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).

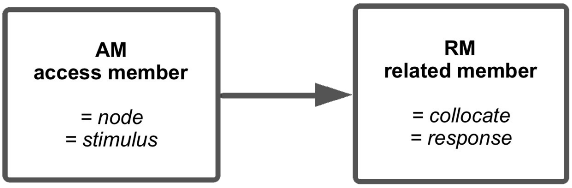

Figure 1: The common structure of collocations and associations.

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.

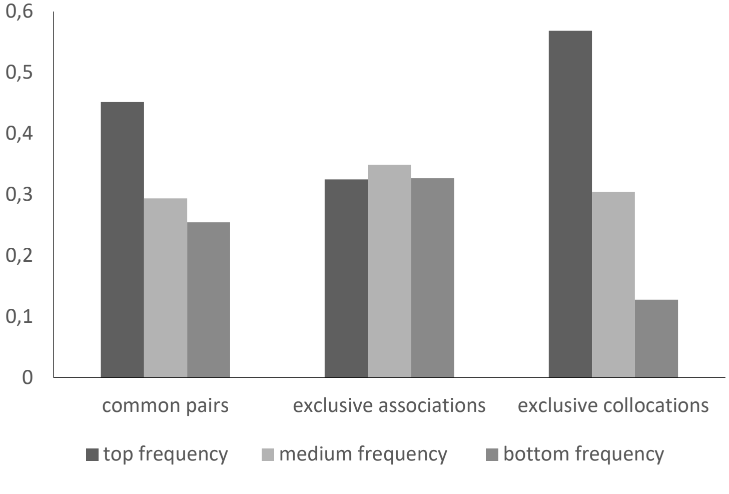

Figure 2: Distribution of data according to AMs' corpus frequency.

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