain a frequent collocation; b) they contain a typical syntactic construction. Some of the collocations provided in the collocational block are also illustrated through sentence examples in the example field.

Not all frequent collocations provided by Word Sketches are included in the final entries in *Mrežnik*. This is because of the difference between statistical collocation, i.e. "any combination of two or more words that is statistically relevant, and a collocation that is deemed relevant for inclusion in a dictionary" (Kosem et al., 2018, p. 991).[13] "Frequent but collocationally unremarkable" (Sinclair, 2002, p. 47) collocations have been excluded from *Mrežnik*. Moreover, due to the nature of *Mrežnik* (standard language dictionary, dictionary for general users, students, and non-native speakers), and especially due to the unrepresentativeness of the existing Croatian corpora, there are many other reasons for excluding statistically relevant collocations from *Mrežnik*: certain collocations are either offensive or inappropriate in polite conversation in standard Croatian, are relevant only to non-standard Croatian, or are not relevant for the general user. It is up to the lexicographers to decide how to select (only) relevant collocations. In addition to choosing suitable candidates, the lexicographers have to decide how and where to indicate collocations, as they can be entered under of the collocational base (semantically more autonomous word) or the collocate (semantically more dependent element), or both.

3.1 Extracting collocations for *Mrežnik*

Collocations for the entries in *Mrežnik* are obtained in two ways:

1. Data is extracted from the corpora using the Sketch Engine web tool (cf. Kilgarriff et al., 2004), which allows the display of lemma/word context through Word Sketches (Kilgarriff and Rundell, 2002, pp. 811–815),[14] which are calculated using

13 Kosem et al. (2018, p. 991) stress that not all statistically relevant collocations are worth 'showing' to dictionary users.

14 "The Word Sketch processes the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb,

the sketch grammar developed for Croatian within the *Mrežnik* project.[15] Collocations can be sorted by absolute frequency or logDice score (typicality of the collocation), per syntactic categories. Searches in Word Sketches can be limited to a selected part of speech, e.g. *lak* can be both a noun ('polish') and an adjective ('easy') in Croatian. Figure 2 shows a part of the Word Sketch for the noun *lak*.

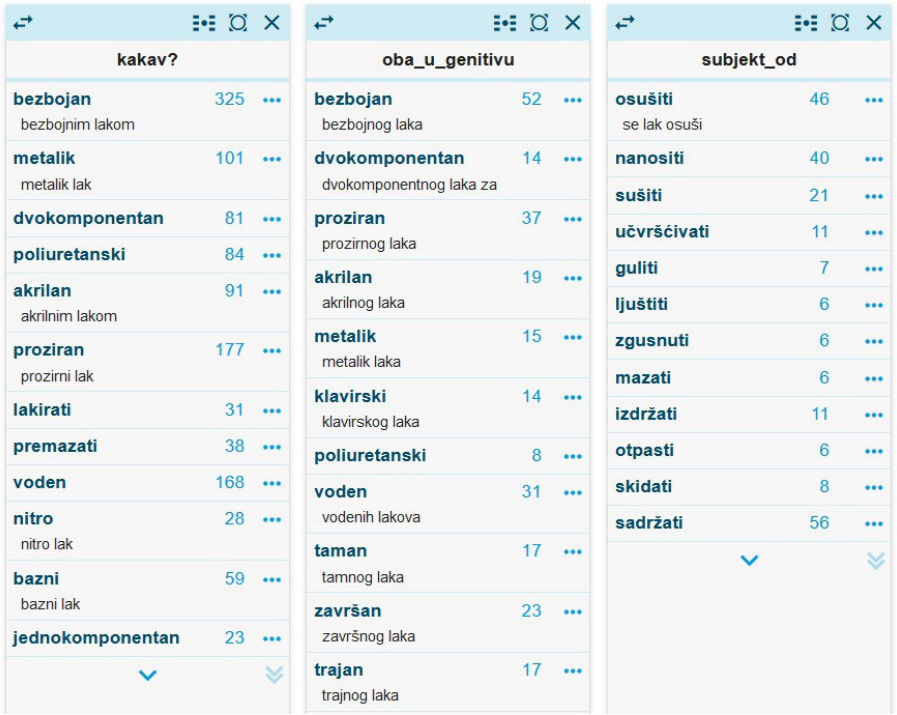


Figure 2: Partial Word Sketch for the noun *lak* ('polish').

The structure is *adjective* + *noun* in the first column, *noun* + *noun* (both in the genitive case) in the second, and (subject)[16] *noun* lak + *verb* in the third.

words that modify the word etc. The words which will be included in the analysis are defined by rules written in the sketch grammar" https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/.

15 The corpora were processed using ReLDI tagger with Word Sketches version 1.4 by Nikola Ljubešić within the *Mrežnik* project. The team members checked Word Sketches and suggested some additions and alterations (cf. Hudeček and Mihaljević, 2018b, pp. 106–107).

16 Although the column is marked as subject, syntactic analysis shows that in many cases the collocate is the object of the collocation, e.g. *nanijeti lak* ('apply polish').

The selected columns are the most typical for the entry *lak*. However, other columns of Word Sketches are analysed by lexicographers as well. Concordances of these collocations are analysed with the option *get a random sample*. A partial concordance of the noun *lak* is shown in Figure 3.

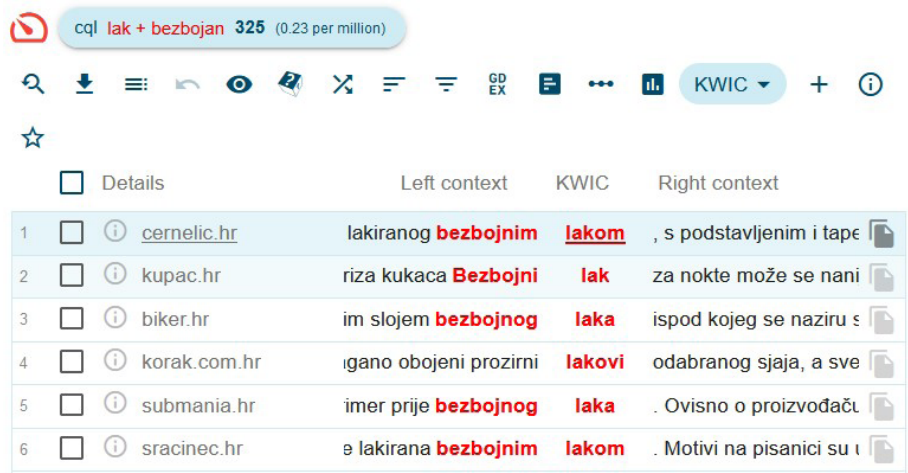


Figure 3: Concordance (random sample) of the noun *lak*.

2. A random sample of approximately 300 examples is checked in the *hrWaC* and *Repository* corpora as some collocations the lexicographers know to be typical are not found via Word Sketches due to the unrepresentativeness of the corpus.

3.2 The collocational block in *Mrežnik*

Mrežnik is compiled in the TLex dictionary-writing system;[17] Figure 4 shows a simple (one meaning and just a few collocations) entry (the particle *čim* 'as soon as') in XML. A frame marks the part showing collocations.

The concept of collocations and their presentation was initially modelled after the example of *elexiko* (Haß, 2005; Storjohann, 2005). Thus, we began developing the model for collocations with the questions introduced in *elexiko* (Klosa, 2015, p. 36; Haß, 2005, p. 118). However, while working with the Croatian corpora, we modified the *elexiko* model in accordance with our language material. Collocations consist of a keyword (the headword or the subentry in our case) and a collocate. The same collocation is often listed in two entries,

17 https://tshwanedje.com/tshwanelex/

```
<Lemma id="8928" Djeca="0" Stranci="0" LemmaSign="čim" Naglaseno="čîm" HomonymNumber="1"
vrsta_rijeci="42" Notes="Lana" prvi_pregled="162" konacni_pregled="0" Modified="2020-04-28 20:42:22"
Created="2018-01-09 14:16:20" ModifiedBy="Lana" CreatedBy="Domagoj">
    <Sense id="119726" SenseNumber="1" stilska_odrednica="0" strucna_odrednica="0">
        <Definicija id="119727" definicija="Čim uvodi komparativ pridjeva ili priloga i naglašuje
        njegovo značenje."/>
        <primjeri id="119728">
            <Primjer id="119729" Primjer="Kraj ronjenja znači skidanje i slaganje opreme, prijenos s
            broda na kopno, a s kopna u ronilački centar, zatim pranje svakog komada opreme u slatkoj
            vodi i njezino vješanje kako bi se ona čim prije osušila."/>
            <Primjer id="126937" Primjer="Novo vodstvo obećalo je očuvati tradicionalne vrijednosti
            tribine, ali i obogatiti je novim multimedijskim sadržajima te nastojati privući čim više
            mlađih pjesnika."/>
            <Primjer id="126938" Primjer=" SEO je proces poboljšavanja mrežne stranice kako bi
            postigla čim bolje rezultate na poznatim tražilicama."/>
            <Primjer id="126939" Primjer="Pokušajte da su slatkiši (slatko na kraju obroka nažalost je
            postalo svakodnevnica) zastupljeni čim manje ili nastojte da budu čim kvalitetniji. "/>
        </primjeri>
        <kolokacije id="119732">
            <Kolokacija id="119733" odrednica="čim + pridjev:" kolokacija="čim bolji, čim
            kvalitetniji, čim veći, čim viši "/>
            <Kolokacija id="126940" odrednica="čim + prilog:" kolokacija="čim bliže, čim
            jednostavnije, čim lakše, čim manje,  čim prije, čim ugodnije, čim više"/>
        </kolokacije>
        <Poveznice id="119757">
            <References id="119758">
                <reflemma lemmaid="119740" type="6">
                    <refsense senseid="119741"/>
                </reflemma>
            </References>
        </Poveznice>
    </Sense>
</Lemma>
```

Figure 4: An entry in XML (particle *čim* 'as soon as').

e.g. *crvena jabuka* ('red apple') under *crven* ('red') and under *apple* ('jabuka'). The structure of the collocational block is divided into two fields. Figure 5 shows the demo version of the collocational block for one of the meanings of the headword *breskva* ('peach').

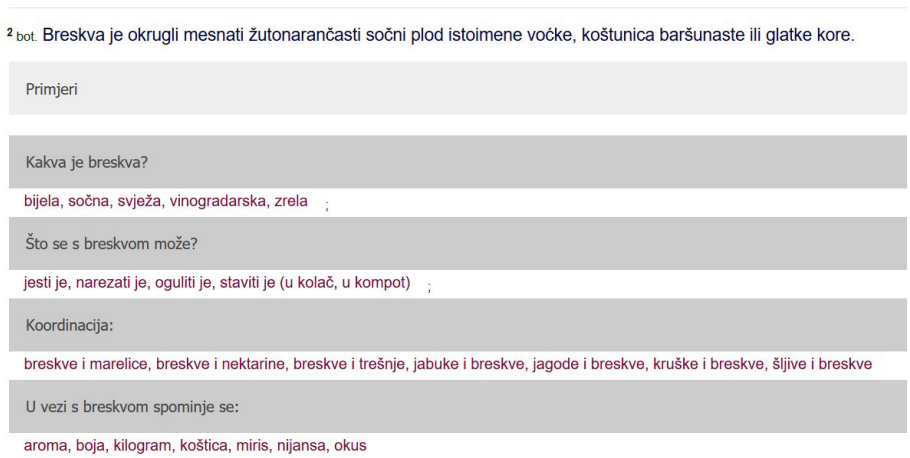


Figure 5: Demo version of the collocational block of the entry *breskva* ('peach').

As is apparent from Figure 5, each collocational field in the collocational block consists of two subfields (determinant and collocates). Determinants can be:

1. Questions, e.g. *Kakva je breskva?* ('What is a peach like?'), *Što se s breskvom može?* ('What can one do with a peach?'). There is a limited number of questions for each word class. However, if needed, the editors can add more questions. These questions usually mirror grammatical relations, e.g. the answer to the question *What is x like?* is typically an adjective, sometimes a noun in the genitive case, and less often the construction *noun + preposition* za ('for') + *noun* or a semi-compound.

2. Introductory phrases, e.g. *Koordinacija:* ('Coordination'), *U vezi s x spominje se*: ('Mentioned in connection with x'), *U imenima* ('In names').

3. Grammatical formula (usually used with grammatical words), e.g. *usklik + imenica u dativu:* ('interjection + noun in the dative').

Kolokacija:

odrednica	Kad putuje?
kolokacija	sutra, svakodnevno, ujutro; iduće ljeto, ovo ljeto, svaki dan, svako ljeto

Figure 6: Comparison of Word Sketch columns and a collocational field (the verb *putovati* 'to travel').

The selected collocates of the headword follow in the second subfield. They are provided in alphabetical order and not by frequency. This is illustrated with a comparison of the Word Sketch columns *kako-kada* ('how-when') and *veznik* ('conjunction') with the collocational field *kada* 'when' in *Mrežnik*, as shown in Figure 6. The *veznik* column includes many words that are not conjunctions, some of which are relevant for this collocational field. The comparison shows which collocates from the Word Sketch have been selected for the presentation in this field in *Mrežnik*. It is an evidence that collocations cannot be extracted mechanicaly from Word Sketches and must be carefully selected by the lexicographer.

Collocates are also occasionally grouped into grammatical and/or semantic groups, with the groups being separated by a semicolon.

3.3 Collocations of different word classes

The editors developed a set of collocational questions, introductory phrases, and/or grammatical formulas for each word class after analysing a sample dataset (cf. Hudeček and Mihaljević, 2018b). These were modified and new questions added if needed. Each word class presented different collocational problems. The collocational questions and introductory phrases always appear in the same order. An overview of typical collocational questions and phrases for each word class is provided in the following sections.

3.3.1 Nouns

Table 1 shows the collocational questions and phrases[18] for nouns.

18 The questions and introductory phrases always appear in the same order, although not all of them are used for every headword. This is why different headwords were used to illustrate the collocations in Table 1.

Table 1: *Collocational questions and introductory phrases – nouns*

Croatian		**English**	
Question or introductory phrase	**Example collocations**[19]	**Question or introductory phrase**	**Literal translation of Croatian collocations**
Kakav je x?	**mašta**: bolesna, bujna, neiscrpna, neobuzdana, pokvarena	What is x like?	**imagination**: sick, vivid, inexhaustible, unrestrained, corrupt
Što x ima?	**list**: bazu, peteljku, plojku, žilice	What does x have?	**leaf**: base, petiole, plate, veins
Što x može?	**Crkva**: osuđivati, priznavati, pozivati, slaviti, učiti, upozoravati	What can x do?	**Church**: condemn, acknowledge, invoke, celebrate, teach, warn
Što se s x može?	**mašta**: buditi je, pobuditi je, razbuktati je,	What can one do with x?	**imagination**: awaken, inflame, kindle
Koordinacija:	**mašta**: mašta i fantazija, mašta i kreativnost, mašta i stvarnost, mašta i volja	Coordination:	**imagination**: imagination and fantasy, imagination and creativity, imagination and reality, imagination and will
U vezi s x spominje se:	**mašta**: plod, proizvod, tvorevina, zaljubljenik	Mentioned in connection with x	**imagination:** fruit, product, creation, lover
U imenima:	**duh**: Koko i duhovi (novel)	In names:	**ghost:** Koko and the ghosts[20]

Descriptive and possessive adjectives that answer the first collocational question *What is x like?* are alphabetized separately and separated with a semicolon, as shown in Table 2.

Table 2: *Collocates of the word mjenjačnica ('exchange office')*

	Croatian	**English**
	mjenjačnica	**exchange office**
	Kakva je mjenjačnica?	What is an exchange office like?
descriptive adjective	obližnja, ovlaštena, povoljna, privatna, zatvorena	nearby, authorized, affordable, private, closed
possessive adjectives	supetarska, trogirska	from Supetar, from Trogir

19 The table provides collocations for one meaning of each headword only. Most of the headwords have more than one meaning.

20 A famous Croatian novel for children.

While modeling the collocational block for nouns, the editors had to answer these questions:

• **which collocations to include.** When choosing collocations, the editors take into account corpus data (provided by Word Sketch) and evaluate the suitability of collocations for inclusion in the collocational block of *Mrežnik*. For example, the collocations *brkata konobarica* ('mustachioed waitress'), *sisata konobarica*, *prsata konobarica* ('large-breasted waitress'), *alkoholizirani maturant*, *pijani maturant* ('drunk secondary-school graduate') would not be considered as suitable collocations for *Mrežnik* although they have the highest logDice score in the column *kakav?* (What is x like?). Namely, any collocations that might insult anybody on the basis of their age, sex, race, sexual orientation, nationality, religion, etc. have been excluded from *Mrežnik* (cf. Hudeček and Mihaljević, 2018b, p. 109).

• **coordination.** The coordination field lists elements connected with *i* ('and'), *te* ('as well as'), *ili* ('or'), *odnosno* ('namely'), and /. Coordination presented the following problems:

1) how to differentiate between the following two cases:

a) *X* belongs simultaneously to two groups connected by a coordinator, e.g. *nogometaš i sportaš* ('a footballer and an athlete'), *nastavnik i pedagog* ('a teacher and an educator'), *profesorica* i *prevoditeljica* ('a teacher and a translator'), *književnica i prevoditeljica* ('a writer and a translator'), *vaterpolist i reprezentativac* ('a water polo player and a member of the national team').

b) Two groups are linked by a coordinator, e.g. *nogometaši i košarkaši* ('footballers and basketball players'), *učenici i nastavnici* ('students and teachers'), *vaterpolisti i košarkaši* ('water polo players and basketball players').

The lexicographer distinguishes between the two groups on the basis of an analysis of Word Sketch and concordances. The solution used in *Mrežnik* is to separate the two groups according to the opposition singular/plural and placing a semicolon between them.

2) How to differentiate between these two cases:

a) the noun refers only to men;

b) the noun (especially in the plural) refers to both men and women.

These two groups were separated by a semicolon and introduced by the introductory phrase *odnosi se samo na muškarce* ('refers only to men') or *odnosi se samo na muške osobe* ('refers only to male persons'). This is illustrated by the coordination of the word *vaterpolist* ('water polo player'). The difference between the two introductory phrases is that the phrase *odnosi se samo na muškarce* is used when it applies only to men (e.g. *liječnik* 'doctor') and the phrase *odnosi se samo na muške osobe* applies to boys as well as men (*nogometaš* 'footballer'). This is shown in Table 3.

Table 3: *Coordination of the noun vaterpolist ('water polo player')*

Koordinacija	Coordination
vaterpolist i reprezentativac; vaterpolisiti i košarkaši; *odnosi se samo na muške osobe*: vaterpolisti i vaterpolistice	water polo player and member of the national team; water polo players and basketball players, *refers only to male persons*: water polo players and water polo players (f.)

In the coordination field, collocations can refer to the same person, e.g. *vaterpolist i reprezentativac* ('water polo player and the member of the national team'), and to two groups of sportsmen, e.g. *vaterpolisti i košarkaši* ('water polo players and basketball players'). The collocation *vaterpolisti i vaterpolistice* ('male water polo players and female water polo players') refers only to *muške osobe* ('male persons').

3) The order of nouns connected by a coordinator had to be determined. After trying out all possibilities,[21] we decided to make the headword the first member of the coordinated phrase. The exception to this are set phrases the order of which is fixed or much more common, e.g. *lijevi i desni* ('left and right').

21 The possibilities were: the collocation is copied in the form it occurs in the Word Sketch with the possibility of repeating the same elements but in a different order (e.g. *vaterpolist i reprezentativac* but possibly also *reprezentativac i vaterporist*), the collocation is listed in the order the elements occur more often but without repeating the elements (e.g. only *vaterpolist i reprezentativac*), the collocation is listed in the order in which the headword appears in the second place (*reprezentativac i vaterpolist*).

4) Collocations are listed in the order of the coordinator used: *i* ('and'), *te* ('and'), *ili* ('or'), *ni* ('neither'), *niti* ('nor'), */*, *odnosno* ('rather'); collocations with each new coordinator are separated by a semicolon.

• **proper names.** Although one can argue that proper names are not collocations, they were included in the collocational field. As proper names occur quite frequently in Word Sketches, this information could be useful for the user,[22] e.g. the word *list* ('leaf') occurs most often in the names of newspapers (*Večernji list*, *Jutarnji list*, etc.); *Večernji list* and *Jutarnji list* have the highest score in Word Sketches for the lemma *list*. After analysing all name categories occurring in Word Sketches, it was decided to include the following categories in the collocational block: place names, names of organizations and events, names of holidays, and names of commemorations.

The occurrence of the headword in names often revealed facts that were commented upon in the pragmatic note, e.g. *jučer* ('yesterday') in the meaning 'past time', *danas* ('today') 'present time', and *sutra* ('tomorrow') 'future time, time to come' are used very often in the names of various events, e.g. *Razvoj turizma u Kninu: danas i sutra* ('The development of tourism in Knin: today and tomorrow'). The term *svjesnost* (and not the synonymous term *svijest*) often occurs in proper names. The word *svjesnost* 'awareness' (as opposed to its (quasi-)synonym *svijest*) occurs most often in the names of days, weeks, or months dedicated to something, often an illness, disability, or disorder, e.g. *Međunarodni dan svjesnosti o mucanju* ('International Stuttering Awareness Day').[23] The Word Sketch Difference for the grammatical relation *noun + preposition o* ('about') + *noun* of the lemmas *svijest* (90,010 occurrences in the corpus) and *svjesnost* (8,621 occurrences in the corpus) is shown in Figure 7. The numbers in the second column indicate the frequency of collocates of *svijest*, while those in the second column indicate the frequency of collocates

22 This is based on our experience in giving language advice. Users sometimes ask for advice in choosing an appropriate name for an event or a document title. This is a question of combining word elements appropriately, not of encyclopedic knowledge. Single-word proper names are never entry words in *Mrežnik*, but they are sometimes provided in the pragmatic note (e.g. the personal names *Jagoda* and *Višnja* in the entries *jagoda* ['strawberry'] and *višnja* ['sour cherry']). This is especially useful in the module for non-native speakers of Croatian.

23 For more on the meaning of the terms *svijest* and *svjesnost*, cf. Vrgoč and Mihaljević (2019).

of *svjesnost,* e.g. *svijest o odgovornosti* ('awareness of responsibility'), *svijest o potrebi* ('awareness of a need') vs. *svjesnost o autizmu* ('awareness of autism'), *svjesnost o mucanju* ('awareness of stuttering').

odgovornost	224	0	•••
potreba	1236	30	•••
vrijednost	346	8	•••
važnost	1299	79	•••
postojanje	230	34	•••
vlastit	479	65	•••
tijelo	115	49	•••
autizam	34	73	•••
mucanje	0	26	•••
antibiotik	0	30	•••

Figure 7: Partial Word Sketch Difference for lemmas *svijest* and *svjesnost*.

Names are introduced by the introductory phrase *U imenima*: ('in names'). The class to which the name belongs (e.g. film, novel, event) is provided in brackets if the name is not self-explanatory, e.g. for *Hrvatski slavistički kongres* ('Croatian Slavic Studies Congress'), the word *kongres* is not provided as an explanation as it is a part of the name; for *Bravo maestro* ('Well done Maestro'), the word *film* is provided in brackets.

• **the grammatical form of collocates.** Although collocational questions and answers are mostly in the singular, sometimes the plural was required. This is the case if the collocation implies more than one person or thing, e.g. *What can x do? okupljati se* ('bring together'). Singular and plural collocates are separated by a semicolon (as shown in Table 3).

• **terminological and stylistic labels.** Terminological and stylistic labels are used in the collocational field in some cases. This is especially true for collocations that are only used in the colloquial style or that do not belong to the standard language. Granger and Paquot (2012, p. 165) stress that non-native writers can be seriously misled by the presentation of collocations, as they are not provided with any help to decide which collocations are the most

appropriate in academic writing. This is often also true for native speakers and in all styles of writing.[24]

Style labels are used when the collocate is stylistically marked or does not belong to the standard language, e.g. one of the answers to the question *Što stomatologinja može?* ('what can a dentist do') is *pokrpati zub* ('mend a tooth') marked by the label *žarg*. ('jargon'), as this collocation does not belong to the standard language.

• **dividing groups of collocates.** Collocates are sometimes grouped and divided by a semicolon according to syntactic and semantic criteria, e.g. the answer to the question *What is x like?* can be an adjective, a compound (consisting of two nouns, sometimes hyphenated) or a phrase that has the structure *headword + noun in the genitive*. These groups are separated by a semicolon as shown in the collocational field *Kakva je čistačica?* of the entry *čistačica* in Table 4.

Table 4: *Collocates of the noun čistačica*[25] *('cleaning lady')*

Kakva je čistačica?	What is a cleaning lady like?
dežurna, obična, školska, vrijedna, zaposlena, x-godišnja; čistačica spremačica, teta čistačica *hip*.	on call, ordinary, school, hardworking, employed, x-year-old; cleaning lady and housekeeper, aunty cleaning lady

Table 4 shows two groups of collocations answering the question *Kakva je čistačica?* ('What is a cleaning lady like?'):

- adjectives, e.g. *vrijedna čistačica* ('hardworking cleaning lady');
- nouns, e.g. *čistačica spremačica* ('cleaning lady and housekeeper'), *teta čistačica* ('aunty cleaning lady').[26]

As the age of a person often occurs in the corpus, this is indicated by the construction *x-godišnji/x-godišnja* ('x-year-old').

24 This statement is supported by our experience in giving language advice, teaching Croatian to students of electrical engineering and journalism (native speakers of Croatian), and editing Croatian texts written by native speakers, as well as by Hudeček (2020) and Blagus Bartolec (2017).

25 For more on masculine and feminine professional nouns in *Mrežnik*, see Hudeček and Mihaljević (2019b).

26 *Teta čistačica* is a hypocoristic way young children address cleaning ladies at school or in kindergarten. This is indicated with the label *hip*. ('hypocoristic').

3.3.2 Verbs

Verbal collocations are very complex as they depend on the syntactic characteristics of the verb (reflexive, impersonal, transitive, intransitive), verbal valence, and the semantic characteristics of the verb. Each semantic class of verbs (e.g. verbs of motion, psychological verbs, etc.[27]) has partly different collocational questions. Collocational questions are divided according to the sentence elements that answer them: subjects, objects, adverbials. Questions are different for imperfective, perfective, and reflexive verbs, as well as for animate and inanimate subjects.

1. Questions for the subject are shown in Table 5.

Table 5: *Collocations denoting the subject*

	Imperfective		**Perfective**	
	Croatian	**English**	**Croatian**	**English**
animate	Tko x?	Who x?	Tko može x?	Who can x?
	trčati: atletičar, konj	**run**: athlete, horse	**upoznati:** polaznik, student	**get to know**: attendant, student,
inanimate	Što x?	What x?	Što može x?	What can x?
	svijetliti: krijesnica, lampa	**shine**: firefly, lamp	**pasti**: bomba, jabuka	**fall**: bomb, apple

2. Questions for the object are shown in Table 6.

Table 6: *Collocations denoting the object*

	Imperfective verbs		**Perfective verbs**		**Reflexive verbs**	
	Croatian	**English**	**Croatian**	**English**	**Croatian**	**English**
Direct object	Što se x?	What is x?	Što se može x?	What can be x?		
	čitati: knjiga, tekst	**read:** book, text	**dati:** glas, odgovor,	**give:** a vote, a response		
Indirect object	Čemu x? Komu x?	To/at whom/what can one x?	Komu se može x?	To/at whom can one x?	Komu se x? Čemu se x?	To/at whom/ what can one x?
	mahati gomili, obožavateljima,	**wave to**: the crowd, the fans	**mahnuti**: konobarici, navijačima	**wave**: the waitress, the fans	**smijati se:** prijatelju, šali	**laugh**: a friend, a joke

27 The semantic classification of verbs is based on the classification made in the project *e-Glava*. More on the project *e-Glava* see Birtić et al. (2017) and on the classification of verbs see Brač and Bošnjak Botica (2015).

