UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

THE USE OF ADJECTIVE-NOUN, VERB-NOUN AND PHRASAL-VERB-NOUN COLLOCATIONS IN ESTONIAN LEARNER CORPUS OF ENGLISH

MA thesis

LENNE TAMMISTE
SUPERVISOR: PILLE PÕIKLIK, PhD

TARTU
2016

ABSTRACT

Vocabulary is one of the most crucial aspects of language learning but a large vocabulary does not always guarantee effective communication. The knowledge of collocations is also important as it improves the fluency and quality of spoken or written language. Unfortunately, learning collocations can be a difficult task because there are no exact rules why some words fit together and others do not. Studies have shown that congruency (i.e. the presence or absence of L1 translation equivalent) and collocate-node relationship (i.e. the type of a collocation) can influence the use of collocations in learner language.

The characteristics of learner language can nowadays be studied by analysing written or spoken learner corpus stored in a computer database. Hundreds of learner corpora have been compiled around the world but in Estonia only few corpora have been built, none of them comprising of texts produced by Estonian learners of English. In 2014, a learner corpus of 127 essays was developed at the English Department of the University of Tartu, which finally made it possible to investigate the use of different collocations in Estonian EFL learners' writing.

The thesis has two main chapters. The first chapter describes the definitions of the term *collocation¸* explains the role of learner corpus in language teaching and gives an overview of previous research conducted. The second chapter describes the empirical study carried out in this thesis, explains the methodology, target collocations and the procedure of collecting the data. The study in this paper adopts a combination of quantitative and qualitative corpus analysis approaches. The *AntConc* Word List and Concordance tools (Anthony 2014) were used in order to extract all adjective-noun, verb-noun and phrasal-verb-noun collocations that were related to the most frequently used nouns in the corpus. The subchapter addressing the results presents the most frequently used adjective-noun, verb-noun and phrasal-verb-noun collocations found in the study, the distribution of collocations based on collocate-node relationship and congruency, and finally it analyses the naturalness of the collocations found in the English language. The last section in the second chapter presents an interpretation of the findings.

TABLE OF CONTENTS

INTRODUCTION

Vocabulary is one of the most important aspects of foreign language learning next to grammar, pronunciation and orthography. At the same time, large vocabulary without the knowledge how words combine does not always guarantee efficient communication. The knowledge of multiword units such as collocations is necessary as it can improve the fluency and quality of spoken or written language (Laufer and Waldman 2011: 648). However, erroneous use of collocations may lead to misunderstandings and signal a lack of expertise and knowledge (Henriksen 2013: 49). The importance of collocations in language learning can be illustrated with a statement by Wray (2002: 143) that "to know a language you must know not only its individual words, but also how they fit together."

Unfortunately, learning collocations can be a very difficult task for foreign language learners. A myriad of studies (Azizinezhad 2011, Brashi 2009, Juknevičienė 2008, Laufer and Waldman 2011, Nesselhauf 2005, Peters 2016, Vuorinen 2013) have shown that it is the productive knowledge where language learners' difficulties with collocations lie. Hill (1999: 5) notes that "students with good ideas often lose marks because they don't know the four or five most important collocations of a key word that is central to what they are writing about." A study conducted by Peters (2016) showed that even controlled productive use of collocations can be a serious challenge. The analysis of National Examination in English held in 2010 showed as well that in the language structures part a task that checked learners' knowledge of collocations had caused the most difficulties (Põder 2010). Therefore, there is a possibility that Estonian learners of English may experience similar difficulties when using collocations in free production, such as speaking orally or writing an essay.

One reason why collocations may cause problems is that there are no exact rules why some words fit together and why others do not. Carter et al. (2007: 59) add that the usual answer to the question why something is expressed in a certain way is "that's just the way we say it”. The influence of mother tongue can also be one of the factors why collocations are often misused. One way to investigate possible mother tongue influences on collocation production in learner language is to analyse collocations from a congruency perspective. Congruency is the presence or absence of a literal first language (L1) translation equivalent. An example of a congruent collocation for Estonian learners of English would be *to see the world* because a word-for-word translation ‘maailma nägema’ would also be acceptable in Estonian. An example of a non-congruent collocation would be *to call a halt* since in Estonian a word-for-word translation of it, ‘kutsuma peatust/seisu’, would seem unnatural and strange.

One possible explanation for L1-based errors in the use of collocations might be that language learners do not notice, as Gyllstad and Wolter (2011: 431) point out, that collocations tend to “vary considerably from language to language”. If the word-for-word meaning of a collocation is transparent, the language learner may not pay enough attention to the differences in L1 and target language collocations. This does not cause comprehension problems but may result in possible unnatural word combinations in the production process. The studies by Nesselhauf (2003) and Gyllstad and Wolter (2011) showed that learners of English made more errors in non-congruent than in congruent collocations. Laufer and Waldman (2011) and Vuorinen (2013) investigated the use of collocations by English learners at three proficiency levels and found that collocational errors were present at all levels, and that L1-based errors continued to be persist even at the most advanced level.

In addition to the mother tongue influence, the type of a collocation or ‘collocate-node relationship’[1] can also be one factor in producing erroneous collocations. Studies by Nesselhauf (2003) and Laufer and Waldman (2011) revealed that learners of English tend to struggle the most with choosing the correct verb in a collocation, such as *make a test* instead of *take a test*. Boers et al. (2014: 55) explain that since the noun in a verb-noun collocation

[1] In Peters (2016), ‘collocte-node relationship’ was used to describe the type of a collocation. In this study, the term ‘collocate-node relationship’ will also be used to refer to the nature of a target collocation, whether it is an adjective-noun, verb-noun or a phrasal-verb noun collocation.

carries the most semantic weight and the verb is very often a high frequency word, the learner may feel that there is no need to pay extra attention to the verb. Because of inflections such as tense, number, aspect and person, verb-noun and phrasal-verb-noun collocations show more variation in morphology as well, compared to adjective-noun or adverb-noun collocations (Peters 2016: 115). Phrasal-verb-noun collocations tend to be especially problematic for learners of English as a foreign language (EFL) because they consist of a verb, a noun and at least one particle, and very often they are also semantically rather opaque.

Since it is clear that collocations tend to cause problems for language learners, it raises a question how collocations should be addressed in language lessons. Brashi (2009: 29) proposes a number of pedagogical implications that can be considered as a framework or a model for teaching collocations to EFL learners: teachers should encourage their students to identify collocations in texts while identifying difficult words, to bear in mind that word-for-word equivalents between L1 and target language are not always appropriate and that creating their own collocations in foreign language can be risky since it may result in unnatural word associations. Peters (2016: 133) and Laufer and Waldman (2011: 666) suggest teachers to make learners aware of the interlingual differences in explicit vocabulary activities.

However, Nation (2001: 325) stresses that collocations do not always deserve attention from a teacher just because they exist. He suggests that only frequently occurring collocations and collocations of frequently used words should be addressed. From a language point of view, Nation (2001: 328) outlines the main reasons why we need research on collocations: research can provide information about high-frequency collocations and also the unpredictable collocations of high-frequency words; research can identify what the most common patterns of collocations are (whether some patterns need special attention while others do not) and finally; research into collocations is useful for dictionaries that help learners deal with low-frequency collocations.

Although numerous studies have been conducted on collocations around the world, there are only few papers written in Estonia that investigate English collocations in foreign language learning. In 2005, Merike Saar defended her MA thesis in EuroAcademy on the topic of noun collocations in English and there is also a diploma thesis on collocations in English by Kersti Kirs, defended at Tartu Teacher Training College in 2001. To date, Estonian learners' use of collocations and especially adjective-noun, verb-noun and phrasal-verb-noun collocations in the English language have not been studied. Therefore, there is a considerable need to investigate this matter. The present study aims to contribute to this growing area of research by exploring Estonian EFL learners' use of collocations in free production.

A suitable option of exploring the characteristics of learner language is to analyse collections of written or spoken texts stored in a computer database. One of the forerunners in this field, Sylvaine Granger, uses the term 'learner corpus' and defines it as "an electronic collection of authentic texts produced by foreign or second language learners" (Granger 2003). Quantitative methods are usually employed in corpus studies. For example, by using suitable software tools, hundreds of words can be extracted automatically from the corpus. In order to allow a deeper insight to the learner language, qualitative approaches have been adopted as well (Nesselhauf 2003, Vuorinen 2013). In these studies, target word combinations have been extracted and analysed manually to be sure that every word or phrase under investigation was recognised and included in the study. However, this method is the most time-consuming.

Although more than a hundred learner corpora of English have been compiled around the world and the interest in exploiting them is growing steadily (Cotos 2014: 203), learner corpus research on Estonian learners' of English is very limited. There are only few corpus studies carried out in Estonia that are based on Estonian learner language of English. In 2015, two MA theses were defended at the University of Tartu: one written by Anna Daniel on the topic of adjectives and adverbs in Estonian and British student writing; and the second by Elina Merilaine who investigated Estonian ESL learners' use of frequency and variability of conjunctive adjuncts. Both authors carried out their studies by using the same learner corpus of English that comprised of 127 essays written in 2014 as a part of the entrance examination to English Language and Literature BA programme. This Estonian learner corpus of English is the object of this study as well.

The main focus of the present study is to investigate Estonian EFL learners' use of adjective-noun, verb-noun and phrasal-verb-noun collocations. It seeks to find out what the most frequently used adjective-noun, verb-noun and phrasal-verb-noun collocations in the corpus are and also examine the role of collocate-node relationship and congruency on collocation production in learners' essays. Considering the large number of target collocations that can be found in the learner corpus, a selection had to be made on which specific word combinations this study will focus. Following the example of Fan (2009), research is carried out by investigating only collocations that are formed with the most frequently used nouns. Detailed information about selecting the nouns and extracting the target collocations is explained in section 2.2.

In terms of research questions, mixing quantitative and qualitative approaches was preferred so that the study could, next to providing basic information concerning the frequencies and distributions of collocations found, also take a closer look at the naturalness of the collocations produced. This means that two research questions in the study require a quantitative approach to identify frequencies and distributions of target collocations. The third research question is of a more qualitative nature and requires the study to analyse the collocations more closely. Hopefully, the results of the study will offer practical information for teachers of English in terms of teaching collocations to Estonian learners.

The three research questions in this study are:

1. What are the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations used in the learner corpus?
2. What is the distribution of collocations found in the learner corpus based on the collocate-node relationship and congruency?
3. To what extent are the collocations produced natural in the English language?

