TRAMES, 2013, **17**(**67**/**62**), 2, 141–158

DISTRIBUTIONAL HYPOTHESIS: WORDS FOR 'HUMAN BEING' AND THEIR ESTONIAN COLLOCATES

Liisi Piits

Institute of the Estonian Language, Tallinn

Abstract. The article was inspired by the Distributional Hypothesis by Zellig Harris, which states that words occurring in similar contexts tend to have a similar meaning. The hypothesis was tested by a comparison of the 10 most frequent collocates of the Estonian words for 'human being'. In the present study, the word *collocate* is used in a neo-Firthian sense, covering all the words that co-occur with the node word the most often. The collocates of the words *inimene* 'human', *mees* 'man', *naine* 'woman', *laps* 'child', *tüdruk* 'girl', *poiss* 'boy', *tütar* 'daughter', *poeg* 'son', *ema* 'mother' and *isa* 'father' were drawn from the context 'three words to the left' of the node word as occurring in the Newspaper subcorpus of the Balanced Corpus of Estonian. The comparison involved the 30 most frequent collocates for each node word. Assuming that a bigger number of shared collocates means a greater semantic closeness, intersections of collocates of the Estonian words for 'human being' were computed. It turned out that antonymous words had the highest number of collocates in common, which indicates that syntagmatic relations of words may also reflect some of their paradigmatic relations. In addition, what may be decisive for the part of speech of collocates, is analysed.

Keywords: collocations, distributional hypothesis, human being, syntagmatic relations, antonymy, Estonian

DOI: 10.3176/tr.2013.2.03

1. Introduction

The famous sentence by the British linguist John Rupert Firth, "You shall know a word by the company it keeps!" draws attention to the fact that the combination of words in phrases and sentences is never random or based on purely syntactic rules; instead, there are special relations between the words, while their way of co-occurrence reveals important information about those relations. The idea was found enlightening by Zellig Harris, also called the last American

structuralist, who based his semantic classification on word distribution, arguing that words occurring in similar contexts have a similar meaning:

"If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C" (Harris 1954:156).

The distribution of an item is understood as a sum of its whole context. This general definition by Harris leaves open what is meant by context and what exactly is the meaning supposed to be reflected by that context. In Harris's own publications, context may mean either phonetic or lexical context, or even the whole text. In terms of the present study, context is understood as left collocations, to be more precise, the three words positioned to the left of the node word. For comparison, some right collocations from an interval of similar length (3 words) have also been presented.

Our analysis deals with the 10 most frequent Estonian words for 'human being' and their 30 most frequent collocates. The words are *inimene* 'human', *mees* 'man', *naine* 'woman', *laps* 'child', *tüdruk* 'girl', *poiss* 'boy', *tütar* 'daughter', *poeg* 'son', *ema* 'mother', and *isa* 'father'. The aspect of frequency is especially relevant in the discovery of collocational relations. The least frequent of the above words was *tütar* with only 612 occurrences. Thus, the list of node words was closed at ten, as less frequent nodes would have meant too much random material among the 30 collocates planned to be presented for each node word. Intuitively, the above ten words should belong to the basic words for 'human being'. True, their psychological salience has not been investigated, but their linguistic salience is reflected in their position in the list of the 130 most frequent Estonian nouns (Kaalep and Muischnek 2002), which is a strong argument for their possible basic word status (for the definition of 'basic word' see Sutrop 2000, Sutrop 2011).

As for meaning, Harris has not given a precise definition of that either. The present study is based on the idea that meaning consists of some smaller aspects or components (for c*omponential analysis* see Fodor and Katz 1963, Nida 1975). A component (for example, 'human being') binds all the relevant words into one *semantic field*. Thus, according to the distributional hypothesis, all words for 'human being' must naturally share a lot of context. However, relations within the semantic field are also of interest. Notably, the words for 'human being' are distinguished by such features as MALE, ADULT, KINSHIP/PARTNERSHIP, GENERATION OLDER THAN SELF and GENERATION YOUNGER THAN SELF. The research question is: Are the values given to those components, i.e. the semantic aspects of words, also manifested in the collocations of those words?

Semantic relations bind word meanings into semantic networks. Ever since Saussure, word semantic relations have been divided into syntagmatic and paradigmatic ones. The former occur between words appearing simultaneously in the same syntagm. For example, in the phrase 'young man' the words 'young' and 'man' are related syntagmatically. Paradigmatic relations, however, occur between words that are mutually substitutable in the syntagm. In the above syntagm, for example, the word 'young' stands in a paradigmatic relationship with the words

'old', 'tall', etc. (Lyons 1977:240–241). Note that Saussure called the paradigmatic relation *associative*, arguing that this is a relationship between the words associated in memory (Saussure 2000:121).

Of paradigmatic relations, the words under discussion mostly display antonymy and hyponymy. In addition, synonymy can sometimes be observed between some less frequent meanings of the words *mees* 'man' and *poiss* 'boy'. According to *Eesti keele seletav sõnaraamat* (2009) /Explanatory Dictionary of Estonian/ the second sense of *mees* is '*abikaasa*' 'husband' and one of the senses of *poiss i*s '*noormees, kavaler*' 'young man, boyfriend'. Similar synonymy can be observed between *naine* 'woman' and *tüdruk* 'girl', as both may denote a partner as well as gender. In addition, *mees* has received a third sense of '*inimene, isik; asjamees, tegelane*' 'person; functionary', which indicates synonymy with 'human being'. However, in the case of the above word pairs the relation of synonymy is not manifested between the most frequent and most important senses of the words in question and thus it escaped analysis this time. The distributional hypothesis has been tested on synonyms by Herbert Rubenstein and John B. Goodenough (1965) and later by Georg A. Miller and Charles Walter (1991), using the same synonym pairs and similar principles. Both studies confirmed the existence of a positive correlation between semantic closeness and similarity of context. In those studies the semantic closeness of the word pairs was determined from questionnaires and the index of closeness was derived from people's assessments. This way, antonyms would probably receive a very low score, as antonymy is a paradoxical semantic relationship, where words with several similar features may look very different at first sight. The hypothesis was considered confirmed if the subjects agreed that the words defined as semantically close were mutually replaceable in a context. Note, however, that mutual replaceability in the same construction also belongs, among other features, to the definition of antonymy: "Two lexical items are antonyms if they are formally substitutable for each other in a construction without resulting in ungrammaticality" (Panter 2012:4). This suggests that antonymous words have a similar context.

