UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

THE FREQUENCY AND VARIABILITY OF CONJUNCTIVE ADJUNCTS IN THE ESTONIAN–ENGLISH INTERLANGUAGE CORPUS

MA Thesis

ELINA MERILAINE
SUPERVISOR: PILLE PÕIKLIK, PhD

TARTU
2015

ABSTRACT

Learner corpus research (LCR) is becoming increasingly popular in linguistic enquiry because it provides insights into learners' authentic language use and can be applied to various aspects of language teaching and learning. In Estonia, corpus-based studies are relatively new and Estonian–English learner corpus has not yet been assembled.

Writing plays a significant role in language learning and improving students' writing skills is one of the basic goals for every teacher of foreign language. The quality of the written text depends on many aspects and one of them is cohesion. Textual cohesion is an essential criterion in text organization. The use of conjunctive adjuncts as cohesive devices plays the integral role in building connections between ideas in text and ensures that the content is comprehensible for the reader.

Based on the explorations of corpus compilation principles, the present thesis provides the Estonian–English Interlanguage Corpus (EEIC) based on the written part of the entrance examination of the Department of English Studies at the University of Tartu and introduces and applies the central techniques in corpus research – the creation of word lists and concordance indexes for quantitative linguistic analyses.

The empirical study of frequency and variability of conjunctive adjuncts from the perspective of Estonian EFL learners' writing has not yet been performed and the current thesis aims at filling the gap. The paper aims to demonstrate how corpus methods can contribute to linguistic research and presents quantitative overview of conjunctive adjuncts that are used by Estonian EFL learners in the compiled corpus.

The Estonian–English Interlanguage Corpus compiled for current thesis as well as the given study will contribute to future research of learner language by providing researchers and teachers with a substantial base for future research – a learner corpus with controlled computerized data, which can be analysed at a range of levels using different software tools.

TABLE OF CONTENTS

1. INTRODUCTION

Despite a relatively short period of existence since the 1980s, the use of computer learner corpus (CLC) has become widely used in linguistic analysis. CLC can be defined as a computerized collection of texts, either written or spoken, that is stored on a computer and used as a sample of the language. Corpus can be approached manually or through specially designed software and the content of texts can be analysed by quantitative and qualitative methods (O'Keeffe et al. 2007: 3–4). The principal research aim for learner corpus is to observe and describe language use of learners. According to Flowerdew (2012: 3), the areas of investigation of learner corpus are manifold, most commonly the purposes for corpus compilation are linguistic, but often these can also be of socio–pragmatic nature. For example, it is possible to collect and compare evidence on learners' language competence and errors, define whether the errors are universal, language or learner–group specific; investigate, observe and describe the overuse or underuse of words, determine whether or to what extent are findings affected by learners' mother–tongue or factors in cultural or educational background (Pravec 2002: 81–83).

According to Geoffrey Leech (1992: 106), a pioneer of corpus linguistics development, CLC forms a distinct discipline – "new research enterprise, /…/ a new philosophical approach to the subject, /…/ an 'open sesame' to a new way of thinking about language". Bowker and Pearson (2002: 9) have given corpus a more specialized meaning and stated that corpus linguistics is an empirical approach or a methodology for studying examples of actual language use. The problem of defining corpus linguistics can be debated from different standpoints, however, as Granger et al. (2002: 4) has stated, "the power of computer software tools combined with the impressive amount and diversity of the language data used as evidence has revealed and will continue to reveal previously unsuspected linguistic phenomena." It is therefore possible to propose that although CLC

does not form a new branch of linguistics or a new theory of language, it provides methodological basis and evidence that has the potential to change perspectives on language (Granger et al. 2002: 4; also see Granger 2012: 1). With the help of statistical operations that computer is able to carry out, it is possible to process a large amount of information instantly and accurately. For example, with the help of the machine–readable corpus, calculation of collocations can be effectively performed – it is possible to examine how words co–occur, what kind of lexical collocates are primarily used and create indexes of the most widely used collocations. Such statistical manipulations of language data would be difficult and time–consuming if not impossible to perform if dealt not electronically, but with a printed matter.

Corpus linguistics belongs to the sphere of applied linguistics, the branch of linguistics that is concerned with the practical applications of language studies, such as language teaching, translation or lexicography. Thus, one of the distinctive and remarkable features of corpus linguistics is that it establishes a possible point of contact between the specialists from various fields of research. The concept of using corpus technology in linguistics emerged with the compilation of the first computerized native–language corpus in the 1960s – The Brown Corpus of Present–Day American English. Among European corpora, the first corpus of British English was launched in 1970 – LOB Corpus (The Lancaster–Oslo/Bergen Corpus). After that, English corpora grew, diversified and already in the 1990s, the first attempts in collecting non–native varieties, more specifically varieties of non–native English learners were made (Granger 1998: 3–4; also see Granger et al. 2002: 5). These corpora became to be referred to as learner corpora and were particularly assembled for description of learners' language – interlanguage.