The thesis has two main chapters. The first chapter focuses on reviewing the literature written on collocations and learner corpus research. It gives an overview of different definitions of the term *collocation*, describes the role of learner corpus in language teaching and the end of the chapter addresses learner corpus studies on collocations carried out in previous years. The second chapter is concerned with methodology and the procedure of collecting the data for this study. The remaining part of the second chapter outlines the results and discusses the value of the findings in the field of learner corpora research.

LITERATURE REVIEW OF COLLOCATIONS AND RESEARCH ON LEARNER CORPUS

1.1 Definitions of the term *collocation*

According to Oxford Advanced Learners' Dictionary, the word *collocation* was first used in Late Middle English, originating from the Latin word 'collocare' which carried a meaning of placing together (from *col* – 'together' and *locare* – 'to place'). The historic meaning of the word connected with placing side by side with something is still prevalent, although *collocation* acquired another meaning from modern linguistics during the 20th century. In linguistics, the term *collocation* refers to the habitual association of particular word with other particular words (Oxford English Dictionary).

Two British linguists – John Rupert Firth and Harold E. Palmer – have been associated with being the first linguists to adopt the term *collocation* in modern linguistics (Piits 2015: 11). Palmer (1931: 4) defined collocations as "succession[s] of two or more words that must be learnt as an integral whole and not pieced together from its component parts or as *comings-together-of-words*". J. R. Firth proposed that *collocation* should be used as a technical term in modern linguistics, and he is also famous for summarising the principle of co-occurrence of words as "You shall know a word by the company it keeps!" (Firth 1957). In recent years, a myriad of alternative definitions have been suggested for the term, since it is used in widely different senses in linguistics and language teaching. Nation (2001: 317) and Henriksen (2013: 30) also admit that determining what should be classified as a collocation has been a major problem. Therefore, it is necessary for any study related to this topic to clarify the scope of collocations (Durrant 2014: 446).

Although a generally accepted definition of collocations is lacking, two main approaches can be identified in defining the term: the 'frequency-based approaches' and ‘phraseological approaches’ (Durrant 2008: 32, Gyllstad 2007: 6, Lukač and Takač 2013: 387, Nesselhauf 2005: 12, Peters 2014: 80, Vuorinen 2013: 15). These two approaches are described as follows:

1. In the ‘frequency-based approach’ collocations are related to frequency and statistics, the aspects of which are mainly investigated by scholars working in the field of computational linguistics and corpus linguistics (Gyllstad 2007: 6). According to Durrant (2014: 446), this approach defines collocations as “sets of words which have a statistical tendency to co-occur in texts” (such as *shrug shoulders* or *drink tea*). The approach goes back to the English linguist J. R. Firth (Nesselhauf 2005: 12), which is why the followers of this tradition are sometimes called the Firthian or Firthians.
2. The ‘phraseological approach’ has largely been influenced by Russian phraseology and it is tightly linked to the fields of language pedagogy and lexicography (Gyllstad 2007: 6, Nesselhauf 2005: 12). According to Durrant and Mathews-Aydinli (2011: 59), in this approach collocations are defined as: 1) word combinations in which one element of the combination does not carry its general meaning (as in *take a step*); and 2) word combinations where some form of restriction is present on which words with similar meanings can be substituted into the phrase (as in *make a decision*, where the verb cannot be substituted with *do* or *produce*, for example). Nesselhauf (2005: 17) notes that the elements comprising the collocation should also be syntactically related, like noun + noun, adjective + noun, verb + noun, etc.

In short, the key aspect in the frequency-based approach is frequency and co-occurrence, in the phraseological approach a degree of semantic or substitutional fixedness. Some authors (Durrant 2008: 32, Gyllstad 2007: 17, Nesselhauf 2005: 18) conclude that there is a fair degree of overlap between the two and sometimes researchers adopt criteria from both traditions. In addition to these two approaches, Durrant and Mathews-Aydinli (2011: 59) propose a third one as well – ‘psycholinguistic approach’, which refers to collocations as word combinations that have psychologically associative links between the elements. They add that this approach clearly overlaps with the previous approaches, since both the frequency of occurrence and semantically restricted word combinations are likely to entail some form of psychological representation as well.

Within these broad approaches, the following subsection presents different definitions that have been suggested for the term ‘collocation’. Sinclair (1991: 170) defines collocations as “the occurrence of two or more words within a short space of each other in a text”, which adopts the frequency-based approach. Cowie, being a typical representative of the phraseological approach, defines collocations by delimiting them from idioms and free combinations (Cowie 1998: 127). Nation (2001: 317) has adopted both frequency-based and phraseological approaches in defining the term, using the term ‘collocation’ to loosely describe any commonly accepted grouping of words into clauses or phrases, including fixed expressions and idioms. From language learning point of view, he suggests regarding collocations as items which have some degree of semantic unpredictability and which frequently occur together.

The definition suggested by Henriksen (2013: 30) draws more attention to the functions of words and the number of elements included, stating that: “Collocations are frequently recurring two-or-three word syntagmatic units which can include both lexical and grammatical words”. Juknevičienė (2008: 2) notes only the criterion of substitutability and distinguishes collocations from free collocations and idioms, defining the term as “word combinations having arbitrary restriction on the commutability of their elements”.

The definition proposed by Laufer and Waldman (2011) adopts the phraseological approach and outlines the most important differences between collocations and other possible word combinations, regarding collocations as: “habitually occurring lexical combinations that are characterised by restricted co-occurrence of elements and relative transparency of meaning" (Laufer and Waldman 2011: 648-649). They (ibid.) add that the restrictiveness of co-occurrence contrasts collocations with free combinations where individual words can be easily replaced, whereas relative semantic transparency, on the other hand, distinguishes collocations from idioms whose meaning is often opaque and less transparent. They (ibid.) conclude that because of the restricted co-occurrence and semantic transparency, collocations are placed "on the continuum between free combinations and idioms".

While a variety of approaches to defining the term 'collocation' has been identified, it is important to explain the definition used in this thesis. The collocation types under investigation in this study are adjective-noun, verb-noun and phrasal-verb-noun collocations in which the collocation elements are syntactically related. On the other hand, word combinations are not classified as free or restricted in this study and free collocations are also included in the investigation. Therefore, *a collocation* in this paper is loosely related to both the frequency-based and phraseological approach when defining the term, referring to all adjective-noun, verb-noun and phrasal-verb-noun word combinations.

1.2 Learner corpus in language teaching

Before the advent of computers, the term 'corpus' was largely associated with just a body of words, such as the writings generated by one author (Carter et al. 2007: 1). Nowadays, the term is generally related to texts stored in a computer. For example, McCarthy (2004: 1) defines a corpus as "a collection of texts, written or spoken, usually stored in a computer database." A computer database allows us to analyse the stored texts in order to find different information, depending on the purposes the corpus was built for. McCarthy (2004: 1) notes that we can get plenty of answers by searching a corpus, such as: finding out the most frequently used words, phrases and tenses in English or differences between written and spoken language. Perhaps the most important advantages of computerised corpora are that by using suitable software tools we can search through corpora reliably and rapidly (Hardie and McEnery 2011: 2) to find out how the language is used in a real context (McCarthy 2004: 1).

Until recently, corpora in language teaching have been related to native speakers' language only (Nesselhauf 2004: 125). Nesselhauf (2004) argues that native speaker corpora are mainly useful because they reveal what native speakers typically say or write but, in terms of language teaching, it is also important to know the difficulties of language learners. However detailed a native corpus may be, it will never tell anything about the difficulties that might be faced by language learners (Granger 2003: 534). In the early 1990s, when academics and publishers started to collect and analyse learner language (Granger 2003), a new phenomenon called 'learner corpus' emerged. To date, more than a hundred learner corpora of English have been compiled and the interest in exploiting them has grown steadily (Cotos 2014: 203). A comprehensive list of learner corpora around the world is provided and managed by Amandine Dumont and Sylviane Granger on the website of Centre for English Corpus Linguistics of Université Catholique de Louvain[2].

Since the data in learner corpora present learners' production skills together with possible mistakes and errors, they may prove highly beneficial for language acquisition researchers, language teachers and publishers. Learner corpus research offers new exciting pedagogical perspectives for a wide range of areas in English language teaching pedagogy: in materials and syllabus design, language testing and classroom methodology (Granger 2003: 542). Nesselhauf (2004: 130) claims that by investigating only experimental data such as grammaticality judgement tasks or choice tasks it does not enable researchers to make conclusions about what learners can spontaneously produce – therefore, it is of great interest to analyse real production data.

[2] Available at https://www.uclouvain.be/en-cecl-lcworld.html.

Some publishers also use learner corpora for error coding so as to collect useful information for dictionary writers and other material compilers to highlight any typical problems (Carter et al. 2007: 17). McEnery and Xiao (2010: 365) point out that in the case of dictionary production for language learners, there used to be a tradition of using invented examples rather than authentic materials because lexicographers had believed that foreign language learners had difficulties with understanding authentic texts and therefore needed to be presented with simple examples. They (ibid.) add that the COBUILD (Collins Birmingham University International Language Database) project broke with that received tradition by using data from native corpora for illustrating learner dictionaries with authentic examples.

As far as the interpretation of the term 'learner corpus' is concerned, there is a degree of uncertainty around the exact definition. Nesselhauf (2004: 127) notes that since learner corpus is a fairly new phenomenon in corpus linguistics, it has not yet been described systematically enough. Sylviane Granger, one of the forerunners in the field, defines the term 'learner corpus' as "an electronic collection of authentic texts produced by foreign or second language learners" (Granger 2003). Granger (2002: 4) uses the term 'computer learner corpora' and suggests adopting another definition of the corpora, proposed by Sinclair (1996):

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT[3] purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.

Sinclair (1996, cited in Granger 2002: 4)

Besides the definition suggested by Sinclair, Nesselhauf (2004: 127) adds that the definition of learner corpora should also include a notion that the text collections are intended for more general use, not only for certain studies. In this thesis, the term 'learner corpus' is used in its broadest sense to refer to systematic computerised collections of both spoken and

[3] SLA – Second Language Acquisition; FLT – Foreign Language Teaching

written texts produced by language learners, despite the fact that in this study only a written corpus is investigated.