In the present discussion, antonymy is used in a wide sense covering all lexical oppositions in general, not just scalar oppositions as found in John Lyons (1977) and David Alan Cruse (1986). The article tries to find out how paradigmatic relations can be manifested in syntagmatic relations, i.e. in collocations. Sahlgren has specified the distributional hypothesis, stating that if a distributional model based on co-occurrence frequencies reflects the syntagmatic relations of words, a comparison of the collocates should reveal information on their paradigmatic relations (Sahlgren 2008:7).

2. Collocation

The first use of the term *collocation* in modern linguistics has been associated with two renowned British linguists: John Rupert Firth and Harold E. Palmer. The

latter used the term *collocation* in 1938. In "A Grammar of English Words" (1949 [1938]:x), Palmer emphasized the language learning aspect of collocations: collocation is a sequence of two or more words that should be learnt as one word. The first to explain the theoretical importance of the term collocation and its role in semantics was Palmer's compatriot J. R. Firth, who called the collocation *an abstraction at the syntagmatic level* as well as *habitual co-occurrence*. Unfortunately this definition by Firth was not too clear and later confusion around the term collocation has only increased. Roughly, the available definitions of 'collocation' can be divided in two groups:

1. In a narrower sense, collocation is understood as an object of phraseology. As such, it is often placed on a scale running from free word combinations to idioms. Auxiliary concepts are *transparency* and *substitutability*. There is no agreement, however, where exactly collocations should be placed on this scale and how transparent or substitutable a collocative phrase should actually be. According to Palmer's definition, for example, the meaning of a collocation cannot be derived from the meanings of its component words. Therefore he recommends that collocations should be taught in the same way as individual words and presented in dictionaries (Palmer 1949 [1938]:x). Nowadays, a similar attitude has been expressed by Stefan Evert (2005:17), who defines 'collocation' as a word combination with such semantic and/or syntactic features that cannot be purely predicted from its individual components and which should therefore be included in dictionaries. Both Palmer's and Evert's definitions are eloquent of the fact that a definition of 'collocation' is largely dependent on the aim of the definition. From the lexicographic and language teaching points of view, collocation is indeed seen more like an object of phraseology.

2. In a wider sense, collocation is understood as the co-occurrence of two or more words, which is observed more than once. Here, frequency is the only relevant parameter, not transparency or substitutability. The most general definition of all comes from the British linguist and initiator of the neo-Firthian approach John Sinclair (2004): "Collocation is the occurrence of two or more words within a short space of each other in a text". John Sinclair was the first to test Firth's ideas on corpus material. During several years spent in Birmingham University he won over a whole group of linguists – Michael Hoey, Susan Hunston, Michael Stubbs, Wolfgang Teubert and Elena Tognini-Bonelli – to share his own understanding of collocational relations (McEnery and Hardie 2012:122). Michael Stubbs (2001:24; 29), for example, has stated that collocation is "a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text". The underlying concept of Stubbs's definition is the frequency of co-occurrence: "collocation is frequent co-occurrence".

In addition to the raw co-occurrence frequency the concept of collocation has to do with association models, which compute collocations so that co-occurrence frequency is combined with individual word frequencies. For a comprehensive overview of various association models see Evert (2005) and Sabine Bartsch

(2004). For the Estonian language, the log-likelihood function has been considered to be the best statistic of word association (Uiboaed 2010). Statistics of word association have been efficiently used to ascertain fixed verbal phrases in Estonian (Muischnek 2005).

Sometimes confusion may be avoided by calling the wider concept of collocations just *co-occurrences*, see, e.g. Evert 2005. For the same purpose, an opposite road has also been taken, making collocations cover all co-occurrences, while those co-occurrences that are observed more frequently than could be presumed from the text frequencies of their member words taken separately, are specified as *significant collocations* (Sinclair 2004). The present article sticks to the term *collocation* used in the wider, neo-Firthian sense of lexical co-occurrence of words.

3. Methods and material

The material was drawn from the Newspaper subcorpus of the Balanced Corpus of Estonian (NBCE), which contained, at the moment of downloading, 5 million text words from the period 1990–2001. The frequencies of the ten most frequent Estonian words for 'human being' occurring in the newspaper subcorpus were as follows:

inimene	'human being'	12277
mees	'man'	7293
laps	'child'	5583
naine	'woman'	4098
ema	'mother'	1504
isa	'father'	1330
poiss	'boy'	1221
poeg	'son'	914
tüdruk	'girl'	853
tütar	'daughter'	612

Besides node word selection, the choice was between a lemmatized and non-lemmatized corpus. Some scholars believe that it is the co-occurrence of word forms, not of mere words, that should be analysed. Sinclair (2004), for example, emphasizes that in certain cases different word forms should be handled as different words, as their collocational distribution may be quite different. Stubbs, who used corpus data in his investigation of different forms of the word *seek,* found that although the word forms *seek, seeking,* and *sought* share 6 of their most frequent collocates, the forms *seeks* and *seek* have only one collocate in common, which is *professional*. *Seeking* and *sought* share no collocates with *seek* at all. Stubbs used the huge Cobuild subcorpus of 200 million words, and yet he suspected that word collocations are closely connected with text type and the collocational peculiarities of *seeks* largely result from the ads section in the

newspaper texts in the corpus (Stubbs 2001:28). Sara Gesuato (2003) has investigated the English collocations of *man* and *woman*, now giving examples of the collocations of *woman* and *women*, now offering generalizations about the context of the lemma *woman*.