3. The questions for adverbial collocations depend on the semantic class of the verb (e.g. motion verbs have different questions than static verbs) and on the adverbial class as shown in Table 7 (only imperfective verbs are shown). Perfective verbs have modified questions, e.g. imperfective verb: *Kako se x?*, perfective verb: *Kako se <u>može</u> x?*; imperfective verb: *Kad se x?*, perfective verb: *Kad se <u>može</u> x?*, etc.

Table 7: *Collocations denoting adverbials*

	Croatian		English	
adverbial	**question**	**example**	**question**	**example**
of manner	Kako (se može) x?	**mahati:** bijesno, nervozno	How can one x?	**wave**: angrily, nervously
of place	Gdje x? (static verbs)	**ljetovati:** u kampu, na Pagu	Where x?	**spend the summer**: in a camp, on Pag
	Kamo x? (verbs of motion)	**putovati:** kući, izvan grada	To where x?	**travel**: home, out of the city
	Kuda x? (verbs of motion)	**putovati**: diljem svijeta, kroz Neum	Which way x?	**travel:** across the world, through Neum
of time	Kad x?	**svijetliti**: noću, trajno	When x?	**shine**: at night, permanently
of reason	Zbog čega x?	**putovat**i: zbog posla, zbog zabave	Why x?	**travel**: for work, for fun
of company	S kim x?	**putovati:** s klubom, s prijateljima,	With whom x?	**travel:** with a club, with friends
of means	Čime se x?	**mahati**; krilima, pištoljem	With what x?	**wave**: wings, a gun
of frequency	Koliko često x?	**putovati**: često, tjedno	How often x?	**travel**: often, weekly

Tables 6 and 7 show the complexity of verbal collocations. Coordination also often occurs with verbs: *voljeti i ljubiti* (love and love/kiss), *voljeti i mrziti* (love and hate).

3.3.3 Adjectives

The most common question introducing adjectives is the question *Što je x?* (What is x?). We list the nouns answering this question in the following order: animate, inanimate, abstract. These three noun groups are divided by

semicolons. Collocational questions and introductory phrases for the adjective *loš* (bad) are provided in Table 8.

Table 8: *Collocational questions and introductory phrases – adjective loš ('bad')*

Croatian	Example	English	Example
Što je loše?	čovjek; navike	What is bad?	person; habits
Koliko je što loše?	jako, iznimno	To what degree is something bad?	very, extremely
Koordinacija:	loš i nekvalitetan; dobar ili loš	Coordination:	bad and of low quality; good or bad

Terminological labels are used only to distinguish between different meanings of the collocate, e.g. the entry *crven* ('red') features the question *Što je crveno?* 'What is red?', the answers to which are e.g. *div* astr. ('giant', astronomy), *karton* sp. ('card', sports), *patuljak* astr. ('dwarf', astronomy), *vjetar* med. ('wind', medicine). Collocates of the adjective *crven* ('red') are given in Table 9.

Table 9: *Collocates of the adjective crven ('red')*

Što je crveno?	What is red?
boja, haljina; div *astr.*, karton sp., krvna zrnca, patuljak *astr.*, vjetar *med.*	color, dress; div *astr.*, card *sp.*, blood cells, dwarf *astr.*, wind *med.*

3.3.4 Adverbs

Collocations of adverbs formed in Croatian from the neutral form of adjectives by conversion (e.g. *jako* formed from the neutral form of the adjective *jak*) are introduced by the questions *Što se može x?* ('What can be done in a x manner?') and *Koliko je što x?* ('To what degree is something x?'). However, in other adverb groups, collocations are introduced by the introductory phrase *uz glagole:* ('with verbs:'), e.g. the adverbs *gdje* ('where'), *kuda* ('where to), *kamo* ('which way'), and *uz prijedloge* ('with prepositions'), e.g. *jako blizu* ('very near'). Table 10 shows the collocational questions and introductory phrases for adverbs on the example of *loše* ('badly').

Table 10: *Collocational questions and introductory phrases for loše ('badly')*

Croatian		**English**	
Question and phrases	**Examples**	**Question and phrases**	**Examples**
Što se može loše?	**loše**: biti plaćen, igrati	What can be done badly?	**badly:** be paid, play
Koliko je što loše?	**loše**: katastrofalno, veoma	To what degree is something bad?	**badly:** disastrously, very
Uz pridjeve:	**besmrtno**: neozbiljan, zaljubljen	With adjectives:	**immortally**: frivolous, in love
Koordinacija:	**loše**: dobro i lose; loše ili nikako	Coordination:	**badly**: well and badly; badly or not at all

3.3.5. Numbers

Collocational questions, introductory phrases, and grammatical formulas differ for cardinal and ordinal numbers, and in Table 11 we provide prototype collocational questions for both groups. Although one can argue that no collocations need be given with numbers and that numbers are not collocational words at all, based on our experience with students and providing language advice, we believe that some prototype collocations with numbers can also be useful (from a semantic and a syntactic point of view), e.g. *prvo mjesto* ('first place'); *sedam patuljaka* ('seven dwarfs'), *sedam dana* ('seven days'); *dvanaest mjeseci* ('twelve months'); *pet do devet* ('five to nine'), *pet na dan* ('five a day'), *pet od šest* ('five out of six'), etc. Table 11 shows collocations of cardinal and ordinal numbers.

Table 11: *Collocational questions and introductory phrases – numbers*

	Croatian		**English**	
	Question	**Example**	**Question**	**Example**
glavni brojevi ('cardinal numbers')	Čega je x?	pet prstiju	What do we have x of?	five fingers
	x + prijedlog + N	pet na dan	x + preposition + y	five a day
	Koordinacija:	pet i šest	Coordination:	five and six
redni brojevi ('ordinal numbers')	Što je x?	peti mjesec	What can be x?	fifth month (May)
	Koordinacija:	peti ili šesti	Coordination:	fifth or sixth

Some collocations with numbers motivated the inclusion of a normative note, e.g. *drugi najbolji* ('second best') is a very common collocation in the corpus but should be replaced by *drugi* ('second') in standard Croatian, as *drugi najbolji* is considered a pleonasm and literal translation from English.

3.3.6 Interjections

Collocations of interjections are mostly introduced with syntactic formulas and the introductory phrases *Koordinacija:* ('koordination') and *U imenima:* ('in names'), as shown in Table 12.

Table 12: *Collocational questions and introductory phrases – interjections*

Croatian		**English**	
glagol + x:	reći bravo	verb + x:	say bravo
x + prijedlog + :	bravo za orkestar	x + preposition + noun:	bravo to the orchestra
x + imenica u vokativu:	bravo dečki	x + noun in the vocative:	bravo (well done) boys
Koordinacija:	ajme i jao	Coordination:	oh my and wow
U imenima:	Bravo Maestro (film)	In names:	Well Done Maestro (film)

3.3.7 Pronouns

Collocational questions depend on the pronoun category (personal pronoun, possessive pronoun, demonstrative pronoun, relative pronoun, etc.). Table 13 shows some collocational questions and introductory phrases for personal and possessive pronouns.

Table 13: *Collocational questions and introductory phrases – pronouns*

	Croatian		**English**	
personal pronouns	**Koordinacija:**	**ja**: (i) ja i ti; (ili) ja ili on/ona	Coordination:	**I/me:** you and I; I or he/she
possessive pronouns	**Što je x?**	**moj:** djetinjstvo, mišljenje	What is x?	**my:** childhood, opinion
	Koordinacija:	moj i tvoj, ti i tvoj...	Coordination:	mine and yours, you and your...
	U imenima:	Naši i vaši (serija)	In names:	Ours and Yours (TV series)

Possessive pronouns have the same question *Što x može*? ('What can x do?') as adjectives. Possessive pronouns sometimes function as nouns, e.g. one of the meanings of the prounoun *naši* ('our'). In this case, collocations can be the same as typical collocations of nouns, e.g. *Što naši mogu?* ('What can ours do?') *biti poraženi, izgubiti, pobijediti, slaviti, trijumfirati* ('be beaten, lose, win, celebrate, triumph').

3.3.8 Conjunctions

Typical collocations of conjunctions are introduced by the phrase *U vezničnim skupinama:* ('in conjunction groups'), e.g. *ali* ('but?'): *ali ipak* ('but still'), *ali isto tako* ('but the same'). Reduplicated conjunctions such as *ili…ili* ('either…or') have syntactic formulas such as *Uz glagole:* ('with verbs') and *Uz prijedloge u prijedložnim izrazima:* ('with prepositions in prepositional phrases'), e.g. *ili dati ili uzeti* ('either give or take'), *ili ostati ili otići* ('either leave or stay'), *ili izvan čega ili unutar čega* ('either outside or inside of something'), etc.

3.3.9 Particles

There is no unique collocational model for particles. The collocational field is adapted to each collocation, e.g. modifiers are introduced by introductory phrases stating the word class which follows the modifier, e.g. *Uz pridjeve:* ('with adjectives') *čim bolji, čim veći* ('as good as possible, as big as possible'), *Uz priloge*: ('with adverbs') *čim bliže, čim jednostavnije* ('as close as possible, as simple as possible').

3.3.10 Prepositions

Prepositions are the only word class for which no collocations are provided in *Mrežnik*, as they are considered a non-collocational word class. The reason for this is that word combinations like *iz daljine* ('from afar'), *iz inata* ('out of spite') are provided in examples under different meanings as shown in Table 14.

Table 14: *Meanings and examples of the preposition iz ('from')*

Croatian		**English**	
Definition	**Example**	**Definition**	**Example**
Iz označuje da tko ili što izlazi ili potječe odakle	Krenuli smo iz Kutine u 6 sati.	Iz ('from') indicates that somebody or something leaves or originates from somewhere.	We left Kutina at 6 o'clock.
Iz označuje da tko ili što pripada određenomu razdoblju.	Crkveni je namještaj uglavnom iz doba baroka i klasicizma.	Iz ('from') indicates that somebody or something belongs to a certain period.	The church furnishings are mostly from the Baroque and Classicist period.
Iz označuje da je što uzrok čemu drugom.	Turci su, nemajući što izgubiti, zaigrali iz inata.	Iz ('out of') indicates that something is the reason for something else.	The Turks, having nothing to lose, played out of spite.

3.4 The role of collocations in determining and differentiating meanings

Work on *Mrežnik* confirms Firth's (1957, p. 11) famous slogan "You shall know a word by the company it keeps". Namely, the meaning of words is "determined by their grammatical and lexical environment (syntagmatic relations like colligation and collocation) as well as by the situation in which they are used (style, pragmatics)" (Altenberg and Granger, 1996, p. 22). Collocations for each word class in *Mrežnik* helped the lexicographers distinguish meanings, provide precise definitions, and list useful pragmatic and normative notes. For example, in the analysis of the antonymous adjectives *dobar* ('good') and *loš* ('bad'), closely connected meanings were defined as shown in Table 15. Other meanings in which these adjectives are not antonymous are not provided in this table. The table only provides collocations answering the question *What is x?*.

Table 15: *Collocates for different meanings of loš ('bad') and dobar ('good')*

	Definition		**Collocates**	
loš ('bad', 'wrong')	Loš je koji ima negativne osobine ili neželjena svojstva.	Bad is that which has negative characteristics.	čovjek, kvaliteta, strana, stvar, vrijeme	person, quality, side, thing, time
	Loš je koji nije onakav kakav treba biti, koji ne ispunjava očekivanja.	Bad is that which is not as it should be, that which does not fulfil expectations.	igra, ocjena, odnos, rezultat, situacija, stanje, start	game, rating, relationship, result, situation, condition, start
	Loš je koji obavještava o nečemu lošem ili najavljuje loše.	Bad is that which reports on something bad or predicts something bad.	najava, vijest, znak	announcement, news, sign
	Loš je koji nije ispravno utemeljen i logičan.	Bad is that which is not correctly founded or logical.	zaključak	conclusion
	Loš je koji ne donosi korist, koji nema rezultate.	Bad is that which does not bring profit or results.	poslovanje, plan, (poslovni) potez	business, plan, (business) move

	Definition		Collocates	
dobar ('good')	Dobar je koji ima pozitivne osobine ili poželjna svojstva.	Good is that which has positive characteristics.	čovjek, igrač, odnos, prijatelj, stvar, vino, vrijeme	person, player, relationship, friend, thing, wine, time
	Dobar je koji je onakav kakav treba biti, koji ispunjava očekivanja.	Good is that which is as it should be, which fulfils expectations	dan, film, igra, momčad, rezultat	day, movie, game, team, result
	Dobar je koji obavještava o nečemu dobromu ili najavljuje dobro.	Good is that which reports on something good or predicts that something good will happen.	najava, vijest, znak	announcement, news, sign
	Dobar je koji je ispravno utemeljen i logičan.	Good is that which is not correctly founded or logical.	ideja, izbor, način, primjer, rješenje	idea, choice, way, example, solution
	Dobar je koji ne donosi korist, koji ima rezultate.	Good is that which does not bring profit or results.	posao, praksa, suradnja	work, practice, collaboration

Collocations led to the identification of new subentries as yet unrecorded in Croatian dictionaries, e.g. *ljubavni trokut, ljubavni četverokut* ('love triangle', 'love rectangle'). Collocations also motivated the lexicographers to introduce new meanings as yet unrecorded in Croatian dictionaries, e.g. two meanings of *phonology* in Table 16. A similar distinction was made in the meanings of *morfologija* ('morphology'), *sintaksa* ('syntax'), *tvorba riječi* ('word formation'), etc.

Table 16: *Collocates for two meanings of fonologija ('phonology')*

Definition		Collocates	
Fonologija je grana gramatike koja proučava glasove kao razlikovne jezične jedinice	Phonology is a branch of grammar concerned with sounds as distinctive units.	dijakronijska, generativna, opća, povijesna	diachronic, generative, general, historical
Fonologija je sustav glasova kao razlikovnih jezičnih jedinica i njihovih međuodnosa.	Phonology is the system of sounds as distinctive units and their interrelations.	čakavska, praslavenska, štokavska	Čakavian, proto-Slavic, Štokavian

Collocations sometimes helped differentiate between meanings of similar words, e.g. the adjectives *maslinin* and *maslinov*. Both of these adjectives are derived from the noun *maslina* ('olive'), have approximately the same meaning *koji se odnosi na maslinu* 'relating to an olive', and are considered synonyms. However, the Word Sketch Difference in Figure 8 shows that most of the collocates of these two adjectives differ.

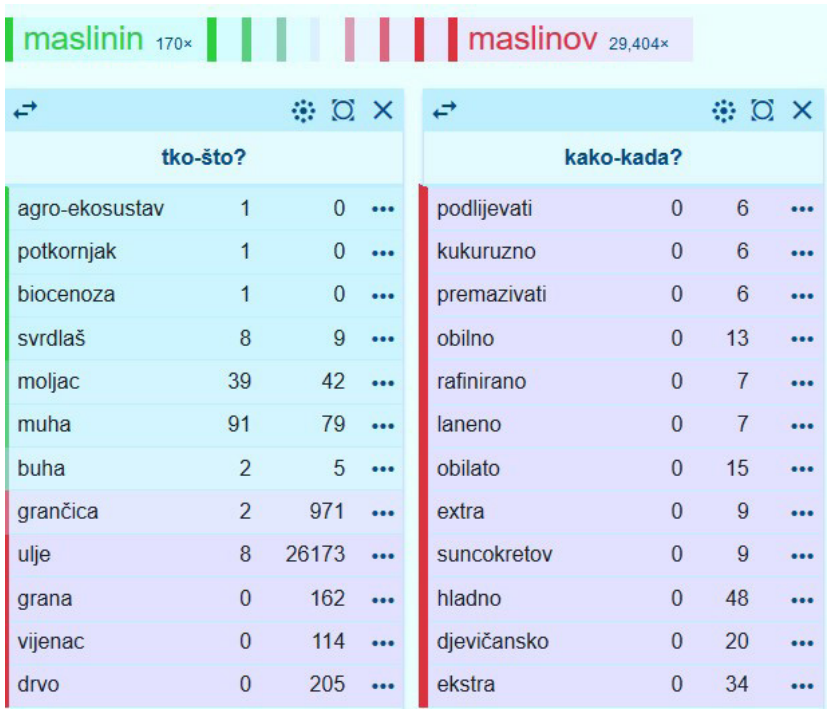


Figure 8: Partial Word Sketch Difference for *maslinin* and *maslinov*.

The adjective *maslinin* (170 occurrences in the corpus) mostly occurs with nouns denoting a parasite: *potkornjak* ('bark beetle'), *svrdlaš* ('borer'), *moljac* ('moth'), *muha* ('fly'), *buha* ('flea'), or with those denoting biological terms *agroekosustav* ('agroecosystem'), *biocenoza* ('biocenosis'). On the other hand, the adjective *maslinov* (29,404 occurrences in the corpus) occurs with nouns denoting parts of the plant, e.g. *grančica* ('twig'), *grana* ('branch'), *drvo* ('tree'), or products made from the plant, e.g. *ulje* ('ulje'), *vijenac* ('wreath'). This resulted in different definitions for these adjectives as shown in Table 17.

Table 17: *Meanings of the adjectives maslinin and maslinov*

Headword	Definition	Collocations	Definition	Collocations
maslinin	Maslinin je koji se odnosi na maslinu.	agroekosustav, biocenoza; potkornjak, svrdlaš, moljac, muha, buha	*Maslinin* is that which relates to olives.	agroecosystem, biocenosis; bark beetle, curculio, moth, flea, fly
maslinov	Maslinov je koji je napravljen od masline.	ulje, vijenac	*Maslinov* is that which is made from olives .	oil, wreath
	Maslinov je koji je dio masline (stabla)	grančica, grana, drvo, list	*Maslinin* is part of an olive tree.	twig, branch, tree, leaf

Similar difference in collocations and meanings can be inferred from the Word Sketch Differences for the adjectival pairs *trešnjin/trešnjev* (adjectives derived from *trešnja* 'cherry'), *višnjin/višnjev* (adjectives derived from *višnja* 'sour cherry'), etc.

3 CONCLUSION

Mrežnik is the first normative born-digital corpus-based dictionary of standard Croatian. It is based on the two existing Croatian corpora, the *Croatian Web Repository* and the *Croatian Web Corpus*, neither of which are representative of the Croatian standard language. This is why other available print and web sources are sometimes consulted[28] and why the approach in the dictionary is corpus-based instead of corpus-driven. This also means that no statistical threshold could be used. For practical lexicographic reasons, multiword expressions in *Mrežnik* are presented in three categories: in subentries, in the collocational block, and in the idiom block. Due to this structure, collocations are defined in a broader sense and include MWEs of grammatical words and proper names, i.e. all relevant data provided by Word Sketch that is not included in a subentry or the idiom block was included in the collocational block.

Each word class, with the exception of prepositions, exhibits different collocational relations and has different collocational questions and phrases. Coordination is the one collocational relation that has the widest range and appears

28 This is especially true for rare words and neologisms not recorded in the corpora, e.g. *koronavirus* (Coronavirus).

in all word classes that display collocational relations. In terms of word classes, verbs show the widest and most complex range of collocational relations.

In dealing with the collocational block in *Mrežnik*, the editors had to answer the following questions: Which collocational questions and introductory phrases should be included for each class or subclass of words?; Which collocations should be included in *Mrežnik*?; When should stylistic labels be included in the collocational block?; When should terminological labels be included in the collocational block?

The analysis of collocations from Word Sketches motivated the lexicographers to form pragmatic and normative notes, which can be helpful to users. This analysis also helped differentiate between meanings or quasi-synonyms, and contributed to the inclusion of new meanings not yet recorded in Croatian dictionaries. The research conducted for the *Mrežnik* project also confirms Michael Rundell's statement: "A high percentage of useful collocations occur in one of four key grammatical relations" (Rundell, 2010). Table 18 contains the four most typical syntactic structures of collocations in *Mrežnik*.

Table 18: *Typical syntactic structures of collocations in Mrežnik*

verb + noun	**maknuti:** posudu, nogu	**move:** a bowl, a leg
adjective + noun	**djevojka:** mlada, slobodna	**girl:** young, single
adverb + verb	**maknuti:** hitno, zauvijek	**remove**: urgently, forever
adverb + adjective	**mali** (comparative *manji*): znatno, jako	**small** (comparative *smaller*): quite, very

Collocations also present a challenge for the gamification of *Mrežnik* (Cf. Mihaljević, 2019a; 2019b), which is in progress at the moment.[29] Games for learning collocations and their relations to different meanings are still in the development phase. The idea is to associate different possible collocates (taken from Word Sketch) to different meanings of a word (taking definitions from *Mrežnik*, e.g. definitions of *kuća* 'house') or to different (similar) words (e.g. *maslinin/maslinov*). Another game provides the collocational question for a

29 Many educational games for children, non-native speakers, and native speakers have been developed. They mostly focus on orthography, morphology, syntax, and on the lexical level. There are also some games for learning special and old alphabets and for learning idioms. Many language games are available at *Hrvatski u igri*.

word from *Mrežnik* and asks players to find some frequent collocates. A sample of the collocational game is shown in Figure 9.

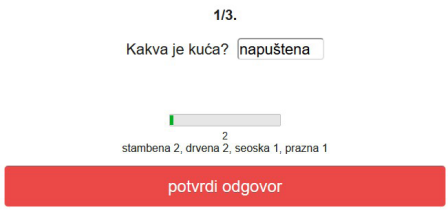


Figure 9: A collocational game (*Kakva je kuća?* 'What is a house like?').

Hopefully, the model used in *Mrežnik* can be useful for other born-digital dictionaries of Croatian and other (Slavic) languages, especially those that do not yet have a born-digital dictionary and a representative corpus of the national (standard) language.

Acknowledgments

This paper was written as part of the research project Croatian Web Dictionary – Mrežnik (IP-2016-06-2141) financed by the Croatian Science Foundation.

REFERENCES

Dictionaries, databases and digital resources

Croatian Collocation Database. Retrieved from http://ihjj.hr/kolokacije/english (1. 2. 2020.)

Croatian Collocation Database. Retrieved from http://ihjj.hr/kolokacije (8. 2. 2020)

Croatian Special Field Terminology – Struna. Retrieved from http://struna.ihjj.hr/en (30. 8. 2019)

Croatian Web Corpus – hrWaC. Retrieved from http://nlp.ffzg.hr/resources/corpora/hrwac/)

Croatian Web Repository Online Corpus. Retrieved from http://riznica.ihjj.hr/index.hr.html

eLexiko. Retrieved from www.owid.de/docs/elex/start.jsp/

Frazemi. Hrvatski u školi. http://hrvatski.hr/frazemi/

Hrvatska školska gramatika. http://gramatika.hr/

Hrvatski jezik. https://hrcak.srce.hr/hrjezik/

Hrvatski u igri. http://hrvatski.hr/igre/

McIntosh, C. (Ed.). (2018). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Sketch Engine Guide. Retrieved from https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/

Other

Altenberg, B., & Granger, S. (1996). Recent trends in cross-linguistic lexical studies. In B. Altenberg & S. Granger (Eds.), Lexis in Contrast. Corpus-based approaches (pp. 3–50). Amsterdam: John Benjamins Publishing Company.

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

Birtić, M., Brač, I., & Runjaić, S. (2017). The Main Features of the e-Glava Online Valency Dictionary. In I. Kosem et al. (Eds.), Electronic lexicography in the 21st century. Proceedings of eLex 2017 Conference, 19–21 September, 2017, Leiden, the Netherlands (pp. 43–62). Brno: Lexical Computing CZ s.r.o.

Blagus Bartolec, G. (2014). *Riječi i njihovi susjedi: Kolokacijske sveze u hrvatskom jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Blagus Bartolec, G. (2017). Glagolske kolokacije u administrativnome funkcionalnom stilu. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 43*(2), 285–309.

Brač, I., & Bošnjak Botica, T. (2015). Semantička razdioba glagola u bazi hrvatskih glagolskih valencija. *Fluminensia, 27*(1), 105–120.

Durkin, P. (Ed.). (2016). *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Firth, J. R. (1957). A synopsis of linguistic theory. *Studies in linguistic analysis,* 1–32.

Granger, S., & Paquot, M. (2012). *Electronic Lexicography*. Oxford: Oxford University Press.

Haß, U. (Ed.). (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. (Schriften des Instituts für Deutsche Sprache). Berlin/New York: de Gruyter.

Hudeček, L., & Mihaljević, M. (2017a). A New Project – Croatian Web Dictionary MREŽNIK. In I. Atanassova et al. (Eds.), *The Future of Information Sciences. INFuture2017, Integrating ICT in Society* (pp. 205–213). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences.

Hudeček, L., & Mihaljević, M. (2017b). Hrvatski mrežni rječnik – Mrežnik. *Hrvatski jezik, 4*(4), 1–7.

Hudeček, L. (2018). Izazovi leksikografske obrade u jednojezičnome mrežnom rječniku (na primjeru *Hrvatskoga mrežnog rječnika – Mrežnika*). In T. Salyha (Ed.), *Visnyk of Lviv University: Series Philology, 69*, 29–38.

Hudeček, L., & Mihaljević, M. (2018a). Croatian Web Dictionary Mrežnik: One year later – What is different? In D. Fišer & A. Pančur (Eds.), *Proceedings of the Conference on Language Technologies & Digital Humanities,* Ljubljana (pp. 106–113).

Hudeček, L., & Mihaljević, M. (2018b). *Hrvatski mrežni rječnik – Mrežnik: Upute za obrađivače*. Retrieved from: http://ihjj.hr/mreznik/uploads/upute.pdf (27. 10. 2019)

Hudeček, L., & Mihaljević, M. (2019a). Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 Conference* (pp. 72–98). Leiden: Lexical Computing CZ s.r.o.

Hudeček, L. (2020). Administrativizmi u rječniku (na primjeru Hrvatskoga mrežnog rječnika Mrežnika). In M. Glušac (Ed.), *Zbornik radova sa znanstvenoga skupa Od norme do uporab 2* (pp. 53 –76). Osijek – Zagreb: Filozofski fakultet Sveučilišta Josipa Jurja Strossmayera u Osijeku – Hrvatska sveučilišna naklada.

Kilgarriff, A., & Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications – a Case Study. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the 10th EURALEX International Congress* (pp. 807–818). Copenhagen: University of Copenhagen.

Kilgarriff, A., Rychlý, 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: Universite de Bretagne – sud.

Klosa, A. (2015). Wortgruppenartikel in elexiko: Einneuer Artikeltyp im Onlinewörterbuch. *Sprachreport Jg, 31*(4), 34–41.

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 XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 989–997). Ljubljana: Ljubljana University Press. Retrieved from https://euralex.org/publications/collocations-dictionary-of-modern-slovene/ (8. 2. 2020)

Mihaljević, J. (2019a). Gamification in E-Lexicography. In P. Bago et al. (Eds.), INFuture 2019: Knowledge in the Digital Age (pp. 155–164). Zagreb: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences.