Granger (2002: 7) describes four basic types of learner corpora: monolingual or bilingual corpus; general or technical corpus: synchronic or diachronic; and written or spoken corpus. She adds that synchronic learner corpora, which describe learner language at a particular point of time, have been created more often than diachronic ones due to the fact that the latter require following learner language for months or even years, and therefore, they are very difficult to compile. Learner corpora also tend to be more written than spoken, since oral data are much more difficult to gather (Granger 2002: 8, Hardie and McEnery 2011: 2). In terms of written learner corpus, they usually contain only argumentative writing texts; other text types, such as descriptive writing, to date have not been studied enough (Cao and Hong 2014: 203) since compiling descriptive data is such a laborious and difficult task.

The findings from learner corpora can be exploited in different areas of language learning and teaching. Some key advantages of learner corpora are listed as follows: the findings can show evidence of any learner under, over- and misuse (Granger 2003: 534); they give researchers an accurate overview of how students are actually using the language (Meyer 2002: 27); and that the findings can be employed in the development of Computer Assisted Language Learning (Carter et al. 2007: 23). Granger (2002: 2) is sure that learner corpora provide a new type of data that help to improve learning and teaching of both second and foreign languages.

Another advantage of learner corpora is that they can be used for both qualitative and quantitative analysis (Carter et al. 2007: 2). In terms of the quantitative methods, the use of suitable linguistic software in computer corpus methodology enables researchers to conduct analyses of extensive learner data (Granger 2002: 2, Nesselhauf 2004: 130). Computer-aided analyses cannot only be used to test different hypotheses but the data may also reveal some undiscovered aspects of learner language and offer new ideas for future research (Nesselhauf 2004: 131). In addition, more comprehensive studies are possible, since many aspects of language can be investigated at once, taking into account the learners' proficiency levels, their mother tongue, text type, the age and sex, the years of acquisition and any other information provided within the corpus (ibid.).

A corpus can also be used to compare varieties of non-native language, or native and non-native languages. Nesselhauf (2004: 125) is sure that comparing learners' language with the language produced by native speakers is the best way to identify possible areas of difficulty that language learners may be struggling with. However, Carter et al. (2007: 28) question the appropriateness of exploiting native speaker models in language learning. They argue that language learners may only be interested in operating in an international context and therefore they should not necessarily be judged against native-user standards (ibid.). By investigating learner corpora of non-native English speakers it can also be identified whether language learners desire to learn more native-like or international type of English, the results of which could be used as a preferred basis for classroom teaching and learning (Carter et al. 2007: 28).

However, learner corpus has its disadvantages as well. Nesselhauf (2004: 130) lists a number of them, such as: 1) they do not enable us to investigate learners' receptive abilities; 2) they cannot be used to answer questions like how certain the learners are about the correctness of their produced language. She also adds that any rare phenomena of language should be better studied experimentally since, if a word or a language structure does not occur in a corpus, it does not automatically mean that the learner does not understand or know how to use it (ibid.). Consequently, learner corpus enables us to investigate only learners' productive skills without any additional information such as how certain the learners are about their accuracy or what other words or phrases they can actually produce.

Another limitation of learner corpora is that corpus compilation is a very laborious task (Hasko 2013: 7, Nesselhauf 2004: 132). In terms of processing tools, there is a desperate need for more sophisticated extraction devices because, at the moment, some search processes can still be only done manually. For example, a computer cannot recognise erroneous word combinations in a certain context. Granger (2004: 138) says that although learner corpus studies have already been recognised by ELT[4] and SLA communities, a wider range of different learner corpora and more elaborate processing are still needed.

McEnery and Xiao (2010: 374) feel that in order to popularise language corpora to more general language teaching context, the corpus linguists' future tasks should be facilitating sufficient access to appropriate corpus resources and offering necessary training for teachers. Nesselhauf (2004: 132) admits that real progress can only be made by collaboration and data-sharing. Therefore, further investigation into learner corpus by collaboration between other researchers is strongly needed.

1.3 Previous learner corpus studies on collocations

Over the past decades, an increasing amount of studies that concentrate on collocations has been conducted in the field of learner corpus research. A frequent research method adopted in these studies has been comparing a learner corpus with a comparable native corpus and identifying errors or patterns of over- and underuse. In most cases, the learner corpora included in the studies consist of texts produced by language learners having the same mother-tongue background (Granger 2012: 132). For example, German learner corpus was used in Nesselhauf (2003, 2005), Finnish in Vuorinen (2013), Lithuanian in Juknevičienė (2008), Swedish in Gyllstad and Wolter (2011, 2013), Polish in Kaszubski (2000) and Chinese in Fan (2009). Most of the studies have investigated written learner corpora but there are some that focus on non-native speech as well (Allami and Attar 2013, Foster 2001).

[4] ELT – English language teaching

Another approach has been employing different testing tools for measuring collocational knowledge. However, as Henriksen (2013: 45) points out, many of these testing instruments have not been validated or piloted extensively. In order to solve this problem, other researchers have carried out considerable work on developing standardised tools. In Gyllstad (2007), two tests were developed, COLLEX and COLLMATCH, which now can be used for testing receptive collocation knowledge. A testing instrument for measuring productive collocational knowledge was developed in Revier (2009). A majority of the studies focus on the product of learning rather than the process of acquisition (Henriksen 2013: 45). One example where the acquisition of collocations has been studied is Peters (2016) which investigated the aspects that can affect the learning burden of learning collocations among Dutch EFL learners. The strength of this study is that both receptive and productive knowledge are measured in different recall and recognition tests.

Very often the studies have focused on a certain type of collocation, investigating, for example, the use of verb-noun collocations (Gyllstad 2007, Juknevičienė 2008, Kaszubski 2000, Koya 2005, Laufer and Waldman 2011, Nesselhauf 2003, 2005, Peters 2009, Vuorinen 2013) or adjective-noun collocations (Siyanova and Schmitt 2008). Some studies have also included different types of collocations such as adjective-noun, adverb-adjective, verb-noun or phrasal-verb-noun collocations (Fan 2009, Peters 2014, 2016). Not all target collocations have always been chosen for investigation, though. Sometimes an analysis is performed only with a selection of word combinations, such as collocations consisting of the most commonly occurring nouns (Fan 2009) or verbs (Juknevičienė 2008, Kaszubski 2000).

The findings have shown that collocations tend to cause various problems for language learners. In terms of verb-noun collocations, selecting the correct verb has caused difficulties the most (Laufer and Waldman 2011, Nesselhauf 2005). Granger (2012: 141) points out that learners have been shown to rely heavily on congruent collocations, which have a translation equivalent in their mother tongue. The findings from studies Fan (2009) and Nesselhauf (2003, 2005) support this view. Mother tongue influences have been evident in erroneous collocations as well. Nesselhauf (2005) and Laufer and Waldman (2011) both found that almost half of the incorrect collocations were influenced by the mother tongue. Furthermore, L1 influence did not decrease with time in Laufer and Waldman (2011) where Hebrew learners from three proficiency levels were examined. On the other hand, these results contradict Vuorinen (2013), in which the negative influence of Finnish language was no longer found at the advanced level.

Although a large amount of studies have been conducted in recent years presenting interesting findings about learning or producing collocations, drawing any general conclusions from them is problematic. Granger (2012: 135) identifies two factors that make it difficult: the studies have examined learners of different mother-tongue backgrounds and language proficiency; and second, the target collocations and methodologies of extracting and analysing them differ greatly. Therefore, the findings are not always directly comparable.

To date, the research on collocations in Estonia has tended to focus on language learning or exploring Estonian collocations related to a certain topic. For example, Piits (2015) examined collocations of the most frequent Estonian words for ‘human-being’. Studies important in terms of language learning have been carried out by Jaanits (2004) and Heinsoo (2010) who both focused on exploring Estonian and Finnish collocations. Estonia as a foreign language for Russian learners has also been investigated recently. Belozerskaja (2013) analysed collocations connected to Estonia that were presented in Estonian language learning materials for Russian learners. Her aim was to explore how Estonia was presented and what kind of conception Russian learners could get through these collocations. Timofejeva (2015) tested Russian learners’ knowledge of Estonian collocations by using the same COLLMATCH and COLLEX testing tools developed in Gyllstad (2007).

As described above, a number of studies have focused on collocations, whether carried out in Estonia or around the world. However, very little is still known of Estonian learners of English and their ability to comprehend or produce English collocations. Although Kris (2001) and Saar (2005) have investigated collocations in English as a foreign language, as was described in the introduction, it is still unknown how Estonian learners of English use different types of collocations. This MA thesis hopes to fill the gap in this area, being a first study to explore Estonian EFL learners’ use of adjective-noun, verb-noun and phrasal-verb-noun collocations. So far this paper has provided a brief overview of different definitions of the term *collocation*, the role of learner corpus in language teaching and recent studies on collocations. The next chapter will describe and discuss the empirical study conducted for this thesis.

2. ADJECTIVE-NOUN, VERB-NOUN AND PHRASAL-VERB-NOUN COLLOCATIONS IN ESTONIAN LEARNER CORPUS OF ENGLISH

The aim of this study is to investigate the use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian EFL learners' essays. The study sets out to find answers to the following research questions: what are the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations used in the learner corpus; what is the distribution of collocations found in the learner corpus based on the collocate-node relationship and congruency; to what extent are the collocations produced natural in the English language. Before describing and discussing the results, a description of the methodology, learner corpus and the procedure of collecting target collocations from the corpus are given in the following sections.

2.1 Methodology

Although quantitative methods are usually employed in corpus studies, taking into account the aims of the thesis, this study adopts a combination of both quantitative and qualitative corpus analysis approaches in order to allow a deeper insight into adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian student writing. This means that in addition to presenting the frequencies of different types of collocations used, their naturalness in the English language will be also examined. Detailed information about the learner corpus, target collocations and the procedure of collecting and dividing them is explained next.

The essays that comprise the learner corpus investigated in this study were written in 2014 as a part of the entrance examination to English Language and Literature BA programme. The requirement for entering the examination was the certificate of secondary education. Additional information about participants' educational background or length of previous study of the English language was not gathered. The aim of the examination was to test applicants' proficiency in reading and writing in English. The participants had to read an excerpt of an academic article about the future of the English language and then write a 200-word essay by explaining their own opinions on the topic. The writing task and the original reading text are provided in Appendix 1.