The richness of Estonian inflection makes the comparison of word form co-occurrence complicated, firstly, because a word may have very many inflectional forms (28 for a noun), and secondly, due to grammatical homonymy. Even if we confine ourselves to the nominative singular forms of *naine* 'woman' and *ema* 'mother' as node words, we get the desired results for *naine,* but in the case of *ema* there is no automatic way of telling apart the collocates of the grammatically homonymous forms *ema* NomSg, *ema* GenSg, and *ema* PartSg.

Therefore I decided to lemmatize, using the morphological analyser *Estmorf* (Kaalep 1997), which automatically lemmatized and disambiguated the whole corpus. However, on several occasions the unlemmatized corpus was consulted for frequency data of different word forms. In the analysis of some collocational relations comparative material for co-occurrence frequencies was drawn from the Internet (www.google.ee; 03.11.2008). Those data are not lemmatized either.

The collocations were computed and the concordance series were analysed by means of the Concord program of the WordSmith Tools 4,0 software, created by Mike Scott (Scott 2004).

Collocate span. One of the problems in the measurement of collocational relations is how far from the node word can a word be positioned to still qualify for a collocate of this node. According to Stubbs there is a consensus that significant collocations are found in the range 4:4, i.e. four words to the left of the search word and four to the right (Stubbs 2001:29). No doubt the collocate span considerably depends on the type of language: Stubbs's conclusions are based on his studies of English, which tends to an analytic way of expression, while Sinclair (2004) notes that for English noun collocations the first left position is almost exclusively filled with the definite article *the*. As Estonian has no articles and the language is altogether more synthetic than English, Estonian material could probably do with a smaller collocate span. There is also a positive correlation between corpus size and the optimum width of the search window. Sahlgren's English-based conclusions read that a larger corpus (British National Corpus) tolerates a slightly wider search window (suggesting 3:3), whereas a smaller corpus (Touchstone Applied Science Associates corpus) yields better results with a smaller (2:2) window (Sahlgren 2006:115). My relevant experience involves a 3:3 span (three words to the left and three to the right). True, the present article is focused on the left collocates, while the right ones have been used just for comparison. In addition, the collocates from the 3:3 range have been weighted, so that the collocate that is closer to the node gets a higher value, relying on the principle that the farther a word is positioned from the node word the less regular is its relation to the latter.

Our list of the 30 most frequent collocates was not meant to include uninflected words, nor the references found in the corpus. For that purpose a *stoplist* of 55

words or letter sequences was used, which cleared the set of collocates of conjunctions (e.g. *kui, et, aga, kuid, sest*), frequent adverbs (e.g. *väga, juba, pärast*), adpositions (e.g. *üle, eest, enne, juurde, vastu, peale, koos*), and corpus references (e.g. *PM, EPL, AJAE, AJA, ML*). The uninflected words were excluded, because their syntactic connectedness with the node word as well as their semantic load is lower than those of inflected words.

4. Left collocates of the Estonian words for 'human being'

The total number of the 30 most frequent collocates of ten words is 300. The total number of different collocates, however, is 96, as several node words have a collocate or more in common. Of those 96, fewer than a half belong just to one search word, while the rest belong to two or more.

A word together with its more frequent collocates can be called a collocate cluster. Like the Harris's (1954) distributional hypothesis referenced above, the more collocates two clusters have in common the higher the probability that their node words are semantically close.

Table 1 shows how many of the 30 collocates of 10 different Estonian words for 'human being' coincide. Moving down the column it can be seen with what words the node word has the highest (bold) number of collocates in common and with what words the collocate intersection is the smallest (grey background).

The first column reveals that the word *ema* 'mother' shares the greatest number of collocates with *isa* 'father' and *naine* 'woman', and the smallest number with *poiss* 'boy' and *tütar* 'daughter'. The words *ema* and *isa* stand in an antonymous relation as contrasted by the gender aspect, but at the same time the two share the parental aspect. The words *ema* and *naine* share two aspects: gender and indeterminate age, while KINSHIP and GENERATION OLDER/YOUNGER THAN SELF are the distinctive aspects.

Table 1. Number of left collocations shared with the node word

	ema 'mother'	*inimene* 'human'	*isa* 'father'	*laps* 'child'	*mees* 'man'	*naine* 'woman'	*poeg* 'son'	*poiss* 'boy'	*tüdruk* 'girl'	*tütar* 'daughter'
ema		15	**20**	16	19	20	15	14	17	14
inimene	15		11	**20**	21	20	12	16	14	11
isa	**20**	11		12	14	16	14	12	14	13
laps	16	20	12		20	19	15	15	15	14
mees	19	**21**	14	**20**		**22**	14	16	17	11
naine	**20**	20	16	19	**22**		15	16	15	12
poeg	15	12	14	15	14	15		13	14	**18**
poiss	14	16	12	15	16	16	13		**19**	11
tüdruk	17	14	14	15	17	15	14	**19**		12
tütar	14	11	13	14	11	12	**18**	11	12	
Total	**150**	**140**	**126**	**146**	**154**	**155**	**130**	**132**	**137**	**116**

According to the second column, the collocate cluster of *inimene* 'human being' is the most reminiscent of that of *mees* 'man'. Of the 30 collocates the two share as many as 21. At the same time, *naine* 'woman' and *laps* 'child' share 20 collocates with *inimene* 'human being', which means a two-thirds coincidence with their collocate clusters. Evidently the very slightly higher similarity between the collocates of *mees* 'man' and *inimene* 'human being' is not indicative of a synonymous relation between the two words, nor does it prove that the Estonian word *mees* is also used in the generic sense of 'man'. Of the ten words chosen for analysis, *laps* 'child', *mees* 'man', *naine* 'woman' and *inimene* 'human being' display an especially high coincidence of their collocates. The reason may lie in that the meaning of the four words is wider than that of the remaining six. The word *inimene* 'human being', for example, has not been specified in gender or age, and nor does it refer to kinship or a family relation. The primary sense of *laps* 'child' is determined by age, while that of *mees* 'man' as well as of *naine* 'woman' is concerned with gender. The remaining six words, however, are defined more specifically, with at least two aspects involved. The words *ema* 'mother', *isa* 'father', *tütar* 'daughter' and *poeg* 'son' refer to both kinship and gender, while *tüdruk* 'girl' and *poiss* 'boy' refer to age as well as gender.