Mihaljević, J. (2019b). Games for Learning Old and Special Alphabets – The Case Study of Gamifying Mrežnik. In R. Bernardi et al. (Eds.), CLiC-it 2019: Italian Conference on Computational Linguistics. Bari: AILC. Retrieved from http://ceur-ws.org/Vol-2481/paper49.pdf (27. 4. 2020)

Mihaljević, M. (2018). Hrvatski mrežni izvori za djecu i strance. In T. Salyha (Ed.), *Visnyk of Lviv University: Series Philology* (69, pp. 75–89). doi: 10.30970/vpl.2018.69.9298

Ordulj, A. (2018). *Kolokacije u hrvatskom kao inom jeziku*. Zagreb: Hrvatska sveučilišna naklada.

Rundell, M. (2010). *Macmillan Collocations Dictionary: from start to finish*. Retrieved from http://www.macmillandictionaries.com/MED-Magazine/October2010/59-MCD-start-to-finish.htm (27. 4. 2020)

Sinclair, J. (2002). Intuition and annotation – the discussion continues. In K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* (pp. 40–59). Göteborg.

Sinclair, J. M. (2004). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Storjohann, P. (2005). elexiko: A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics, 34,* 55–82.

Vrgoč, D., & Mihaljević, M. (2019). Jesmo li svjesni situacije? Terminološka raščlamba naziva *situational awareness* u vojnome kontekstu. *Strategos, 3*(1), 7–42.

KOLOKACIJE V HRVAŠKEM SPLETNEM SLOVARJU *MREŽNIK*

Cilj projekta *Hrvaški spletni slovar – Mrežnik* je izdelati brezplačni, enojezični, enostaven, hipertekstni, izhodiščno digitalno in korpusno zasnovan slovar standardnega hrvaškega jezika. V *Mrežniku* imajo kolokacije pomembno vlogo. Na začetku projekta so kolokacije in njihova predstavitev temeljile na projektu *elexiko,* kasneje pa je bil na podlagi korpusnih analiz koncept nekoliko prilagojen. V prispevku predstavimo model vključevanja kolokacij pri iztočnicah različnih besednih vrst. Hkrati izpostavimo pomembnejše tematike, povezane s kolokacijami v *Mrežniku*, kot so: metode luščenja kolokacij, vloga kolokacij pri ločevanju med pomeni in prepoznavi novih pomenov, uporaba stilnih in terminoloških oznak pri navajanju kolokacij ter odnosi med kolokacijami in normativnimi in pragmatičnimi informacijami, razlagami in podgesli.

Ključne besede: kolokacije, hrvaški jezik, e-slovar, *Mrežnik*, izvirno digitalni slovar

UPDATING THE DICTIONARY: SEMANTIC CHANGE IDENTIFICATION BASED ON CHANGE IN BIGRAMS OVER TIME

Sanni NIMB, Nicolai HARTVIG SØRENSEN, Henrik LORENTZEN
Society for Danish Language and Literature

Nimb, S., Hartvig Sørensen, N., Lorentzen, H. (2020): Updating the dictionary: semantic change identification based on change in bigrams over time. Slovenščina 2.0, 8(2): 112–138

DOI: https://doi.org/10.4312/slo2.0.2020.2.112-138

We investigate a method of updating a Danish monolingual dictionary with new semantic information on already included lemmas in a systematic way, based on the hypothesis that the variation in bigrams over time in a corpus might indicate changes in the meaning of one of the words. The method combines corpus statistics with manual annotations. The first step consists in measuring the collocational change in a homogeneous newswire corpus with texts from a 14 year time span, 2005 through 2018, by calculating all the statistically significant bigrams. These are then applied to a new version of the corpus that is split into one sub-corpus per year. We then collect all the bigrams that do not appear at all in the first three years, but appear at least 20 times in the following 11 years. The output, a dataset of 745 bigrams considered to be potentially new in Danish, are double annotated, and depending on the annotations and the inter-annotator agreement, either discarded or divided into groups of relevant data for further investigation. We then carry out a more thorough lexicographical study of the bigrams in order to determine the degree to which they support the identification of new senses and lead to revised sense inventories for at least one of the words Furthermore we study the relation between the revisions carried out, the annotation values and the degree of inter-annotator agreement. Finally, we compare the resulting updates of the dictionary with Cook et al. (2013), and discuss whether the method might lead to a more consistent way of revising and updating the dictionary in the future.

Keywords: corpus statistics, bigrams, dictionary update, semantic change, Danish

1 INTRODUCTION AND MOTIVATION

The Danish Dictionary (DDO) was originally edited from 1994 to 2003 based on studies of Danish word senses in corpus texts from 1983-1992, in total 40 million tokens (cf. Norling-Christensen and Asmussen, 1998). It was initially published in print 2003-2005 and at the time it described the senses of 66,000 lemmas (cf. Lorentzen, 2004). Since 2009 it has been available online at ordnet.dk/ddo, and in recent years the main focus has been to update it with new lemmas. Today, 25 years after the first editorial work was carried out, the dictionary covers 100,000 lemmas, and time has come to update the earliest edited ones by supplying them with new senses, new fixed expressions, new collocations, and also new citations. After the first published version of the dictionary, this has only been done sporadically, as a result of user suggestions and whenever the lexicographers observed new ways of using a word in the language. When it comes to citations, the dating of these in the dictionary can be used as an indicator since entries with only older ones probably need an update. The editorial staff is currently going through all senses which are only illustrated with a citation from the 1980s. However, presenting more updated citation information would also be relevant in many other cases, but these are hard to find systematically, as are those cases where there is a need for new collocations or even more importantly, for a slightly different sense description or even a new sense, maybe in the form of a fixed expression. Our aim is to be able to supply the current practice building on suggestions from users and editorial observations with a more systematic approach across the whole vocabulary, based on corpus statistics.

2 METHOD

It is a well-established fact that collocational change might indicate sense change (Tahmasebi et al., 2018; Pollak et al. 2019; Traugott, 2017). For instance, Pollak et al. (2019) compare automatically extracted collocations from computer-mediated communication (such as blogs and social networks) with those from a general language reference corpus and discover not only topic/genre-related new words, but also new meanings of previously lexicographically described vocabulary. In contrast to this, the present paper is based on the comparison of sets of automatically extracted collocations from corpora which are similar in composition and genre, but which instead cover

different timespans. We describe a method where the collocational change in these corpora is used as input for lexicographers in their search for new meanings of already included vocabulary in a dictionary. We initially calculate the statistically significant variation in bigrams in a corpus and create a dataset of those that are estimated to be new in Danish texts. Independently of each other, two lexicographers judge whether, at a first glance, the bigrams indicate the need for a semantic revision of the lemmas involved, and if so, should it be 1) in the form of a defined sense or fixed expression, or 2) in the form of a collocation added to an existing sense with no need of explanation? Afterwards, the lemmas represented by the bigrams which were marked as 1) or 2) either by one or both lexicographers are more thoroughly inspected, leading to a revision in the dictionary when required, otherwise not. The judgments of the data are based on a set of internal guidelines to be followed by editors of the dictionary when new lemmas, senses and fixed expressions are to be added.

In this paper, we study and discuss the relation between annotation value (1 or 2), inter-annotator agreement and the final type of update to be carried out. We conclude that especially when the annotators agree that the bigram is semantically relevant, but disagree upon which exact type of semantic change it indicates, we find many new senses. Finally, we compare our findings with Cook et al. (2013).

In the next section we describe the statistical method that we estimate to be suitable for our purpose, as well as the computational creation of the dataset.

3 CREATING THE DATASET

Since 2005, the Society for Danish Language and Literature has collected newswire data of roughly the same size daily. The newswire corpus consists of 20 to 40 million tokens for each year, 512 million running words in all. It consists of articles that are randomly selected from major Danish newspapers each day (due to license restrictions the corpus is not publicly available, but see korpus.dsl.dk/resources.html for other Danish corpora from DSL that are).

The homogeneous data type, the relatively even distribution, and the sufficiently long time-scale make this corpus ideal for investigating our

hypothesis. If lexical data in the form of a token or e.g. a bigram has not occurred at all in the initial period of the text collection, but occurs regularly in the more recent corpus texts, it might indicate that it is a neologism or, in the case of bigrams, either a new expression in the language, or a new way of using one (or more) of the words involved. We have previously used this method to identify potential new single lemmas for DDO, but have never evaluated the method formally. We divided the corpus by year, and selected all tokens which do not appear at all in the first 3 years, 2005-7, but appear frequently during the remaining 11 years. The set of tokens was checked by a lexicographer who removed proper nouns and errors, and now it is used as input to lexicographers in the task of supplying DDO with new lemmas. However, it has not been studied to which degree these lemma candidates do end up being included as new lemmas in the dictionary. This paper describes the same method carried out on bigrams, but takes it a step further. In this case not just one, but two lexicographers check and annotate the output data independently of each other. Furthermore we also check how useful the remaining manually selected part of the data turns out to be when it comes to the concrete task of updating the dictionary, and study the relation between the initial annotations and the usefulness. The updates that we decide upon are either carried out immediately or listed as future tasks in the editorial process of keeping the dictionary up to date.

Once again, we use the corpus text collection divided by year, and now collect all the bigrams which do not appear at all in the first three, but appear with a certain frequency during the next 11 years. Our method is easily reproducible.

1. We calculate the statistically significant bigrams for the complete newswire corpus 2005 - 2018 (~ 512 million tokens), see [3.1] below for details;
2. We divide the corpus into 14 sub-corpora, one for each year;
3. We count the occurrences of the bigrams for each sub-corpus, i.e. each year, separately;
4. We make a dataset of all bigrams that meet the following two requirements:

a. The bigram does not occur in the first three years, 2005, 2006, and 2007, 3 being the lowest number of years that we felt would prevent accidental gaps in the distribution of the bigram.

b. The bigram occurs at least 20 times in the following time period of 11 years,

(--> frequency ~20/400 million = 0.00000005).

The output of the process is a dataset of 745 bigrams considered to be new in Danish. These bigrams are listed and used as input for the manual annotation task.

3.1 Calculating the statistically significant bigrams

In order to calculate the statistically significant bigrams we developed a small Python script using the Phrases module of the Gensim package (Řehůřek and Sojka, 2010; Řehůřek, 2020). We used the so-called *original scorer* algorithm based on the bigram scoring function developed by Mikolov et al. (2013) for calculating the bigrams.

The bigrams are calculated using the formula:

$$score = (count(w_i, w_j) - m) * count(vocab) / count(w_i)*count(w_j)$$

where $count(w_i, w_j)$ is the frequency of the bigram, count(vocab) is the size of the vocabulary, $count(w_i)$ is the frequency of the first word, $count(w_j)$ is the frequency of the second word, and m is the minimum frequency of the bigrams.

We chose the minimum frequency of bigrams to consider (m) to be 5 and we chose the threshold of 7 for significant bigrams. This threshold was chosen based on manual inspection in order to select only the most significant bigrams without letting too much noise into the dataset. This threshold removes arbitrary, ad-hoc bigrams like *nævne nogle* ('mention some', score 3.9) and *skal betale* ('must pay', score 1.2), but keeps wanted bigrams like *offentlig institution* ('public institution', score 8.8) and *monopolagtige tilstande* ('monopoly-like conditions', score 385.0). However, any fixed threshold must of course be expected to give some unfortunate results. In our case we find that some bigrams that are clearly non-collocational are included in the dataset (e.g. *stormer flyet*, 'raid the plane', score 7.3), and some excellent

ones are excluded (e.g. *stor betydning*, 'great importance, score 6.8). We have not investigated the perfect threshold for this experiment, but it is clearly a task we wish to perform.

4 MANUAL ANNOTATION OF THE DATASET

We established the following five questions for the manual annotation task. The categories we chose are closely related to the type of information described in the dictionary which is to be updated with new semantic information.

1. Is the bigram likely to represent a new sense of one of the words, possibly in the form of a fixed expression, to be included in the dictionary?
2. Is it instead more likely to represent a new collocation, both words being transparent in sense?
3. Is the bigram (part of) a proper noun? For example the title of a Danish movie *Den skaldede frisør* (English title: Love is all you need), or a Danish tv-program *Den store bagedyst* (corresponding to the English program: The Great British Bake Off).
4. Is it a grammatical construction, for example *anno 2013* ('in the year 2013'), *arvelovens paragraf (X)* ('section (X) of the Inheritance Act').
5. Is it not at all relevant to include in the dictionary? *Eurozonens tredjestørste* ('the third largest of the Eurozone', *din smartphone* ('your smartphone').

The first 2 categories are particularly important in the semantic update task. In Figure 1, the DDO entry *design* is shown, and here we see how the two categories are used. Category 1 refers to defined senses in the dictionary which can be expressed as either a main sense or subsense (1., 1.a and 1.b in Figure 1), or in the form of a multiword unit where the lemma is included, initiated by the headline 'Faste udtryk' ('Fixed expressions') in the figure illustrated by *intelligent design* ('intelligent design'). Category 2 refers to the use of bigrams (or trigrams) as examples of how the word combines with other words in this sense, e.g. *industrielt design* ('industrial design') and *italiensk design* ('italian design'). We have chosen to call only these example bigrams 'collocations' in this paper. Others use the term 'collocations' differently. In a similar

work, Pollak et al. (2019) use it in a broader sense, corresponding to the entire set of bigrams that they operate with, due to the fact that this only contain noun lemmas and their collocates. They operate with only bigrams containing noun lemmas in the dataset. Only their term 'collocationally new collocations', which is used to define one of the 7 core categories among their initially extracted collocations, correspond to what we call 'collocations'.

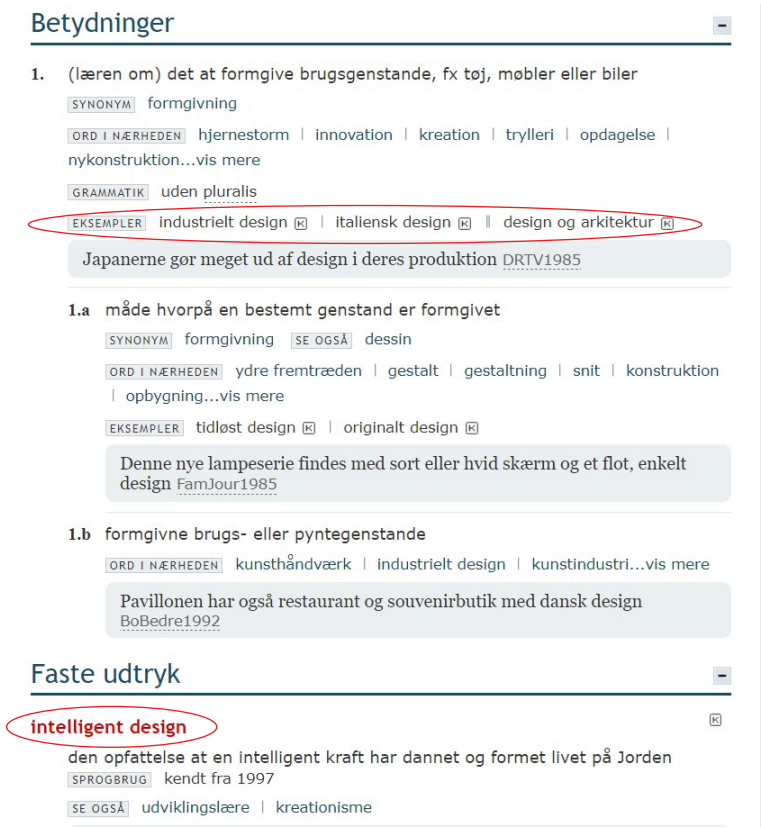


Figure 1: The noun lemma *design* in DDO.

Two of us, both experienced lexicographers, annotated the output of 745 bigrams independently of one another with one of the 5 categories listed above. We both have a good knowledge of the lexical content of the DDO, and are very familiar with the task of updating the dictionary with new lemmas, senses etc. Table 1 shows an extract of one of the two independently annotated lists of bigrams.

Table 1: *The list of bigrams with frequency information and annotation, one annotator*

Bigram	Frequency	Annotation
amerikanske=internetgigant	23	2
amerikanske=jobmarked	32	5
amerikanske=medicinalselskab	57	5
amerikanske=whistleblower	74	5
analyserer=kulturelle	123	5
anbefalinger=fordeler	94	5
andengenerations=bioethanol	32	2
anno=2012	124	4
anno=2013	111	4
anno=2015	113	4
anno=2017	103	4
annoncerede=ordrer	26	5
antisemitiske=hændelser	25	2
anvendte=billedmateriale	422	5
arabiske=forårs	45	1
arabiske=opstande	21	2
arabiske=revolutioner	30	2
arktiske=kyststater	26	2
arktiske=stater	46	2

To compare our annotation task with similar work carried out by Pollak et al. (2019), they instead initially annotated a dataset manually (not double-annotated) in only three categories (p. 190): 'non-relevant data' (corresponding to 4 and 5 in our task), 'proper words and abbreviations' (corresponding to 3 in our task), and finally 'core results', which correspond to our categories 1 and 2. Afterwards the 'core results' in their study were annotated by two linguists (again not double-annotated) into 7 more specific categories, some of which are related to their specific interest in non-standard vocabulary and therefore not relevant to our case. But their 4 categories: 'lexically', 'collocationally', as well as 'semantically new vocabulary', and finally 'terminology', are all covered by the content of our first 2 categories: 'new sense or fixed expression' or 'new collocation'.

Pollak et al. (2019) apparently do not double-annotate the data, and as we shall see, the double annotation is in our case an important part of our method,

and likewise plays an important role in the analysis and conclusions. Nor do Pollak et al. (2019) investigate to which degree the annotated data in each case entails an update in a practical lexicographic project, and what exact type of update that ends up being carried out on the basis of each bigram in the dictionary. Our study allows us to compare on the one hand the annotations and the inter-annotator agreement, on the other hand the different types of resulted updates, and to draw some conclusions based on the combinations.

The output of the annotation task that we carried out – two lists with 745 annotated bigrams – was subsequently compared in order to calculate the inter-annotator agreement. The results are discussed in the next subsection.

4.1 Inter-annotator agreement and relevant data

The overall inter-annotator agreement was 85% in the annotation task described above. However, there was almost 100% agreement between the two lexicographers on whether the data was unlikely to influence the semantic description in the DDO (the categories 3, 4 and 5, covering proper nouns, grammatical constructions or simply not relevant information to include in a dictionary). This data, 1/3 of the statistically significant bigrams, was therefore discarded as non-relevant for further lexicographic inspection, a share which corresponds roughly to the 37,4% of the extracted data which was found irrelevant in the Slovene study (Pollak et al., 2019, p. 191). The high inter-annotator agreement indicates that the task of discarding non-relevant bigrams from the automatically extracted list could probably have been carried out by just one experienced lexicographer.

The bigrams said to belong to either category 1 or 2 by both lexicographers, and thus likely to influence the semantic description of one of the lemmas (or both), constituted 482 bigrams, corresponding to 2/3 of all statistically significant bigrams. These were selected as highly relevant for a more thorough lexicographic inspection.

4.2 Frequency

Our choice of a frequency criteria of 0.00000005 seems suitable for our purpose of finding enough data to initiate a more systematic update process of the dictionary. A large part, namely more than 1/3 of the new bigrams, had a

frequency between 20 and 30 (of 400 million tokens), and most of them, 3/4, had a frequency lower than or equal to 50. If the initial criteria on frequency had been raised from 20 to 50, we would only have obtained 1/4 of the relevant data that was found. It might even pay off to also check bigrams with a frequency between only 10 and 20 in the corpus, since more than a third of the relevant bigrams had 30 or less occurrences.

5 LEXICOGRAPHIC INSPECTION OF THE BIGRAMS AGREED UPON TO BE RELEVANT DATA

Figure 2 illustrates how the 745 statistically significant bigrams are overall distributed in non-relevant and relevant ones as described above and, maybe more importantly, how the relevant 2/3 (482 bigrams) are further divided into three groups: two groups with those where the lexicographers agreed upon the type of semantic update (both chose category 1, or both chose category 2) and one where they disagreed (the one chose category 1, the other chose category 2), or put differently, agreed upon it to be either category 1 or 2 (and not any of the categories 3, 4 or 5).

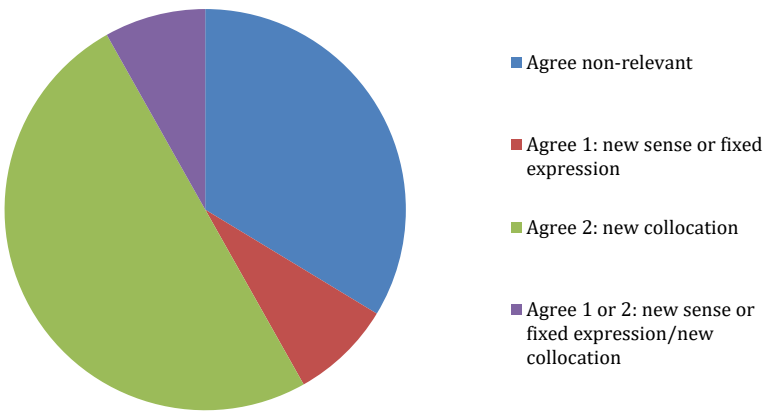


Figure 2: Double annotation of 745 statistically significant bigrams results in 4 groups: one with bigrams agreed upon as being non-relevant, one with bigrams agreed upon to represent 1) a new sense or fixed expression, one with bigrams agreed upon to represent 2) a new collocation, and finally one where the one annotator chose 1) new sense or fixed expression, and the other chose 2) new collocation.

By dividing the relevant bigrams in this way we obtain a distinction between the relatively clear cases (the first two groups where the annotators agreed

upon the type of update) in opposition to the more unclear, albeit relevant cases (the third group where the annotators disagreed on the type of update). Interesting data concerning sense change tends to hide in the unclear data, as we shall see in section 6.3.

Our next step was to thoroughly inspect the bigrams from all three groups with the purpose of updating one or maybe even both lemmas in the dictionary with new semantic information. As an example, the multiword expression *fri fagskole* ('free vocational school', a new type of educational institution in Denmark) was added to the noun entry of *fagskole* ('vocational school') based on the bigram *frie fagskoler* ('free vocational schools'). The collocation *streame musik* ('to stream music') was inserted in the verb entry *streame* ('to stream') based on the identical bigram *streame musik*, and the collocation *nordisk køkken* ('Nordic cuisine') was added to the noun entry of *køkken* ('cuisine') based on the bigram *nordiske køkkens* (genitiv: 'of the Nordic cuisine').

It turned out that the updates would not only consist in a new sense, fixed expression or collocation, but also a slightly changed definition, or an added citation illustrating the bigram. In some cases the lemma was even updated in more ways than one, e.g. the bigram *intelligente løsninger* ('intelligent solutions') entailed both a new collocation as well as a slightly changed definition in the adjective entry *intelligent*, which now includes the new digital and computerized aspect of the sense.

Other bigrams turned out to be of less relevance than originally expected during the initial annotation task when they were more thoroughly inspected. E.g. the bigrams *forbyde burkaer* ('to ban burkas', reflecting a political debate) and *levende myrer* ('live ants', a much debated dish at the famous Danish restaurant, Noma) did not entail any revision of entries in the dictionary, estimated to be connected to very specific former events, and therefore, from a linguistic and lexicographic point of view, less relevant to include in the DDO today.

After having closely studied 189 bigrams and the corresponding two lemmas in the dictionary, we ended up deciding upon 103 semantic updates to be carried out in the dictionary. However, 300 bigrams from the collocation group have not yet been thoroughly analysed, but based on our studies of 1/5 of the

group, we estimate the total amount of bigrams leading to an update to be approx. 41% of all the bigrams annotated to be relevant (category 1 or 2), and thereby 27% of the initial dataset of automatically extracted and calculated bigrams. This will be discussed further in the next section, where we will study the relation between the annotations carried out and the resulting types of updates, and draw conclusions on how to profit in more than one way from the double annotation of the bigrams.

6 THE RELATION BETWEEN TYPE OF ANNOTATION AND TYPE OF RESULTING UPDATE IN THE DICTIONARY

In Table 2, the number of updates (some of which are not yet carried out but listed as future editorial tasks), are presented in relation to the annotated data.

Table 2: *Bigrams divided into three groups depending on inter-annotator agreement*

482 relevant bigrams (of 745 statistically significant bigrams)	**Agree 1**: 55 bigrams. Both annotators agree: new sense or fixed expression	**Agree 2**: 367 bigrams. Both annotators agree: collocation	**Agree 1 or 2**: 60 bigrams One annotator: collocation Another annotator: new sense or fixed expression
Number leading to update	All inspected 49 lead to update	1/5 inspected (a sample of 74 bigrams) 24 lead to update (estimate full set: ~120)	All inspected 30 lead to update

Note. For each group, the number of bigrams leading to an update is given.

The same data is illustrated in Figure 3. When at least one of the annotators estimate the bigram to represent a new sense or new fixed expression, the data very often turns out to be useful in the process of updating previously described lexicographical vocabulary with new semantic information, as illustrated by the first and last columns.

Furthermore, and perhaps quite surprisingly, Figure 3 also clearly shows that when both annotators agree that a bigram constitutes a new collocation, the bigram quite often does not result in any update at all.

Apart from studying the amount of updates made up by the bigrams of each annotation group, it is also interesting to find out what kind of updates the

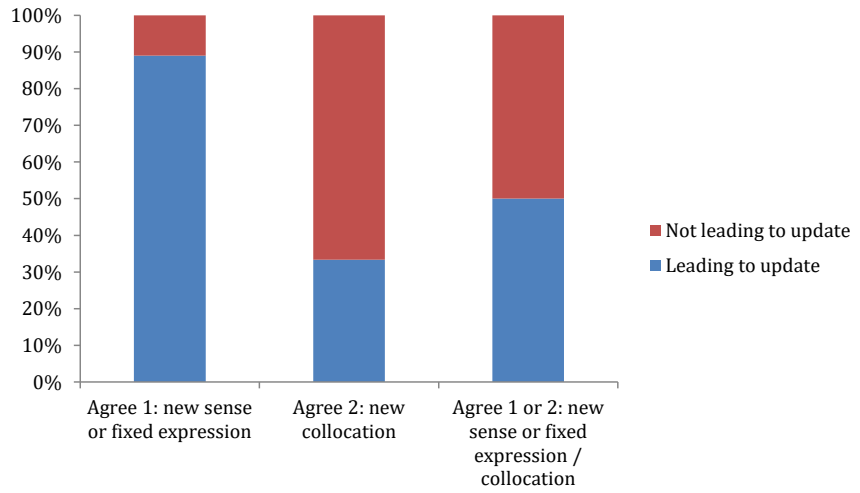


Figure 3: The figure illustrates how often the each of the three groups of relevant bigrams contained data which was useful in the task of updating the dictionary.

three different groups typically entail. Table 3 presents the number of specific updates in relation to the type of annotation.

Table 3: *Bigrams leading to updates and the types of updates that they entailed related to annotations*

Type of annotation leading to update → **Type of update**	**Agree 1**: Both annotators: new sense or fixed expression = 49	**Agree 2**: Both annotators: collocation = 24 of sample (estimation full set ~ 120)	**Agree 1 or 2**: One annotator: collocation. The other annotator: new sense or fixed expression = 30	**Estimated total number of updates** = 200
new lemma	22	2 (full set ~10)	2	34
fixed expression	19	0	8	27
new sense	1	3 (full group ~15)	7	23
changed definition	3	0	4	7
collocation	4	11 (full group ~ 55)	10	69
new citation	0	8 (full group ~40)	0	40

Note. The table also presents the estimated total number of updates entailed by the extracted dataset of bigrams.