One aspect of the entrance examination to English Language and Literature BA programme has to be clarified. Some applicants did not have to complete the entrance examination if they fulfilled the following requirements: had scored at least 95 points out of 100 in the National Examination in English in 2014 or in previous years; had a Certificate in Advanced English (CAE) at least level C1; had a Certificate of Proficiency in English (CPE); had scored 100 points (the maximum) in the Test of English as a Foreign Language (TOEFL); or had scored at least 7 points in The International English Testing System (IELTS). Therefore, this learner corpus does not contain essays written by all applicants who wanted to enrol in the BA programme.

In order to convert the essays into a learner corpus, each essay was typed in electronically and then checked by two postgraduate students. During the typing process, the essays were not edited. Illegible words were highlighted but not removed and mistakes, including spelling mistakes, were not corrected. Only the titles were removed from the essays. The corpus contains one essay per participant. Although the total number of participants in the examination was 132, the learner corpus includes 127 essays, since two participants did not hold Estonian citizenship and three participants did not write the essays at all. The participants without Estonian citizenship were excluded from the study to ensure that the learner corpus consisted of texts produced by people from a similar educational and language background. The importance of participants' connection with Estonian language is also described further in the next subchapter addressing target collocations and the procedure of collecting the data.

Additional information about the essays was stored separately. The corpus was provided with a documentation file where each essay was given a code, and if necessary, the codes could have been deciphered in order to link the texts to the authors. The documentation file with additional information was not visible during the corpus study. To maintain confidentiality, the researcher had access only to the texts and the code numbers the essays had been assigned.

The main features of the Estonian ESL learner corpus can be described as follows:

- the corpus consists of 127 essays;
- the corpus consist of 24, 457 tokens[5];
- the average length of an essay is 193 words (the length of the essays varies from 60 to 320 words);
- participants' mother tongue is not specified;
- all participants have Estonian citizenship;
- the age of the participants varies from 18 to 35, with an average age of 19
- among 127 participants there are 88 females and 39 males;
- reference tools were not allowed but the participants were provided with a source text (see Appendix 1).

The corpus had to go through a 'data massaging' process before it was ready for using it in corpus analysis software programs. The essays were originally stored in one Microsoft Word document, each essay with its identification code on a separate page. Corpus analysis software programs usually require plain text formats and therefore, each essay had to be converted into plain text files. First, all identification codes, unnecessary spaces and lines, also comments made in the essays during the typing process were removed by executing the

[5] According to Microsoft Office Word word count tool.

Find and Replace tool in MS Word. Then, the Macro tool[6] was used to convert the Word document into 127 separate plain text files. After inserting the necessary code, the Macro tool split the Word document into 127 different Word files and converted them into plain text files that were now ready to be loaded into corpus analysis software programs.

2.2 Target collocations and the procedure of collecting the data

The target collocations and the specific procedure how to investigate the collocations found in the essays were chosen by partly adapting the methods used by Nesselhauf (2005), Fan (2009) and Peters (2016). In Nesselhauf (2005), only verb-noun collocations were included in the study but the learner corpus used there was rather large (318 essays, comprising of 154,191 words). Since the Estonian ESL learner corpus is much smaller, it was decided to broaden the scope of the study. By following the example of Peters (2016), adjective-noun and phrasal-verb-noun collocations were added to the list of target collocations. In Peters (2016), where three different types of collocations were under investigation, the term 'collocate-node relationship' was used to describe the differences in target collocations. The noun in each collocation was called a 'node' and depending on the collocation type, the noun or 'node' had a 'collocate': an adjective, verb or a phrasal verb. In this study, the term 'collocate-node relationship' will also be used to refer to the nature of a target collocation, whether it is an adjective-noun, verb-noun or a phrasal-verb-noun collocation.

It has to be clarified that this study did not investigate all adjective-noun, verb-noun or phrasal-verb-noun collocations found in the essays because it would have meant finding and analysing thousands of different word combinations. Because of this, the scope of the study had to be limited and by following the example of Fan (2009), it was decided to focus on the

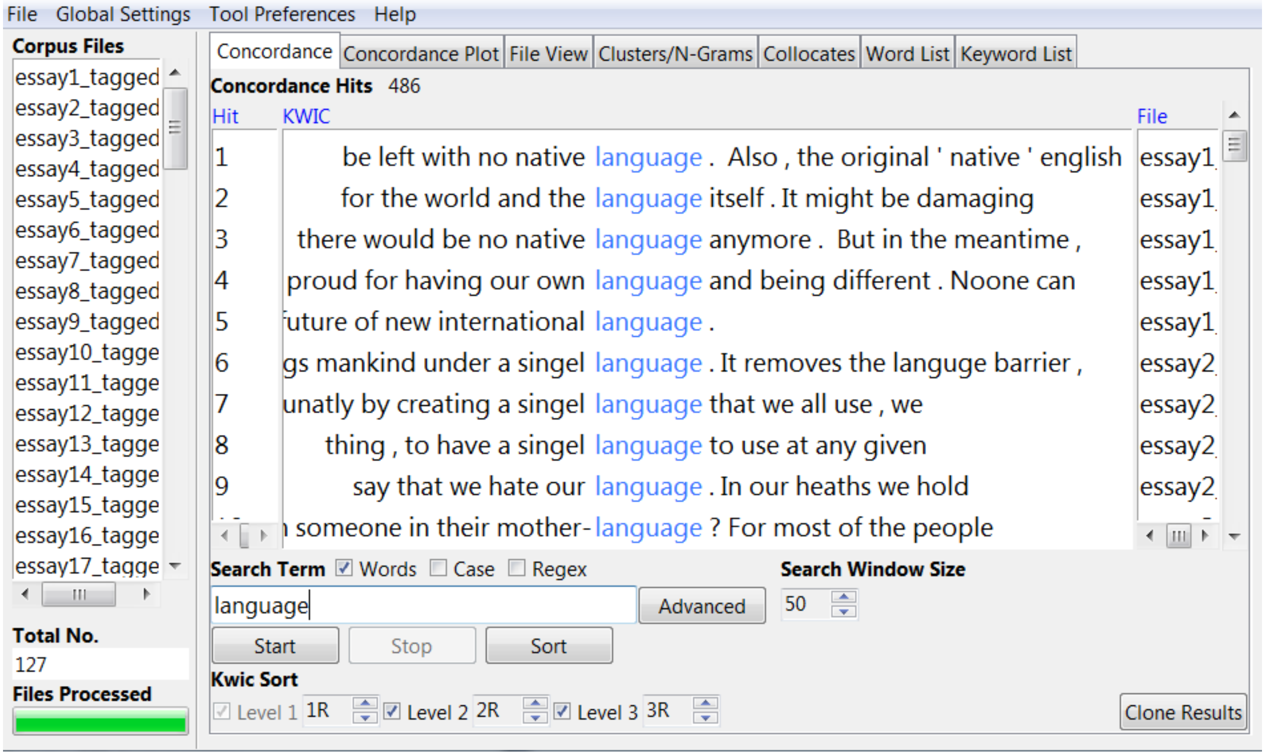
[6] Macro tool is a useful program in Microsoft Word which is able to perform complex tasks by using series of commands and instructions.

most frequently used nouns and their possible adjective or verb collocates. In order to discover the most frequent nouns in the corpus, the *AntConc* (Anthony 2014) Word List tool was used. It presented the list of most frequent words, of which the nouns were manually extracted. To enlarge the scope of the study, both singular and plural forms were included. For example, the word *language* was used 486 times and *languages* 248 times, occurring together 734 times in the corpus. In the same way all other nouns and numbers of their singular or plural occurrences were found and written down. Not all frequent nouns were added to the list, though. Words like *example*, *opinion* or *conclusion* occurred over 20 times in the corpus but they were only used in fixed contexts (*in my opinion*, *for example*, *in conclusion*) and because of this, these nouns were left out of the study. The list of ten nouns that were finally chosen for the study is presented in Table 1.

Table 1. Ten nouns chosen for the study and the number of their occurrences.

Language, languages	734	Consequence, consequences	123
Standard, standards	254	Culture, cultures	65
People	247	Word, worlds	46
Country, countries	189	Rule, rules	44
World, worlds	139	Problem, problems	37

The next step in the data collection procedure was to extract all adjective, verb and phrasal verbs that were surrounding the nouns chosen for the investigation. This was done in *AntConc* program by executing the Concordance Tool which can produce concordance lines in a KWIC (KeyWord in Context) format after entering the search term in the search box. Concordance lines were generated with each noun which enabled to discover the adjective, noun or phrasal-verb collocates that surrounded them. The commonly used range of items shown to the left and to the right of the search term is four or five (Adolphs 2006: 52). As shown in Figure 1, in this study the default setting of *AntConc* was used so that each noun was surrounded by five items to the left and to the right.



Corpus Files: essay1_tagged, essay2_tagged, essay3_tagged, essay4_tagged, essay5_tagged, essay6_tagged, essay7_tagged, essay8_tagged, essay9_tagged, essay10_tagge, essay11_tagge, essay12_tagge, essay13_tagge, essay14_tagge, essay15_tagge, essay16_tagge, essay17_tagge

Total No. 127

Files Processed

Concordance Hits 486

```
Hit  KWIC                                                                                   File
1         be left with no native language .  Also , the original ' native ' english         essay1
2         for the world and the language itself . It might be damaging                      essay1
3      there would be no native language anymore .  But in the meantime ,                   essay1
4    proud for having our own language and being different . Noone can                      essay1
5    uture of new international language .                                                 essay1
6    gs mankind under a singel language . It removes the langue barrier ,                  essay2
7    unatly by creating a singel language that we all use , we                             essay2
8        thing , to have a singel language to use at any given                              essay2
9          say that we hate our language . In our heaths we hold                            essay2
     someone in their mother- language ? For most of the people
```

Search Term ☑ Words □ Case □ Regex

language

Search Window Size 50

Kwic Sort

☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 1. Concordance lines of the word *language* in the Estonian EFL learner corpus.

Although the essays had been POS-tagged using the *TagAnt* (Anthony 2015) software and it would have been possible to run more specific searches such as giving the Concordance Tool a command to find all adjectives before the word *language* (by entering "*_JJ language"[7] in the search box), this option was not selected. The concordance results showed that the automatic POS-tagger had not been successful in identifying parts of speech in erroneous sentences and because of this, it was decided that all adjective, verb and verb-noun collocates were extracted manually.