The third column shows that the collocates of *isa* 'father' have the biggest intersection with those of *ema* 'mother' and the smallest with those of *inimene* 'human being'. The word *laps* 'child' shares the smallest number of collocates with *isa* 'father' and the biggest with *inimene* 'human being' and *mees* 'man'. Columns five and six show that *mees* 'man' and *naine* 'woman' mutually share the largest context, while the smallest number of common collocates bound them with *tütar* 'daughter'. The word *poeg* 'son' has the greatest number of collocates in common with *tütar* 'daughter' and the smallest with *inimene* 'human being'. The words *poiss* 'boy' and *tüdruk* 'girl' mutually share the greatest number of collocates, while their smallest shared context binds them with *tütar* 'daughter'. The word *tütar* 'daughter' has a majority of collocates in common with *poeg* 'son', whereas its common context with any of the other words, especially with *inimene* 'human being', *mees* 'man' and *poiss* 'boy', is very small indeed.

Our analysis of the collocation clusters revealed that usually the node word shares most of its left collocates with antonymous words: *isa* 'father' mutually with *ema* 'mother', *mees* 'man' mutually with *naine* 'woman', *tütar* 'daughter' mutually with *poeg* 'son', and *poiss* 'boy' mutually with *tüdruk* 'girl'. However, the above tendency is characteristic only of the left collocates, not the right ones. Does the similarity of the contexts of antonymous words confirm or refute the distributional hypothesis? The answer depends on our definition of antonymy. As was mentioned in the introduction, antonymy can be considered a paradoxical semantic relation, because at first sight the senses involved differ radically, yet on second thoughts the opposite senses have more aspects in common than not. The opposition works between words of one and the same lexical field. Those words differ by one semantic component, while the values of the rest of the components coincide. When we compare the semantic components of the antonym pair *naine*

'woman' and *mees* 'man' with those of the pair *naine* 'woman' and *ema* 'mother', we find that the former pair are antonyms distinguished by the MALE feature, while the latter has two distinctive features – KINSHIP and GENERATION OLDER THAN SELF – that distinguish between the main senses of the pair members. The similarity is vividly demonstrated by association tests, where the first association given in response to the stimulus word is, in a predominant number of cases, its antonym. For the stimulus words *man* and *boy,* for example, more than three-thirds of the subjects responded first with their respective antonyms (Panter 2012:1).

Lynne Murphy (2006) has found that antonymy is not just a lexical paradigmatic relation, but it is also manifested on the syntagmatic level, as antonyms occur in several constructions: *X and Y, both X and Y, X or Y, X and Y alike* etc. For example, all of the 100 most frequent tokens of the construction *X and Y alike* contain antonym pairs, the most frequent of which is *men and women alike* (Panther 2012:10–11). The material used in the present study does not deal with analogous constructions with antonyms, but a comparison of the collocates showed that the 30 most frequent collocates of *mees* 'man', *naine* 'woman', *ema* 'mother', *isa* 'father', *tütar* 'daughter', *poeg* 'son', *tüdruk* 'girl' and *poiss* 'boy' invariably contained an antonym of the node word. Hence, antonymy was also found to be manifested on the syntagmatic level.

A hyponymic relationship can be observed between *inimene* 'human being' as a hyperonym and all the rest of the words analysed, which are all hyponyms of *inimene*. Hyponymy was not manifested in any particular way in the collocations. No hyponym displayed a particularly large coincidence between the collocate clusters of the hyperonym and those of its own. However, the words under discussion are bound into a lexical field by a hyponymic relationship, which may well be manifested just in the large intersection observed in this field. This cannot be proved, however, without additional material about the distribution of words outside this lexical field.

Next, I compared the relations among all of the ten collocate clusters in order to find out which word for 'human being' shares the greatest number of collocates with the nodes under discussion. It turned out that *naine* 'woman' and *mees* 'man' were the words with the largest average percentage (57%) of collocates shared with the rest of the node words. Hence we may conclude that *naine* 'woman' and *mees* 'man' are the central words of the 'human being' lexical field, displaying the most similarity with the other node words on the syntagmatic level.

The runners-up are *ema* 'mother' and *laps* 'child' with 56% and 54%, respectively, of shared collocates. Interestingly, *isa* 'father' with its 14 shared collocates occupies the last but one position in this list. The lowest percentage of shared collocates (43%) belong to *tütar* 'daughter'. One of the reasons may be the low corpus frequency (612) of the word. This might have caused the occurrence of some random words among the 30 most frequent collocates, which would not have been there if the node word had been more frequent.

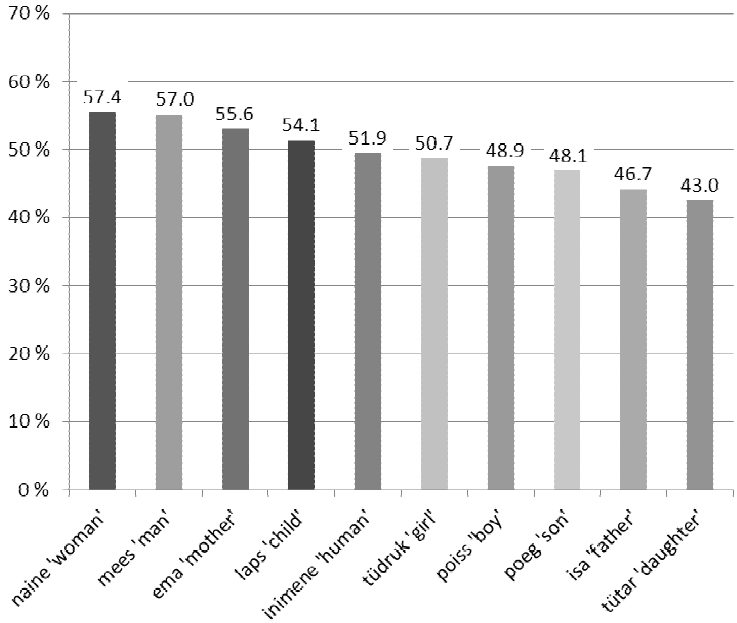

Figure 1. Percentage of left collocations shared with the node word.

