Simon Krek/Polona Gantar/Iztok Kosem

EXTRACTION OF COLLOCATIONS FROM THE GIGAFIDA 2.1 CORPUS OF SLOVENE

Abstract This paper describes a method for extracting collocation data from text corpora based on a formal definition of syntactic structures, which takes into account not only the POS-tagging level of annotation but also syntactic parsing (syntactic treebank model) and introduces the possibility of controlling the canonical form of extracted collocations based on statistical data on forms with different properties in the corpus. Specifically, we describe the results of extraction from the syntactically tagged Gigafida 2.1 corpus. Using the new method, 4,002,918 collocation candidates in 81 syntactic structures were extracted. We evaluate the extracted data sample in more detail, mainly in relation to properties that affect the extraction of canonical forms: definiteness in adjectival collocations, grammatical number in noun collocations, comparison in adjectival and adverbial collocations, and letter case (uppercase and lowercase) in canonical forms. The conclusion highlights the potential of the methodology used for the grammatical description of collocation and phrasal syntax and the possibilities for improving the model in the process of compilation of a digital dictionary database for Slovene.

Keywords Collocations; discovering collocations in corpora; digital collocation database

1. Introduction

Large text corpora and tools for their processing created over the last three decades have enabled the development of various methods for the automatic extraction of multi-word units from corpora, mainly for the purpose of compilation of dictionary resources, for natural language processing tools, and for the development of various language applications.[1]

Multi-word expression extraction procedures typically exploit a mechanism that recognises sequences of lexical units on the basis of their morpho-syntactic annotation in the corpus and statistical measures that determine co-occurrence values. The most recognized and established model, especially in the field of lexicography, is the word sketch model in Sketch Engine, which operates on the basis of a word sketch grammar (Krek/Kilgarriff 2006; Krek 2015; Gantar 2015; Kosem et al. 2018) and a lemmatized and POS-tagged corpus.[2] Our aim, however, was to devise a methodology for extracting collocation data from the Gigafida corpus that upgrades the existing system and is based on the assumption that collocation candidates can be successfully extracted from a syntactically parsed corpus, using labelled dependency relations and morphological features of the heads and the dependents in the dependency tree.

In this paper we describe a methodology for automatic extraction of collocations from the Gigafida 2.1 corpus, based on definitions of syntactic relations within a phrase, also taking into account some statistical parameters. First, we present the extraction procedure and the

[1] elexiFinder web service yields 423 results in six languages for the search "collocation": http://finder.elex.is/intelligence?conditions=3-wikidata:Q1122269-collocation&percentile=100&dataType=news&-dataType=video&tab=items&type=articles&articlesSortBy=date (last access: 25-03-2022).

[2] Sketch grammar for Slovene (Krek in Kilgarriff 2006) was used in the Communication in Slovene project (Krek 2015) in the creation of the Slovene Lexical Database (Gantar 2015) and the Collocation Dictionary of Modern Slovene (Kosem et al. 2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

database of extracted collocations (Krek et al. 2021). We evaluate the extracted data based on quantitative and qualitative linguistic analyses. In conclusion, we highlight the potential of the methodology and the resulting open data for the grammatical description of collocations and phrase syntax, as well as the possibilities for improving the model in the construction of a digital dictionary database for Slovene.

2. Automatic extraction of collocations from the corpus

In this section we describe the formal description of collocation structures in an XML file (2.1), which is the core part of the new collocation extraction methodology. The most important part of the description is included in the definition of collocation structures (2.2), which consists of a description of the components of a collocation, the syntactic relations between them, and the various constraints on a) the identification of the components in the corpus, and b) the extraction of the final canonical forms of the collocation. In the last part of the section (2.3), we describe the automatic extraction procedure of collocations from the corpus, based on the proposed system.