When all word combinations were extracted from the corpus, they were sorted into two different groups. As it was mentioned in the introduction, the influence of mother tongue is one of the main factors why collocations are often misused and especially incongruent collocations tend to cause problems for language learners. Following the approach of Peters (2016), the extracted word combinations were divided into two groups based on their

[7] POS-tagger softwares use different tagging symbols. In *TagAnt*, for example, _JJ is a tag for an adjective, _NN for a noun, etc (Anthony 2015).

congruency. According to Nesselhauf (2005: 221), a congruent collocation is a word combination that can be given a word-for-word translation in L1. A non-congruent collocation, on the contrary, does not have an exact translation equivalent in learner's mother tongue (ibid.). Although in this study the participants' mother tongue was not specified, based on the knowledge that all participants held Estonian citizenship it was still decided to choose Estonian as the reference language. An example of how the collocations were divided by congruency and collocate-node relationship can be seen in Table 2.

Table 2. Target collocations divided by congruency and collocate-node relationship.

	Congruent collocations	**Non-congruent collocations**
Adjective-noun collocations	*different nations*	*powerful languages*
Verb-noun collocations	*learn languages*	*call a halt*
Phrasal-verb-noun collocations	*agree with people*	*look into the problem*

Word combinations were divided only by congruency and collocate-node relationship and other divisions, such as categorising the word combinations according to the level of restriction were not made. Therefore, this study also includes free collocations which are formed purely based on their semantic suitability (such as *read a book*), not only restricted collocations that carry a degree of semantic or substitutional fixedness (such as *do homework* instead of *make homework*). All adjective-noun, verb-noun or phrasal-verb-noun collocations, whether free or restricted, were investigated.

To answer the question whether the adjective-noun, verb-noun and phrasal-verb-noun collocations found were natural in the English language, different online dictionaries (Oxford Learner's Dictionaries, Online Oxford Collocation Dictionary of English) and word corpora (British National Corpus and Corpus of Contemporary American English) were consulted. Following the example of Nesselhauf (2005), collocations were judged acceptable if they occurred in identical form in the dictionaries. If the word combination was not found in the dictionaries, it was looked up in the word corpora. Word combinations were judged acceptable if they occurred in identical form in at least five written texts in the British National Corpus (BNC) or in the Corpus of Contemporary American English (COCA). Although in Nesselhauf (2005) only the BNC corpus was used, in this study the COCA corpus was included as well in order to cover both British and American language use. The advantage of COCA over BNC is its size (520 million words compared to 100 million words in BNC), which makes it the largest freely-available corpus of English. In order to divide the word combinations by congruency, an Estonian word corpus etTenTen (2013) was also used. The latter consists of 260,559,829 words and different types of texts, including forums and blogs found on the internet. The corpus was accessible in the Sketch Engine corpus software interface[8]. It also needs to be clarified that verb-noun combinations with *to be* and other combinations such as 'noun-verb' combinations where the noun functioned as subject not an object, were excluded from the study.

Another aspect that must be explained is that the participants were provided with the reading article as the source text (see Appendix 1) while writing the essays. Although reference tools were not allowed, the source text could still offer some help to the learners. Thus, the participants' word choice may have been influenced by the original reading text. Since the main focus of this study was to investigate Estonian ESL learners' use of collocations and therefore, vocabulary in general terms, it was a big concern how to distinguish any possible negative influences from the source text. Some essays even contained entire sentences copied directly from the source text. It was decided to extract all adjective-noun, verb-noun and phrasal-verb-noun collocations from the source text and if any of them were found in the essays, these word combinations were highlighted. Collocations found in citations were completely left out. Having introduced the methodology and procedure adopted in this study, the next subchapter focuses on describing the results.

[8] Available at https://www.sketchengine.co.uk/ .

2.3 Results

This subchapter presents the results of the study. It begins with presenting the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations found in the learner corpus, after which it will move on to describe how the collocations were distributed based on the collocate-node relationship and congruency. At the end of the subchapter, attention is paid to the naturalness of the collocations produced in the learner corpus. Before proceeding to examine the findings of the study, a general description of the collocations analysed in the corpus study is given.

Altogether, 988 adjective-noun, verb-noun and phrasal-verb noun-collocations were found that were connected with the ten nouns selected for this study (*language*, *standard*, *people*, *country, world*, *consequence*, *culture, word, rule*, *problem*). On average, a participant produced five adjective-noun, two verb-noun and less than one phrasal-verb-noun collocation in an essay. At least two adjective-noun collocations occurred in each essay while nine participants did not produce any verb-noun or phrasal-verb-noun collocations at all. The largest number of collocations in an essay was 23.

2.3.1 The most frequent collocations produced

Since the essays were all written on the same topic – the future of the English language, it was not a surprise that most frequently used nouns and their collocates all revolved around the given topic. Not surprisingly, some word combinations presented in the source text and the task description also appeared in the essays. Table 3 presents the list of the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations found in the essays. Collocations highlighted in grey are those that were given in the reading article and the task description.

Table 3. The most frequent adjective-noun and verb-noun collocations.

Rank	Adjective-noun collocations		Verb-noun or phrasal-verb-noun collocations	
1	new standard	192	have a language	34
2	positive consequence	49	speak a language	29
3	negative consequence	48	learn a language	25
4	native language	41	use a language	17
5	other language	41	have consequences	16
6	different country	19	know a language	9
7	different language	17	different cultures	8
8	smaller language	15	communicate with people	7
9	official language	13	teach a language	7
10	foreign language	12	have a standard	6
11	international standard	11	have rules	6
12	new word	11	forget a language	5
13	common language	10	bring consequences	4
14	new language	10	lose a language	4
15	other country	10	make languages	4
16	own language	9	communicate with countries	3
17	different culture	8	corrupt a language	3
18	foreign country	8	help people	3
19	small country	8	hold on to languages	3
20	business language	6	keep languages	3
21	beautiful language	5	solve problems	3
22	local language	5	travel to a country	3
23	main language	5		
24	original language	5		

The most frequently occurring adjective-noun collocations were *new standard* (192 occurrences), *positive consequences* (49), *negative consequences* (48), *native language* (41), *other language* (41). The most frequent verb-noun or phrasal-verb-noun collocations found were *have a language* (34), *speak a language* (29), *learn a language* (25), *use a language* (17), *have consequences* (16). The most frequent phrasal-verb-noun collocations were *communicate with people* (7), *communicate with countries* (3) and *hold on to languages* (3). Although it was suspected that word combinations related to the topic of the essay might occur often, the most surprising aspect of the result was that *new standard* stood out so strikingly. It was used almost five times more frequently (192 occurrences) than the rest of the collocations, followed by *positive consequences* (49) and *negative consequences* (48). On the other hand, *new standard* and *positive/negative consequences* were all word combinations presented in the task description, which influenced the participants to use these collocations more often.

Another interesting aspect is that nearly all the most frequent collocations were produced with the noun *language.* It might have been suspected, as *language* was also the most frequently used noun in the corpus, but it was still interesting to observe that it was a common node word in both adjective-noun and verb-noun collocations. On the other hand, collocations with the noun *world* were produced less often since not a single word combination listed in Table 3 was formed with *world.* Phrasal-verb-noun collocations tend to be used less often, since nearly all most frequent word combinations consisting of a verb are verb-noun collocations.

Table 4. The most common adjectives, verbs or phrasal verbs found in collocations.

Rank	Adjectives		Verbs or phrasal verbs	
1	new	204	have	66
2	different	57	speak	29
3	other	53	learn	24
4	international	51	use	18
5	positive	49	know	14
6	negative	48	communicate with	10
7	native	41	teach	9
8	smaller	22	forget	6
9	foreign	20	lose	6
10	official	13	make	5
11	common	10	bring	4
12	own	9	take over	4
13	main	7	affect	3
14	business	6	become	3
15	mother	6	corrupt	3
16	national	6	develop	3
17	beautiful	5	follow	3
18	local	5	help	3
19	old	5	hold on to	3
20	original	5	keep	3

Table 4 shows the list of the most frequent collocates the nouns were surrounded by. As the word combination *new standard* was produced so often, it was not a surprise to find the adjective *new* as the most common adjective as well. The adjectives *positive* and *negative* had only been used together with the noun *consequence*, as the number of their occurrences match with the number of collocations presented in Table 3. Among verbs, the most common collocate is *have,* which is not a surprise as it is an auxiliary verb and belongs to the group of the most frequently used verbs in the English language. The verbs coming next (*speak*, *learn*, *use*, *know*) might indicate the connection of the node *language*. Again, the number of occurrences related to phrasal verbs is low: only three phrasal verbs make it to the list of twenty most common verbs that collocate with the nouns selected for investigation.

2.3.2 Distribution of collocations based on collocate-node relationship and congruency

As Table 5 shows, there were 672 adjective-noun collocations, 265 verb-noun collocations and 51 phrasal-verb-noun collocations extracted from the essays which means that the most frequently used word combination in the corpus was an adjective-noun collocation. Word combinations formed with verbs were produced less often, as verb-noun collocations comprised 26% and phrasal-verb-noun collocations only 5% of all combinations found.

Table 5. Collocations divided by collocate-node relationship.

Collocate-node relationship	Number of occurrences	Percentage
Adj+N	672	68%
V+N	265	26.8%
Phr+V+N	51	5.2%
Total	*988*	*100%*

One possible explanation why adjective-noun collocations were produced more often is that sometimes participants had used more than one adjective before a noun (such as *new* *official language*) and in these cases, the word combinations were split into two different adjective-noun collocations: *new language* and *official language*. On the other hand, even if many adjective-noun combinations were divided into multiple different collocations, the number of adjective-noun collocations was still remarkably high, comprising more than two-thirds of all collocations found, which indicates that adjective-noun collocations tend to occur more often than verb or phrasal-verb-noun collocations.

Table 6 presents the distribution of collocations based on congruency, i.e. whether the word combinations found could have been translated into Estonian or not. The results show the tendency to produce more congruent collocations and less non-congruent collocations.

Table 6. Collocations divided by congruency.

Collocate-node relationship	**Congruent collocations**	**Non-congruent collocations**	**Percentages**
Adj+N	619	53	92%/ 8%
V+N	247	18	93%/ 7%
Phr+V+N	36	15	71%/ 29%
Total	*902*	*86*	*91%/ 9%*

Altogether, the proportions of all word combinations were accordingly 91% of congruent collocations and 9% of non-congruent collocations. When taking into account the collocate-node relationship as well, the proportions generally remain the same. A slight change of proportions is present only among phrasal-verb-noun collocations where 29% of all phrasal-verb-noun collocations were non-congruent. In general, it can be said that congruent collocations tend to be more dominant in learners’ language use than non-congruent collocations. Possible reasons and explanations why non-congruent collocations occurred so rarely across the corpus will be addressed in the discussion section. A more detailed table of how congruent and non-congruent collocations distributed in terms of specific nouns is provided in Appendix 2.