The graph in Figure 2 visualizes the mutual relations of the words denoting 'human being'. The longer the distance between two words, the smaller the share of the collocates the two had in common.

The graph clearly demonstrates the central position of *naine* 'woman' and *mees* 'man', as well as the most distant position of *tütar* 'daughter'. The relationships of paradigmatic and syntagmatic relations are visualized by the nearest neighbour position of antonomous words.

The intersection of collocate clusters of words for 'human being' is impressive, covering about half of the collocates. It would be easy to believe that this is a fine proof of the distributional hypothesis. However, if we take an outgroup term such as, e.g. *auto* 'car', which is frequent, but certainly does not have the sense of 'human being', and its 30 most frequent collocates, we can also find a relatively large coincidence with the collocates of the above discussed words. For example, the 30 most frequent collocates of *auto* 'car' share as many as 8 collocates with all our node words for 'human being'.

4.1. Collocates

The collocates shared by all of the discussed node words for 'human being' were *üks* 'one/a', *kaks* 'two', *kolm* 'three', *mina* 'I', *tema* 'he/she', *see* 'this/the', *olema* 'be', *oma* 'his/her'. All the nodes except *tütar* 'daughter' shared the collocate *saama* 'get'. Seven node words shared the collocates *eesti* 'Estonian', *noor* 'young', *pidama* 'should' and *teine* 'other/second'. The collocates shared by

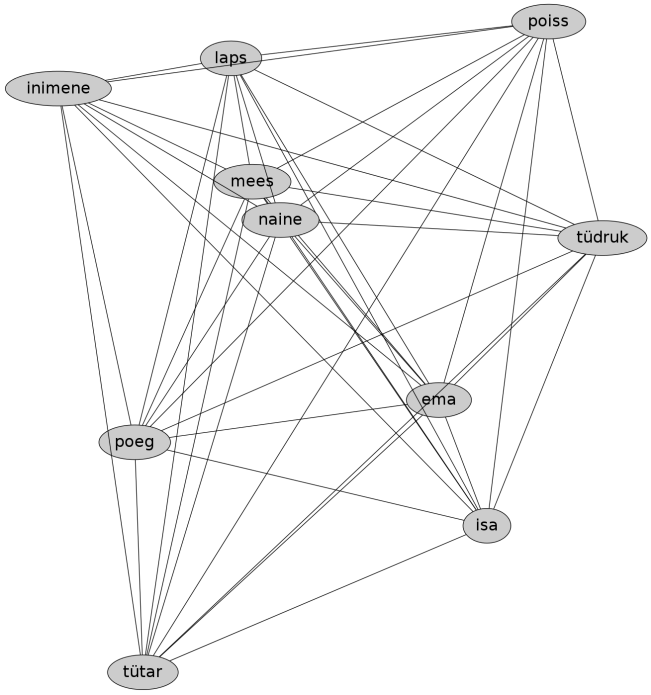

Figure 2. Relations between the words *inimene* 'human being', *mees* 'man', *naine* 'woman', *laps* 'child', *tüdruk* 'girl', *poiss* 'boy', *tütar* 'daughter', *poeg* 'son', *ema* 'mother' and *isa* 'father' as reflected in the intersections of their left collocates. The length of the lines is inversely proportional to the number of shared collocates.

six node words were *laps* 'child', *vanem* 'parent/older/elder', *tegema* 'make/do', *tulema* 'come', *ütlema* 'say/tell'. Five nodes shared the collocates *ise* '(one)self', *aasta* 'year', *naine* 'woman/wife', *rääkima* 'talk/tell/speak', *sina* 'you', *väike* 'little/small'. The collocates *ema* 'mother', *iga* 'every', *kõik* 'all', *minema* 'go', *uus* 'new' were shared by four nodes. Three words for 'human being' shared the collocates *elav* 'lively', *isa* 'father', *kodu* 'home', *mõni* 'some', *neli* 'four', *poeg* 'son', *poiss* 'boy', *tütar* 'daughter', *vene* 'Russian (adj.)', *võima* 'can/may', and *võtma* 'take'. The collocates *aastane* 'a year old', *aeg* 'time', *elama* 'to live', *hea* 'good', *ilus* 'beautiful', *klass* 'grade', *noorem* 'younger', *pere* 'family', *puue* 'disability', *selline* 'such', *sündima* 'be born', *teismeline* 'teenaged', *tüdruk* 'girl', *vana* 'old', *viimane* 'last' were shared by merely two nodes. The remaining 41 collocates co-occurred with just one node word.

Although the less frequent collocates are certainly very interesting, their individual analysis is beyond the present article. The following analysis addresses only the eight left collocates shared by all of the ten node words. Those collocates are certainly not specific to words denoting human beings. The numerals *üks* 'one', *kaks* 'two' and *kolm* 'three' reflect the general necessity of counting items and defining them numerically. The Estonian word *üks* is not only a numeral for

'one', but it also functions as a pronoun in positions where some other languages use the indefinite article. The role of the definite article is fulfilled by the pronoun *see*. As the analysis is based on pure co-occurrence, without special consideration for separate word frequencies, the most frequent Estonian lemma *olema* 'be' is also found among the left collocates of all the node words.

In addition, the list of collocates shared by all of the ten nodes, included the personal pronouns *mina* 'I', *tema* 'he/she' and *oma* 'his/her'. Note that after running the Estmorf program for lemmatization, the frequency readings of the singular personal pronouns *mina* 'I' and *tema* 'he/she' had come to include those of the respective plural pronouns *meie* 'we' and *nemad* 'they'. Bypassing the plural forms I compared different forms of the lemma *mina* 'I' as presented in the unlemmatized corpus. The aim was to find out which words for 'human being' prefer the nominative forms *mina* and *ma* 'I' and which prefer the genitive forms *minu* and *mu* 'my' among their three left collocates. In order to level the differences arising from the node words having different frequencies the raw frequencies were converted to relative ones by dividing the frequency of co-occurrence with the frequency of the node word.