2.1 Formal description of collocations

For the purposes of the new methodology, it was necessary to define more precisely the term "collocation", which is described in Gantar/Krek/Kisnik (2021). In defining the morpho-syntactic structure, we started from the previously defined grammatical relations in the Word Sketch tool for Slovene (Krek 2015). Starting with the POS-tagging annotation level, we added a syntactic parsing level, where we defined dependency syntactic relations within a collocation. Statistical and frequency data were considered both at the level of the lemma and the collocation as a whole, which has been shown to be an appropriate procedure in previous automatic extractions of collocations from the corpus (Gantar/Kosem/Krek 2016). At the same time, frequency data were also taken into account when determining the "representation" or the output form of the extracted collocation, i. e. the form in which the collocation should be included in the dictionary. In the process of creating a new formalism for collocation extraction, most of the collocation structures included in the Slovene Lexical Database were translated from the Sketch Engine grammar into the new formalism. The new method differs from the existing Sketch Engine methodology (Kilgarriff et al. 2014) in the following aspects:

- instead of the Corpus Query Language (CQL) used in the Sketch Engine, which mainly takes into account POS-tagging annotation, the new method uses its own system to define constraints on any level of annotation, from morphology (parts-of-speech and their properties), syntactic dependency relations, concrete lexical items, and any other types of annotation that can be used for other purposes, e. g. semantic roles, semantic types, word senses, etc.;
- in the new system, verbal structures are explicitly separated in terms of negation (expressed by a negation particle or by a verb) and reflexiveness (expressed by a free verb morpheme or a reflexive pronoun);
- unlike in the Sketch Engine system, identification numbers and (syntactic structure) names do not differ according to whether the starting point is the first or the second collocator in the collocation;

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

– the human-readable syntactic structure names directly reflect the characteristics of the individual collocation components in terms of their parts-of-speech and grammatical properties, according to the Multext-East/JOS annotation system;
– in addition to the constraints (restrictions) enabled by the CQL system, it is also possible to specify which of the forms of each component found in the corpus should be used in a specific collocation, according to the options offered within the pre-defined canonical collocation form in a specific structure (representation).

The total number of collocation structures in the DDD system is (currently) 82, of which, counting by collocator pair, there are six that include negation, 25 with reflexive verb structures, and 26 combinations with prepositions. Like in the Sketch Engine, collocators belong to four (content) word types: nouns, verbs, adjectives, and adverbs.

For human-readable codes, we use a short combination of the included morphosyntactic categories and features according to the MTE/JOS system (Erjavec et al. 2010, 2011). This is important for linguistic use, while the identification number is important for computational use.

In Table 1 below, we provide an example of a selection of ten structures, the first five according to the number of collocations extracted, the remaining five for the purpose of displaying the tags in all nine columns/categories.

ID	CODE	EXAMPLE	TRANSLATION	1	2	3	4	5	6	7	8	9	COL No.
34	p0-s0	svetovno prvenstvo	world championship		p0						s0		720,605
53	s0-s2	direktor podjetja	company director		s0						s2		518,199
70	s0-gg	raziskava pokaže	research shows		s0						gg		385,018
23	gg-s4	podpisati pogodbo	to sign a contract		gg						s4		270,965
15	gg-d-s5	imeti v mislih	to have in mind		gg			d			s5		235,771
30	p0-vp-p0	domač in tuj	domestic and foreign		p0		vp				p0		32,127
77	s1-gp-s1	nogomet je šport	football is a sport		s1					gp	s1		26,520
72	s0-l-gg	trditev ne drži	the claim is not true		s0				l		gg		19,400
95	l-gg-zp-ggn	ne uspeti se uvrstiti	fail to qualify	l	gg	zp					ggn		479
94	gg-zp-ggn-zp	odločiti se vrniti se	decide to return		gg	zp					ggn	zp	5

Table 1: Collocation structures according to the categories represented and the number of cases extracted

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To create an algorithm for the automatic extraction of collocations from the corpus, we have created a formalism that includes all the necessary information in XML format. This allows for later adaptation, addition, or reduction of structures in further extractions processes. We describe the formalism in more detail below.