2.3.3 The naturalness of collocations produced

Some examples of different adjective-noun, verb-noun and phrasal-verb-noun collocations found in the corpus are presented next. Most adjectives, verbs and phrasal verbs that collocated with the most frequent nouns in the corpus were considered acceptable in English, as only 22 word combinations were classified as unnatural.

For example, the noun *language* usually collocated with ‘official’, ‘foreign’, ‘common’ and ‘universal’. In terms of verbs or phrasal verbs, collocations were most commonly produced with ‘have’, ‘learn’, ‘speak’, ‘use’ or ‘know’. Only the adjectives ‘small’ and ‘strong’ were judged unnatural, since *small languages* and *strong language* were not found in the word corpus in the same context as in the essays. Instead of writing *small languages,* a word combination such as *minority languages* would have been a better choice. *Powerful language* would have been suited better instead of s*trong language,* since the latter is usually referred to as offensive or rude speech in English.

The noun *problem* most commonly collocated with ‘big’ or ‘biggest’ or ‘solve’ and ‘cause’, which are all acceptable combinations in English. An adjective-noun collocation *understanding problems* (where ‘understanding’ functioned as an adjective in the sentence) was not counted as a natural collocation, since another word choice (such as *comprehension problems*) would have been better in this context. In addition, one participant had used the phrasal verb ‘bring up’ in the meaning of to reveal or expose problems. *Bring out* would have been more natural in this context, since a common meaning of *bring up* in English is related to looking after a child until it is an adult.

Acceptable and most common collocates of *rule* were ‘new’, ‘different’, ‘have’, ‘follow’, and ‘know’. However, queries in BNC and COCA corpora did not show that the verb ‘study’ is a natural collocate for *rules.* Examples of *teach with rules* was also not found in the corpora and therefore regarded as an unnatural word combination. The most common adjective that collocated with *word* was ‘new’, other adjectives (‘archaic’, ‘meaningless’, ‘old’, ‘specific’, etc.) were used only once. All of them were categorised as acceptable except for ‘error-fulled’ which was not found in any of the dictionaries nor the word corpora. In terms of verbs that collocated with *word* there were few instances of *to loan words* and *to lend words* where the verb ‘borrow’ would have been more suitable in the contexts of the sentences they occurred. The noun *standard* collocated the most with ‘new’, ‘international’, ‘different’ and ‘have’. It was once used together with a verb *make* but according to the dictionaries and word corpora, a combination *to make a standard* was not found. Dictionaries suggested other verbs such as ‘set’ or ‘establish’ for *standard* in the given context.

In addition to the unnatural combinations found in the essays, a number of collocations contained other types of deviations as well such as mistakes in spelling, articles or singular and plural forms. Some examples of them are presented below:

- ***effect*** *countries* (correct: *affect countries)*
- ***techical*** *world* (correct: technical world)
- ***concider*** *standards* (correct: consider standards)
- ***trovel*** *countries* (correct: *travel countries)*
- ***grammer*** *rule* (correct: *grammar rule*)
- ***communycate*** *with people* (correct: *communicate with people*)
- *have* ***a*** *new* **standards** (correct: *have a standard* or *have new standards*)
- *travel in* ***a*** *different* **countries** (correct: *travel in different* countries/*travel in a different country*)
- ***world wide*** *problem* (correct: *worldwide problem*)
- ***singel*** *language* (correct: *single language*)

The examples of unnatural or misspelt collocations presented above show why a certain degree of manual work is needed in a corpus study because they can confuse automatic parts-of-speech taggers and cause unreliable results during the extraction process. This is especially crucial in studies that use a small corpus because data misinterpretation caused by automatic extraction can tremendously influence the nature of the results. Manual extraction is more time-consuming but still an effective way to discover deviant word combinations.

After having described the main results, the final part of the thesis continues with discussing the key findings of the study. It will analyse the results in more detail and offer possible reasons for the findings by comparing the results with previous literature and research. It will also explain the limitations of the study and what important issues should be investigated in future research.

DISCUSSION AND CONCLUSION

This MA thesis aimed to investigate an important aspect of language learning – the use of collocations in Estonian EFL learners' writing. Since research on learner corpus studies has revealed that collocations of different types can pose problems for language learners and there was very little known of Estonian EFL learners' ability to produce collocations, there was a considerable need to investigate this topic. The present study examined the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations produced in the Estonian learner corpus of English.

The study began by reviewing the literature in Chapter 1. It gave a brief overview of the definitions of the term *collocation*, describing different approaches and contexts the term has been used in. Then the role of learner corpus in language teaching was explained and the last section of Chapter 1 addressed previous studies conducted on collocations in learner corpus research, describing the main approaches and findings revealed in the papers. Chapter 2 gave an overview of the empirical study. It began by describing the methods adopted, the learner corpus studied and the procedure of collecting and analysing the target collocations. The last section of Chapter 2 presented the results of the study. The next section provides answers to the research questions which were the following: 1) what are the most frequent adjective-noun, verb-noun and phrasal-verb-noun collocations used in the learner corpus; 2) what is the distribution of collocations found in the learner corpus based on the collocate-node relationship and congruency; 3) to what extent are the collocations produced natural in the English language.

The first research question sought to identify the most frequently used adjective-noun, verb-noun and phrasal-verb-noun collocations that were formed with the ten nouns selected for this study: *language, standard, people, country, world, consequence, culture, word, rule, problem*. Altogether, 988 adjective-noun, verb-noun and phrasal-verb-noun collocations were found. The most frequently occurring adjective-noun collocations were *new standard, positive consequences, negative consequences, native language, other language*. The most frequent verb-noun or phrasal-verb-noun collocations found were *have a language, speak a language, learn a language, use a language* and *have consequences.* The most frequent phrasal-verb-noun collocations were *communicate with people, communicate with countries* and *hold on to languages*. Since this study was not comparative in its nature and therefore the frequencies of the most common collocations produced in the learner corpus were not compared with a native corpus, the study does not offer evidence for any learner under- or overuse. However, the study can still provide some information which adjective-, verb- or verb-noun-collocations are used the most frequently when writing an essay on English languages.

As the essays were all written on the same topic, it was assumed that the use of vocabulary would be quite similar. Not surprisingly, all of the collocations found were connected with the topic the participants had to write about – the future of the English language. Some collocations were even the same that were given in the source text or task description, many of them used more than five times more frequently than the next most common collocations (such as *new standard*). On the other hand, there is a possibility that the participants could have used similar word combinations with or without the help of the source text. Even if the number of word combinations that were also given in the source text or task description was high (*new standard* with 192 occurrences, *positive* and *negative consequences* with 49 and 48 occurrences, etc.), when taking into account the number of different collocations produced, only few most frequent collocations were provided in the examination materials: seven adjective-noun collocations out of 24 most frequent Adj+N collocations and only one verb-noun collocation (*have consequences*) out of 22 V+N collocations were given in the source text or task description (see the highlighted word combinations in Table 3 from section 2.3.1).

The second research question tried to find out how the collocations distributed in the learner corpus based on collocate-node relationship and congruency. The study revealed that by far the most frequently used collocation type was an adjective-noun collocation (68%), verb-noun collocations counted for 26.8% and phrasal-verb-noun collocations only 5.2% of all target collocations found. Even if taking into account those nouns that collocated with more than one adjective, the number of adjective-noun collocations was still strikingly high, comprising more than two-thirds of all collocations found across the corpus. Evidence that adjective-noun collocations tend to seem easy for language learners was shown in Peters (2016): in each test where the participants had to recall different types of collocations, the highest scores were found for the adjective-noun collocations. One reason why phrasal-verb-noun collocations were not used so frequently can be that the participants tend to avoid using phrasal verbs and resort to one-word verbs, which was demonstrated in Schmitt and Siyanova (2007). Their study offered another explanation as well, that phrasal verbs are often not transparent in meaning and therefore, are usually avoided in language production.

The general distribution of collocations based on congruency showed that the majority (91%) of all word combinations were congruent and could be directly translated from Estonian. This finding supported the results from other studies that language learners tend to use congruent collocations, rather than non-congruent collocations, as Granger (2012: 141) and Henriksen (2013: 36) both declare that language learners tend to rely heavily on congruent collocations. In Peters (2016), congruent collocations were also more likely to be produced in recall tests than non-congruent ones, no matter what the collocate-node relationship was. However, it should be mentioned that drawing a line between a congruent and non-congruent collocation was sometimes difficult, even with the help of the Estonian etTenTen (2013) corpus. In these cases, a decision was made based on the author’s knowledge of collocational naturalness in the Estonian language. There is a possibility that some word combinations classified as non-congruent in this study may be classified as congruent in others. However, since of all the collocations produced, 91% were congruent and only 9% non-congruent, the general proportions would probably still remain the same even with slight differences in percentages.

The last research question addressed the naturalness of the collocations produced in the corpus. From a pedagogical perspective, it was encouraging to see that the majority of adjectives, verbs and phrasal verbs which collocated with the most frequently used nouns was considered natural in English. Only 22 collocations out of 988 word combinations were classified as unnatural and almost half of the unnatural collocations were related to spelling or article mistakes. Since the number of unnatural collocations was so low, any general conclusions could not be made, only some examples of the deviant collocations found were given and explained. It has to be noted that the participants wrote essays in an examination situation and therefore, the learners were probably under pressure and in some cases maybe even had to finish their writing in a hurry. This may have been one reason for spelling mistakes such as *communycate with people* or *trovel countries*. On the other hand, this is only a presumption which cannot be supported with evidence because, as Nesselhauf (2004: 130) reminds us, written learner corpora cannot be used to answer questions such as how certain the participants of the study were about the correctness of their language. This notion leads the discussion to different limitations of this study, which are addressed next.