A comparison of the relative frequencies of the nominative forms *mina+ma* 'I' and the genitive *minu+mu* 'my' revealed that while the former occurred practically evenly among the collocates of all the node words for 'human being', the genitive pronouns were the most frequent among the collocates of words expressing partnership: *isa* 'father', *ema* 'mother', *poeg* 'son', *tütar* 'daughter' (see Fig. 3).

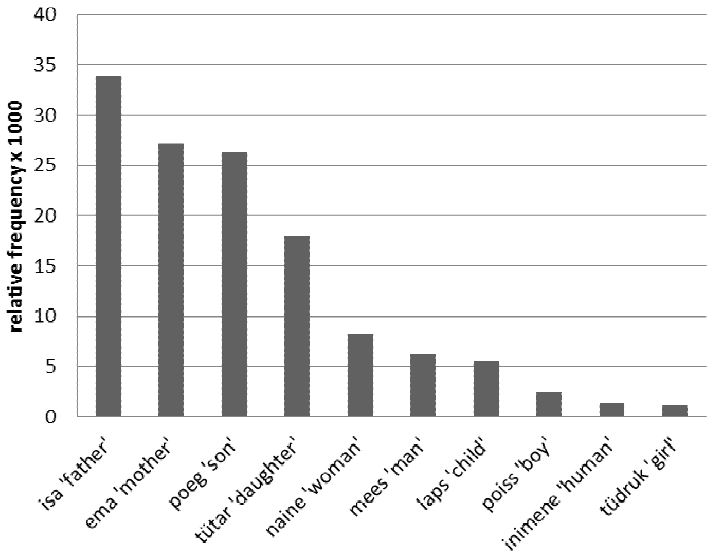

Figure 3. The relative frequency of the genitive forms *minu+mu* 'my' among the left collocates of Estonian words for 'human being' in the NBCE corpus.

The genitive form of the first person personal pronouns is used in the first left (L1) position to refer to the kinship relations of the 'I'. The highest relative frequency is characteristic of the collocation *minu isa* 'my father'. Although the corpus frequency of the word *ema* 'mother' is higher than that of *isa* 'father', the absolute frequency of the collocation *minu isa* 'my father' is slightly higher than that of *minu ema* 'my mother', while the relative frequency (the ratio of the frequency of the collocation to that of the node word) shows a wider difference. The data suggest that first-person speakers tend to speak more of their fathers than of their mothers. The collocations *minu ema* 'my mother' and *minu poeg* 'my son' are almost equally frequent, whereas daughters are mentioned slightly less frequently in journalistic texts.

One of the reasons why the genitive form *minu* 'my' occurs among the collocates of the four words *isa* 'father', *ema* 'mother', *poeg* 'son' and *tütar* 'daughter' is their shared semantic component of +KINSHIP. The next frequent group consisted of the words *naine* 'woman/wife', *mees* 'man/husband' and *laps* 'child'. Each of the three also has a meaning connected with +KINSHIP OR PARTNERSHIP. However, for these three words the semantic component of family relation is not primary in journalese. The fewest co-occurrences with the genitive form of the first person pronoun were registered with *inimene* 'human being', *poiss* 'boy' and *tüdruk* 'girl'. The word *inimene* lacks a sense enabling a positive value to be assigned to the component KINSHIP OR PARTNERSHIP, as well as an attributive use of a personal pronoun in the genitive form. Only one of the 16 occurrences of the collocate *minu+mu* 'my' occupied the first left position L1, the rest occurred either in L2 or L3. Although *tüdruk* 'girl' and *poiss* 'boy' can also mean 'girlfriend' and 'boyfriend', respectively, thus denoting partners in a couple relationship, in journalese this sense of the words must be less popular, which is reflected in the incidence of the personal pronoun.

The material drawn from the Newspaper subcorpus was compared with that obtained from the Internet during a day spent with Google's search engine. The frequency of the co-occurrence of the ten most frequent words for 'human being' with the genitive forms *minu* and *mu* 'my' was computed for that material (see Fig. 4).

Comparing the Figures 3 and 4 it strikes the eye that in the Internet the relative frequency of the occurrence of *minu* and *mu* 'my' among the collocates of words for 'human being' is considerably higher than in the NBCE. Moreover, in the corpus material *minu* and *mu* occurred in all of the three left positions investigated, whereas the respective Internet pronouns were only counted in L1 (L2 and L3 were not studied). One of the reasons for the higher web frequency may be that in the Internet there is more of first-person communication, especially considering the numerous web forums.

As for the general picture, however, it is rather similar for the two sources. In the web materials, too, the most frequent collocation in the relevant set is *minu isa* 'my father', which is followed by *minu ema* 'my mother'. The collocation *minu laps* 'my child' is more frequent than in the newspaper subcorpus, but the rest

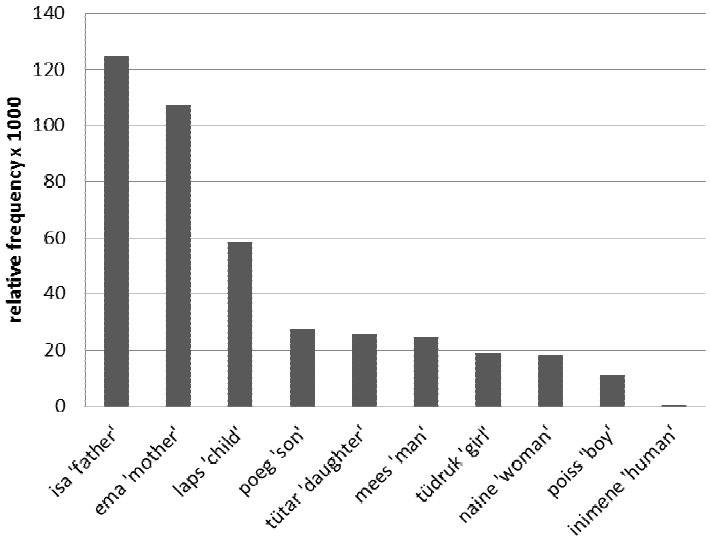

Figure 4. The relative frequency of *minu+mu* 'my' among the left collocates of Estonian words for 'human being' in the Internet.

follow suit in a more or less the same order and with only slight differences in frequency. The collocation *minu inimene* '*my person*' brings up the rear. Thus, a comparison of the two sets of data reveal similar tendencies, namely, the genitive attribute *minu* 'my' is the most often used together with KINSHIP words, the following category comprises words that can also refer to family relations. The least frequent co-occurrence belongs to the word *inimene*, whose generic meaning of 'human being' cannot be associated with kinship, family or couple relations.