2.2 Definition of syntactic structures

The individual syntactic structure is defined by the <syntactic_structure> element, which provides three mandatory attributes. These include:

- structure ID: @id,
- a human-readable code of the structure: @label,
- structure type: @type.[3]

The structure definition relies on specific tag sets and corpus tagging systems, so at the first level under structure in the <system> element we define the tagging system which we will follow. This contains the @type attribute, whose value defines the selected tagging system. In the context of the project where this procedure was developed, we applied the JOS or Multext-East tagging system to the morphosyntactic and syntactic tagging of the Gigafida 2.1 corpus, both at the morphosyntactic and syntactic levels.[4]

Within the specific labelling system, we further define three distinct groups of information:

- the individual words or elements that make up a collocation – the components,
- links between elements at the syntactic level – dependency tree,
- constraints and other information needed to extract collocations – structure definition.

2.2.1 Components

The components are defined in the <components> element containing the @order attribute. This may contain the values "fixed" and "variable". This attribute specifies whether the automatic parsing of the structure and the extraction of the components takes into account the order of the components as specified in the structure definition, or whether the predominant order as it is found in the corpus is taken into account – i. e. the order of the components of a particular collocation that represents the majority of the sentences in the corpus. An example of a structure where the sequence is variable is the adverb-verb phrase r-gg (ID 43), where the output of the collocation will vary according to the typical occurrence of the two elements or the adverb semantic group, e. g. *ostati doma* /to stay at home/ (gg-r) vs. *veliko pomeniti* /to mean a lot/ (r-gg).

All components are listed in the <component> (sub)elements, containing several attributes:

- the component identification number: @cid,
- the component human-readable code: @label,

[3] In this paper we consider 82 structures belonging to the type="collocation". The other types are: type="single" for single-word lexemes and type="other" for multi-word units.

[4] New grammar of contemporary standard Slovene: sources and methods.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

– component type: `@type`,

– component status: `@status`.

The `@label` attribute repeats information from the entire structure code, but only references the part that defines the specific component. The `@type` attribute defines the core of the components and can contain two values: "core" and "other". Core components are the actual components included in the collocation structure, which are defined in the collocation code and are also included in its output. Components marked with "other" are used in cases where, in order to correctly identify a collocation in a specific structure (in a corpus sentence), we need to define additional elements which are either mandatory or prohibited. Components defined with the value "other" in the `@type` attribute must therefore also contain the `@status` attribute, where the values "obligatory" and "forbidden" are allowed. The former specifies that the component must necessarily be in the sentence in which the collocation is found, even if the component itself is not included in the output as part of the collocation. The second value has the reverse role: a component with the status value "forbidden" defined in the structure must not be present in the corpus sentence.

2.2.2 Syntactic relations

The next major unit of structure description is the `<dependencies>` element, which defines the syntactic relations between components. Three attributes (`@from`, `@to`, `@label`) are mandatory in the `<dependency>` (sub)elements, the number of which must correspond to the number of components. An additional (optional) attribute `@order` is possible:

– origin of the dependency tree link (MTE/JOS): `@from`,

– the target of the dependency tree link (MTE/JOS): `@to`,

– link identifier (MTE/JOS): `@label`,

– order of the linked components: `@order`.

The last attribute `@order`, with allowed values "to-from", "from-to" or the default value "any", determines whether the two components associated with this dependency link must be in a specific word order in the sentence or not. In the case of the ID 34 structure given above, the use of the `@order` attribute means that the adjectival modifier must precede the nominal head in the corpus sentence, in order for the collocation to be recognised as corresponding to this structure. The hash character (#) used as a value in the `@from` and `@label` attributes denotes that we do not want to restrict the dependency head of this component, or its label. Therefore, it replaces any origin or label of the link.

2.2.3 Restrictions and output

The most extensive part of the formal description of the structure is the `<definition>` element, with the `<restriction>` element defining the constraints for each component when querying the corpus, and the `<representation>` element defining the variables of the extracted collocations. The latter contains only components that are defined as core and are actually included in the collocation output.

The `<restriction>` element contains a `@type` attribute that specifies at which annotation level the restriction information will be found. Currently, the values "morphology" and "lexis" are used. The first value specifies that the constraints will refer to the POS-tagging

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

annotation level in the corpus. The second value denotes that when identifying a component, we restrict ourselves to specific occurrences, either at the level of the word form or the lemma as found in the corpus.