First of all, the size of the corpus investigated was small. For example, whereas the learner corpus in Nesselhauf (2005) comprised of 318 essays and 154,191 words, this learner corpus consisted only of 127 essays and 24,457 words. In addition, all essays were written on a single topic which influenced the use of vocabulary. Therefore, the learner corpus was not representative enough to claim any conclusive findings about all Estonian EFL learners’ use of collocations. Another limitation is that since only the most commonly occurring nouns were chosen for investigation, the study did not examine all nouns and their collocates found in the corpus. This means that many word combinations were left out of the picture and the study took the risk of overlooking some probable and important aspects of collocations produced. Furthermore, a major limitation of written corpus studies is that it can only be used to examine productive skills without any information what other words the participants can actually produce. Therefore, even if this study revealed that non-congruent collocations were used sparsely, it does not automatically mean that the learners could not produce them if they were asked to.

The study also did not follow common methodological approaches adopted in similar researches, such as comparing a learner corpus with native corpora or making distributions of collocations according to the level of restriction involved. Since the findings of previous studies have revealed that native speakers are able to produce more natural and varied collocations, a comparative study of native and non-native speaker corpora was not conducted in order to avoid finding answers that are already somewhat predictable. For the same reasons, the collocations were not categorised based on their restrictedness because free collocations are shown to be produced more often than restricted collocations in previous learner corpus studies (Nesselhauf 2005, Vuorinen 2013).

To remedy these limitations, additional research is needed to develop a full picture of Estonian EFL learners’ use of collocations. In future investigations, it might be possible to conduct comparative studies by comparing an Estonian learner corpus with a comparable native corpus or to investigate other types of collocations such as adjective-adverb collocations. In terms of verb-noun collocations, the focus can be, for example, on exploring collocations formed with high-frequency verbs (*make, do, have, take, give*) considering that *do* and *have* are primary verbs with the highest frequencies in English, whereas *make, take* and *give* are lexical verbs that display the average cross-corpus frequency (Kaszubski 2000:

2). Other and very interesting approaches are, for example, to examine the effectiveness of different teaching methods, measure the learning burden of learning English collocations or to test receptive and/or productive knowledge of collocations by adopting testing tools developed by Gyllstad (2007) and Revier (2009).

In spite of the limitations listed, the present study is still significant in at least two major respects. First, it explored for the first time the use of different types of collocations in Estonian learner corpus of English, analysing real language production in a certain context. Second, the results of the study gave a general overview of the target collocations: 1) adjective-noun collocations tend to be produced more often than verb- and phrasal-verb-noun collocations; 2) a majority of the collocations produced are congruent rather than non-congruent; 3) the collocations produced are mostly natural in the English language. The findings also reaffirm important implications for language teaching that Nation (2001: 328) has outlined: the study provided information about the most frequent collocations and collocations of the most frequently used nouns, and it also identified the most common patterns of collocations. The results revealed that verb-noun and phrasal-verb-noun collocations were used sparsely and therefore might be paid more attention to in the classroom compared to adjective-noun collocations. In addition, it would be helpful if teachers of English encouraged their students to use more non-congruent collocations in order to improve the level of language production.

Taking into account the small size of the learner corpus and the fact that the essays were written on a single topic, the corpus was not representative to confirm any general findings about all Estonian EFL learners’ use of collocations. On the other hand, the study still could fill an important gap in learner corpus research on collocations. It was the first study to explore the use of different types of collocations in Estonian learner corpus of English, providing language teachers and future research with useful information about the most frequently used adjective-noun, verb-noun and phrasal-verb-noun collocations and their distributions in terms of collocate-node relationship and congruency. The study also collected new information about the Estonian learner corpus of English, which has been previously investigated only on two topics: adjectives and adverbs (Daniel 2015) and conjunctive adjuncts (Merilaine 2015). Hopefully, this study was only the first step in examining collocational use in Estonian learner corpus of English and the interest of investigating this matter even further will grow in the future.

REFERENCES

Adolphs, Svenja. 2006. *Introducing Elextronic Texy Analysis. A Practical Guide For Language And Literary Studies.* New York: Taylor & Francis e-Library.

Allami, Hamid and Elahe Movahediyan Attar. 2013. The effects of teaching lexical collocations on speaking ability of Iranian EFL learners. *Theory and Pracice in Language Studies*, 3: 6, 1070-1079.

Anthony, Laurence. 2014. AntConc (Version 3.4.4w). Computer Software. Waseda University, Tokyo, Japan. Available at http://www.laurenceanthony.net/, downloaded February 10, 2016.

Anthony, Laurence. 2015. TagAnt (Build 1.2.0). Computer Software. Available at at http://www.laurenceanthony.net/, downloaded April 2, 2016.

Azizinezhad, Masoud, Masoud Hashemi and Sohrab Dravishi. 2012. Collocation a neglected aspect in teaching and learning EFL. *Procedia – Social and Behavioral Sciences*, 12: 31, 522-525.

Belozerskaja, Aleksandra. 2013. *The Concept of Estonia in Schools with Russian Teaching Language Through Collocations in Estonian Language Textbooks.* Unpublished MA thesis. Narva College of University of Tartu, Narva, Estonia.

Boers, Frank, Averil Coxhead, Murielle Demecheleer, Stuart Webb. 2014. Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, 18: 1, 54-74.

Brashi, Abbas. 2009. Collacability as a problem in L2 production. *Reflections on English Language Teaching*, 8: 1, 21-34.

British National Corpus. Available at https://www.sketchengine.co.uk/, accessed April 1, 2016.

Cao, Feng and Huaqing Hong. 2014. Interactional metadiscourse in young EFL learner writing: a corpus-based study. *International Journal of Corpus Linguistics*, 19: 2, 201-224.

Carter, Ronald, Michael McCarthy and Anne O'Keefe. 2007. *From Corpus to Classroom: Language Use And Language Teaching.* New York: Cambridge University Press.

Cotos, Elena. 2014. Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26: 2, 202-224.

Cowie, Anthony P. 1998. *Phraseology: Theory, Analysis and Applications.* Oxford: Clarendon Press.

Daniel, Anna. 2015. *The Use of Adjectives and Adverbs in Estonian and British Student Writing: A Corpus Comparison*. Unpublished MA thesis. University of Tartu, Tartu, Estonia.

Dumont, Amandine and Sylviane Granger. Learner corpora around the world. Available at https://www.uclouvain.be/en-cecl-lcworld.html, accessed April 16, 2016.

Durrant, Philip and Julie Mathews-Aydinlıi. 2011. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes,* 30: 1, 58-72.

Durrant, Phillip. 2008. *High-frequency Collocations and Second Language Learning*. Unpublished PhD thesis, University of Nottingham, Nottingham.

Durrant, Philip. 2014. Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics* 19:4, 443–477.

etTenTen Corpus. 2013. Available at https://www.sketchengine.co.uk/, accessed April 2, 2016.

Fan, May. 2009. An exploratory study of collocational use by ESL students – A task based approach. *System,* 37: 1, 110-123.

Firth, John Rupert. 1957. *Papers in linguistics, 1934-1951*. Oxford: Oxford University Press.

Foster, Pauline. 2001. Rules and routines: A consideration of their role in the task based language production of native and non-native speakers. In Martin Bygate, Peter Skehan and Merrill Swain (eds). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*, 75-93. Harlow. Longman.

Granger, Sylviane and Magali Paquot. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32: 130-149.

Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In Sylviane Granger, J. Hung and S. Petch-Tyson (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3-33. Amsterdam/Philadelphia: Benjamins.

Granger, Sylviane. 2003. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37: 3, 538-546.

Granger, Sylviane. 2004. Computer learner corpus research: current status and future prospects. *Language and Computers*, 52: 1, 23-145.

Gyllstad, Henrik and Brend Wolter. 2011. Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32: 4, 430-449.

Gyllstad, Henrik and Brent Wolter. 2013. Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, 35: 3, 451-482.

Gyllstad, Henrik. 2007. *Testing English Collocations. Developing Receptive Tests for Use with Advanced Swedish Learners*. Lund: Medya-Tryck.

Hardie, Andrew and Tony McEnery. 2011. *Corpus Linguistics. Method, Theory and Practice.* Cambridge: Cambridge University Press.

Hasko, Victoria. 2013. Capturing the dynamics of second language development via learner corpus research: a very long engagement. *Modern Language Journal*, 97: S1, 1-10.

Heinsoo, Heinike. 2010. Adjektiivide tajumine ja õpetamine. *Lähivõrdlusi*. *Lähivertailuja*: 19, 120-136.

Henriksen, Birgit. 2013. Research on L2 learners' collocational competence and development – a progress report. In Camilla Bardel, Christina Lindqvist and Batia Laufer (eds). 2013. *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, 29-56. Eurosla Monographs Series, 2.

Hill, Jimmie. 2000. Revising priorities: from grammatical failure to collocational success. In Michael Lewis (ed). *Teaching Collocation: Further Developments in the Lexical Approach*, 47-69. Hove: Language Teaching Publications.

Jaanits, Kadri. 2004. *Leksikaalsetest kollokatsioonidest soome ja eesti keeles.* Unpublished MA thesis. University of Tartu, Tartu, Estonia.

Juknevičienė, Rita. *2008. Collocations with high-frequency verbs in learner English: Lithuanian learners vs native speakers. Kalbotyra, 59: 3, 119-127.*

Kaszubski, Przemyslaw. 2000. *Selected Aspects of Lexicon, Phraseology and Style in the Writing of Polish Advanced Learners of English: A Contrastive, Coprus-Based Approach.* PhD thesis. Poznàn: Adam Mickiewicz University. Available at http://www.staff.amu.edu.pl/~przemka/rsearch.html#PhD, accessed April 28, 2016.

Kirs, Kersti. 2001. *Collocation in English.* Unpublished diploma thesis. Tartu Teacher Training College, Tartu, Estonia.

Koya, Taeko. 2005. *The Acquisition of Basic Collocations by Japanese Learners of English.* Unpublished doctoral dissertation. Waseda University, Japan. Available at http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/5285/3/Honbun-4160.pdf, accessed 10 April, 2016.

Laufer, Batia and Tina Waldman. 2011. Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning*, 61:2, 647-672.

Lukač, Morana and V. Pavičić Takač. 2013. How word choice matters: An analysis of adjective-noun collocations in a corpus of learner essays. *Jezikoslovlje* 14: 2-3, 385-402.

McCarthy, Michael. 2004. *Touchstone – From Corpus to Course Book*. Cambridge: Cambridge University Press.

McEnery, Tony and Richard Xiao. 2010. What corpora can offer in language teaching and learning. In Ed Hinkel (ed). *Handbook of Research in Second Language Teaching and Learning (Vol. 2)*, 364-380. London/New York: Routledge.