We also estimate how many updates the dataset will lead to when the total set of annotated data is thoroughly studied. Around 27% of the automatically

extracted bigrams lead to an update, which constitutes around 41% of the bigrams annotated as relevant for the semantic revision of the dictionary by both lexicographers. A little over 1/3 of the updates take the form of a new collocation in the dictionary, 1/4 take the form of a new senses or fixed expression, equally distributed. 1/5 is in the form of new citations, and almost 1/5 are new lemmas. See Figure 4.

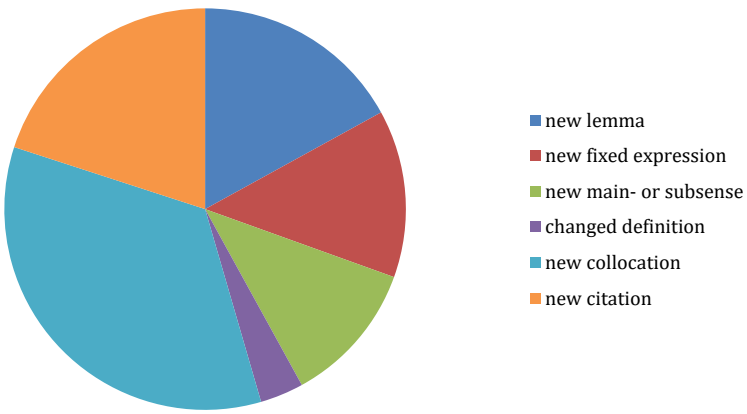


Figure 4: The share of the different types of updates entailed by the information on extracted bigrams.

In the next 3 subsections, we will go into detail with the data from each group.

6.1 Agree 1: Both annotators agree that it is a new sense, maybe in the form of a fixed expression

The two lexicographers agreed that a rather small, but valuable part of the semantically relevant bigrams represented a new sense or fixed expression. Here we find the most useful data when it comes to updating the already included lemmas in the dictionary, since almost all of it leads to revisions when the bigrams and the two corresponding dictionary entries are thoroughly inspected. See Figure 5.

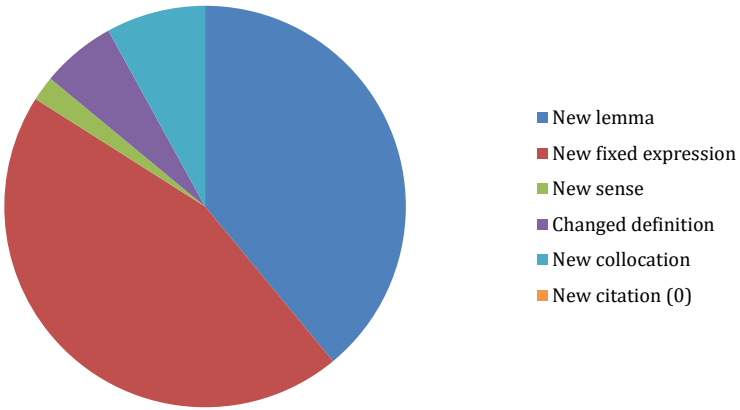


Figure 5: The distribution of different types of semantic updates entailed by the group of bigrams agreed to be a new sense or fixed expression by the two annotators.

Somewhat surprisingly, almost half turned out to constitute new lemmas based on an English multiword expression (e.g. *urban farming*, *augmented reality*). Danish neologisms are highly influenced by English, and loans from multiword expressions are often written in one word when they are included in Danish dictionaries, due to Danish spelling rules (*street food* → *streetfood*, *game changer* → *gamechanger*), if not, simply constituting a lemma entry spelled in two word. Pollak et al. (2019, p. 192) also deal with such loan words from English.

A substantial part of the bigrams in the group leads to a new fixed expression in the dictionary as foreseen by the annotators. In contrast to this, only very few led to the addition of a new main sense or subsense. More frequently they led to a change in existing definitions of the lemmas so that they now include the new phenomena described by the bigram. This was the case of the adjective *præhospital* 'prehospital' (based on the bigram *regionens præhospitale*), and *funktionel* ('functional'), based on the bigram *funktionelle lidelser* ('functional diseases'), see also other examples and a comparison with Cook et al. (2013) in section 7. Another rather small part led to new collocations in the entries. It is worth noticing that only among the bigrams in this group do we find the cases where the semantic information they represent had already been included in the dictionary, discovered during recent editorial work carried, for example due to user suggestions. In fact this goes for 12% of the updates, and most of

them are fixed expressions which apparently attract the attention to a much higher extent than new senses and collocations.

6.2 Agree 2, inter-annotator agreement: collocations

Now we turn to the other part of the relevant bigrams in which the type of update was agreed upon by the two lexicographers, in this case judged to be new collocations by both. This part constitutes the largest group of the relevant data by far, namely ¾ (367 bigrams), and we have not inspected all of them yet. Here we find bigrams like *tørrede tranebær* ('dried cranberries'), *syriske borgerkrig* ('Syrian civil war'), *klimatiske udfordringer* ('climate challenges'), and *brystforstørrende operation* ('breast enlargement surgery'). In our investigation, we have previously only studied one fifth (74 bigrams) in detail, however we estimate this to be a sufficient number to enable us to draw some conclusions. We have compared them with the current lexical description of the two lemmas in the dictionary and also studied the occurrences in the corpora. As seen in Figure 5 above, only one third of the studied ones lead to an update of the dictionary. Many of them turn out to be very topical, time-limited and related to specific political or economic events in recent years. Therefore they are discarded in the final analysis and not integrated in the dictionary. One example of this is the bigram *amerikanske droneangreb* ('American drone strikes').

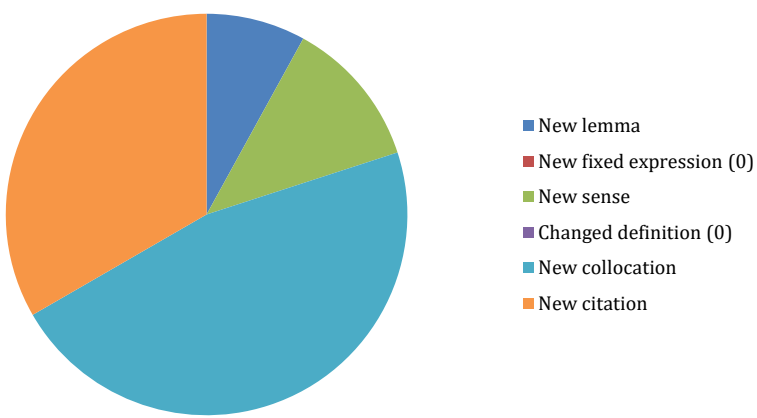


Figure 6: The distribution of updates entailed by the group of bigrams agreed to be collocations (category 2) by the two annotators.

Figure 6 illustrates how those of the category 2 bigrams that did result in an update are distributed when they are to be implemented in the dictionary. Almost half of them are added in the form of a collocation as also foreseen by both lexicographers, i.e. *trådløs opladning* ('wireless charging') which has been added to the adjective *trådløs*, *politiets vagtchef* ('police officer on call') which has been added to the noun *vagtchef* ('officer on call'), *ulovlig overvågning* ('illegal surveillance') which has been added to the noun *overvågning (*'surveillance'), and *kriseramte banker* ('crisis-stricken banks') which has been added to the adjective *kriseramt* ('crisis-stricken'). But Figure 6 also reveals that quite a lot of the bigrams that were estimated to be collocations in the first place instead have led to the adding of a new citation representing the bigram. It is worth noticing that only this group of bigrams (agreed upon to be collocations by both lexicographers) leads to this type of update in the dictionary. This suggests the future use of the same method in the task of updating citations in the dictionary, as a supplement to the criteria we use at the moment where we only look at entries with old citations from specific magazines. Another interesting fact about the updates based on the collocation group is that none of the data had already been discovered and included in the dictionary by other editors in the period since the bigrams were extracted for our experiments, indicating that this type of information, which is in fact highly needed in order to keep the dictionary content up to date at a more general level, would probably have been overlooked without the statistical investigation of bigrams.

However, the group of collocations also contains the highest amount of inapplicable data. It contains a lot of time-limited bigrams which according to the editorial guidelines of the DDO are not relevant to include in the dictionary. This is due to the fact that we are dealing with bigrams extracted mainly from newspapers. From a structural point of view, they are of course typical collocations: adjective + noun, verb + object etc., which is also why the two lexicographers easily agreed upon their status as such at first hand, but from a more pragmatic point of view they are not, and we should probably have been aware of this problem from the beginning. We can also conclude that very few bigrams in this group led the lexicographers on the track of new senses or new lemmas. One rare example is the loanword *big data* based on the English

multiword expression. The lemma *data* is already part of the DDO which is why both lexicographers annotated it as a new collocation. However, since it is a term and a direct new loan pronounced in English it has instead to be included at lemma level in the dictionary.

6.3 Agree 1 or 2: inter-annotator disagreement whether it is a collocation or rather a new sense, maybe in the form of a fixed expression

The third and last part of the data selected for further lexicographic inspection consists of 60 bigrams that the two lexicographers agreed to be highly relevant. They disagreed, however, upon how to include them in the dictionary structure. While one annotator estimated that the bigram was most likely to represent a new sense or fixed expression, the other believed that it was more likely to represent a new collocation. In fact, only half of the bigrams in this group entailed a dictionary update. See Figure 7 for the distribution of the different types of updates.

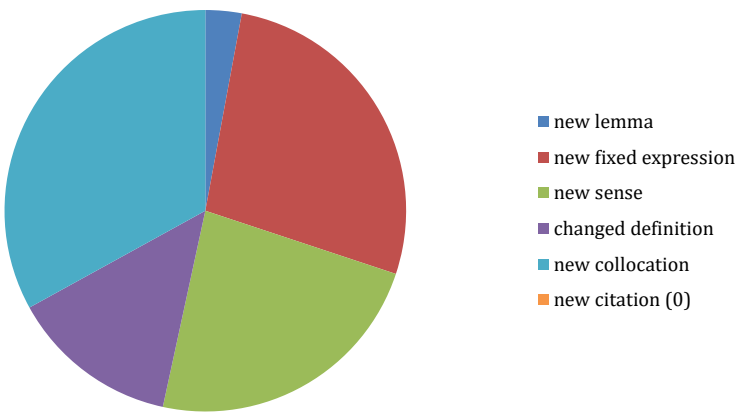


Figure 7: The distribution of updates entailed by the bigrams agreed to be relevant. However the annotators disagreed upon whether the bigram represented a new sense or fixed expression, or rather a collocation.

The vast majority of those which entailed an update did so in the form that was suggested by either one or the other annotator, more or less equally distributed. For the first time, we find quite a lot of new senses and not only fixed expressions. One third of the bigrams were included as collocations (e.g. *bæredygtig omstilling* ('sustainable conversion', *mentalt helbred* ('mental health')),

almost another third as a fixed expression (*bibelske dimensioner* ('biblical proportions'), *pædagogiske assistenter* ('teaching assistents', new job title)), and, particularly interesting, one quarter in the form of a new main sense or subsense. E.g. the new subsense of the noun *boble* ('bubble') discovered from the bigram *glas bobler* (lit. 'glass of bubbles' – i.e. 'a glass of sparkling wine, e.g. champagne') was included in the dictionary, and the adjective *mobil* ('mobile') is planned to be provided with a new sense triggered by the bigrams *mobile bredbånd* and *mobilt internet* ('broadband/internet via a cellular phone').

Some of the bigrams will result in several changes. In the case of the new concept *selvkørende bil* ('self-driving car') which is also a part of the new data described in Pollak et al. (2019, p. 193), the definition of the adjective entry *selvkørende* needs to be changed in DDO, as does the entry of *bil* ('car'). The entry will be extended with a new fixed expression with its own definition.

It is worth noticing that this group of bigrams is the one reveals the largest amount of new senses by far. Several bigrams lead to the inclusion of a new main sense or subsense in the dictionary. Many also entail the need of a changed definition for one of the lemmas. For instance, a revision of the definition of *digital* ('digital') is needed due to the bigram *digital dannelse* ('digital code of conduct/digital education'), likewise a revision of the definition of *cannabis* ('cannabis; marijuana') was needed due to the bigram *medicinsk cannabis* ('medicinal marijuana'). We also found one new lemma in the group, the adjective *æresrelateret* ('honor-related'), due to the bigram *æresrelaterede konflikter* ('honor-related conflicts'). This lemma would also be discovered by single lemma extraction methods, but since it very often occurs together with *konflikter* in our data, this should be added as collocational information when the new lemma is included and edited.

Among the discarded data in the group were bigrams that had only been frequent for a short period of time (based on the study of the occurrences in our corpus), others were considered to be terminology which is not suitable for inclusion in the dictionary. As in the case of the agreed collocations, it's worth noticing that no lexical information discovered from our study of this group of bigrams had been registered in the dictionary by other editors since the data was extracted, and it would probably have been hard to discover without the use of statistical methods.

6.4 Conclusions on annotation and resulting updates

Our computational measure of the appearance of new bigrams in homogenous newswire corpora combined with double annotations of the output dataset and the entailed updates of the dictionary allow us to draw a number of conclusions.

6.4.1 How useful was the automatically calculated dataset?

First of all, we can conclude that quite a lot, i.e. approx. 1/4, of the automatically extracted dataset leads (or will lead) to a resulting update in the dictionary, while 3/4 do not. In comparison, Pollak et al. (2019) find a little less "lexically, collocationally, or semantically new data that can be considered in the process of updating existing lexical resources for Slovene" (p. 197), namely 21.6%. The initial annotation by two lexicographers made it possible to discard many bigrams in the extracted dataset in an efficient and not very time-consuming way. The data that the lexicographers selected as most likely to be relevant turned out to be useful when more thoroughly inspected and compared to the content of the dictionary entries in almost half of the cases. Had the initial annotation task been carried out on the basis of more detailed and elaborated guidelines, we could probably have avoided even more 'noise' (bigrams not leading to any updates after all), for example the many time-limited bigrams. The automatic extraction of the bigrams can maybe also be tuned in a way so that such time-limited data is better avoided in the first place, and not even included in the output dataset. Pollak et al. (2019) also propose that the automatic extraction procedure should include language recognition in the preprocessing step in order to identify and remove the English bigrams from the list. However, this would entail that several new loan words would not have been discovered and included in the DDO.

6.4.2 New lemmas

We found far more lemma candidates in the dataset than expected, namely 4%, due to the fact that many English multiword expressions are to be integrated in the dictionary at lemma level. This is in line with the results of Pollak et al. (2019).

6.4.3 Fixed expressions

A little over 4% of the initial dataset ended up being included in the dictionary in the form of fixed expressions. They constitute 14% of the updates carried out. From our investigations, we can see that when a bigram is recognized by

two lexicographers as a fixed expression, it very often holds true, and it almost surely will influence the semantic description of one or both lemmas that are part of the bigram in one way or another. Very few bigrams that had been annotated as a fixed expression by both lexicographers led to no update at all, so if you want to make sure you find relevant data for the updating task of a dictionary, then this a way to go. Furthermore we can conclude that when two lexicographers agree that a bigram is not a fixed expression but rather a collocation, we can also be sure that it is not. Fixed expressions also seem to be the easiest to discover without applying any systematic method, since around 1/6 of them had already recently been included in the dictionary.

6.4.4 New main senses and subsenses

We found quite a lot of new senses via the dataset. Around 3% of the automatically extracted bigrams led us to this information, and among the annotated relevant data one in every 20 bigrams revealed a new sense. Pollak et al. (2019) find a bit more (4.9% of the extracted data), but they state that many are found in non-standard colloquial language (p. 193), which might explain the higher amount – this type of language is not included in our corpus texts. Due to the method of double annotation, we discovered that new senses tend to hide between the more ambiguous data where the lexicographer is not so sure whether the bigram represents a sense or a fixed expression that needs to be explained to the dictionary user, or whether it is rather a collocation with transparent meanings of both words. However, new senses can also be found among bigrams which when presented to the lexicographers in the first place, were estimated to be merely collocations of already included senses in the dictionary. In contrast, new fixed expressions were in fact found only when both annotators estimated the bigram to be either a new sense or a fixed expression.

6.4.5 Collocations

Bigrams resulting in updates in the form of a collocation constitute 9% of the extracted data, and almost half of those that were annotated as category 2 by both lexicographers, also turned out to lead to a new collocation in the dictionary. Thereby they constitute the cases in which inter-annotator agreement is very high and at the same time they most often corresponded to the type of resulting update Pollak et al. (2019) find a higher percentage of ‘collocationally

new collocations' in their extracted data (13.3%, p. 193), but the many collocations that we chose not to include in the dictionary after a more thorough investigation probably explains the difference. In contrast to the DDO update guidelines, Pollak et al. (2019) propose that such data should not necessarily be left out of dictionaries: "trending vocabulary that is often bound to specific political and social events", should instead be included in digital dictionaries. They advocate for "a faster and more fluid lexicography that focuses not only on the stable and established, but also on the changeable and variable aspects of language – which is where language users often need assistance" (p. 200). We find that the inclusion of such data would probably entail an ongoing and maybe time-consuming control with the already lexicographically described vocabulary in the DDO in order to be sure to avoid lexical information that has become outdated.

Since two thirds of the collocation bigrams did not lead to any updates, we can conclude that when two lexicographers independently of one another agree that a bigram is a collocation, it is much less likely to represent useful data for the semantic update of a dictionary than if at least one of them consider it a new sense or fixed expression as described above.

6.4.5 Citations

Many collocations were included in the form of a citation when the data was thoroughly inspected, and we are in fact pleased to have discovered a more systematic way of updating this part of the dictionary information across lemmas.

7 RESULTS COMPARED WITH PREVIOUS RESEARCH

In this section we compare our study with a similar project presented by Cook et al. (2013). They used a reference corpus from 1995 and a focus corpus from 2008 to identify new elements to be included in an English learner's dictionary (Macmillan). In their paper, they use three categories:

1. the uninteresting findings, which are mostly due to the many news stories in the corpus; certain items exhibit a sudden spike and then they disappear and never turn up again; one example of this is the word *junta* referring to the regime in Myanmar that would not accept humanitarian help from the outside world after a disastrous cyclone that caused

many deaths; another example is the word *candy* that popped up because some Chinese candy had been contaminated with melamine;

2. much more interesting are the cases where a dictionary entry should be changed in some way, it needs 'tweaking'; for instance the existing entry for *cleric*, which only referred to clerics typical of the Church of England, but in the 2008 corpus, clerics are often Muslim and this should be reflected in the entry; the example *video* is obvious: in the 1990s a video would be a video tape of the VHS type, but nowadays it is typically a digital recording of images and sounds distributed via online media;

3. the third category is cases where new senses should be included in specific entries in the dictionary, for instance the verb *to search* (= 'do a web search'), and *text* as in *text messaging*, *send someone a text* or *text someone*, a technology that was not yet available in 1995.

Let us take a look at our findings using more or less the same categories as Cook et al. (2013) We have a high number of irrelevant findings, which we first categorized as collocations without deciding if they would lead to an actual change in the entries for the two words (cf. Section 6.2). The high amount of newspaper texts in our corpus accounts for findings related to specific events and political discussions; *tibetansk flag* ('Tibetan flag') for instance refers to a demonstration where Danish police unlawfully removed a Tibetan flag so that it would not be seen by the Chinese president who was visiting Copenhagen.

As is the case for Cook et al. (2013) we have changed (tweaked) several dictionary entries, for instance *cannabis*, where the collocation *medicinsk cannabis* ('medicinal marijuana') shows that cannabis may also be used for medical purposes nowadays; or *intelligente løsninger* ('intelligent solutions'), which indicates a new nuance in the meaning of *intelligent* involving digital functions and computers - so this has been added to the definition (cf. Section 6.3).

The entirely new senses include the word *digital*; the current entry describes the situation in the 1980s and 1990s when you would distinguish between a digital watch and an analogue one; of course, this is not up to date and the entry *digital* needs a new sense that will account for collocations like *digitale indfødte* ('digital natives') and *digital mail*.

A fourth category not mentioned by Cook et al. (2013) is new fixed expressions. As mentioned in section 5.4 this category is very salient in the list of bigrams and we have decided to include several of these. The most significant one is probably *sociale medier* ('social media'), which had already been discovered by other methods and added to the dictionary; other interesting examples are *assisteret reproduktion* ('assisted reproduction'), *cirkulær økonomi* ('circular economy') and *brændende platform* ('burning platform', i.e. a difficult situation that urgently needs taking care of); the expression refers to a fire on an oil platform in 1988 which resulted in many deaths.

A fifth category contains new lemma candidates, mostly of English origin; many of the English bigrams in the list may be included in our dictionary, either as headwords consisting of two words (*pulled pork*) or as a solid compound like *komfortzone* ('comfort zone' in English); even a pragmatic phrase like *oh, my god* and its abbreviation *omg* are lemma candidates if you take into account how common the phrase has become in everyday Danish, and the same goes for other English phrases that have been included in the DDO in recent years, such as *you name it*, *whatever*, and *take it or leave it*.

8 FINAL CONCLUSIONS AND PERSPECTIVES

In this final section we make a brief evaluation of our study: what are the overall pros and cons of this method and of our approach? On the upside, it provides the editors of the DDO with very useful input for updating senses, definitions, collocations, etc. In fact, the editors are so happy with it that the plan is to repeat the bigram calculation regularly, for instance every three years. It is also very encouraging that the material supports updates that have already been made - quite reassuring for a corpus-based dictionary. The material is a necessary supplement to other methods used by the dictionary editors to keep track of lexical and semantic change, like user suggestions, other corpus-linguistic data and good old editorial observations since it guarantees a systematic check across the entire vocabulary.

A drawback, of course, is that manual filtering is indispensable, but the good news is that one experienced lexicographer can fulfill the first phase (discarding non-relevant bigrams), whereas it takes two (or more) lexicographers to annotate the rest reliably and eventually make the actual changes in the

dictionary. An important lesson from the experience is that a very large proportion of the bigrams consists of topical (time-limited) examples, which is due to the composition of the corpus (mostly newspaper material). Other types of corpus texts are too scarce for the time being, and this is a task that the dictionary staff intends to work on in the future, keeping in mind, however, that a homogeneous data type as well as an even distribution of text types over time is absolutely necessary in order to obtain good results with the statistical method that we have described in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions and careful reading of the manuscript. We would also like to thank our colleague Jonas Jensen for useful feedback and for proofreading the article.

REFERENCES

Dictionaries

DDO = *Den Danske Ordbog* [The Danish Dictionary]. Retrieved from https://ordnet.dk/ddo (17. 2. 2020)

Macmillan = *Macmillan English Dictionary*. Retrieved from https://www.macmillandictionary.com/ (17. 2. 2020)

Corpora

Korpus.dsl.dk = *Language Technology Resources for Danish*. Retrieved from https://korpus.dsl.dk/resources.html

Other

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., & Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference* (pp. 49–65). Tallinn, Estonia.

Lorentzen, H. (2004). The Danish Dictionary at large: Presentation, Problems and Perspectives. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 285–294). Lorient, France.

Mikolov, T., Sutskever, I, Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems 26*. Retrieved from https://arxiv.org/abs/1310.4546

Norling-Christensen, O., & Asmussen, J. (1998). The Corpus of The Danish Dictionary. *Lexikos (Afrilex Series) 8*, 223–242.

Pollak, S., Gantar, P., & Arhar Holdt, Š. (2019). What's New on the Internetz? Extraction and Lexical Categorization of Collocations in Computer-Mediated Slovene. In *International Journal of Lexicography, 32*(2), 184–206.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46–50). Valletta, Malta: University of Malta.

Řehůřek, R. (2020). *models.phrases – Phrase (collocation) detection*. Retrieved from https://radimrehurek.com/gensim/models/phrases.html (17. 2. 2020)

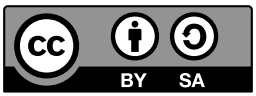
Tahmasebi, N., Borin, L., & Jatowt, A. (2018). Survey of Computational Approaches to Lexical Semantic Change [Preprint at ArXiv 2018]. Retrieved from https://arxiv.org/abs/1811.06278

Traugott, E. C. (2017). *Semantic Change*. Oxford Research Encyclopedias [Online publication]. doi: 10.1093/acrefore/9780199384655.013.323

POSODABLJANJE SLOVARJA: PREPOZNAVANJE SEMANTIČNIH SPREMEMB NA PODLAGI DIAHRONIH SPREMEMB BIGRAMOV

V prispevku preizkusimo metodo sistematičnega posodabljanja Danskega enojezičnega slovarja z novimi semantičnimi podatki o obstoječih lemah. Metoda temelji na hipotezi, da so diahrone spremembe bigramov v korpusnih podatkih lahko pokazatelj sprememb pomena ene od besed v bigramu. Pri metodi kombiniramo korpusno statistiko z ročnim označevanjem. V prvem koraku izmerimo kolokacijske spremembe v homogenem korpusu novic za 14-letno obdobje (2005 do 2018), tako da izračunamo vse statistično pomembne bigrame. Te bigrame potem preverimo v novi različici korpusa, razdeljenega na podkorpuse, pri čemer vsak podkorpus zajema obdobje enega leta. Nato izluščimo vse bigrame, ki se nikoli ne pojavijo v prvih treh letih, se pa pojavijo vsaj 20-krat v naslednjih 11 letih. Na podlagi tega postopka dobljenih 745 bigramov, ki jih obravnavamo kot potencialno nove v danskem jeziku, označita dva označevalca. Bigrami so glede na rezultate označevanja in ujemanje označevalcev bodisi izločeni bodisi razvrščeni v skupine glede na relevantnost za nadaljnjo obravnavo. Sledi temeljitejša leksikografska analiza, s katero določimo, do kakšne mere gre za nove pomene besed in posledično potrebo po spremembi pomenske členitve pri vsaj eni od besed v bigramu. Poleg tega analiziramo tudi povezavo med potrebnimi popravki, oznakami in odstotkom ujemanja označevalcev. V zadnjem delu prispevka primerjamo slovarske posodobitve s pristopom, ki so ga izvedli Cook idr. (2013), in podamo razmisleke o tem, ali tovrstna metoda lahko predstavlja doslednejše popravljanje in dopolnjevanje slovarskih gesel.

Ključne besede: korpusna statistika, bigrami, posodabljanje slovarja, semantične spremembe, danski jezik



A COMPARISON OF COLLOCATIONS AND WORD ASSOCIATIONS IN ESTONIAN FROM THE PERSPECTIVE OF PARTS OF SPEECH

Ene VAINIK, Maria TUULIK, Kristina KOPPEL
Institute of the Estonian Language

Vainik, E., Tuulik, M., Koppel, K. (2020): A Comparison of collocations and word associations in Estonian from the perspective of parts of speech. Slovenščina 2.0, 8(2): 139–167

DOI: https://doi.org/10.4312/slo2.0.2020.2.139-167

The paper provides a comparative study of the collocational and associative structures in Estonian with respect to the role of parts of speech. The lists of collocations and associations of an equal set of nouns, verbs and adjectives, originating from the respective dictionaries, is analysed to find both the range of coincidences and differences. The results show a moderate overlap, among which the biggest overlap occurs in the range of the adjectival associates and collocates. There is an overall prevalence for nouns appearing among the associated and collocated items. The coincidental sets of relations are tentatively explained by the influence of grammatical relations i.e. the patterns of local grammar binding together the collocations and motivating the associations. The results are discussed with respect to the possible reasons causing the associations-collocations mismatch and in relation to the application of these findings in the fields of lexicography and second language acquisition.