The `<representation>` element defines variables in the output of the extracted collocations. These will also be found in the `<feature>` element, but with different attributes. The `@rendition` attribute defines the type of information to be used in the output. The values "lemma" and "word_form" specify that we will use either a lemma or one of the word forms of the component as found in the corpus. The value "lexis" in the `@rendition` attribute means that we will use an element that we may not have found in the corpus, but we want it in the output anyway, in place of the component from the collocation. To make this element concrete, we use the `@string` attribute with a chosen string of letters, which represents the actual output in the collocation. An example of such use is negation structures, where we want to control the output of the negation particle *ne* "not", even though the variant *ni* "not" would be more common in the corpus, or the negated forms of the verb *biti* (to be), e. g. *nisem* (I am not).

Furthermore, in the `<feature>` element, the `@selection` attribute (in combination with the `@rendition` attribute) determines which of the possible word forms found in the corpus should be chosen for the output in the collocation. The possible values in the `@selection` attribute are: "all", "msd", or "agreement". The first ("all") means that we include all forms of the component found in the corpus. This is useful, for example, in the case of reflexive pronouns, which have the possible forms *se* and *si* in different combinations, and if both are found in the corpus, both are also rendered in the collocation – *izogibati se/si pogovoru* (avoid the conversation).

The value "msd" in the `@selection` attribute is used in cases where we want to specify which of the forms found in the corpus is chosen for the output, according to its morphosyntactic properties. Individual properties in the same element are defined by combining the property and its value, e. g.

```
<feature selection="msd" case="nominative"/>
```

This means that we want the algorithm to output the (most common) nominative form of the word it found in the corpus.

The value "agreement" in the `@selection` attribute is used in case we want the extracted form of a component to agree in certain properties with the same properties defined in another component, which is defined in the `@msd` and `@head_cid` attributes. The first attribute defines the properties to be matched, the second one refers to the component ID containing the properties to be considered for matching. For example:

```
<feature selection="agreement" msd="gender+number+case"
head_cid="2"/>
```

The example specifies that the two components must agree in gender, case, and number.

The elements described above (in combination with the categories, properties and values in the chosen tagging system) define all 82 collocation structures that we used to extract a total of over 4 million collocations from the Gigafida 2.1 corpus, which we describe below.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Extraction of collocation data from the Gigafida corpus 2.1

For the automatic extraction of collocation candidates, we used the Gigafida 2.0 corpus (Krek et al. 2020), published in 2018. The upgrades from the previous version include, among others, improvements in lemmatisation and POS-tagging, the removal of non-standard texts, and the inclusion of underrepresented and more recent texts. The Gigafida 2.1 version of the corpus, which was used for collocation extraction, also includes an additional level of syntactic parsing, semantic role labelling, and named entity recognition.

The final collocation database (Krek et al. 2021) contains 4,002,918 collocations, automatically extracted from the corpus, based on the definition of 82 collocation structures. The minimum frequency of extracted units in the database is 10. The output is divided by structure into 81 files in tabular format, with comma as a separator (CSV format). The number of files in the database is lower than the number of structures because structure ID-97 (l-gg-zp-ggn-zp, *ne bati se pokazati se* 'not to fear not to show') did not produce results with collocations above a frequency of 10. All collocations are assigned with the following information in 26 columns (Table 2).

Col	Column heading	Description
1	Structure_ID	structure identification number
2	C1_Lemma	lemma of the first component
3	C1_Representative_form	word form of the first component (according to the structure definition)
4	C1_RF_msd	morphosyntactic description for the word form of the first component
5	C1_RF_scenario	scenario for the word form output of the first component
6	C1_Distribution	number of different collocations containing the C1 component lemma (within the structure)
7	C1_lemma_structure_frequency	number of corpus sentences with collocations containing the C1 lemma (within the structure)
8	C2_Lemma	SAME INFORMATION FOR COMPONENTS C2/3/4/5
...	...	...
21	Colocation_ID	collocation identification number
22	Joint_representative_form_fixed	output of the canonical form of the collocation (according to the structure)
23	Joint_representative_form_variable	a list of the most frequent forms of collocation (according to word order)
24	Frequency	frequency of collocation
25	logDice_core	collocation strength calculation (logDice)
26	Distinct_forms	number of different forms of collocation