4.2. Parts of speech of the left collocates

Besides the semantic components of the node word the occurrence of the collocates depends on the part of speech of the node word as well as on whether the collocate is found to the right or to the left of the node word, as a collocational relation usually exists between words that are syntactically related as well (Jackson and Amvela 2000:131). In the given case the node words are nouns, and their left collocates differ from the right ones due to syntactic constraints, if to nothing else. The dependence of word co-occurrence on structural causes was first discussed by Harris in his article "Co-occurrence and transformation in linguistic structure" (1957).

Analysing the parts of speech of the left collocates under scrutiny, note that uninflected words – adverbs, conjunctions and particles – were left out. For the part-of-speech distribution of the total of 300 collocates see Fig. 5.

Nearly one in every four collocates is a pronoun, the same percentage (24%) applies to verbs, while nouns cover 23% of the collocates. The adjectives (17%) and numerals are less numerous (17% and 12%, respectively). Note that the

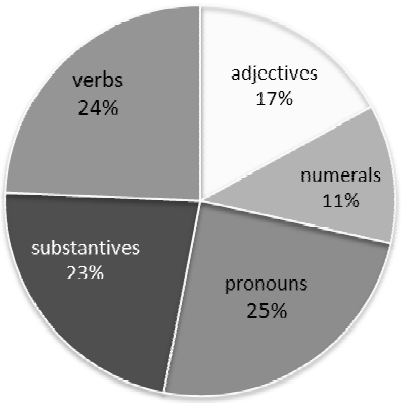

Figure 5. Distribution of 300 left collocates across parts of speech.

part-of-speech structure of collocations depends on the syntactic rules of word order and so are the left collocates analysed subject to syntactic constraints. For comparison, see the respective graph for the right collocates (Fig. 6).

In comparison with Figure 5 we can see how the percentage of adjectives, numerals and pronouns has decreased at the expense of verbs and substantives.

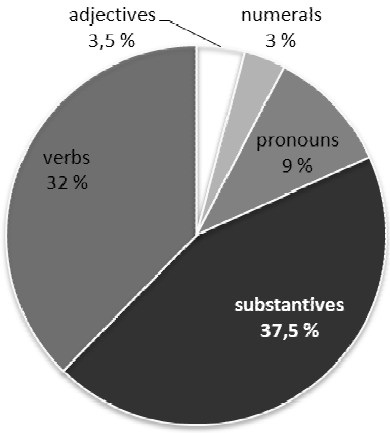

Figure 6. Distribution of 300 right collocates across parts of speech.

5. Conclusion

A comparison of the collocates of 10 Estonian words for 'human being' shows that nearly half of the collocates of any node word coincide with those of some other node word. For some pairs the coincidence is even higher: the words *mees*

'man' and *naine* 'woman', for example, share nearly three-fourths of their collocates. This result supports the distributional hypothesis, although the reason of the coincidence is left open: it is not clear whether the contexts coincide because all the node words share the semantic component 'human being' or because they are all nouns. The part of speech of the node word and the collocation span place their own constraints on the contextual options. Collocations also reflect the syntactic structure in that the left collocates contain considerably more (attributive) adjectives, numerals and pronouns than the right ones.

An analysis of the mutual relations within the lexical field of 'human being' revealed that the central position is occupied by *naine* 'woman' and *mees* 'man', which have the most in common with the rest of the words on the syntagmatic level. The word *ema* 'mother', which was third in the collocate coincidence list, had a larger intersection with kinship words than the former two. The words *inimene* 'human being', *laps* 'child', *mees* 'man/husband' and *naine* 'woman/wife' make up a relatively closely knit core, where the mutual coincidence of the collocate clusters is high. The reason may be that, in comparison with the remaining six words, the primary sense of the four is wider. The word *inimene* 'human being', for example, is not specified for gender, age, kinship or family relation. The primary sense of *laps* 'child' is age-specified and both *mees* 'man' and *naine* 'woman' are, in their primary sense, specified just for gender. The remaining six words, however, have been specified more narrowly, considering two or three aspects: the words *tüdruk* 'girl' and *poiss* 'boy' refer to both age and gender. The words *ema* 'mother', *isa* 'father', *tütar* 'daughter' and *poeg* 'son' refer to kinship, gender and to a higher or lower generation than self.

Within a lexical field, individual semantically distinctive collocates were generally not analysed. Only the co-occurrence of the genitive forms *minu* and *mu* 'my' of the personal pronoun *mina* 'I' with different words for 'human being' was investigated. (The pronoun *mina* 'I' belonged to the collocates of every node word involved.) An analysis of the data drawn from the unlemmatized corpus and the Internet revealed that the incidence of the genitive personal pronoun *minu* 'my' among the collocates distinguished the words with the semantic component +KINSHIP OR PARTNERSHIP, from those without it.

It was found that any semantic component taken separately did not affect the general distribution. There was no evidence, for example, of a greater coincidence among the collocate clusters of the words referring to the female gender. A comparison of individual collocates within the lexical field would probably have led to some collocates that are indeed characteristic of node words referring to a certain gender, but their role in the general distribution was nil. On the contrary, it was found that the highest coincidence occurred between the collocate clusters of words referring to opposite genders. All antonomous word pairs such as *mees-naine* 'man-woman', *poeg-tütar* 'son-daughter', *poiss-tüdruk* 'boy-girl', *ema-isa* 'mother-father', whose members differed by the value of the MALE component, had the greatest number of collocates in common. This result confirmed Sahlgren's argument that a comparison of the syntagmatic context (collocations)

of words may also enable conclusions on their paradigmatic (antonomous) relations. If antonymy is interpreted as a similarity relation (which is quite natural considering that the opposite value applies to just one semantic component out of several) we can agree with Harris in that similarity between words is manifested in their contextual coincidence.