Keywords: collocations, associations, parts of speech, lexicography, Estonian language

1 INTRODUCTION

Both the terms *collocation* and *word association* designate an implicit bond between words[1]. Whether the collocations and associations are basically the same or represent different kinds of lexical and/or mental organisation is a question that has intrigued researchers for some time already (for an overview see Deyne and Storms, 2015). In the present paper we do not intend to answer the question theoretically and once and for all but aim to bring forth the tendencies that occur in the Estonian language in that regard. The existing literature about comparisons of associations and collocations covers data of Indo-European languages so far (mostly English, see overview in Kang, 2018; German as in Shulte im Walde et al, 2008; and Russian as in Sinopalnikova, 2004). Some evidence from genetically different language groups would hopefully bring more insights into the field. We take the advantage of having two relevant data sources published by the Institute of the Estonian Language in 2019; the Dictionary of Estonian Word Associations (DEWA)[2] and the Estonian Collocations Dictionary (ECD)[3]. On this basis we aim to provide a systematic comparison of the collocations and associations, also by paying special attention to the parts of speech (PoS).

PoS analysis is relevant because of two reasons. Firstly, Estonian is a Finno-Ugric language that belongs to the agglutinating-flective typological class. The PoS categorisation in Estonian relies on multiple factors: semantics, morphological inflection, syntactic behaviour and pragmatics (Paulsen et al., 2019). Estonian is characterised by well-formed morphosyntactic structure, among other features. This implies that a word's behaviour in speech (and text) is expected to be predetermined by its implicit PoS, which can further affect the structure of collocations derived from the texts. To which extent the word associations retrieved from memory follow the determined-by-the-PoS structure of text production is an interesting question. Secondly, there is a

1 By the term *word association* we refer to a concept used in applied linguistics and psycholinguistics (e.g. Deyne and Storms, 2015; Fitzpatrick et al., 2015). We do not use *word association* in the general sense of the term that would cover also patterns of relatedness of the words in text (e.g. Church and Hanks, 1990).

2 http://www.eki.ee/dict/assotsiatsioonid/

3 http://www.eki.ee/dict/kol/, collocations are also presented in https://sonaveeb.ee/ (Koppel et al., 2019a).

tradition of classifying word associations according to their PoS homogeneity/heterogeneity principle, which has also been applied to the Estonian data (Toim, 1980). Thus, the PoS categories are expected to affect both the collocational and the associative structure of Estonian.

We assume that the Estonian data can contribute to the overall theoretical discussion by elaborating the role that PoS play in the formation of implicit bonds that the collocations and word associations tend to explicate. We consider that there is also some practical importance to elaborating the overlap vs non-overlap of collocations and word associations. So far, the practical interest in the topic has relied on the expectation that the (relatively low-cost) procedures of text mining for collocates would replace the high-cost psycholinguistic testing needed for establishing the relations comprising the mental lexicon (see, e.g. the Word Association Network[4] or Church and Hanks, 1990). We propose applicability also in the fields of lexicography and language teaching.

In this paper we will give a brief theoretical background, introduce the principles of material selection and carry out a systematic comparison of associations and collocations, paying special attention to the role of PoS categories. The paper ends with a discussion about the reasons of the mismatch between collocations and associations in our data and about applicability of the results.

2 COLLOCATIONS AND ASSOCIATIONS

We refer to *collocation* as a frequent and meaningful combination of content words with other lexical and grammatical units (see, e.g. Firth, 1957). As such, collocations can be detected by computational analysis of a large text corpus by means of corpus query systems (CQS), one of which is Sketch Engine (Kilgarriff et al., 2004; Kilgarriff et al., 2014)—a CQS widely used among lexicographers in Europe. For automatic extraction of the ECD database (Kallas et al., 2015), the Sketch Engine function Word Sketch (Kilgarriff et al., 2010; Kallas, 2013) was used. Word Sketch is a one-page summary of a word's grammatical and collocational behaviour, and it displays collocations of a given keyword (or a node), grouped together according to their grammatical relation (e.g. adjectives as modifiers).

4 Retrieved from https://wordassociations.net/en/about (24. 11. 2019)

Collocation has a structure of a *node* and its *collocate*. Nodes refer to the words that are being looked at (e.g. *dog*) and collocates refer to words with which they form collocations (e.g. *barks → dog barks*; *bites → dog bites*; *friendly → friendly dog*) (see Sinclair, 1966; Roth, 2013). Any given node occurs in a number of collocations and has a number of collocates. The role of node vs. collocate depends on the perspective. For example, looking from the perspective of the noun *dog* as a node, the dog can *bark*, *bite* and *sniff;* looking from the perspective of the verb *bite* as a node, the *dog* acts as a collocate, as also *bugs*, *mosquitoes* and *spiders*.

We refer to *word association* in the psycholinguistic sense of the term. The notion originates in the context of testing people (WAT[5]) for their first and spontaneous responses to a range of verbal stimuli (for the origins of the method, see Galton, 1879; Jung, 1910; for the peak of popularity see e.g. Rosenzweig, 1961; Kiss et al., 1973; Postman and Keppel, 1970; Deese, 1965, and for current understanding see e.g. Nelson et al., 2000, and Deyne and Storms, 2015). The word association can be, thus, defined as a person's lexical response to a lexical stimulus, e.g. if one says *cat* the reply might be *dog*, or if the stimulus would be *bread* the response could be *butter*. *Stimulus* and *response* are the basic structural components of word association.

The responses may vary over the respondents (e.g. *bread* may evoke *butter* but also *breakfast* etc.). Thus, one stimulus can have a list of responses and the same response can occur with a number of stimuli (e.g. *bank→money* and *to waste→money*). The collections of responses summed up over a number of respondents (at least one hundred, usually) and elicited to a certain range of stimuli are called *association norms* (see e.g. Kent et al., 1910; Postman and Keppel, 1970; Nelson et al., 2004; Schulte im Walde and Borgwaldt, 2015).

The idea to compare the set of recurrent collocates of a word in texts (i.e. in actual usage) with the same word's associations elicited in the psycholinguistic tests (i.e. revealing the structure of memory) is not new (see De Deyne and Storms, 2015, for an overview). Despite the fact that the comparative research into collocations and associations has shown somewhat controversial results (De Deyne and Storms, 2015; Kang, 2018), a general agreement holds about

5 WAT is an abbreviation for Word Association Test, see https://dictionary.apa.org/word-association-test (14. 4. 2020).

the moderate overlap of the two (e.g. Fitzpatrick, 2007; Durrant and Doherty, 2010). It is difficult to provide a general quantitative measure because of the variation in the methodologies and in the statistics used (Kang, 2018).

One of the variables affecting the outcome of the comparison seems to be the inclusiveness of the lists of associations and collocations. The longer the span of text from which the collocations are extracted (e.g. in Kang's (2008) study the span is one paragraph, in Schulte im Walde et al. (2008) ±20 words), the longer the list of collocations and the greater the probability of coincidence with some of the salient associations. Thus, a limit set upon the data may restrict the probability of discovering the coincident pairs. For example Scott and Tribble (2006) searched for the matching pairs among the ten strongest associations and hundred first collocations of a keyword—a fact that might have reduced the outcome. Mollin (2009), on the other hand, strived for maximum-size inclusivity and compared the full range of associations of 30 randomly chosen keywords from EAT[6] with their collocations in BNC[7] (100 million words). Despite the inclusiveness of data (20,003 pairs altogether), only 626 (3%) were found to be common to both datasets.

It has been proposed that the partly controversial results of previous studies that compare collocations and associations may be due to the fact that collocations were misleadingly considered as emerging from the texts being treated as »a bag of words« (De Deyne and Storms, 2015), i.e. by ignoring the grammatical relations and syntactic structures that give the flow of language its natural texture. On the other hand, the previous studies have reached the conclusion that "...the word association task, as a special method of elicitation, is not of the same kind as the natural task of language production..." (Mollin 2009, p. 197) and hence the difference between associations and collocations.

A closer look at the structures represented by collocations and associations is a question of qualitative analysis. In that respect, word associations—if not mere clangs—have been interpreted traditionally as either belonging to a paradigmatic or syntagmatic class of relations (see e.g. Fitzpatrick, 2007; De

6 The Edinburgh Associative Thesaurus (see Kiss et al., 1973).

7 See Leech and Smith (2000).

Deyne and Storms, 2015). An example of a paradigmatic relation would be *red* (stimulus) → *blue* (response). They are both members of the category 'colour terms' and are cohyponymous with each other. Both are adjectives and could be substituted with each other in a text with no grammatical inconsistency because they occur in the same syntactic role (attribute). The relations of synonymy and antonymy are other typical members of the class of paradigmatic relations. An example of the syntagmatic relation would be *red* (stimulus) → *umbrella* (response). In this case, the stimulus is an adjective, and the response is a noun. The relation attributes the quality designated by the adjective to the thing designated by the noun. There is no way to substitute the two with each other in the text; they form a noun phrase together, whereas their syntactic roles are different (attribute and head noun).

Collocations are extracted from the running flow of text and represent, supposedly, syntagmatic rather than paradigmatic relations. The latter can occur in the flow of text, exceptionally, in the case of coordinated constituents (like listings of the members of the same category or pairs of equal and/or alternative constituents).

Theoretically, thus, we can expect some similarities in the qualitative structure of the collocations and associations to occur too. Homogeneity versus heterogeneity (in terms of PoS) of the relations can be a revealing factor in this respect.

3 THE STUDY

Collocations and associations are similar by structure as pairs of words despite the difference in their origin (corpus query procedures versus psycholinguistic testing). Both collocations and associations consist of two structural members and asymmetry laid upon them: one of the two members that is in focus as a keyword is always an »access member« (AM) and the other is the »related member« (RM). These two are called »stimulus« and »response« in the case of word associations and »node« and »collocate« in the case of collocations (See Figure 1). In present analysis we will use the term *access member* (AM) to refer both to the stimuli (of associations) and nodes (of collocations). We use the term *related member* (RM) both in case of referring to responses (of associations) and to the collocates (of collocations).

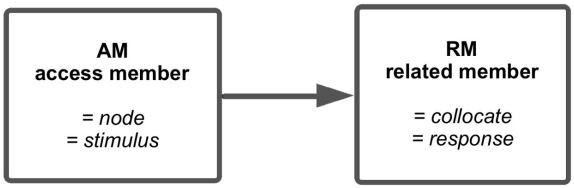


Figure 1: The common structure of collocations and associations.

The goal of the study is to carry out a systematic comparison of collocations and associations in Estonian and to outline the role of PoS. Our expectations, resulting from the theoretical background, contain both quantitative and qualitative aspects and are as follows:

i) Relying on the studies of other languages, we expect an overlap in the range of collocations and associations. We are interested in the proportion of that overlap and whether there are differences with respect to PoS (nouns, adjectives and verbs). For example, is there a combination of PoS that is particularly favoured among the overlapping pairs?

ii) We expect that syntagmatic relations prevail in the case of collocations and that paradigmatic relations make the most of the associations, while we do not know what to expect concerning the intersection of the two. We intend to discover the role of grammatical relations in the overlap.

iii) We assume that the RMs with top positions in the ranking will dominate among the common pairs while the non-overlapping pairs will include RMs with a relatively low ranking. We are interested in whether this holds for all PoS.

3.1 Material and method

As mentioned in the Introduction, we rely on the newest and best organized data available: the Estonian Collocations Dictionary (ECD) and the Dictionary of Estonian Word Associations (DEWA). The dictionaries represent, respectively, collocations extracted from the latest available text corpus (see Kallas et al., 2015, for how the database was generated) and the latest and topical associations gathered (Vainik, 2018). More detailed description of the data sources is presented in Table 1.

Table 1: *Overview of the two data sources*

Dictionaries	DEWA	ECD
General description	Monolingual online dictionary for general public, compiled in 2016-2018	Monolingual online dictionary for (advanced) learners, compiled in 2014—2018
Coverage	1,300 headwords (stimuli), 300 responses per stimulus on average, No of recurring pairs 37,602	9500 headwords, No. of collocations 300,887
Organization of material	The responses are listed according to their decreasing frequency	Collocations are listed according to their decreasing corpus frequency and grouped by collocate's PoS
Distribution of AMs by PoS	Nouns: 68%, Adjectives: 13%, Verbs: 6.3%, Other: 11.7%	Nouns: 64%, Adjectives 16%, Verbs 17%, Adverbs 3%
Presentation mode of AMs and RMs	Base forms: nouns and adjectives in the nominative singular case, verbs in *ma*-infinitive	As lemmas or in their most frequent grammatical form
Method of compilation	A citizen science project with more than 400 participants. See description in Vainik (2018)	Semi-automatic; using Sketch Engine for the extraction of collocations from the Estonian National Corpus 2013 (463 million words)

In ECD, the node (AM) and the collocate (RM) are presented as lemmas (e.g. *sõbralik koer* (friendly-ADJ-SG-NOM dog-SG-NOM) 'friendly dog') or in a particular inflectional word form (e.g. *koer haugub* ('dog-SG-NOM barks-PERS-PRS-IND-SG3-AFF') 'dog barks'), showing the collocations in their correct grammatical form. In the database of ECD, however, the base forms of both the AM and RM are also available. This makes the systematic comparison of the two data sources possible.

In both of the databases, the AMs and RMs are accompanied by their PoS-tags and statistics about the frequency and salience (ECD) / strength (DEWA) of the connection. These pairs of AM and RM are the main object of comparison in this study. Additional information is available about the grammatical relations in the ECD. These relations are a product of the corpus query system Sketch Engine in which a grammatical relation represents a category that displays collocates with the same relation to the search word (e.g. modifiers of a noun or objects of a verb) (see Kallas, 2013, for more details).

The coverage of the two sources differs almost ten times with respect to the number of AMs. The overlap of keywords in two dictionaries is 1102, which makes 11.6% of ECD and 85% of DEWA. For the purpose of the study we made a selection that contains 90 AMs present in both dictionaries and is balanced in two ways: by PoS and by corpus frequency[8]. The procedures were as follows: the list of shared keywords was ranked according to decreasing frequency, and equal proportions (N = 10) of adjectives, nouns and verbs were retrieved from the top, from the bottom and from around the middle of the frequency list. This step was taken in order to avoid the possible side effects of varying frequency of AMs across PoS (e.g. that nouns would appear to be more frequent, generally, than verbs or adjectives). The selection of AMs was not based on any semantic criterion.

The data for comparison (pairs of AMs and RMs) were retrieved from the databases of ECD and DEWA by queries containing equal sets (N = 30) of adjectives, nouns and verbs in the search list. The procedure resulted in data tables containing full lists of collocations (N = 4743) and associations (N = 8138), which were further filtered for the recurrent (F > = 2) connections. Subsequently, the two lists were compared automatically in order to find the cases where both the AMs and RMs coincided. We refer to those coincidental cases as *common pair*s in the following sections, while the non-coincidental collocations and associations of those 90 AMs are referred to as *exclusive* collocations and associations, respectively. Our method of comparing full lists of recurrent associations and collocations strives for accounting for the maximum of the potential overlap.

3.2 Results

3.2.1 Comparison in general terms

One of the main results of this study is the list of the common pairs (N = 582). The intersection makes 23.4% of the list of recurrent associations (N = 2488) and 14.9% of the list of recurrent collocations (N = 3903). The diverging parts are much greater than the coincidental ones. The proportions of exclusive associations and collocations are 76.6% and 85.1%, respectively. The average number of common pairs per AM is 6.53 (StDev = 3.41). Some examples of

8 See https://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=en (retrieved 22. 1. 2020).

AMs with the highest number (16—10) of common pairs are *laps* 'child', *kirjutama* 'to write', *tundma* 'to feel, to know', *uskuma* 'to believe', *mõistlik* 'sensible', *töö* 'work', *rõõmus* 'joyful', etc. The AMs with only one or two common pairs are *petma* 'to deceive', *meelitama* 'to flatter', *raiskama* 'to waste', *raudtee* 'railway', etc. It is remarkable that only one word out of 90 AMs (the verb *hämmastama* 'to astonish') had no common pairs at all.

The number of collocations (types) is moderately correlated (r = 0.67) with the AMs' general corpus frequency, while in the case of the associations, there is no such correlation (r = 0.1). Figure 2 illustrates this tendency. Three sets of data are compared (the common pairs, the exclusive collocations and the exclusive associations) and data is provided about their distribution across the groups of corpus frequency (see section 3.1.). It appears that the AMs with high corpus frequency enjoy a moderate dominance among the common pairs, whereas there is no such dominance in the case of exclusive associations. On the other hand, the AMs with the highest corpus frequency strongly dominate in the pool of exclusive collocations.

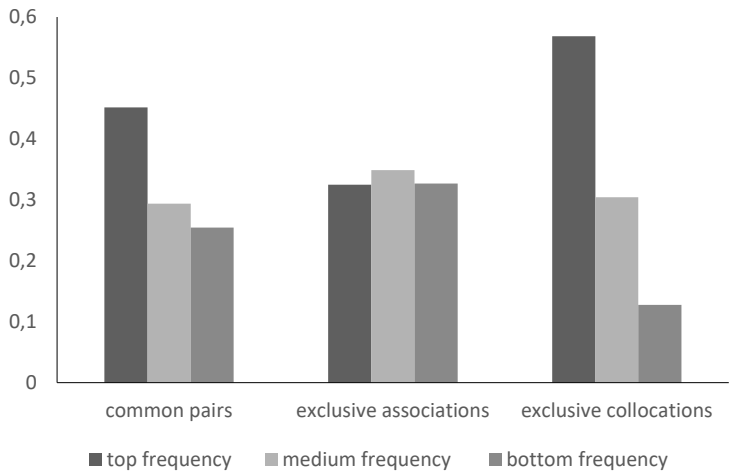


Figure 2: Distribution of data according to AMs' corpus frequency.

3.2.2 Comparison in terms of parts of speech

There is an intriguing division of the leading role between the PoS as AMs. Adjectives comprise a larger proportion in the pool of common pairs (see Table

2). There seems to be greater consensus with respect to attributing qualities in both associations and collocations. Some examples of such consensual adjectives are *mõistlik* 'sensible', *abivalmis* 'helpful', *vajalik* 'necessary', *rõõmus* 'joyful', *märg* 'wet', etc. Nouns comprise a larger proportion in the case of exclusive collocations (e.g. *töö* 'work, job', *aeg* 'time', *aasta* 'year', *asi* 'thing', etc.) and verbs tend to prevail in the case of exclusive associations (e.g. *meelitama* 'to flatter', *solvuma* 'to be offended', *vaidlema* 'to argue', *vihastama* 'to anger', *käskima* 'to give an order', etc). One can notice that the verbs that describe emotion-evoking processes have most diverging associations.

Table 2: *Distribution of PoS among the AMs*

AMs	Test words	Common pairs	Exclusive collocations	Exclusive associations
Adjective	33.30%	**38.14%**	30.66%	31.29%
Noun	33.30%	30.07%	**37.22%**	30.72%
Verb	33.30%	31.79%	32.13%	**37.99%**
Total (N)	90	582	3340	1953

The distribution of RMs follows neither the equal proportions of the test words nor the slightly diverging proportions of the AMs. Table 3 demonstrates that nouns comprise the biggest proportion of RMs among both the common and exclusive pairs. In the case of exclusive collocations, the prevalence can be observed to a lesser degree, and, in addition, some other PoS (mostly adverbs) emerge as RMs.

Table 3: *Distribution of PoS among the RMs*

RMs	Test words	Common pairs	Exclusive collocations	Exclusive associations
Adjective	33.30%	21.48%	16.26%	17.46%
Noun	33.30%	**62.54%**	**42.93%**	**61.19%**
Verb	33.30%	14.26%	23.44%	15.16%
Adverb		1.37%	15.21%	1.54%
Others			2.16%	4.66%
Total (N)	90	582	3340	1953

The prevalence of nouns among RMs can be explained in a few ways. The most obvious explanation is that the proportion of nouns in the lexicon generally

is larger (see e.g. Hudson, 1994)—a fact that gives this PoS an advantage in making any kind of relationships. Another explanation is that nouns serve in diverging functions with respect to forming relationships. An RM-noun can occur in a paradigmatic relation with an AM-noun (e.g. they form pairs of synonyms, antonyms and cohyponyms, which are both elicited in WATs and do co-occur in the texts). An RM-noun can also participate in syntagmatic relations, for example being the head of a phrase (e.g. *house* (N) in a phrase *big house*) or emerge as an argument of a verb e.g. *house* (N) in a phrase *building a house*. Relations similar to the syntagmatic one can also motivate word associations: for example, in the case of a well-known verb (such as *to build*) being a stimulus, the »typical objects« of the activity designated by the verb (such as *house*, *home* or *garage*) can often occur as responses.

The third possible explanation is that it is not only nouns as PoS which prevail among the RMs but perhaps certain specific nouns revealing the most important topics. It occurs that some nouns do indeed recur (e.g. *inimene* 'man, human being', *elu* 'life', *toit* 'food', *raha* 'money', *ema* 'mother', *laps* 'child', *vanem* 'parent'). These seem to represent important and recurrent aspects of sustainable life. In the case of exclusive collocations, the most frequent RM-nouns are *hulk* 'amount', *osa* 'part', *rahvas* 'people', *töö* 'work', *aeg* 'time', and *riik* 'state', which are more abstract by nature and perhaps represent the aspects and values related to social organisation[9]. The recurrent RM-nouns among the exclusive associations are: *mees* 'man, male person', *pood* 'shop', *riided* 'clothes', *pidu* 'party', *kodu* 'home', etc. These seem to represent the domestic sphere of life. Such a hint towards a division of topics in memory and language usage is worth further investigation. This observation is striking considering that our 90 test words were selected without any consideration of the semantics.

Homogeneity versus heterogeneity of stimulus and response in terms of PoS has been taken as a heuristic of the paradigmatic and syntagmatic relations, respectively. A pair is considered to be homogenous while both the AM and RM are of the same PoS and heterogeneous while they are different in respect

9 The words with meanings '*people*', '*work*' and '*time*' reveal that these notions are topical, and thus, valued in the public sphere. The word with meaning '*state*' points directly to the institution of social organisation and the words '*amount*' and '*part*' give a hint of the importance of »book-keeping« of the goods in a society.

of PoS (Toim, 1980). Table 4 presents the distribution of homogenous and heterogenous pairs. It appears that the exclusive associations (and apparently the associations in general) include more homogenous relations. This finding seems to be in line with the claims that »the word class of the stimulus word plays a role in that it causes the same word class to be over proportionally represented in the responses to it« (Mollin, 2009, p. 196). Whether the percentage from roughly 10 to 25 is overproportional depends on the perspective.

Table 4: *Distribution of the homogenous and heterogenous AM→RM pairs*

	AM→RM	Common pairs	Exclusive collocations	Exclusive associations
Homogenous	N →N	**18.90%**	**10.39%**	**24.63%**
	A→A	**13.75%**	3.44%	**9.78%**
	V→V	**9.97%**	2.99%	**12.70%**
Heterogenous	N→A	7.39%	**11.80%**	2.00%
	N→V	3.78%	**14.79%**	1.08%
	A→N	**23.54%**	**14.40%**	**17.15%**
	A→V		5.66%	1.38%
	A→D		7.16%	0.26%
	V→A	0.34%	1.02%	3.28%
	V→N	**20.10%**	**18.14%**	**17.97%**
	V→D		7.99%	0.72%
Total (N)		582	3340	1953

Note. N = Noun, A = Adjective, V = Verb, D = Adverb. Proportions larger than or close to 10% are in bold. The combinations with some other PoS, which are diverging and marginal or ambiguous, are not presented in this table.

The most prevalent group in the analysed dataset is N→N relation among the exclusive associations. The relation is also relatively stronger among the common pairs. The second most prevalent type of relation is heterogeneous A→N, which is the leading pair among the common pairs. The third prevalent type, V→N, occurs also in the range of the common pairs. All three most prevalent patterns have a noun in the position of RM. It is also worth mentioning that the common pairs lack heterogenous relations where nouns are not involved (e.g. A→V, A→D and V→D). These patterns seem to occur only among collocations. Exceptionally, there are some pairs with the structure V→A (e.g. *maitsma→hea* 'to taste→good', *tundma→mõnus* 'to feel→pleasant').

Taken together, the homogenous relations make up a larger proportion among the exclusive associations (47.11%) and common pairs (42.61%), while their proportion is much lower in the case of exclusive collocations (16.83%). The latter tend to demonstrate a heterogeneous PoS structure and thus reveal syntagmatic relations. This is quite expected, realising that collocations are derived from texts, which are syntactically arranged, while associations are driven from people's memory where such an arrangement cannot be taken for granted. It is still interesting that the biggest overlaps between associations and collocations occur among heterogeneous relations: A→N and V→N. Apparently, the syntagmatic (or syntagmatic-like semantic) relations play a role also in the memory and/or in the strategies of association elicitation.

3.2.3 Distribution of grammatical relations

In this section we provide a closer look at the distribution of grammatical relations that motivate the different types of AM→RM pairs. Information about grammatical relations derives from the ECD database.

As stated in Section 2, collocations in ECD are presented according to their grammatical relation in order to make it easier for the learner to acquire them and put them directly into use in their correct grammatical form. The grammatical relations illustrate what word pairs most typically occur in texts written by native speakers. Grammatical relation represents a category which displays collocates with the same relation to the search word (e.g. modifiers of a noun or objects of a verb).

Even though associations do not reveal grammatical relations directly—both stimulus and response are presented in base form in DEWA—we can take the corresponding grammatical relations in ECD as indicators of the potential grammatical relations motivating the emergence of certain associations.

The distribution of grammatical relations among both the common pairs and exclusive collocations is given in Table 5, and the most salient grammatical relations are discussed below.

Table 5: *Comparative distribution of grammatical relations between the common pairs and exclusive collocations*

Grammatical relation	**Common pairs (%)**	**Exclusive collocations (%)**	**Example(s)**	**AM→RM**
and/or	**33.68**	7.04	*kuud ja* ***aastad*** 'months and **years**', *ilus ja* ***uus*** 'beautiful and **new**', *kirjutama ja* ***lugema*** 'to write and **read**'	N→N A→A V→V
modifies	**23.54**	**13.83**	***pikk*** *tee* '**long** road'	A→N
object	9.79	5.93	*valu* ***tundma*** 'to **feel** pain'	V→N
adverbial_ semantic case	7.90	**15.50**	***restoranis*** *sööma* 'to eat in a **restaurant**'	N→V
adj_modifier	7.04	**10.45**	*vasak* ***käsi*** 'left **hand**'	N→A
genitive_modifies	5.15	4.04	***lapse*** *ema* '**child's** mother'	N→N
subject	2.75	4.88	*ülemus* ***käsib*** 'the boss **commands**'	V→N
subject_of	2.06	2.69	***sõjavägi*** *marsib* '**army** is marching'	N→V
genitive_modifier	2.06	1.95	*kassi* ***saba*** 'cat's **tail**'	N→N
object_of	1.55	3.83	***saba*** *liputama* 'to wag a **tail**'	N→V
adv_modifier	1.37	**15.21**	tohutu **suur** 'enormously **big**, koos **mängima** '**to play** together'	A→D V→D
	[...]	[...]		
Total (N)	**582**	**3340**		

Note. N = Noun, A = Adjective, V = Verb, D = Adverb. In examples AMs are highlighted in bold.