Table 2: Types of data in the collocation database for each collocation structure

In the following section, we describe some basic data about the extracted collocations and some of the more important advantages of the new method.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Linguistic aspects of collocation database description

In the third section, we discuss selected linguistic topics of interest for the analysis of extracted collocations, including (in)definite forms in adjectival collocations (3.1), grammatical number (dual/plural vs. singular) in nominal collocations (3.2), degree (base vs. comparative and superlative) in adjectival and adverbial collocations (3.3), and uppercase vs. lowercase script (3.4).

For the purpose of linguistic evaluation, cumulative data for collocation candidates were extracted for 88 lemmas, with a minimum frequency of at least two occurrences. The number of collocations considered was, therefore, higher than the number of occurrences contained in the database for these lemmas, where the frequency limit is 10. Given the previous extraction methodology, it is mainly the representational part of the definition that is of interest for the evaluation, which is described in more detail below. The possibility to control the collocation output means that we can allow variability in the selected collocation elements, which reflects the actual situation in the corpus for specific collocation candidates. In the case of the 82 structures selected, the variability was allowed at the level of:

- definite (or indefinite) nominative forms of adjectives in the masculine singular
- grammatical number in collocations with nouns
- degree in collocations with adjectives and adverbs
- word forms in collocations written in lower or upper case.

The findings are described in more detail for these four categories (cf. Pori/Kosem 2021).

3.1 Definiteness in adjectival collocations

The new method makes it possible to highlight more adequately the relation between definite and indefinite forms of adjectives as they appear in real usage. We extracted the first 30 collocational candidates, ranked by logDice and filtered by:

- a morpho-syntactic description (the adjectival element must exhibit the following properties: masculine, singular, nominative);
- the difference between the attributed corpus lemma (which, according to lexicon convention, is always in the indefinite form, if it exists) and the extracted form of the adjective;
- a corpus frequency of at least 10 occurrences (the limit used in the collocation database;,
- the occurrence of each component in at least two collocations.

As an example: the indefinite form of the adjective *akuten* 'acute' (masculine, nominative, singular) is *akuten*, and the definite form is *akutni*.

As expected, the candidates with the definite form are often terms in a specific field, e. g. *etilni alkohol* 'ethyl alcohol', *akutni sindrom* 'acute syndrome', *avtomatični stabilizator* 'automatic stabiliser', *akutni hepatitis* 'acute hepatitis', etc. Similarly, they include names of animals and plants: *kodrasti pelikan* 'Dalmatian pelican', *kodrasti ohrovt* 'curly-leaf kale', *dolgoživi bor* 'Great Basin bristlecone pine', etc.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The definite form of the adjective is also used to express a number of fixed phrases or expressions that are both in terminological use in a particular field and also part of the general vocabulary, e. g. *tuji jezik* 'foreign language', *letni dopust* 'summer holiday', *materni jezik* 'mother tongue', *solatni bife* 'salad buffet', *samopostrežni bife* 'self-service buffet', *kolektivni dopust* 'collective holiday', *neplačani dopust* 'unpaid leave', etc.

The method used in previous extractions (cf. Krek 2006) only allowed lemmas for adjectival elements as outputs in similar structures, which led to the export of "unnatural" collocations, e. g.: *tuj jezik* 'foreign language', *leten dopust* 'summer holiday', *solaten bife* 'salad buffet', etc., which was basically due to choices made by the creators of the tag set and the output of the POS-tagger for Slovene.

In addition to clear-cut choices, there are two cases where both indefinite and definite forms are acceptable, since the adjective can be understood as expressing either a species or a property: *bakren kotliček* 'copper kettle, *dobrodelen bazar* 'humanitarian bazaar'. However, even in these cases, the predominance of the definite form in the corpus data suggests that this form might be more suitable as a dictionary headword. It can be concluded that when it comes to the choice of adjectival (in)definite forms, allowing for variability produces the intended results.