Acknowledgements

This work was supported by the Estonian Research Council project "Modelling intermodular phenomena in Estonian"; no SF0050023s09 and Estonian Science Foundation grant No 7998. Author thanks prof. Urmas Sutrop for his helpful comments.

Address:
Liisi Piits
Institute of the Estonian Language
Roosikrantsi 6
10119 Tallinn, Estonia
E-mail: liisi.piits@eki.ee

References

Bartsch, Sabine (2004) *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Tübingen: Gunter Narr Verlag.

Cruse, David Alan (2001) *Lexical semantics*. Cambridge University Press.

Evert, Stefan (2005) *The statistics of word coocurrences: word pairs and collocations.* PhD dissertation, IMS, University of Stuttgart.

Eesti keele seletav sõnaraamat. 2nd ed. 6 vols. [Explanatory Dictionary of the Estonian language.] Tallinn: Eesti Keele Sihtasutus, 2009.

Firth, John Rupert (1957) *Papers in linguistics, 1934–1951*. Oxford University Press.

Fodor, Jerry A. and Jerrold J. Katz (1963) "The structure of a semantic theory". *Language* 39, 2, 170–210.

Gesuato, Sara (2003) "The company women and men keep: what collocations can reveal about culture". In *Proceedings of the corpus linguistics 2003 conference*, 253–262. Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, eds. UCREL, Lancaster University. Available at <http://ucrel.lancs.ac.uk/publications/cl2003/papers/gesuato.pdf>. Acsessed on 26.02.2013.

Harris, Zellig Sabbatai (1954) "Distributional structure". *Word. Journal of the linguistic circle of New York.* 10, 2–3, 146–162.

Harris, Zellig Sabbatai (1957) "Co-occurrence and transformation in linguistic structure". *Language* 33, 3, 283–340.

Jackson, Howard and Etienne Zé Amvela (2000) *Words, meaning and vocabulary: an introduction to modern English lexicology*. London and New York: Continuum.

Kaalep, Heiki-Jaan (1997) "An Estonian morphological analyser and the impact of a corpus on its development". *Computers and the Humanities* 31, 115–133.

Kaalep, Heiki-Jaan and Kadri Muischnek (2002) *Eesti kirjakeele sagedussõnastik*. [Estonian Frequency Dictionary.] Tartu: Tartu Ülikooli kirjastus.

Lyons, John (1977) *Semantics*. 2 vols. Cambridge: Cambridge University Press.

McEnery, Tony and Andrew Hardie (2012) *Corpus linguistics: method, theory and practice*. New York: Cambridge University Press.

Miller, Georg A. and Walter Charles (1991) "Contextual correlates of semantic similarity". *Language and Cognitive Processes* 6, 1, 1–28.

Muischnek, Kadri (2005) "Eesti keele verbikesksed püsiühendid tekstikorpuses". [Estonian multi-word expressions in a text corpus.] *Emakeele Seltsi aastaraamat* (Tallinn) 51, 80–106.

Murphy, Lynne (2006) "Antonymy as lexical constructions: or, why paradigmatic construction is not an oxymoron." In *Constructions all over: case studies and theoretical implications. Constructions SV*, 1–8. Doris Schönefeld ed. Available online at <www.constructions-online.de>. Accessed on 01.04.2013.

Nida, Eugene Albert (1975) *Componential analysis of meaning: an introduction to semantic structures*. The Hague: Mouton.

Palmer, Harold E. (1949 [1938]) *A grammar of English words*. London, New York, and Toronto: Longman.

Panther, Klaus-Uwe and Linda Thornburg (2012) "Antonymy in language structure and use". In *Cognitive linguistics between universality and variation*, 159–186. Mario Brdar, Milena Žic Fuchs, and Ida Raffaelli, eds. Newcastle upon Tyne: Cambridge Scholars.

Rubenstein, Herbert and John B. Goodenough (1965) "Contextual correlates of synonymy". *Communications of the ACM* [Association for Computing Machinery] 8, 10, 627–633.

Sahlgren, Magnus (2006) *The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD dissertation. Stockholm University.

Sahlgren, Magnus (2008) "The distributional hypothesis". *Rivista di Linguistica* (Italian Journal of Linguistics) 20, 1, 33–53. Special issue *From context to meaning: distributional models of the lexicon in linguistics and cognitive science*.

Saussure, Ferdinand de (2000) Course in general Linguistics. Charles Bally and Albert Sechehaye, eds. London: Duckworth.

Scott, Mike (2004) *"WordSmith Tools version 4": computer program. Oxford: Oxford University Press*.

Sinclair, John McHardy (2004) *Trust the text: language, corpus and discourse*. Ronald Carter, ed. London and New York: Routledge.

Stubbs, Michael (2001) *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Sutrop, Urmas (2000) "Basic terms and basic vocabulary". In. *Estonian: typological studies* 4, 118–145. Mati Erelt, ed. Tartu.

Sutrop, Urmas (2011) "Mis on põhivärvinimi, põhitase ja põhitaseme objekt?". [What is basic colour term, basic level and basic level object?] In *Värvinimede raamat,* 39–46. [Book on colour names.] Mari Uusküla and Urmas Sutrop, eds. Tallinn: Eesti Keele Sihtasutus.

NBCE = *Newspaper subcorpus of the balanced corpus of Estonian*. Available online at <http:// www.cl.ut.ee/korpused/grammatikakorpus/ajalehekirjeldus>. Acsessed on 25.08.2008.

Uiboaed, Kristel (2010) "Statistilised meetodid murdekorpuse ühendverbide tuvastamisel". [Statistical methods for phrasal verb detection in Estonian dialects.] Eesti *Rakendus-lingvistika Ühingu Aastaraamat. Estonian Papers in Applied Linguistics* 6, 307–326.