Table 5 shows that the *and/or* relation is the most frequent one, forming about 1/3 of all common pairs. This is because this homogeneous relation is not specific to any PoS. The *and/or* relation represents semantic relations like synonyms (*tähtis ja oluline* 'significant and important'), antonyms (*kerge või raske* 'easy or difficult') and cohyponyms (*ema ja laps* 'mother and child'), which are paradigmatic in nature. The remarkable intersection between associations and collocations shows that paradigmatic relations are not only restricted to memory but occur as coordinated constituents of a clause at the syntactic level of expression too.

The second most frequent grammatical relation among the common pairs is the *modifies* relation between AM-adjectives and RM-nouns. It is a syntagmatic relation of attribute and its head. The intersection shows that, apparently, qualities tend to make well-established connections to their typical carriers both in memory and written language use. This relation also comprises the third largest proportion of the exclusive collocations, revealing the wealth of attributive constructions in the texts.

When we look at exclusive collocations, the distribution of grammatical relations is different as no prevalent ones occur. The most frequent one is *adverbial_semantic case* between AM-nouns and RM-verbs, which captures adverbials that are nouns in semantic case forms[10] (e.g. inessive, adessive, comitative etc, as in ***restoranis*** *sööma* 'to eat in a **restaurant**', *inimestega suhtlema* 'to communicate with **people**', *naisesse armuma* 'to fall in love with a **woman**'). This grammatical relation contributes to the N→V type of PoS patterns, which is rather low among the common pairs and almost missing among the exclusive associations.

The second most frequent grammatical relation *adv_modifier*[11] between AM-verbs, AM-adjectives and RM-adverbs captures adverbs that modify verbs (*koos* ***mängima*** '**to play** together') and adjectives (*tohutu* ***suur*** 'enormously **big**'). This type represents the V→D and A→D PoS patterns that were missing among the common pairs and exclusive associations (see Table 4). The third most frequent grammatical relation (*modifies;* A→N) coincides with the second most prevalent one among the common pairs (see comments above).

Table 5 also shows that in some cases a specific PoS pattern can be motivated by more than one grammatical relation. One of those is N→N, to which two grammatical relations—in addition to the *and/or* relation—also contribute: *genitive_modifies* and *genitive_modifier*. The latter two represent the possessive construction as seen from two perspectives. In the case of the *genitive_modifies* relation, the AM-noun GEN (e.g. *lapse* 'child's') is modifying RM-noun NOM (e.g. *ema* 'mother') (*lapse ema* '**child's** mother'); in the case of *genitive_modifier*, AM-noun NOM (e.g. *saba* 'tail') is modified by RM-noun

10 Estonian is a morphologically rich language that uses semantic cases, whereas English, for example, uses prepositions.

11 Adverb as a modifier.

GEN (e.g. *kassi* 'cat's') (*kassi saba* 'cat's **tail**'). Another PoS pattern, possibly motivated by multiple grammatical relations, is N→V. There are two grammatical relations that—in addition to the a*dverbial_semantic case* discussed above—contribute to this syntagmatic pattern: *subject_of* and *object_*of. The same syntagmatic relation is reflected in V→N patterns *object* and *subject*, again, as from the other perspective.

In sum, there are indeed certain types of grammatical relations that are favoured both among collocations and associations. These are the paradigmatic *and/or* relation, which subsumes different PoS, and the syntagmatic relation *modifie*s, which holds between an adjective and its head noun.

3.2.4 Comparison in terms of ranking

Our data sources (ECD and DEWA) are similar in respect to presenting the RMs of a given AM in a decreasing order of frequency (see Table 1 in section 3.1.). The rank of a RM reflects its position in an ordered list and as such it is an approximate indicator of the (relative) strength of the relation. Rank 1 indicates the strongest relation in a given list, rank 2 the second strongest, etc. Equal rank of two RMs indicates their equal frequency in a given list.

It must be taken into account that the dictionaries differ, too, not only in their coverage of headwords (see Table 1) but also with respect to the number of RMs presented. The average number of different RMs (F > = 2) associated with an AM in ECD was 43.4 (StDev = 27.2), while in DEWA the average was 27.6 (StDev = 7.9). This indicates more variation, generally, in the length of the lists of collocations rather than of associations, which further affects the ranking. The mean rank of collocations, in general, is 28.4 (StDev = 23.10) while the mean rank of associations, in general, is 8.6 (StDev = 3.5).

We hypothesised that the RMs in top positions in the ranking would dominate among the common pairs, while the non-overlapping pairs would include RMs with a relatively lower rank. If this is the case, there should be a difference in the mean ranks of the common pairs as compared to the sets of exclusive associations and collocations.

The results of the comparison are presented in Table 6. The set of common pairs is characterised by the mean ranks in both DEWA and ECD, and those

two should be compared to the means of the exclusive associations and collocations, respectively. It is indeed the case that the mean ranks of the common pairs are smaller than the mean ranks of exclusive associations and collocations.

The means are rather even across the PoS, except for the mean for the collocations of adjectives among the common pairs, which is lower (16.29) than the mean for the collocations of verbs and nouns. This could mean that adjectives as AMs are selected for stronger collocative relations. Another explanation could lie in the fact that adjectives are provided with shorter lists of collocates in ECD compared to verbs and especially nouns. The longer lists of AM-nouns in ECD are reflected in their larger mean rank (37.43) among the exclusive collocations.

Table 6: *Comparison of the mean ranks across the common pairs vs exclusive associations and collocations*

	Common pairs		Exclusive associations	Exclusive collocations
AM	DEWA	ECD		
Adjective	6.79	16.29	9.25	25.21
Noun	6.69	20.35	8.89	37.43
Verb	7.17	21.07	9.00	26.00
All	**6.88**	**19.03**	**9.04**	**30.01**

It is still not the case that all of the strongest relations (with ranks 1—5) will appear among the common pairs. There is actually a great deal of variation in the ranks among the common pairs—StDev in DEWA = 3.8 and StDev in ECD = 18.7— and, on the other hand, the exclusive lists of associations and collocations also contain strong relations (with the ranks 1—5), which are not mutually present.

There were, for example, only few common pairs that shared the first rank both among associations and collocations: *beež→pruun* 'beige→brown', *kana→muna* 'hen→egg', *lahutama→abielu* 'to separate→marriage', *laps→väike* 'child→small', *lugema→raamat* 'to read→book', *naine→mees* 'woman→man', *tantsima→laulma* 'to dance→to sing', *võidupüha→paraad* 'independence day→parade'.

Examples of the strongest exclusive associations (rank = 1) include: pairs of the most obvious antonyms (*meeldiv→ebameeldiv* 'pleasant→unpleasant', *vasak→parem* 'left→right'), pairs of an attribute and its typical carrier (*oranž→apelsin* 'orange→orange', *triibuline→sebra* 'striped→zebra'), pairs of synonyms (*sõjavägi→armee* 'army→army', *ostukeskus→pood* 'shopping centre→ shop') and many more. These kinds of pairs are interpretable as strong relations in the memory, which are, at the same time, not represented as collocations in the language usage. It seems that the words are either mutually closing out or too obvious by semantics to be used in a close proximity while talking or writing. It has also been proposed that the strongly associated pairs which do not occur in the corpus reflect the world knowledge rather than the information that needs to be expressed in context (Schulte im Walde et al., 2008, p. 19).

Examples of the strongest collocations (rank = 1) missing from the associations include: the grammatical relations a*dv_modifier* (see section 3.2.3.), e.g. *mõnus→väga* 'pleasant→very', *mängima→hästi* 'to play→well', *uskuma→siiralt* 'to trust→sincerely'; the grammatical relation *modifies*, e.g. *emotsionaalne→seisund* 'emotional→state', *odav→tööjõud* 'cheap→workforce'; the grammatical relations *predicate_adj_translative_of*, e.g. *selge→tegema* 'clear→to make' < *selgeks tegema* 'to make it clear', *hapu→minema* 'sour→go' < *hapuks minema* 'to clabber', etc. One of the reasons that the exclusive collocations also include a number of high-ranking collocations is the fact that the set consists mostly of word pairs with the top frequency AMs (see Figure 1), which have the potential to make more frequent connections.

4 DISCUSSION

The main result of our study revealed (section 3.2.1) that the coincidental part of AM→RM relations is much lower than the divergent parts of exclusive AM→RM relations. This finding is well in line with previous studies of English (Mollin, 2009). The overall proportion of our common pairs (582) makes 9% of the total set of recurrent associations and collocations and fits quite well with Mollin's 3%. However, the proportion of coincidental pairs in our study is three times bigger. We can give two reasons for this difference. Firstly, Estonian as a morphologically rich language does not exploit function

words widely to indicate grammatical relations. The presence of *content word→function word* collocations that were missing among associations was one of the main arguments for the collocation association mismatch in Mollin's study of English. Secondly, the lists of associations in Estonian data were elicited by ca. 300 respondents (Vainik, 2018) while Mollin (2009) used the data of EAT, which contains responses of 100 undergraduate students (Kiss et al., 1973). The bigger number of respondents leads to longer lists of recurrent associations, which increases the probability of coincidence with some of the collocations.

4.1 The association-collocation mismatch

It was mentioned above that ECD is a much richer source of information both in terms of coverage of the headwords and the number of collocates presented. This is a quantitative factor inducing an overflow of collocations resulting inevitably in a larger proportion of mismatches on the side of collocations. There are also some qualitative factors affecting the incompatibility of the outcome.

One of the factors is the nature of the data that stems from the method of data gathering. The material presented in ECD is influenced by the size and character of the corpora on which it is based (Kallas et al., 2015; Koppel et al., 2019b). The material in DEWA, on the other hand, is influenced by the number of respondents, by the selection of the stimuli, etc. (see Vainik, 2018) and, apparently, also by following the common strategies of association elicitation by respondents (see Clark, 1970).

The nature and quality of the corpus influence, for example, which word pairs would emerge as more salient in ECD. In section 3.2. we mentioned that the RMs of the exclusive collocations revealed more abstract concepts related to the aspects and values of social life (e.g. *regionaalne* 'regional', *riiklik* 'national', *koostöö* 'collaboration'). This might easily be because of the more official register brought forth by the content of the corpus, which includes an abundance of official documents and texts. One can also notice vocabulary related to certain specific fields like sports (e.g. *märg rada* 'wet track', *naiste turniir* 'women's tournament') and weather forecasting (*märg lumi* 'wet snow'). Another aspect that may reduce the number of coinciding AM→RM relations is the fact that the semi-automatically gathered material of ECD was controlled

manually, and collocations pointing to obvious idioms and proverbs were deliberately excluded[12].

There are also some systematic characteristics of the material in DEWA that may have caused its partial incompatibility with the collocations. One of them is the form of the stimuli, which is presented in the base form, i.e. the nominative singular case (in the case of declinable words) (see section 3.1.). For example, if an adjective is presented to the respondent in the nominative singular case, then the answers tend to be substantives (i.e. the head nouns of attribute phrases e.g. *märg→pesu* 'wet→laundry') or antonyms, i.e. adjectives related to the *and/or* relation, e.g. *märg→kuiv* 'wet→dry'). In the texts, on the other hand, one finds inflected adjectives in collocations (e.g. *viimaseks* [adjective-SG-TRANSL] *jääma* [verb-INF] 'come in last', *märjaks* [adjective-SG-TRANSL] *saama* [verb-INF] 'to get wet'), which represent the grammatical relation *predicate_adj_translative_of*. Such combinations do not emerge as responses in the WAT test.

Another reason for formal incompatibility might be due to the association stimuli being given in singular, which influences the form of responses. Therefore, the cases in which a collocation is frequent but where AM is in plural, e.g. *kohalikud valimised* 'local elections', are not found among the common pairs. Another notable form-related difference is the scarcity of comparative forms among associations. There were common collocations found in the corpus which contained comparative adjectives (e.g. *suurem laps* 'older child') that did not occur in associations.

In section 3.2.2. (Table 4) we highlighted that adverbs were almost missing from the RMs in the case of associations and were totally absent in the case of the common pairs. The reason for the lack of adverb word pairs is likely due to both semantics as well as word order in Estonian. For example, since adverbs are placed before adjectives in the sentence, then in the case of adjective stimuli, the response is probably less likely to be the preceding word than the following one. The general semantics of the adverbs as a PoS also plays a role. One can speculate that adverbs, though frequent collocates in corpora, are often semantically emptier as they mostly function as intensifiers (e.g. *tohutult*

12 Such a decision was related to the policy of the portal Sõnaveeb, to avoid duplicating the information (Koppel et al., 2019a).

(D) *suur* (A) 'enormously big') or modifiers (e.g. *peamiselt kohalik* 'mainly local', *enamasti kohalik* 'mostly local', etc.). Such adverbs express the extent of a quality rather than a true relation between two content words, and are thus less likely to occur in the WAT tests. People prefer to give lexical rather than function words as responses (Clark, 1970, p. 283).

In conclusion, the constituency of corpora as well as form, word order and semantics all play a role in creating the difference between associations and collocations.

4.2 Practical implications

We foresee applicability of the knowledge about common pairs of collocations and association in lexicography and language teaching. In both fields, a strategy of prioritisation is needed because of the everlasting demand for efficiency in the condition of a rich flow of information. Mimicking deliberately the structure of a native speaker's mental lexicon would be one possible strategy of prioritisation when presenting the material in web dictionaries and supporting materials targeted at learners.

In that respect, one could formulate a tentative principle, "the first relations first", while deciding where to start learning from or to which type of constructions to pay the most attention. If a dictionary, language portal or teaching material contains a lot of collocations, associations can offer an alternative strategy to corpus frequency in deciding which ones should be given priority. For example, the collocations dictionary is very sizable (e.g. some frequent nouns can have over 100 collocates) and can be difficult for a learner to absorb. The supporting information about the presence of these relations in the native speaker's mental lexicon would be a valuable key for the first approximation. Common pairs, as the more focal relations, could be marked for learners by adding key-symbols, for example.

In ECD, collocations are presented as constructions in order to make it easier for the learner to use them and include them into their active vocabulary. Based on the findings of this analysis, we could suggest that the paradigmatic relations represented by the *and/or* relation and the syntagmatic relation of attribution (the grammatical relation *modifies*) should also be given special attention when compiling materials for language teaching.

From the perspective of PoS, one could infer that the combinations A+N and V+N seem to be more central in the mental lexicon than, for example, combinations including verbs, adverbs and adjectives.

One can consider applicability of the results also in relation to writing dictionary definitions in dictionaries where familiarity for the user is strived for. In such cases associations could play a major role. For example, if at certain words or group of words paradigmatic relation is found more relevant, providing synonyms/antonyms next to or as part of the definition would be useful[13]. It has been also suggested that associations reveal information about domain information and relevance of the senses for the ordinary speakers (Sinopalnikova, 2004). This should be even more true about the association-collocation overlap.

5 CONCLUSION

The main goal of the present paper was to systematically compare word associations and collocations in Estonian in order to achieve some new insights regarding the role of PoS. We assumed that Estonian as a language with a well-developed morphosyntactic structure would reveal some constructions that may favour the occurrence of certain PoS combinations. The analysis was based on a representative selection of test words (N = 90) and their related items from two recent dictionaries, ECD and DEWA.

The results revealed an overlap of 14.9% of all collocations and 23.4% of all associations related to the test words. We interpreted the common pairs (N = 582) as a similarity of collocations and associations and the exclusive pairs as a mismatch.

With regard to the PoS, it was discovered that adjectives tend to make proportionally more common pairs than nouns and verbs. There was a well-established combination of adjectives and nouns recurring that was explained as being motivated by the attributive grammatical relation *modifies*. It also appeared that adjectives tend to make somewhat stronger collocations, which is a topic that needs further study. We tentatively concluded that there is a remarkable consensus concerning attributing qualities in both memory and language use.

13 We thank our anonymous reviewer for this idea.

It was also discovered that, regardless of the PoS of the headword/stimulus, there occurred proportionally more nouns as collocates/responses among the common pairs. The biggest overlaps between associations and collocations were found among heterogeneous relations comprising different PoS: in addition to the A→N relation mentioned above, the relation V→N was salient. Apparently, the syntagmatic (or syntagmatic-like semantic) relations play a role not only in texts but also in the semantic memory and/or in the strategies of association elicitation. Interestingly, the common pairs lacked heterogenous relations when nouns were not involved, which reveals also the tendency for nouns to recur as the related members.

The *and/or* relation was found to be the dominant grammatical relation among the common pairs because it subsumes different PoS and expresses paradigmatic relations (e.g. synonymy, antonymy, cohyponymy). On the other hand, a totally different grammatical relation (*adverbial_semantic case*) was found to prevail among the exclusive collocations. This is obviously because Estonian is a morphologically rich language that uses semantic cases, whereas English, for example, uses prepositions.

The most frequent combination of PoS was the homogenous N→N combination, which was prevalent among the exclusive associations. Although the *and/or* relation seems a convenient and plausible motivation, our analysis showed that other grammatical relations like *genitive_modifies* and *genitive_modifier* contribute to this prevailing pattern too.

As the non-coincidental part of collocations and associations was large—85.1% and 76.6%, respectively—we also paid attention to discussing some possible reasons for the systematic mismatch. Besides the quantitative disproportion of collocations, we proposed such qualitative factors as the constituency of the corpus, a form of stimuli, word order and semantics playing a role.

In sum, we can see several reasons, both quantitative and qualitative, that may cause the mismatch between associations and collocations. It is still remarkable though that these reasons seemingly do not rule out completely the similarities between associations and collocations. We interpret the similarity as revealing a set of core connections that are actively upheld while people think, talk and write texts in Estonian. The core connections seem to share a

structure that can be described in terms of the PoS fitting into certain recurrent grammatical relations.

Acknowledgements

This study was supported by the Estonian Research Council grant PSG227.

REFERENCES

Dictionaries

DEWA = Vainik, E. (2019). *Eesti keele assotsiatsioonisõnastik* [Dictionary of Estonian Word Associations]. doi: 10.15155/3-00-0000-0000-0000-07DF6L

ECD = Kallas, J., Koppel, K., Paulsen, G., & Tuulik, M. (2019). *Eesti keele naabersõnad 2019* [Estonian Collocations Dictionary]. doi: 10.15155/3-00-0000-0000-0000-0823EL

Other

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Clark, H. H. (1970). Word associations and linguistic theory. In J. Lyons (Ed.), *New horizons in linguistics* (pp. 271–286). Baltimore, Maryland: Penguin.

De Deyne, S., & Storms, G. (2015). Word associations. In Taylor (Ed.), *The Oxford Handbook of the Word (Oxford Handbooks)* (p. 471). OUP Oxford: Kindle Edition.

Deese, J. (1965). *The Structure of Associations in Language and Thought*. Baltimore: The Johns Hopkins Press.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? *Investigating the thesis of collocational priming*. *Corpus Linguistics and Linguistic Theory*, *6*(2), 125–155.

Firth, J. R. (1957). 'Modes of Meaning'. *Papers in linguistics 1934–1951*, 190–215. Oxford: Oxford University Press.

Fitzpatrick, T. (2007). Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics, 17*(3), 319–331.

Fitzpatrick, T., Playfoot, D., Wray, A., & Wright, M. J. (2015). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics, 36(1)*, 23–50. doi: 10.1093/applin/amt020

Galton, F. (1879). Psychometric experiments. *Brain, 2*(2), 149–162. doi: 10.1093/brain/2.2.149

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language, 70*(2), 331–339.

Jung, C. G. (1910). The association method. *The American Journal of Psychology*, *21*(2), 219–269. doi: 10.2307/1413002

Kallas, J. (2013). *Eesti keele sisusõnade süntagmaatilised suhted korpus-ja õppeleksikograafias* [Syntagmatic Relationships of Estonian Content Words in Corpus and Pedagogical Lexicography]. Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32. Tallinn: Tallinna Ülikool. Tallinn: *Tallinn University, Dissertations on Humanities Sciences*.

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E, Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. 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, 11–13 August, 2015, Herstmonceux Castle, United Kingdom* (pp. 11–13) Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Kang, B. M. (2018). Collocation and word association: Comparing collocation measuring methods. *International Journal of Corpus Linguistics, 23*(1), 85–113.

Kent, G. H., & Rosanoff, A. J. (1910). A study of association in insanity. *American Journal of Insanity, 67*(1–2), 37–96.

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the XI Euralex International Congress* (pp. 105–116). Lorient: Université de Bretagne Sud.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., & Tiberius, C. (2010). A quantitative evaluation of word sketches. *Proceedings of the XIV Euralex International Congress, 6–10, July 2010*, *Leeuwarden* (pp. 372–379). Ljouwert: Fryske Academy.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken & R. W. Bailey (Eds.), *The Computer and Literary Studies* (pp. 153–165). Edinburgh: University Press.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019a). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October, 2019, Sintra, Portugal* (pp. 434–452). Brno: Lexical Computing CZ, s.r.o.

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V., & Michelfeit, J. (2019b). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference, 1–3 October, 2019, Sintra, Portugal* (pp. 763–782). Brno: Lexical Computing CZ, s.r.o.

Leech, G., & Smith, N. (2000). *Manual to accompany the British National Corpus (Version 2) with improved word class tagging*. Lancaster: UCREL. Retrieved from http://ucrel.lancs.ac.uk/bnc2/bnc2postag manual.htm

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory, 5*(2), 175–200. doi: 10.1515/CLLT.2009.008

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, 28 (6), 887–899. doi: 10.3758/BF03209337

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. doi: 10.3758/ BF03195588

Postman, L., & Keppel, G. (1970). *Norms of Word Association*. New York NY: Academic Press.

Rosenzweig, M. R. (1961). Comparisons among word-association responses in English, French, German, and Italian. *The American Journal of Psychology, 74*(3), 347–360. doi: 10.2307/1419741

Roth, T. (2013). Going Online with a German Collocations Dictionary. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (Eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 Conference, 17–19 October, 2013, Tallinn, Estonia* (pp. 152–163). Retrieved from http://eki.ee/elex2013/proceedings/eLex2013_11_Roth.pdf

Schulte im Walde, S., Melinger, A. Roth, M., & Weber, A. (2008). An empirical characterisation of response types in German association norms. *Research on Language and Computation 6*(2), 205–238.

Schulte im Walde S., & Borgwaldt, S. (2015). Association Norms for German Noun Compounds and their Constituents. *Behavior Research Methods 47*(4), 1199–1221.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam/Philadelphia: John Bejamins. doi: 10.1075/scl.22

Sinclair, J. (1966). Beginning the Study of Lexis. In C. E. Bazell et al. (Eds.), *In Memory of J. R. Firth* (pp. 410–430). London: Longman.

Sinopalnikova, A. (2004). Word Association Thesaurus as a Resource for Building WordNet. *Proceedings of the 2nd International WordNet Conference,* Brno, Czech Republic (pp. 199–205).

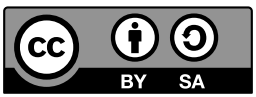
Toim, K. (1980). Estonian word association norms for the Kent-Rosanoff test. Problems of cognitive psychology [Труды по психологии. Проблемы когнитивной психологии]. *Tartu Riikliku Ülikooli Toimetised*, *522*, 60–76.

Vainik, E. (2018). Compiling the Dictionary of Word Associations in Estonian: from scratch to the database. *Eesti Rakenduslingvistika Ühingu aastaraamat*, *14*, 229–245. doi: 10.5128/ERYa.1736-2563

PRIMERJAVA KOLOKACIJ IN BESEDNIH ASOCIACIJ V ESTONŠČINI Z VIDIKA BESEDNIH VRST

V prispevku predstavimo primerjalno študijo kolokacijskih in asociacijskih struktur v estonščini s poudarkom na vlogi besednih vrst. Z namenom, da bi ugotovili prekrivne in različne strukture, opravimo analizo seznamov kolokacij in asociacij za enako število samostalnikov, glagolov in pridevnikov, ki jih najdemo tako v Kolokacijskem slovarju estonskega jezika kot v Slovarju besednih asociacij v estonskem jeziku. Rezultati pokažejo, da med asociacijami in kolokacijami prevladujejo samostalniki. Prekrivne strukture lahko deloma pojasnimo z vplivom gramatičnih relacij oz. slovničnih vzorcev, ki povezujejo kolokacije in motivirajo asociacije. Rezultate ovrednotimo tudi z vidika morebitnih razlogov za neujemanja med asociacijami in kolokacijami, v zaključku pa podamo razmisleke o izrabi rezultatov študije na področjih leksikografije in poučevanja tujih jezikov.

Ključne besede: kolokacije, asociacije, besedne vrste, leksikografija, estonski jezik



Slovenščina 2.0, 2020 (2)

THE ATTITUDE OF DICTIONARY USERS TOWARDS AUTOMATICALLY EXTRACTED COLLOCATION DATA: A USER STUDY

Eva PORI, Jaka ČIBEJ, Špela ARHAR HOLDT
Faculty of Arts, University of Ljubljana

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

Pori, E., Čibej, J., Kosem, I. and Arhar Holdt, Š. (2020): The attitude of dictionary users towards automatically extracted collocation data: a user study. Slovenščina 2.0, 8(2): 168–201.

DOI: https://doi.org/10.4312/slo2.0.2020.2.168-201

The paper is based on a survey conducted within the framework of the basic research project *Collocations as a Basis for Language Description: Semantic and Temporal Perspectives* (KOLOS; J6-8255). It presents a qualitative analysis of a user evaluation of the interface of the *Collocations Dictionary of Modern Slovene* (CDS). It discusses an alternative perspective—the user's point of view—on problematic aspects of individual dictionary features, which require further lexicographic analysis and discussion. The collocations user study presents a model of the process of user evaluation; its findings are significant primarily for determining problems encountered by users. They also serve as a useful basis for methodology improvements in future, comparable lexicographic user studies and analyses.

K**eywords:** collocations dictionary, responsive dictionary, user evaluation, attitude towards errors, dictionary interface

1 INTRODUCTION

In the digital world, a dictionary is increasingly becoming a network of dynamic shifts between different language information and resources, as well as a testing ground for various contemporary conceptual lexicographic approaches. The concept of a "responsive dictionary"—a dictionary characterised by its capacity to respond to the dynamics of language development and include the interested language community in the development of language resources in a methodologically transparent manner (Arhar Holdt et al., 2018)—first came to fruition (both in Slovenia and internationally) with the *Thesaurus of Modern Slovene*.[1] The responsive dictionary was created as a reaction to the language needs and desires of the modern community of users. The innovative characteristics of the Thesaurus, such as open-access, flexibility, and interconnectedness, provided an alternative to already established dictionary forms. The unique character of *The Collocations Dictionary of Modern Slovene,*[2] the second example of a responsive language resource and the topic of this paper, introduced a new dynamic in Slovene lexicography: its basic design follows the original concept of a responsive, linear (but not only) lexicographic structuring, bends established lexicographic surfaces and both shifts and transcends traditional lexicographic patterns.

In addition to coming up with an alternative dictionary form, modern lexicography has increasingly recognised the undeniable value of dictionary users. Despite the growing interest of international lexicographers in user studies, in Slovenia the field remains understudied and overlooked. This is why the

1 The Thesaurus of Modern Slovene was published in March 2018 and was compiled automatically. It contains 105,473 headwords and 368,117 synonyms with links to the Gigafida Corpus of Written Standard Slovene; it is freely accessible at: https://viri.cjvt.si/sopomenke; the database is freely accessible at CLARIN.SI under the CC BY-SA 4.0 licence: Krek, Simon; et al., 2018, Thesaurus of Modern Slovene 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166.