3.2 Grammatical number in noun collocations

For noun components, most structures allow for variability in grammatical number. This means that the choice of the singular, dual or plural form of a noun is left to the observed corpus frequency, regardless of the grammatical case or other properties. We analysed the first 30 collocations from the set of 88 headwords where the plural form was extracted for (any) noun. We sorted them by logDice and filtered them by exhibiting the plural property of the noun, with the frequency of at least 10 and with at least three occurrences in the corpus.

Collocations that indicate phraseology are quickly noticeable, e. g. *briti* [*norce*] [*PL*] (*iz koga/česa*) 'to make a fool of (someone/something)', *brusiti* (*si*) [*kremplje*] [*PL*] 'to sharpen (one's) claws – to prepare for an (aggressive) action'. In principle, plural forms can be expected to be justified in these cases, but these units have a logic of their own and, in most cases, considerable variation can also be expected. The remainder can be divided into three categories – collocations where the plural form is a) justified or necessary, e. g. *drama s* [*talci*] [*PL*] 'hostage drama'; b) unjustified or incorrect, e. g. [*kotli*] [*PL*] *na biomaso* 'biomass boilers'; c) perhaps more common, but one would expect the dictionary form to be singular, e. g. *brinove* [*jagode*] [*PL*] 'juniper berries'. Categories a) and b) are correctly represented in most cases. The largest group is c), where one might expect the singular form to be more likely, but the plural form is neither wrong nor "annoying".

We also examined the extracted dual grammatical number forms on a slightly smaller set. In the 88 selected headwords, there are no eligible dual forms at the top of the collocation set (sorted by logDice). If we look at an extended set of extracted dual forms from the whole collocation database, it is possible to find cases where dual forms would be justified, especially in the case of paired (human-animal) organs or in similar pairing situations, e. g. [*ledvici*] [*DU*] *odpovesta* 'kidneys fail', *uiti med* [*nogama*] [*DU*] 'to escape between the legs', *enojajčni* [*dvojčici*] [*DU*] 'identical twins', etc. We can conclude that despite the predominance of the plural (or dual) form shown in the corpus, the existing criterion for the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

extraction of the plural form (more than half) is mostly not justified. Statistical criteria for narrowing down the extraction of plural forms to category (a) from the above analysis remains a task for further work.

3.3 Degree in adjectival and adverbial collocations

In the case of adjectives and adverbs, variability is also checked at the level of degree – i. e. if the corpus in a particular collocation is dominated by the comparative and superlative forms, as opposed to the base form, which is also the default form of the lemma in adjectives and adverbs. Relatively few collocations were extracted for the 88 headwords that necessarily require the use of the comparative or superlative form. Most of these are related to adjectival base forms that are rarely used (e. g. *blizek* 'close') or where there is a marked semantic difference between the two forms. For example, *blizka bolnišnica* 'a nearby hospital', *ljuba kitara* 'favourite guitar'. In almost all the other cases, it would seem that the comparative and superlative forms would not be strictly wrong, but the output would be problematic if the collocation with the base form were ignored due to the majority of the two non-base forms. Analysis suggests that it would be more appropriate to consider comparative and superlative forms in the extraction only in cases where the base forms are not found in the corpus at all.

3.4 Upper or lower case

For all extracted components, we also allow for variation at the level of upper and/or lower case. This gives us insight into their dominant occurrence in the corpus and has provided interesting results. We analysed 30 of the most frequent collocations for 88 headwords, where one of the components (the dominant one) is written in uppercase or in capital letters. We sorted the collocations by absolute frequencies from the Gigafida 2.1 corpus and filtered them by the number of forms at least 3.

As expected, names of institutions, publications, and geographical names are dominant on the list, e. g. *ljubljanska Drama* 'name of a theatre from Ljubljana', *Slonokoščena obala* 'Ivory Coast – a country in Africa'. Understanding the use of upper or lower case is useful, particularly because it clearly indicates that the extracted collocation is not part of the general vocabulary; these are mainly proper names that we do not want to include in dictionary databases or the analysis of collocation data.