Merilaine, Elina. 2015. *The Frequency and Variability of Conjunctive Adjuncts in the Estonian-English Interlanguage Corpus.* Unpublished MA thesis. University of Tartu, Tartu, Estonia.

Meyer, F. Charles. 2002. Corpus analysis and linguistic theory. In Charles F. Meyer (ed). *English Corpus Linguistics. An Introduction*, 1-29. Cambridge: Cambridge University Press.

Nation, Paul. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24: 2, 223-242.

Nesselhauf, Nadja. 2004. Learner corpora and their potential for language teaching. In John McHardy Sinclair (ed). *How to Use Corpora in Language Teaching*, 125-166. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nesselhauf, Nadja. 2005. Collocations in a Learner Corpus. In Elena Tognini-Bonelli (ed). *Studies in Corpus Linguistics (Book 14)*. Amsterdam: John Benjamins Publishing Company.

Online Oxford Collocation Dictionary of English. Available at http://oxforddictionary.so8848.com/, accessed February 25, 2016.

Oxford English Dictionary. Available at http://www.oed.com/, accessed May 1, 2016.

Oxford Learner's Dictionaries. Available at http://www.oxfordlearnersdictionaries.com/, accessed February 25, 2016.

Palmer, E. Harold. 1931. *First interim report on vocabulary selection*. Tokyo: Kaitakusha.

Palmer, E. Harold. 1933. *Second interim report on English collocations*. Tokyo: Kaitakusha.

Peters, Elke. 2009. Learning collocations through attention-drawing techniques: a qualitative and quantitative analysis. In Andy Barfield and Henrik Gyllstad (eds.). *Researching Collocations in Another Language: Multiple Interpretations*, 194-207. Basingstoke: Palgrave Macmillan UK.

Peters, Elke. 2014. The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research,* 18: 1, 75-94.

Peters, Elke. 2016. The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, 20: 1, 113-138.

Piits, Liisi. *Collocations of the Most Frequent Estonian Words for 'Human Being'*. Unpublished PhD thesis. University of Tartu, Tartu, Estonia.

Põder, Irma. 2010. *Inglise keele riigieksamist 2010*. Available at http://www.ekk.edu.ee/vvfiles/0/REA_inglise%20keel.pdf, accessed February 28, 2016.

Revier, Robert Lee. 2009. Evaluating a new test of *whole* English collocations. In Andy Barfield and Henrik Gyllstad (eds). *Researching Collocations in Another Language: Multiple Interpretations*, 125-138. Basingstoke: Palgrave Macmillan UK.

Saar, Merike. 2005. *Noun Collocations in English and Estonian*. Unpublished MA thesis. EuroAcademy, Tallinn, Estonia.

Schmitt, Norbert and Anna Siyanova. 2007. Native and non-native use of multi-word vs one-word verbs. *IRAL*, 45:2, 119-139.

Schmitt, Norbert and Anna Siyanova. 2008. L2 learner production and processing of collocation: a multi-study perspective. *The Canadian Modern Language Review*, 64: 3, 429-458.

Sinclair, John. 1991. *Corpus. Concordance.Collocation.* Oxford: Oxford University Press.

Sinclair, John. 1996. Preliminary recommendations on corpus typology. *Eagles*. Available at http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html, accessed February 6, 2016.

The Corpus of Contemporary American English. Available at http://corpus.byu.edu/coca/, accessed April 1, 2016.

Timofejeva, Veronika. 2015. *The Use of Estonian Collocations by Russian-Speaking Learners*. Unpublished MA thesis. University of Tartu, Tartu, Estonia.

Vuorinen, Anni. 2013. *Finnish Foreign Language Learners' Use of English Collocations.* Unpublished MA thesis. University of Jyväskylä, Jyväskylä.

Wray, Alison. 2002. *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Appendix 1. Entrance Examination 2014: Task Description and the Source Text

Read the passage from Guy Cook's *Applied Linguistics* (2008, pp. 26–28) and complete the tasks that follow. Your answers should be logically structured and use appropriately academic and grammatically correct English.

English and Englishes

Whereas, in the past, English was but one international language among others, it is now increasingly in a category of its own. In addition to its four hundred million or so first-language speakers, and over a billion people who live in a country where it is an official language, English is now taught as the main foreign language in virtually every country, and used for business, education, and access to information by a substantial proportion of the world's population.

This growth of English, however, also has some paradoxical consequences. Far from automatically extending the authority of English native speakers, it raises considerable doubts about whose language English is, and how judgements about it can be made. It may even – as we shall see shortly – make us reconsider not only our definition of 'English native speaker', but also whether this term is as significant in establishing norms for the language as is usually supposed.

As we observed at the beginning of this chapter, it is usual for speakers of a language, while welcoming the learning of it by others, to feel a sense of ownership towards it. In the case of smaller and less powerful languages, limited to a particular community in a particular place, this is both unexceptional and unremarkable. Once however, a language begins to spread beyond its original homeland and the situation changes and conflicts of opinion begin to emerge. Thus, even until surprisingly recently, many British English speakers regarded American English as an 'impure' deviation, rather as they might have regarded non-standard forms within their own islands. While such feelings of ownership are to be expected, they quickly become untenable when speakers of the 'offspring' variant become, as they are in the USA, more numerous and more internationally powerful than speakers of the 'parent'.

With any language which spreads this backwash effect is inevitable, and the justice of the process seems incontrovertible. There is a similar relationship between South American and Castilian Spanish, and the Portugueses of Brazil and Portugal. Yet despite the inevitability of this process, there is still possessiveness and attempts to call a halt. Few people nowadays would question the legitimacy of different standard Englishes for countries where it is the majority language. We talk of standard American English, standard Australian English, standard New Zealand English, and so on. Still contested by some, however, is the validity of standards for countries where, although English may be a substantial or official language, it is not that of the majority. Thus there is still opposition, even within the countries themselves, to the notion of Indian English, Singapore English, or Nigerian English. Far more contentious, however, is the possibility that, as English becomes more and more widely used, recognized varieties might emerge even in places where there is no national 'native- speaker' population or official status. Could we, in the future, be talking about Dutch English, or Chinese English, or Mexican English?

The Indian scholar Braj Kachru describes this situation as one in which English exists in three concentric circles: the inner circle of the predominantly English-speaking countries; the outer circle of the former colonies where English is an official language; and the expanding circle where, although English is neither an official nor a former colonial language, it is increasingly part of many people's daily lives. At issue is the degree to which the English in each of these circles can provide legitimate descriptions and prescriptions. The rights of the outer circle are now reasonably well established. What, though, of the English used in the expanding circle? Could a new standard international English be emerging there, with its own rules and regularities, different from those of any of the 'native Englishes'?

3. According to Cook, it is likely that a new standard of international English will emerge. What might be some of the consequences (both positive and negative) of this for English as well as other languages? Provide reasons for your opinion. (Write an answer of approximately 200 words on your answer sheet.)

Appendix 2. Distribution of collocations based on congruency

	Collocate-node relationship	Congruent collocations	Non-congruent collocations	Total
Language	Adj+N	181	48	229
	V+N	147	9	156
	Phr+V+N	2	0	2
Standard	Adj+N	212	1	213
	V+N	14	0	14
	Phr+V+N	2	2	4
Country	Adj+N	55	1	56
	V+N	7	0	7
	Phr+V+N	16	1	17
World	Adj+N	6	1	7
	V+N	7	0	7
	Phr+V+N	6	4	10
Consequence	Adj+N	109	1	110
	V+N	20	1	21
	Phr+V+N	1	4	5
Culture	Adj+N	14	1	15
	V+N	9	0	9
	Phr+V+N	5	3	8
Word	Adj+N	18	0	18
	V+N	17	5	22
	Phr+V+N	0	0	0
Rule	Adj+N	13	0	13
	V+N	15	2	17
	Phr+V+N	3	1	4
Problem	Adj+N	11	0	11
	V+N	11	1	12
	Phr+V+N	1	0	1

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Lenne Tammiste

The use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian learner corpus of English / Omadussõna-nimisõna, tegusõna-nimisõna ja ühendverb-nimisõna kollokatsioonide kasutus inglise õppijakeele korpuses

Magistritöö

2016

Leheküldede arv: 54

Käesoleva magistritöö eesmärk on kirjeldada omadussõna-nimisõna, tegusõna-nimisõna ja ühendverb-nimisõna kollokatsioonide kasutust inglise-eesti õppijakeele korpuses. Töös uuritakse, millised on enim kasutusel olevad kollokatsioonid, kuidas kollokatsioonide sagedused jagunevad lähtudes kollokatsioonitüübist ja otsetõlke võimalusest eesti keelde, ning kuivõrd loomulikud on kasutatud kollokatsioonid inglise keeles.

Magistritöö teooriapeatükis antakse ülevaade kollokatsiooni mõiste erinevatest definitsioonidest, õppijakeele korpuse kasutusvõimalustest keeleõppes ning hilisematest kollokatsioonidele keskenduvatest korpusuurimustest. Töö empiirilises osas kirjeldatakse inglise õppijakeele korpuse loomise etappe ning metodoloogiat, kuidas vastavad kollokatsioonid keelekorpusest välja otsiti ja ning jaotati. Lisaks antakse ülevaade, kuidas otsustati kollokatsioonide loomulikkuse üle inglise keeles.

Tulemustes kirjeldatakse, millised olid enimlevinumad omadussõna-nimisõna, tegusõna-nimisõna ja ühendverb-nimisõnade kollokatsioonid inglise õppijakeele korpuses. Samuti selgub, et kõige sagedamini kasutati omadussõna-nimisõna kollokatsioone ning sõnaühendeid, mida oli võimalik sõnasõnalt eesti keelde ümber tõlkida. Vaid väikest osa kollokatsioonidest peeti ebaloomulikuks või leidus kollokatsioonides kirja- ja artiklivigu. Enamik kasutatud kollokatsioonidest peegeldasid siiski loomulikku keelekasutust inglise keeles. Käesoleva magistritöö peamiseks tugevuseks on selle uudsus – tegemist on esimese korpusuurimusega Eestis, mille fookuseks on kollokatsioonide kasutamine inglise-eesti õppijakeeles.

Märksõnad: kollokatsioonid, korpuslingvistika, õppijakeel, õppijakorpused

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Lenne Tammiste,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

The use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian learner corpus of English,

mille juhendaja on Pille Põiklik,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 20. mail 2016