2 The *Collocations Dictionary of Modern Slovene* was published in October 2018 and is based on automatically extracted data. It contains 35,989 headwords, 7,717,561 collocations, and 36,736,168 examples from the Gigafida Corpus of Written Standard Slovene; it is freely accessible at: https://viri.cjvt.si/kolokacije; the database is freely accessible at CLARIN.SI under the CC BY-SA 4.0 licence: Kosem, Iztok et al., 2019, *Collocations Dictionary of Modern Slovene CSD 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

present study examines the role of user reception and contribution to the upgrades and improvements of dictionaries. The idea of a responsive dictionary recognises the user as an active co-creator of (digital) language resources, as well as a critical evaluator of the features offered. The results of an open discussion between linguists and users represent a useful starting point for further analysis of the design of dictionaries, and, in the present case, of the general role of the collocations dictionary as a responsive dictionary within the field of lexicography.

The present study focuses on the users' attitudes towards automatically extracted collocation data, especially in relation to specific features introduced into lexicography by responsive dictionaries. In their initial phase, responsive dictionaries are automatically compiled and relatively quickly published for public use; alongside linguists, the language community then gradually helps improve and clean the data. The *Collocations Dictionary of Modern Slovene* was also immediately made available to the public, i.e. in the initial, unprocessed stage containing noise or errors. The design of the dictionary interface, however, featured options to eliminate these shortcomings (data evaluation and cleaning), information about the linguistic completeness of the entry, and other similar features (Kosem et al., 2018c). The present study was interested in specific groups of users and their attitudes towards the present state of the dictionary, their opinion on its responsiveness (which includes automatic compilation, gradual upgrades, and user involvement), and their response to particular types of existing errors in the data. The user evaluation is intended to serve as a basis for identifying problematic areas, as well as less problematic areas in need of improvement, and will play a key role in the improvement of the collocations dictionary interface.

The paper begins by presenting the method of user evaluation of the *Collocations Dictionary of Modern Slovene 1.0*. This is followed by an analysis of the three thematic segments of the user evaluation, i.e. the three-part design of the evaluation interview. A representative case (proper nouns) demonstrates user perspective on (non-)problematic features of data and the dictionary interface. The conclusion summarizes the key findings of the study and examines the suitability of the applied method as a model for user evaluation in similar lexicographic user studies.

2 METHODOLOGY

2.1 Research Framework

In lexicography, user research has a tradition reaching back to the 1960s (e.g. Barnhart, 1962; Householder, 1967), but the research area was firmly established later in the 1980s and 1990s (e.g. Tomaszczyk, 1979; Hartman, 1987; Atkins, 1998; Nesi, 2000). The emergence of the digital medium in the 2000s offered a vast array of new methodological possibilities (e.g. Bergenholtz and Johnsen, 2013; Müller-Spitzer, 2014; Lew and De Schryver, 2014). More recently, existing approaches were also critically evaluated and surpassed (Bogaards, 2003; Tarp, 2009; Lew, 2015; Kosem et al., 2018a).

Despite growing opportunities for user involvement, Slovene lexicography has been relatively slow in developing an interest in user studies. This is why, as mentioned in previous research (Rozman, 2004; Stabej, 2009; Logar, 2009; Gorjanc, 2017), Slovene lexicography has a glaring lack of data in relation to user habits, needs, capacities, and preferences. Over the past few years, important steps have been taken, such as the development of a user typology (Arhar Holdt et al., 2016), the research of user needs in relation to selected language problems (Čibej et al., 2016; Arhar Holdt et. al, 2017), the participation in an international study on user attitudes to general monolingual dictionaries (Kosem et al., 2018a, 2018b), and the development of methodologies for user inclusion and tracking within the framework of a responsive dictionary (Arhar Holdt et al., 2018).

The present study contributes to the available array of tried and tested methodologies (a comprehensive overview of existing methodologies is provided in Welker, 2013a, 2013b) with the addition of user evaluation based on the guided think-aloud method. Think-aloud protocols have been described by Tarp (2009, p. 287) as:

> The informants are invited to freely express which reflections and problems they have during the consultation process [while working with a specific dictionary (author's note)]. These »thoughts« are tape-recorded and subsequently transcribed and written down in protocol form. [...] [This method] gives the researcher an idea of the users' way of working as well as what is happening during the process, what users are looking for, what they think they are looking for, and which problems they face when trying to find and interpret the relevant data. A number of research projects performed with this method have provided valuable

results, among others Wingate (2002) who did research into the usefulness of various types of definitions in learners' dictionaries, and Thumb (2004) who focused on the users' different look-up strategies and the problems they faced during the process.

We used the basic idea of the method, but adapted it to serve the purposes of a straightforward evaluative approach: the participants were presented with the dictionary; while they were using it, an interviewer was actively involved, suggesting queries and guiding the "thinking" with a set of prepared questions. Both the audio and the participants' interaction with the screen were recorded. However, only the audio was transcribed and analyzed (as the "protocol" itself was guided and thus comparable).

2.1 Research Goals and Sample Structure

The primary aim of the study was to determine the participants' opinion on the advantages and disadvantages of the *Collocations Dictionary of Modern Slovene* and responsive dictionaries in general, and to find ways of improving its user-friendliness. It was our intention to examine whether adult speakers of Slovene – particularly those with linguistic background or keen linguistic sensibility – know how to use, read and interpret the Collocations Dictionary of Modern Slovene, despite the fact that the dictionary featured raw, automatically extracted data. Our focus was on determining the participants' attitudes towards:

- automatic data compilation and errors;
- continuous dictionary upgrades and updates;
- possibility of user inclusion or contribution;
- innovative interface functions.

Following the typology of potential dictionary users (Arhar Holdt et al., 2016), the study included four distinct target groups of participants: translators and proof-readers; teachers of Slovene as a first language; teachers of Slovene as a second or foreign language; and lexicographers. The selected sample covers different scenarios of potential use, which allows the joined feedback on the dictionary to be perceived as more representative. Teachers were included to evaluate the didactic value of the dictionary, primarily its usefulness for teaching vocabulary to students. Translators can benefit significantly from

knowing what collocations and colligations are typical for a given word, while proofreaders need straightforward normative information to support their decisions. Finally, the group of lexicographers was included to identify whether and how their views differ from the opinions of actual dictionary users, e.g. whether as the creators of the dictionary, they perceive its pros and cons similarly to other groups, and whether they propose similar steps for further development than other groups.[3]

Table 1: *Structure of the participant sample*

GROUP	Affiliated institutions	Region	Age	Professional experience
10 teachers of Slovene as L1	SŠ Ravne na Koroškem II. gimnazija v MB Ekonomska šola (+gimnazija) Ljubljana	Ljubljanska Podravska Koroška Gorenjska	30–50	10–30 years
10 teachers of Slovene as L2 / foreign language	Centre for Slovene as a Second/Foreign Language (Faculty of Arts, University of Ljubljana)	Hungary Czech Republic Štajerska Ljubljanska Primorska	30–50	10–30 years
10 translators / language editors (proofreaders)	SLG Celje self-employed independent cultural employee	Primorska Dolenjska Savinjska Gorenjska Ljubljanska	30–50	10–30 years
10 lexicographers	CJVT UL FDV UL FF UL self-employed	Ljubljanska Štajerska	30–50	10–20 years

The study included 40 participants. As seen in Table 1, the participants were primarily between 30–50 years of age, with 10–30 years of work experience; they originated from different Slovene regions or—in the case of teachers of Slovene as a second or foreign language—from abroad. The call for participation was circulated widely through various means of communication

3 Students of Slovene as an L1 and learners of Slovene as an L2 did not participate in this step of the study. We chose to focus on adult professional users to make the best of the time and resources available within the project. Compared to the selected user groups, students are more easily accessible and after the project, the study can be continued to include both them as well as other potentially relevant user groups.

(such as mailing lists). The participants responded voluntarily, which needs to be taken into account in the interpretation of the results: the sample consists of participants who are relatively familiar with innovative, digital, and responsive language and dictionary resources, as they use them in their everyday work.

2.2 Evaluation Interview: Design

The evaluation interview was carefully planned and pre-tested on a group of researchers, i.e. linguists and research colleagues assuming the roles of interviewees. Our method was selected in order to enable identification of relevant data communicated in various ways by the interviewee, with minimal interviewer influence; its aim was to detect problems encountered by the interviewee while attempting to complete a specific task—working with a dictionary, on particular dictionary entries. To facilitate internal processing and analysis of acquired data, the participants were guaranteed full anonymity and asked for prior written consent for the recording of their screen and voice.

The approximately 30-minute long evaluation interview was based on a prepared three-part questionnaire (Appendix 1). During the first part of the session, the participants were asked—while thinking aloud—to click randomly in the dictionary and to query entries of their own choice. In this way, they could familiarize themselves with the Collocations Dictionary and form a first impression. At the same time, they were encouraged to spontaneously express their thoughts, feelings, and emotions and report whether they encountered, sensed or noticed any problems. Attention was primarily focused on the participant's capacity to recognize the range of functions and their possible combinations provided by the Collocations Dictionary (visual information on entry completeness, sense menus, various filters, such as frequency filter (showing only either rare or frequent words), or ordering by alphabetical order; collocate clustering, information on collocation relevance, examples of use, links to the Gigafida corpus and other dictionaries, etc.). In this way, we primarily examined attitudes towards functionality, intuitiveness, and user-friendliness of the dictionary.

The second segment of the interview involved working with specific headwords; the participants were guided and tested to determine whether they

recognized the (non-)problematic nature of particular entries. We were interested in their ability to interpret raw data, the amount of problems or errors detected, the nature of these errors, and the levels of distraction posed by the errors. The evaluation included three types of dictionary entries; prior to conducting interviews, we created a list of existing data errors for each entry and thus anticipated the participants' potential observations.

a) An example of a non-problematic and lexicographically fully examined entry, albeit highly polysemous and thus collocationally diverse: the noun *belina* 'whiteness'.

b) An example of an entry with only few potentially problematic collocates: the noun *pivo* 'beer'.

c) Two examples of more problematic entries, with the difficulties expressed either on the level of collocation structure or headword: the noun '*klop'*, where most of the collocates are erroneous due to homonymy (*klóp* 'bench', *klòp* 'tick'); and the verb *usesti* (*se*) 'to sit (oneself) down', which appears in inadequate structures due to the absence of the reflexive pronoun *se*.

Table 2: *A list of identified errors for the noun headword* pivo *on the levels of collocates or headwords, syntactic structures and collocations*

Problem	Example in Slovene	Translated example
Errors on the level of collocates or headwords		
The collocate was incorrectly lemmatized.	*plata piva* instead of *plato piva*	'plate of beer [cans]' instead of 'box (lit. plateau) of beer [cans]'
The collocate or headword should be in a specific inflected form (such as plural or comparative).	*drag od piva* instead of *dražji od piva*	'[expensive] than beer' instead of '[more expensive] than beer'
The collocation did not include the verb morpheme *si/se*.	*nacejati s pivom* instead of *nacejati se s pivom*	'to guzzle beer' [missing *se* morpheme]
Errors on the level of syntactic structures		
The collocate was tagged with an incorrect part-of-speech.	*pivo pite* instead of *pivo piti*	'beer of pie' instead of 'to drink beer'
The verb collocate should appear in the negative form.	*piti piva* instead of *ne piti piva*	'to drink beer' instead of 'to not drink beer' [missing negative particle]

Problem	Example in Slovene	Translated example
Errors on the level of collocations		
The collocation is nonsensical as it makes no sense if taken out of context or without additional elements.	*pivo k ustom* instead of *dvigniti kozarec piva k ustom*	'beer to the mouth' instead of '[to raise a glass of] beer to the mouth'
The headword appears next to a syntactic structure in the genitive plural or is a plural noun; the collocation makes no sense without an additional, quantitative element.	*pivo po tolarja* instead of *pivo po 300 tolarjev*	'beer for tolar' instead of 'beer for 300 tolars'

The third and final segment of the interview examined the participant's opinion on the general usefulness of the dictionary, its digital form (continuous upgrades) and their assessment of its look.

2.3 Transcription and Annotation

The annotation of interviews with the participants was done on the transcriptions of audio recordings, which were completed by four students of linguistics. The transcription followed a set of clear guidelines; one of the key guidelines was that the transcription should not be reduced to summarizing, but should instead record the conversations as faithfully as possible, with linguistic adaptation and standardization only permissible on the morphological level.

The annotation process followed the general thematic structure of the questionnaire (Appendix 1). A set of annotation guidelines was prepared, containing a list of available tags, their descriptions, and several examples from the transcriptions. Four annotators were familiarized with the guidelines and assigned 10 transcriptions each. The annotation was made in a local installation of Taguette (Rampin et al., 2019), an open-source online platform for collaborative text annotation (Figure 1). Taguette is an example of computer-assisted qualitative data analysis software (CAQDAS), the aim of which is to facilitate a systematic analysis of unstructured or half-structured data, particularly transcriptions of interviews. It enables multiple annotators to collaboratively annotate each transcription. Relevant text segments are marked either top-down (i.e. the annotators are presented with a set of tags to use during annotation) or bottom-up (i.e. the annotators mark

relevant information with their own tags, which can be easily grouped in the end to achieve the final annotation scheme). There are two main advantages of this approach to qualitative data analysis: a) tagging the transcriptions can provide a quantifiable overview of the data (e.g. the frequency of the tags reveals the most frequently discussed topics, issues, and recurring patterns in the analyzed texts); and b) Taguette is designed in a way that allows segments related to a specific feature to be exported to a separate file, essentially combining all related segments from different transcriptions into a single document. This allows for a more thorough analysis of a specific issue across all participants or participant groups.

Because the interviews in our research were semi-structured and focused on specific features of the *Collocations Dictionary of Modern Slovene*, we elected to follow a top-down approach and prepared a limited tagset for the annotators to use. The higher the frequency of the annotation, the more prevalent or topical the discussed argument in the user group. On the other hand, less frequently annotated topics might indicate that the user either has not noticed a feature or found it less important compared to others.

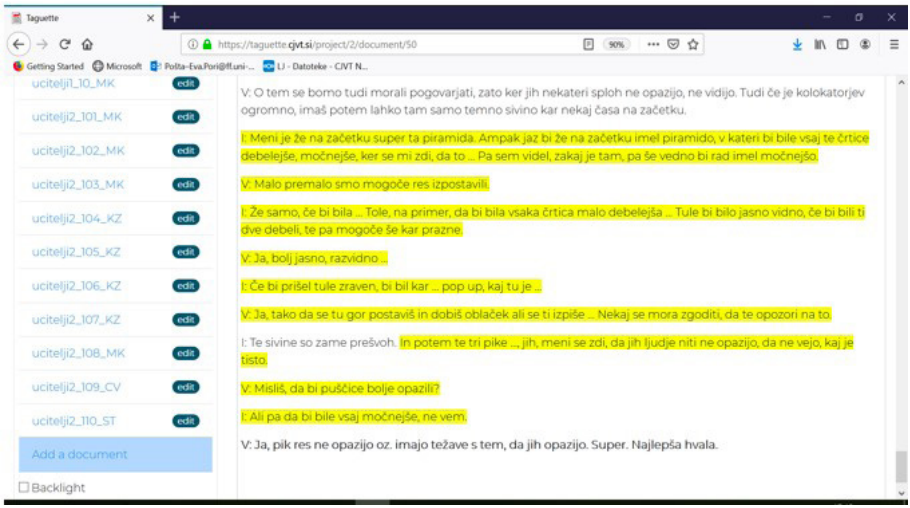


Figure 1: A screenshot of the Taguette annotation platform.

2.4 Annotation Results

The annotation typology (shown in Table 3, along with the total frequency of each tag) consists of 4 main categories[4] with multiple subcategories. The table also presents the general attitude towards a specific feature indicating whether the participating evaluators expressed more arguments *pro* or *contra*. These labels are discussed in more detail in Section 3.

Table 3: *Frequency of annotations by thematic blocks of the interview*

Category	Frequency	General attitude
General features		
Automatic compilation *Segments related to the participants' opinion on the fact that the dictionary was compiled automatically*	27	PRO
Dictionary usefulness *Segments related to the usefulness of the dictionary*	112	PRO
Look and design *Segments related to the overall look and design of the dictionary*	37	PRO
Digital form *Segments discussing the fact that the dictionary is digital-only*	69	PRO
Interface		
Entry phase indicator *Segments discussing the phase indicator pyramid symbol in the dictionary*	69	PRO
Sense indicators *Segments discussing the menu that enables the semantic disambiguation of collocates*	43	PRO
Three dot icon *Segments discussing the three-dot icon that leads to the list of all collocations with a specific syntactic structure*	32	PRO
Filter (frequency) *Segments discussing the function that allows the collocates to be filtered by corpus frequency*	43	PRO
Filter (alphabetical) *Segments discussing the function that allows the collocates to be sorted alphabetically*	14	PRO
Filter (relevance) *Segments discussing the function that allows the collocates to be sorted by relevance*	4	PRO

4 The fourth category – Participant suggestions – was included in the typology as a catch-all category for any user suggestions that did not fit in any of the other (more finegrained) categories. These segments were also annotated in the transcriptions.

Colour scale for relevance *Segments discussing the fact that collocates are colour-coded by relevance*	56	PRO
Collocate clusters *Segments discussing the function to display automatically generated collocate clusters*	39	PRO
Links to Gigafida *Segments discussing links to the Gigafida corpus of Slovene*	39	PRO
Other links *Segments discussing other links in the dictionary*	14	PRO
Corpus examples *Segments discussing corpus examples included in the dictionary*	44	PRO
Other resources *Segments discussing other resources*	12	PRO
Navigation menu *Segments discussing the navigation menu that allows the user to filter collocation by syntactic structure*	82	PRO
User votes *Segments discussing the option for users to up- or downvote collocations*	78	PRO/ CONTRA
Noise in dictionary data		
Errors (definite form of adjectives) *Segments discussing the lack of definite forms in adjectival collocations*	6	PRO/ CONTRA
Errors (homonyms) *Segments discussing errors with homonymous headwords*	63	CONTRA
Errors (proper nouns) *Segments discussing proper nouns included in the dictionary*	62	PRO/ CONTRA
Errors (prepositions) *Segments discussing errors with prepositions*	5	PRO
Errors (comparative form of adjectives) *Segments discussing the lack of obligatory comparative forms of adjectives*	13	PRO/ CONTRA
Errors (reflexive pronoun) *Segments discussing the lack of the reflexive pronoun in collocations containing inherently reflexive verbs*	61	PRO/ CONTRA
Errors (missing collocation element) *Segments discussing the lack of additional collocation elements in multi-word collocations*	59	PRO/ CONTRA
Errors (negative form) *Segments discussing the lack of negative forms in collocations that require the presence of a negative particle*	17	PRO
Errors (other) *Segments discussing other errors related to noise found in the dictionary*	136	PRO/ CONTRA
Participant suggestions		
Different participant suggestions regarding the potential improvements of the dictionary	215	

3 DATA ANALYSIS OVERVIEW

The initial overview and analysis of categorized opinions included all the structural and thematic segments covered by the evaluation interview (Appendix 1): examining the intuitiveness of the dictionary interface, the participants' attitudes towards errors and selected general features of the dictionary. All the assessed categories mentioned above were divided into groups according to predominant opinion on their adequacy (the category is marked by PRO) or inadequacy (the category is marked by CONTRA) (Table 3).[5] We were interested in determining the areas in which the participants agreed or disagreed. This data is relevant for identifying problematic and less problematic categories, and for further improvements of the dictionary interface.

An example of an opinion[6] marked by PRO:

> [1] *"Fantastic! In my opinion, digitalization is the only way of coming up with useful dictionaries."* [teacher of Slovene as a second/foreign language, on the digitalization in lexicography]

An example of an opinion marked by CONTRA:

> [2] *"I'm put off by mistakes, because I find this slows down my work considerably."* [translator, on automatic noise in dictionary data]

3.1 Evaluating Features of the User Interface

The first part of the interview involved the participant exploring the dictionary features in a free and unstructured manner. The aim was to evaluate the intuitiveness of the user interface, e.g. the entry phase indicator (pyramid icon), the presence or absence of sense indicators (sense menus), the three-dot icon for accessing specific syntactic structures, etc.

As shown in Table 3, the participants from all groups described all the selected features as positive (PRO): they rated them as excellent, highly useful,

5 For time and resource constraints, we leave the exact distribution of PRO and CONTRA opinions for a future paper on this subject, in which we also intend to analyze the distribution of annotations between users and user groups.

6 In order to facilitate reading, all the participant statements were edited to conform to standards of written language. Where the provided context makes it difficult to discern what the statement (or part of the statement) refers to, an explanation or the concrete referent was added in angular brackets – [ ].

functional and intuitively designed dictionary elements. The participants highlighted the clarity of use and the practicality of individual filters, the inclusion of sense indicators, visual indicators of entry completeness, and especially the links to corpus examples, i.e. the use of collocations in actual language use:

[3] *"These examples, to me, they're the best thing about this, because I really, really missed them, yes. There's very few of them in SSKJ [the General Monolingual Dictionary of Standard Slovene], but here you can really... In fact, a single entry gives you a lot of information. That's great, you can really find whatever it is that you need—a really useful thing, this."* [teacher of Slovene as a first language, on the relevance of corpus examples]

[4] *"Straight away, I find this pyramid icon great. But I would have a pyramid, from the outset, where all these lines would be thicker, stronger."* [lexicographer, on the entry phase indicator]

[5] *"I find this great. This thing where everything is sorted according to meaning... Especially for our foreign learners, so they can limit themselves to this, to this single meaning."* [teacher of Slovene as a second/foreign language, on sense menus in the dictionary]

None of the participants expressed arguments against any of the features. However, we have identified a common suggestion (across all participant groups) for improvement relating to the visual upgrade of the pyramid icon, i.e. the icon should be more noticeable and its function clarified. Divergent opinions (PRO/CONTRA) were noted with regards to the possibility of user involvement. All the participants see the option of up- or downvoting the collocations as a useful and welcome feature; proof-readers and translators, however, pointed out that they often lack time for doing so, whereas the teachers expressed concern about the feature being used by non-competent users:

[6] *"I have very mixed feelings about this. If the idea is that this is only intended for more advanced users, then this is a great option. But if I think of showing this to the children in primary school and then they would click away and play a little, I think they could really spoil this situation here."* [teacher of Slovene as a first language, on the dictionary's voting feature]

[7] *"Yes, I definitely find this great. I often notice these mistakes in a lot of places, and others notice them, too, when I'm reading online news, and I notice things being misspelled. But I can't be bothered to register only to bring attention to the mistake. I mean, if I could do it, I suppose I would, sometimes. So I think it's great that this here is made in such a way that the user can immediately point out a mistake."* [teacher of Slovene as a second/foreign language, on the convenience of not having to register to provide user votes]

3.2 Evaluating Data Error Distraction

The second part of the interview, which focused on examining the participants' attitudes to various types of errors, demonstrated that the participants—judging by their response to test entries and their self-reports on previous, oftentimes daily dictionary use—mostly do not seem to notice them. In fact, they seemed to first become aware of the errors only during their participation in the user study, after being guided in their work on specific entries (*belina*, *pivo*, *klop* and *usesti* (*se*), i.e. after being systematically queried whether they noticed any errors and asked about the extent of their disruption.[7]

Prompted by the interviewer, the participants evaluated specific types of errors, such as the absence of the reflexive pronoun *se* in the verb headword, errors due to homonymy, the inclusion of proper nouns in the dictionary, etc. As seen in Table 3, the most distracting type of error occurs due to homonymy and was mostly independently detected by the participants. In the headword *klop*, homonymy results in most of the collocates being wrong (*greti klôpa* 'to keep a tick warm' – instead of *greti klóp* 'to keep a bench warm', *guliti klôpa* 'to wear out a tick'– instead of *guliti klóp* 'to wear out a bench', *sesti v klôpu* 'to sit on a tick' – instead of *sesti v klopí* 'to sit on a bench').[8]

The participants also had mixed opinions (PRO/CONTRA) on the inclusion of proper nouns in the dictionary. Due to the diversity of opinions on this issue and some very interesting results, we examine the issue in more detail in Section 4. The participants marked all the other shortcomings (i.e. types of errors) with CONTRA, and mostly did not notice them independently during their work with dictionary entries, as mentioned above:

7 It should be noted that the above was not true for the group of lexicographers—unlike the other participants, who encountered such errors for the first time, the lexicographers were well acquainted with the dictionary. Namely, the group of lexicographers included many of the original authors involved in the diverse stages of the building of the collocations dictionary (data processing, user interface design, and other processes of development).

8 Homonymy-related problems can occur because of incorrect morphosyntactic tagging and/or problems in post-processing. One particular issue of corpus data is that lemmas are form-based, so differently-pronounced headwords with the same form will be combined under the same lemma. The problems become particularly noticeable when such a word (as a headword or a collocate) features in the grammatical structure in a case that is not nominative.

[8] *"I don't know, I wasn't really distracted... If you hadn't told me, I wouldn't even have noticed. I think that as soon as I saw it, I somehow already imagined the correct meaning and then got the meanings of the sort I was thinking about."* [teacher of Slovene as a first language, on the]

[9] *"These are mistakes of the kind where the petty Slovene mind, which would rather criticise than help or praise, could say: there, I knew it, I found a mistake right away."* [translator, on dictionary errors]

[10] *"Because, for instance, we've been using it [the Collocations Dictionary] now [in class], we've had a look at quite a number of things, at least those that were in the texts, and we haven't found a single mistake, not a single problematic thing. So, I think, well, you really have to try hard to find a page where something bothers you. To the point that you find the page useless."* [teacher of Slovene as a second/foreign language, on the scarcity of errors in the Collocations Dictionary]

[11] "Because the user knows in advance [to expect mistakes], I don't think it's a problem, no. Because then, even someone who is learning Slovene, they know not to trust it blindly. So I think that even in this stage, this phase, this resource is really valuable." [teacher of Slovene as a second/foreign language, on the usefulness of the Collocations Dictionary]

3.3 Evaluating General Features of the Dictionary

In the final part of the interview, the participants evaluated the general features of the collocations dictionary, such as its automatic compilation, digital-only form, and look/design.

As shown in Table 3, all the above features were positively evaluated by all the participant groups. The reasons were mostly unanimous. The participants find the Collocations Dictionary a clear and coherent resource, with relatively clearly recognizable functions; translators and proof-readers see it as an invaluable resource; the teachers consider it an extremely useful one (both for the preparation of didactic exercises and for classroom use, e.g. to check the adequacy of phrases, find expressions typical for newspapers, works of fiction, etc.); its strengths are its authenticity, the interconnectedness of its language data, and the relative ease of use in comparison to corpora. Its look and the distribution and density of data are clear and user-friendly, whereas its digital-only form, which enables continuous upgrades and updates, is functional, indispensable and a necessary precondition for work in modern times.