4. Conclusion

In this paper we describe a new procedure for extracting collocation candidates from a chosen corpus. The new formalism for collocation extraction takes into account various levels of corpus annotation, for which it uses its own (generic) system to define constraints at any level of annotation, ranging from POS-tagging and grammatical properties of word forms, syntactic (dependency) relations, concrete lexical items, and other levels of annotation, e. g. semantic roles, semantic types, etc. To automate the extraction process – in addition to constraints that take into account any annotation level in the corpus – the new system also allows us to specify which of the forms of each component found in the corpus should be included in a specific collocation, according to possibilities limited by the canonical collocation form in a specific collocation structure.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the second part of the article, we highlighted some of the variability in collocation extraction that the new system allows. This includes: the relationship between the definite and indefinite forms of the masculine singular nominative in adjectives; the singular, dual or plural forms of the nouns; the degree (comparative, superlative) of the adjective and adverb; the capitalisation of all elements of collocations. Our analysis shows that the possibility to manage the extracted forms is useful, but in most cases the threshold should be raised or the parameters further defined to take these phenomena into account when extracting collocations.

5. Future work

The main priorities for future consideration are:

1) Upgrading collocation structures from binary to extended collocations. In the existing 82 syntax structures, only binary collocations are considered. In some cases, it may be useful to include additional elements in collocations. While keeping the basic binary collocation, it would be beneficial to mention the additional element explicitly. For example: *govoriti jezik* --> *govoriti* [*angleški, francoski,* ...] *jezik* 'to speak a language --> to speak [English, French, ...] language'. The system is already set up in a way that allows existing structures to be combined into a more complex set that also takes into account the identification of extended collocations.

2) Taking into account statistics on distribution by corpus source or genre. It is possible to add various metadata from the corpus, such as textual distribution (the number of different texts in which a collocation appears) or distribution by source, to the statistics attributed to extracted collocations in the existing system. Similarly, the temporal dimension can be taken into account, meaning that we also take into account the distribution by year, which is not offered by current statistics.

3) More precise specification of the parameters for the form of the collocation output: as the analysis has shown, the possibility of managing the output of collocation forms is an important mechanism that helps to automatically extract more natural collocation forms. It is possible to build on the existing mechanism and create more precise specifications about when additional properties are taken into account and when not.

4) Consideration of other levels of annotation: semantic tagging of corpora (named entity recognition, semantic types, semantic frames, word sense disambiguation, wikification, etc.) has made significant progress, especially with the introduction of new technologies – deep neural networks. This means that future work should also take into account the next – semantic – level of annotation, which is likely to yield even better results, especially when considering clustering collocations and mapping them to corresponding dictionary senses.

References

Erjavec, T./Krek, S./Fišer, D./Ledinek, N. (2011): Project JOS: linguistic annotation of Slovene. Institut Jožef Stefan, Odsek za tehnologije znanja. http://nl.ijs.si/jos/ (last access: 23-03-2018).

Erjavec, T./Krek, S./Arhar, Š./Fišer, D./Ledinek, N./Saksida, A./Sivec, B./Trebar, B. (2010): Oblikoskladenjske specifikacije JOS V1.1 2010-03-07. http://nl.ijs.si/jos/msd/html-sl/index.html (last access: 23-03-2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gantar, P. (2021): Zapis frazeoloških enot v Leksikonu večbesednih enot za slovenščino. In: Arhar Holdt, Š. (ed.): Nova slovnica sodobne standardne slovenščine: viri in metode. Ljubljana, pp. 198–230.

Gantar, P. (2015): Leksikografski opis slovenščine v digitalnem okolju. Znanstvena založba Filozofske fakultete. http://www.ff.uni-lj.si/Portals/0/Dokumenti/ZnanstvenaZalozba/e-knjige/Leksikografski.pdf (last access: 23-03-2018).

Gantar, P./Gorjanc, V. (2015): Obrazilo -en/-ni v slovarski obravnavi pridevnikov. In: Smolej, M. (ed.): Slovnica in slovar – aktualni jezikovni opis. Ljubljana, pp. 233–241.

Gantar, P./Kosem, I./Krek, S. (2016): Discovering automated lexicography: the case of Slovene lexical database. In: International Journal of Lexicography 29 (2), pp. 200–225.

Gantar, P./Krek, S./Kosem, I. (2021): Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. In: Kosem, I. (ed.): Kolokacije v slovenščini. Ljubljana, pp. 15–41.

Gantar, P./Krek, S./Krsnik, L. (2021): Strojno berljiv Vezljivostni leksikon slovenskih glago-lov. In: Arhar Holdt, Š. (ed.): Nova slovnica sodobne standardne slovenščine: viri in metode. Ljubljana, pp. 259–297.

Gorjanc, V./Gantar, P./Kosem, I./Krek, S. (eds.) (2015): Slovar sodobne slovenščine: problemi in rešitve. Ljubljana.

Kilgarriff, A./Baisa, V./Bušta, J./Jakubíček, M./Kovář, V./Michelfeit, J./Rychlý, P./Suchomel, V. (2014): The Sketch Engine: ten years on. In: Lexicography ASIALEX 1.

Kosem, I./Krek, S./Gantar, P./Arhar Holdt, Š./Čibej, J./Laskowski, C. (2018): Kolokacijski slovar sodobne slovenščine. In: Fišer, D./Pančur, A. (ed.): Zbornik konference Jezikovne tehnologije in digitalna humanistika. Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, September 20–21, 2018. Ljubljana, p. 133.

Kosem, I./Gantar, P./Krek, S./Arhar Holdt, Š./Čibej, J./Laskowski, C./Pori, E./Klemenc, B./Dobrovoljc, K./Gorjanc, V./Ljubešić, N. (2019): Collocations dictionary of modern Slo-vene KSSS 1.0. Ljubljana. https://www.clarin.si/repository/xmlui/handle/11356/1250 (last access: 23-03-2018).

Krek, S. (2015): Leksikografska orodja za slovenščino: slovnica besedlnih skic. In: Gorjanc, V./Gantar, P./Kosem, I./Krek, S. (ed.): Slovar sodobne slovenščine: problemi in rešitve. Ljubljana, pp. 358–378.

Krek, S./Kilgarriff, A. (2006): Slovene word sketches. In: Erjavec, T./Žganec Gros, J. (eds.): Language technologies. Proceedings of the 9th International Multiconference Information Society IS 2006, Ljubljana, 9–10 October 2006. Ljubljana: Institut "Jožef Stefan", p. 62.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 202: Twelfth International Conference on Language Resources and Evaluation, Marseille, May 11–16, 2020. Marseille, p. 3340.

Krek, S./Gantar, P./Kosem, I./Dobrovoljc, K./Arhar Holdt, Š./Čibej, J./Laskowski, C./Klemenc, B./Krsnik, L. (2021): Frequency lists of collocations from the Gigafida 2.1 corpus. Ljubljana. http://hdl.handle.net/11356/1415 (last access: 23-03-2018).

Pori, E./Kosem, I. (2021): Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. In: Kosem, I. (ed.): Kolokacije v slovenščini. Ljubljana, pp. 43–77.

Ramisch, C. (2020): Computational phraseology discovery in corpora with the MWE-TOOLKIT. In: Corpas Pastor, G./Colson J-P. (ed.): Computational Phraseology. Amsterdam/Philadelphia, pp. 111–134.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ramisch, C./Savary, A./Guillaume, B./Waszczuk, J./Candito, M./Vaidya, A./ Barbu Mititelu, V./ Bhatia, A./Iñurrieta, U./Giouli, V./Güngör, T./Jiang, M./Lichte, T./Liebeskind, Ch./Monti, J./Ramisch, R./Stymne, S./Walsh, A./Xu, H. (2020): Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In: Markantonatou, S./McCrae, J./Mitrović, J./ Tiberius, C./Ramisch, C./Vaidya, A./Osenova, P./Savary, A. (eds.): Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona (Online), December 2020. Barcelona, p. 107.

Acknowledgements

This paper was written within the framework of the research project New Grammar of Modern Standard Slovene: sources and methods (J6-8256) and the programme groups Slovene Language – Basic, Contrastive and Applied Research (P6-0215) and Linguistic Resources and Technologies for the Slovene Language (P6-0411), funded by the Slovenian Research Agency.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.