[12] *"I believe that these two dictionaries [the Thesaurus of Modern Slovene and The Collocations Dictionary of Modern Slovene] are the best thing that has happened to Slovene in the past few years, I really do. And the people are infinitely, truly grateful, for having these resources."* [proof-reader and translator, on their attitude toward responsive dictionaries]

[13] *"So I really enjoyed it today when we could show this to the foreign learners: 'Here, this is the entire selection [of collocates]. There are some things that are not in accordance with the orthography manual, and a newspaper proof-reader might correct a lot of things, but you encounter all of this in every-day language. Everything you see here is real-life language.' So it's great that these dictionaries exist and offer so many options. Because this is what foreigners often experience: 'Well, I heard someone say this on the street, but where can I check if it's OK?" And then, with Fran [Slovene dictionary portal] or, I don't know, the orthography manual, well, there's nothing there. For a foreign learner there's not enough headwords in there. It's much easier to browse through this than it is directly through corpora. I find this dictionary much more user-friendly than corpora."* [teacher of Slovene as a second/foreign language, on the usefulness of the Collocations Dictionary for foreign learners]

[14] *"It's nice and user-friendly, because it's so clean and clear and there's enough space, the page isn't crowded. Yes, I like it and those shades of grey aren't too conspicuous, it's clear, well, I like it. Here, the titles are nicely listed, so you know what you're looking for, down here you get the collocations, great. So I find it ... Well, I'd just like to say well done, really, great."* [teacher of Slovene as a second/foreign language, on the user-friendliness of the Collocations Dictionary]

[15] *"I don't find the fact that it's in digital-only form a disadvantage at all. It's an advantage, really, because it takes less time to access it and precisely because you can correct it, update it, improve it. Because if this wasn't the case, then you could wait forever for such a dictionary, and in the meantime expressions go out of use, or maybe not out of use, but new things come along, the language develops and so the dictionary would be left behind."* [translator, on the advantages of a digital-only dictionary form]

3.4 Participants' Improvement Suggestions

While evaluating specific interface features, the participants also suggested several improvements on their own initiative. The suggested improvements included adding information on the collocate or collocation frequency, the option to export data, the addition of accents and pronunciation to headwords (especially homonymous headwords). The bulk of suggestions was primarily concerned with the option to click on the headword in order to return to the initial page, the visual upgrade of specific interface elements, such as upgrading the frequency filter with a color scheme or a color code, making the pyramid icon more graphically pronounced by enlarging it, using intense colors or stripes, including a short headline, description, etc.

4 QUALITATIVE CASE ANALYSIS: PROPER NOUNS

In this section, we describe a qualitative analysis of the participants' attitude towards the inclusion of proper nouns. *The Collocations Dictionary of Modern Slovene 1.0* includes proper nouns as collocates, but not as headwords.[9]

While *the Collocations Dictionary* was under development, lexicographic discussions frequently highlighted the problematic nature of proper nouns. Because they refer to a single, specific referent, they are semantically specific and often bring into question the relevance of the dictionary entry. A typical example of this includes headwords which necessitate a longer sequence enumerating collocates of the same type, e.g. geographical proper nouns: *prestolnica* [*Slovenije*, *Štajerske*, *Rusije*] 'the capital of [Slovenia, Styria, Russia]', *bivati v* [*Sloveniji*, *Rusiji*, *Ukrajini*] 'to live in [Slovenia, Russia, the Ukraine]', or adjectives derived from proper nouns: [*slovenski*, *angleški*, *nemški*, *češki*] *jezik* '[Slovene, English, German, Czech] language', etc. Aside from data overload, the inclusion of proper nouns may also lead to difficulties by adding potentially recognizable personal names (personal data), trademarks, etc. On the other hand, their complete exclusion may lead to omitting an important segment of vocabulary which, statistically speaking, conforms to collocation criteria (type, frequency, occurrence).

The complexity of this issue and its possible solutions were reflected in the results of the participants' evaluation. Most participants supported the inclusion of proper nouns in the dictionary (see Table 3). However, all the participant groups identified reasons both for and against the inclusion. This was especially pronounced in the group of lexicographers, where all the participants listed reasons both for and against the inclusion. Table 4 gives an overview of the above discussed opinions within individual groups.

9 However, it should be noted that the Collocations Dictionary does include headwords derived from proper nouns which, in Slovene, begin with lower-case initials (as opposed to many foreign languages in which the opposite is often the case). The dictionary thus contains e.g. adjectives derived from proper nouns, such as *slovenski* 'Slovene', *angleški* 'English', *nemški* 'German', etc.

Table 4: *An overview of participant attitudes (PRO, CONTRA, PRO/CONTRA) towards inclusion of proper nouns across individual groups*

	PRO	CONTRA	PRO/CONTRA
Teachers of Slovene as L1	9	0	1
Teachers of Slovene as L2	9	1	0
Translators, proof-readers	6	3	1
Lexicographers	0	0	10

4.1 Attitude of Teachers of Slovene as a First Language

The majority of teachers of Slovene as a first language (Table 4) had a positive attitude towards the inclusion of proper nouns, especially for the following reasons:

- the students find them more illustrative and concrete;
- they pique the interest of students and promote intellectual and cognitive processes;
- their specificity is attractive and intuitive, which is reflected in increased study motivation of the student and, consequently, in a more flexible understanding and adequate language use.

While giving a positive evaluation of the inclusion of proper nouns because of their ability to illustrate and convey a more specific example of language use, one of the teachers expressed doubts regarding the benefits of including trademarks (e.g. *Laško pivo*, a Slovene beer brand) and questioned their contribution towards understanding word use.

4.2 Attitude of Teachers of Slovene as a Second/Foreign Language

Almost all teachers of Slovene as a second language (Table 4) find the inclusion of proper nouns important because they give useful information on the morphological characteristics of a particular part-of-speech category, such as declension patterns or the use of prepositions with proper nouns (a frequent problem for foreign learners, e.g. *potovati na* [*Hrvaško, Kitajsko*] 'to travel to [Croatia, China]', but *potovati v* [*Evropo, Azerbajdžan*] 'to travel to [Europe, Azerbaijan]'. There was a suggestion to exclude specific types of proper nouns, such as personal names and surnames.

As seen in Table 4, only one of the teachers was of opposed to proper nouns. The teacher pointed out several proper nouns incorrectly spelled with a lower-case initial letter (*večernji list* 'evening newspaper' instead of *Večernji list* 'Evening Newspaper'; *smučati v dolomitih* 'to ski in the dolomites' instead of *smučati v Dolomitih* 'to ski in the Dolomites'), which might cause difficulties for students trying to learn the language. An incorrectly spelled proper noun may mislead a foreign learner who is incapable of recognizing or disambiguating language mistakes; it can provide misleading information on orthography and the role of particular part-of-speech categories and their inflections in phrases and syntactic structures. The above examples may misinform the learner about the proper form and use of the deadverbial adjective (*večernji* instead of *večerni*) or the correct use of the common noun (*Dolomiti* as the Italian mountain range instead of *dolomiti* as a mineral).

4.3 Attitude of Proof-Readers and Translators

6 out of 10 participating proof-readers and translators gave reasons in favour of the inclusion of proper nouns (Table 4). Much like the teachers of Slovene as a first language, they recognised the quality of intuitiveness arising from the concreteness of proper nouns: the collocation *klop Reala* 'the bench of Real [Madrid]' or *klop Liverpoola* 'the bench of Liverpool' may be more illustrative and meaningful than *klop prvoligaša* 'first league bench', where the lack of context may make it difficult to determine that this is a football club.

On the other hand, a smaller number of proof-readers and translators—3 out of 10—argued against the inclusion, especially in relation to trademarks (e.g. *Illy kava* 'Illy coffee', *Laško pivo* 'Laško beer'), since they find this degree of specificity meaningless and unnecessary. Furthermore, one of the participants had a mixed opinion, since they believe that the decision regarding the inclusion of proper nouns in the dictionary depends primarily on the type of proper noun and the relevance of the information conveyed by the proper noun.

4.4 Attitude of Lexicographers

As already mentioned above, all the participating lexicographers expressed arguments both for and against the inclusion (Table 4), which is to be expected considering the fact that they see the dictionary not only from the

perspective of the user, but also as content developers and originators of lexicographic concepts.

The arguments for the inclusion were related to semantically relevant proper nouns; the participants stressed that not all proper nouns are equally semantically relevant (*kranjski Janez 'John Doe' – Janez Novak*; *delati se Francoza* 'lit. to pretend to be a Frenchman, meaning *to feign ignorance' – Francoz* 'Frenchman'). Proper nouns were also considered a valuable source of information on the most typical ways of addressing people, with the caveat that the specific personal name in and of itself is not that relevant (*dragi Janez* 'dear Janez' – *dragi* + [personal name]); the key information here is the discourse category.

The arguments against the inclusion were related to longer sequences of collocates of the same type, since this type of information is distracting and does not enhance user experience. This is the case for the selected entries *klop* and *pivo*, where there is a longer sequence enumerating adjectives derived from proper nouns: [*češko*, *belgijsko*, *angleško*, *dansko*] *pivo* '[Czech, Belgian, English, Danish] beer' or geographical proper nouns (e.g. names of cities): *klop* [*Celja*, *Maribora*, *Kopra*, *Gorice*] 'the bench of [Celje, Maribor, Koper, Gorica]'.

4.5 Participants' Suggestions for Dictionary Improvements

The participants suggested two solutions on the topic of inclusion and presentation of proper nouns in the dictionary.

The proof-readers and translators suggested an introduction of a special button for hiding the proper noun candidates; this would give them the option to choose whether to use it and thus make querying the dictionary more efficient. Their work is related to the specific nature of various text types and vocabulary, the variety of topics subject to intense linguistic research, as well as time as one of the key components, which is why this group believes that the dictionary should adjust to the needs, wishes, and expectations of its target users as much as possible.

Lexicographers proposed a solution of grouping collocates belonging to the same semantic type under a semantic label (e.g. football, hockey, basketball > sport; dog, cat, hamster > (domestic) animal). This would improve the

visibility of collocational behaviour of the word and ease browsing through (long) lists of collocates.

5 DISCUSSION AND METHOD ASSESSMENT

The user evaluation of *The Collocations Dictionary of Modern Slovene 1.0* identified the participants' attitudes towards its features, which were grouped in three discrete segments in the research interview. The user evaluation was, to a great degree, positive. In the first segment of the interview, the participants evaluated as positive (i.e. relevant for the dictionary and useful) all the features that they independently recognized. In the guided part of the interview (during which they worked with selected entries), the participants expressed reservations about some (but not necessarily all) data errors, especially mistakes arising as the result of homonymy and ambiguous word inflections. Opinions also differed with regards to the (non-)inclusion of proper nouns (as seen in Section 4). The third and final segment of the interview asked the participants to evaluate general dictionary features; here, also, their opinion was unanimously positive.

The analysis of the participants' attitudes towards errors has demonstrated that even in their initial stage (during which they still contain mistakes), responsive dictionaries represent an invaluable tool—this was a common opinion across all participant groups taking part in the study. In order to understand this degree of positive or permissive attitudes towards data errors, we need to keep in mind that before the publication of the *Collocations Dictionary of Modern Slovene*, collocation data for Slovene had not been readily available. To a great extent, the participants' enthusiasm is thus a reflection of the newly opened possibilities offered by the dictionary—it is, therefore, safe to conclude that the participants prefer easy accessibility over fully clean data. The evaluation further demonstrated that: a) it is vital that dictionary users are alerted to the presence of errors with the pyramid icon, which indicates the phase of entry completeness; and b) given the presence of context, the possibility of accessing examples, and links to the Gigafida corpus, it is possible for the users to resolve any ambiguities.

In terms of dictionary shortcomings, special attention should be given to the most "vulnerable" user groups, i.e. teachers of Slovene as a first language and

teachers of Slovene as a second/foreign language. Teachers bear the responsibility of choosing the sources used in the classroom with students who as language learners are somewhat less qualified to independently identify and resolve data ambiguities in the manner described above. Didactic use demands precise and unambiguous information, so that the teacher does not lose time by having to correct errors. On the other hand, the teachers themselves found the dictionary to be very useful and of great help, especially as a starting point for exercises, a tool for enriching vocabulary, for checking the correctness and adequacy of phrases; for writing fiction and poetry, for discussing collocations, using idioms, newspaper language, etc. They were excited by the authenticity of the language, the interconnectedness of different resources, and especially by the possibility to observe language as a natural phenomenon across all segments of its use.

What is important is that the study made it clear that many of the characteristics that were deemed problematic by linguists are not necessarily problematic for the users—this was seen, for instance, in the discussion of the participants' attitudes towards the inclusion of proper nouns. Contrary to our expectations, the particpants found proper nouns to be interesting and illustrative despite referring to a specific referent. Whereas the lexicographers' main concern was that the inclusion may result in overcrowding the dictionary (e.g. in cases where the headword is followed by a long, enumerating sequence of collocates of the same type), the participants found such concreteness more intuitive.

The evaluation identified areas of the dictionary and its interface which the participants find adequate and those that need to be re-examined, improved and further assessed. In this sense, the study achieved its main goal and the selected method proved to be successful. Even though collecting, recording and categorizing evaluation data is extremely time consuming, the transcribed opinions offer insight into problems and solutions that significantly contribute to concepts proposed by dictionary developers. The evaluation study has resulted in a number of positive findings, but also revealed possibilities for improving the methodology in case of further, comparable studies.

One of the positive aspects of the study was its multi-stage design (i.e. interviews – transcription – annotation – analysis): on the one hand, it enabled a

careful and thorough planning of the entire process of the study; on the other, it increased the time needed to realize individual tasks. The study took place between May and September 2019, with the time span depending on several outside factors: the availability and flexibility of the participants, their willingness to co-operate, collaboration with students, and unforeseen technical difficulties. Apart from demonstrating the need to plan for a longer time span, our experience has also shown the following:

- in order to secure participation, it is very important to adopt a personal approach, including personal correspondence, willingness to record sessions in the participants' place of work, etc.;
- collaboration with students demands careful and consistent monitoring of their work, including providing clear and understandable guidelines and a detailed examination of the transcriptions and annotations;
- a methodological process reliant on the use of recording software and equipment and the use of a digital dictionary should take into account potential technological difficulties and provide for adequate data backup.

6 CONCLUSION

The user evaluation of the *Collocations Dictionary of Modern Slovene* has proven to be a highly efficient way to detect (non-)problematic dictionary features and represents a solid foundation for further attempts to improve and upgrade the interface to make it more user-friendly and functional. It presents a model for evaluation and identification of user problems; the gathered results reveal areas for potential methodological improvements and are thus useful for similar lexicographic user studies and analyses.

The findings of the study indicate that the methodology of automatic extraction of lexical data has indeed reached the levels where such data can be immediately presented to the users, something that has been often claimed by authors such as Kilgarriff et al. (2013) and others. Nonetheless, what the study also shows is that the presentation of such data matters, i.e. features are needed that alert the users to the different stages of data validation and that enable data manipulation/filtering. Part of the reason for this need lies in the

quantity of automatically extracted data which always exceeds the quantity after human clean up and selection.[10]

As envisaged when preparing the study, the user feedback obtained will be used in the preparation of the next version of the *Collocations Dictionary of Modern Slovene*. First and foremost, we need to acknowledge that no radical changes are needed; to some extent, the aspects of data quality and quantity, as well as clarity of presentation, need to be addressed. For example, we plan to introduce additional options to filter collocates, such as an option to hide proper nouns (as opposed to removing them from the dictionary completely), hiding or downgrading semantically less relevant collocates, and viewing a selection of top collocations (or collocate clusters) regardless of their syntactic structure. In terms of visual improvements, the pyramid icon will be made more conspicuous. In cases where the distribution of collocations over syntactic structures is uneven, structures with more collocations will receive more space in the display. Moreover, an option for downloading entries will be added.

As evidenced by the results of the study, user groups differ in their attitude towards the inclusion of proper names, which makes it difficult to propose universal answers for this issue. Solutions that introduce a choice for the user (as the on/off buttons), seem to be a way to go for such cases. Nonetheless, one feature that seemingly requires a rethink is the option of user participation; to this end, we are already testing other approaches such as gamification, which may help us clean the dictionary data even faster and less obtrusively than existing voting method in the dictionary. And gamification, in combination with improvements to the automatic data extraction method, will make the dictionary even more »responsive«.

Acknowledgments

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.

10 This is also the rationale behind the pyramid icon – wider at the bottom in the initial stages, and narrower at the top when the entry is completed.

P6-0411, *Language Resources and Technologies for Slovene*). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 731015. The research was conducted within the framework of the CA160105 eNetCollect COST Action. The authors would also like to thank Bojan Klemenc for his assistance in setting up the local installation of Taguette, all the users of the *Collocations Dictionary of Slovene*, and the annotators who participated in the transcription/annotation campaign: Jan Gajski, Tjaša Jelovšek, Saša Jenko Pahor, Manja Kraševec, Manja Ocepek, Chiara Vianello and Karolina Zgaga.

REFERENCES

Arhar Holdt, Š., Kosem, I., & Gantar, P. (2016). Dictionary user typology: the Slovenian case. In T. Margalitadze & G. Meladze (Eds.), *Lexicography and linguistic diversity. Proceedings of the XVII EURALEX International Congress, 6–10 September, 2016* (pp. 179–187). Tbilisi: Ivane Javakhishvili Tbilisi State University.

Arhar Holdt, Š., Čibej, J., & Zwitter Vitez, A. (2017). Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography, 30*(3), 285–308.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., & Robnik Šikonja, M. (2018). Thesaurus of modern Slovene: by the community for the community. In J. Čibej et al. (Eds.), *Lexicography in global contexts. Proceedings of the XVIII EURALEX International Congress, 17–21 July, 2018, Ljubljana* (pp. 401–410). Ljubljana: University Press, Faculty of Arts.

Atkins, B. T. S. (Ed.). (1998). *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.

Barnhart, C. L. (1962). Problems in Editing Commercial Monolingual Dictionaries. *International Journal of American Linguistics, 28*(2), 161–181.

Bergenholtz, H., & Johnsen, M. (2013). User Research in the Field of Electronic Dictionaries: Methods, First Results, Proposals. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp.

556–568). Berlin/New York: Walter de Gruyter.

Bogaards, P. (2003). Uses and users of dictionaries. In P. van Sterkenburg (Ed.), *A practical Guide to Lexicography* (pp. 26–33). Amsterdam in Philadelphia: John Benjamins.

Čibej, J., Gorjanc, V., & Popič, D. (2016). Analysing translators' language problems (and solutions) through user-generated content. In T. Margalitadze & G. Meladze (Eds.), *Lexicography and linguistic diversity. Proceedings of the XVII EURALEX International Congress, 6–10 September, 2016* (pp. 158–167). Tbilisi: Ivane Javakhishvili Tbilisi State University.

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

Hartman, R. R. K. (1987). Four Perspectives on Dictionary Use: A Critical Review of Research Methods. In A. P. Cowie (Ed.), *The Dictionary and the Language Learner* (pp. 11–28). Tübingen: Niemeyer.

Householder, F. W. (1967). Summary Report. In F. W. Householder & S. Saporta (Eds.), *Problems in lexicography* (pp. 279–282). Bloomington: Indiana University Publications.

Kilgarriff, A., Husak, M., & Jakubíček, M. (2013, October). Automatic collocation dictionaries. Presented at eLex 2013 conference, Tallinn, Estonia. Retrieved from https://youtu.be/b3KyhPBeoLU

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S. et al. (2018a). The image of the monolingual dictionary across Europe: Results of the European survey of dictionary use and culture. *International Journal of Lexicography*. doi: 10.1093/ijl/ecy022

Kosem, I., Wolfer, S., Lew, R., & Müller-Spitzer, C. (2018b). Attitudes of Slovenian language users towards general monolingual dictionaries: an international perspective. *Slovenščina 2.0: empirical, applied and interdisciplinary research 6*(1), 90–134. Ljubljana: University Press, Faculty of Arts. Retrieved from https://revije.ff.uni-lj.si/slovenscina2/article/view/8142/8467

Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š., Čibej, J., & Laskowski, C. (2018c). Collocations dictionary of modern Slovene. In J. Čibej et al. (Eds.), *Proceedings of the XVIII EURALEX International Congress, 17–21 July, 2018, Ljubljana* (pp. 989–997). Ljubljana: University Press, Faculty of Arts. Retrieved

from https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Kosem, I. et al. (2019). *Collocations Dictionary of Modern Slovene KSSS 1.0*. Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1250

Lew, R., & De Schryver, G. M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography, 27*(4), 341–359.

Lew, R. (2015). Research into the Use of Online Dictionaries. *International Journal of Lexicography, 28*(2), 232–253.

Logar, N. (2009). Slovenski splošni in terminološki slovarji: za koga? In M. Stabej (Ed.), *Infrastruktura slovenščine in slovenistike. Obdobja 28* (pp. 225–231). Ljubljana: Znanstvena založba Filozofske fakultete.

Müller-Spitzer, C. (Ed). (2014). Using Online Dictionaries. *Proceedings of the XVIII EURALEX international congress*. Berlin, Boston: De Gruyter Mouton.

Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer Verlag.

Rampin, R., Steeves, V., & DeMott, S. (2019). *Taguette* (Version 0.8). Zenodo. doi: 10.5281/zenodo.3246958

Rozman, T. (2004). Upoštevanje ciljnih uporabnikov pri izdelavi enojezičnega slovarja za tujce. *Jezik in slovstvo, 49*(3–4), 63–75.

Stabej, M. (2009). Slovarji in govorci: kot pes in mačka? *Jezik in slovstvo*, *54*(3–4), 115–138.

Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos, 19(1)*, 275–296.

Thumb, J. (2004). Dictionary Look-up Strategies and the Bilingualised Learner's Dictionary. *Lexico-graphica (Series Maior 117)*. Tübingen: Max Niemeyer.

Tomaszczyk, J. (1979). Dictionaries: Users and Uses. *Glottodidactica 12*, 103–119.

Welker, H. A. (2013a). Methods in Research of Dictionary Use. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 540–547). Berlin, New York: Walter de Gruyter.

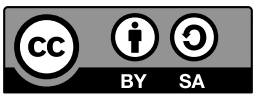
Welker, H. A. (2013b). Empirical Research into Dictionary Use since 1990. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 531–540). Berlin, New York: Walter de Gruyter.

Wingate, U. (2002). The Effectiveness of Different Learners Dictionaries: An Investigation into the Use of Dictionaries for Reading Comprehension by Intermediate Learners of German. *Lexicographica (Series Maior 112)*. Tübingen: Max Niemeyer.

ODNOS UPORABNIKOV DO AVTOMATSKO PRIDOBLJENIH KOLOKACIJSKIH PODATKOV: UPORABNIŠKA RAZISKAVA

Prispevek izhaja iz uporabniške raziskave, izvedene v okviru temeljnega raziskovalnega projekta Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (KOLOS; J6-8255). Prikaže analizo uporabniške evalvacije vmesnika *Kolokacijskega slovarja sodobne slovenščine* (KSSS). Z nekoliko drugačnega gledišča – skozi uporabniški aspekt pokaže, kje in katera so problematična mesta posamezne slovarske kategorije, ki so potrebna nadaljnje leksikografske obravnave in diskusije. Kolokacijska uporabniška študija predstavlja model procesa uporabniškega evalviranja, ugotovitve, ki jih prinaša, pa bodo predvsem relevantne za detekcijo uporabniških problemov, pa tudi za izboljšavo metodologije, kar bo predvsem koristno za primerljive leksikografske uporabniške raziskave in analize.

Keywords: kolokacijski slovar, odzivni slovar, uporabniška evalvacija, odnos do napak, slovarski vmesnik



APPENDIX 1: EVALUATION QUESTIONNAIRE

First segment: Free use of the dictionary

During the first interview segment, the participants are asked to browse the dictionary freely while thinking aloud. This allows them to form the first impression and get the general sense of the dictionary.

Second segment: Guided work with dictionary headwords

In the second part of the interview, the participants are guided by the interviewer to click on a number of headwords that were pre-selected according to a carefully designed set of criteria. The participant is thus familiarized with the various functions offered by the resource.

The participant is presented with the following headwords:

belina 'whiteness' – a non-problematic entry that has already been finalized by lexicographers

> How do you find this headword? Is it in any way problematic? Do you notice any errors? Can you identify the various functions available (e.g. the entry phase indicator, sense menus, collocate clusters), the possibility of using various filters, the option to contribute to the dictionary by rating collocations?

pivo 'beer' – an entry with potentially problematic collocates

> Do you notice that the noun/adjective (collocate or headword) is not in the expected inflected form? Does this motivate you to refer to the corpus examples provided? Are you bothered by this type of errors (semantic nonsense)?
>
> **A selection of the identified errors (on the levels of collocate/headword, collocation structure or collocation):**
>
> - The collocate is incorrectly lemmatized: *plata piva* 'plate of beer' instead of *plato piva* 'box of beer [cans], lit. plateau of beer cans'
> - The collocate/headword should appear in a specific inflected form (e.g. comparative, plural): *drag od piva* 'expensive than beer' instead of *dražji od piva* '[more expensive] than beer'
> - The headword appears next to a collocate tagged with wrong part-of-speech: *pivo pite* 'beer of pie' instead of *pivo piti* 'to drink beer'
> - The verb collocate of the noun headword does not appear in the negative form (as required by the genitive case of the headwod): *piti piva* 'to drink beer' instead of *ne piti piva* 'to not drink beer'

- The collocation makes no sense out of context or without additional elements: *pivo k ustom* 'beer to the mouth' as in *dvigniti kozarec piva k ustom* 'to raise [a glass of] beer to the mouth'
- The headword is either a plural noun or appears next to a syntactic structure in the genitive plural; as such, the collocation makes no sense without an additional, quantitative element: *pivo po tolarja* 'beer for tolar' instead of *pivo po 300 tolarjev* 'beer for 300 tolars'

klop 'bench' or 'tick' – a homonym that has not been disambiguated in the dictionary

Do you find anything about the entry distracting? Did you identify the word as a homonym (words having the same spelling but different meanings)? Do you find the ambiguity distracting? Are you distracted by proper nouns as collocates? Do you find that there are too many errors?

usesti (se) 'to sit (oneself) down' – an inherently reflexive verb which is missing the obligatory *se* pronoun in the dictionary

[The participant first enters the word into the search window; the interviewer observes their reaction and then continues with the questions.]

Did you notice the absence of the *se* pronoun? (*or* Does the lack of reflexivity (*usesti se)* bother you? Do you find that there are too many errors?

Third segment: General dictionary features

Automatic compilation

[The questions are meaningfully incorporated into the discussion about specific headwords.]

In its initial stage, this resource is compiled completely automatically. This is why, as you may have noticed, it also includes information that should not be here. Do you feel there is too much noise or that there are too many errors? Do you find this distracting? Why (not)?

This resource enables dictionary entry tracking and provides information on the phase of entry completeness, generated by clicking on the pyramid icon. Did you notice this? How do you find this?

This resource was compiled automatically and as such was made freely and openly accessible as soon as it was compiled. Do you prefer free and open resources with raw data or payable sources with clean data?

This new form of language resource allows for continuous upgrades and updates; the development team can include new collocations and headwords, the users can vote on collocation candidates, etc. Do you prefer static, unchangeable resources, or are there any advantages to a dictionary that can change over time?

Changes also mean that the dictionary is never fully complete and is continuously developing. How do you feel about that?

User inclusion

[Questions are meaningfully incorporated into the discussion about specific headwords.]

Did you notice it was possible to contribute to the dictionary as a user (i.e. up- or downvote collocates/collocations)?

Do you find user involvement positive or negative?

Once the user up- or downvotes a collocation, their rating immediately appears on the page. How do you feel about this?

Do you find the resource stimulating enough to contribute to it yourself? Would you provide your votes in the dictionary? Why (not)?

What would motivate you to contribute to the compilation of the dictionary? What would additionally motivate you to do so?

Do you have any reservations about user inclusion? *[The participant is given the space to respond first; they are then asked to discuss whether they see user inclusion as shifting the burden of responsibility onto the users by means of crowdsourcing; whether this constitutes taking advantage of the user; whether they are concerned about the potential lack of experience or professionalism in users; whether user judgement may in fact improve the quality of the dictionary, etc.]*

Digital-only form

This resource has no printed version. Is that a problem or do you find its digital-only form an advantage?

Interface

Interface problems

[The interviewer asks specific questions]

Do you find the dictionary useful? What do you like most about it?

What are the main reasons you wouldn't use this dictionary?