The notion of interlanguage was initially proposed by Larry Selinker (1972), who claimed that interlanguage is a language system used by the L2 learners, which is

influenced by their L1. Interlanguage is neither the system of the native language nor that of the target language; instead it forms a transitional state from L1 to the L2 that is evident during the process of the second language acquisition (Song 2012: 778). Interlanguage features are rule-governed and systematic, therefore studying interlanguage variability among learners is necessary for providing theoretical basis and implications for efficient classroom instruction (Song 2012: 781). In this way, storing, processing and investigating language with the help of learner corpus creates an important link between the two previously disparate fields of corpus linguistics and interlanguage research (Granger et al. 2002: 4).

Computerized database of the language, produced by foreign language learners serves as a reliable and representative model of interlanguage and allows focusing on theoretical and pedagogical issues to make assumptions about the needs of learners (Pravec 2002: 81). It has been proposed (Granger 1998, 2002; also O'Keeffe et al. 2007) that the compilation and analysis of learner corpora can be particularly relevant from the pedagogical perspective in regard to teaching/learning materials design or curriculum development, because learner corpora enable to gain insights not only into learners' authentic language use but also to the mechanism of foreign/second language acquisition. Thus, CLC contributes directly not only to foreign/second language (EFL/ESL) research and second language acquisition (SLA) research, but in addition to foreign language teaching (FLT), by helping to create and improve teaching methods and EFL tools – pedagogical materials and learning applications (Granger 2002: 4–6; also see Granger 2012: 2).

The first learner corpus that was compiled in academic setting to make specific learner-language oriented investigations was ICLE (International Corpus of Learner English), launched in the 1990s. ICLE presents the collection of essays from ESL/EFL

learners from different native language backgrounds and provides an empirical resource for large-scale comparative studies in the field of learner language (Pravec 2002: 83; also see Flowerdew 2012: 169–170). Today, in addition to academic corpora, it is also possible to find numerous profit-oriented (commercial) learner corpora. The most popular of them being LLC (Longman Learner Corpus) and CLC (Cambridge Learner Corpus) that aim at creating new practical materials (dictionaries, grammar reference books, workbooks) of different proficiency levels for students and teachers.

Learner corpora thus comprise a relatively new and rapidly growing field of linguistic research. New corpora are assembled worldwide and it is difficult to be fully informed of all the various projects, therefore an exhaustive list of learner corpora and their research objectives is clearly beyond the scope of the current chapter and thesis in general. The major existing learner corpus projects have been presented and discussed in detail in surveys conducted by Norma Pravec (2002) and Yukio Tono (2003) and the most comprehensive, regularly updated list of learner corpora assembled in the world today can be found on the CECL (Centre of English Corpus Linguistics) webpage[1], coordinated by the University of Louvain in Belgium.

In Estonia, the field of learner corpus linguistics is relatively young. Numerous native–language corpora have been assembled and are under the coordination of The Centre of Estonian Language Resources (CELR)[2] that organizes the digital resources of Estonian language, such as digital dictionaries, corpora – both text and speech and various language databases. There are three institutions that belong to the consortium and provide corpus research in Estonia – the University of Tartu, the Institute of Cybernetics at the Tallinn University of Technology and the Institute of the Estonian Language. At the present moment there is only one learner corpus in Estonia that directly aims at studying

[1] Available at http://www.uclouvain.be/en-cecl-lcworld.html (5.03.2015)
[2] Available at http://keeleressursid.ee/en (28.02.2015)
[2] Available at http://keeleressursid.ee/en (28.02.2015)

learner language from the perspective of the learners' needs. The Estonian Interlanguage Corpus (EIC) assembled at the chair of General and Applied linguistics at Tallinn University presents the collection of written texts (state examinations) mainly produced by the Russian learners of Estonian as a second or foreign language (Eslon and Metslang 2007: 105, 116).

The principal tool applied in EIC is the concordance programme that allows searching and extracting various linguistic (lexical or syntactic) occurrences from the learners' texts, arranging the found occurrences (in alphabetical order or frequency lists) and performing error analyses (Eslon and Metslang 2007: 105–106). Statistical error analyses provide researchers with the most common and relevant problems that the learners of Estonian language encounter and these results serve as a basis for writing or revising grammar reference books, textbooks, dictionaries or pedagogic materials for students and teachers (Eslon and Metslang 2007: 116). EIC has not only contributed to the research of Estonian as a second/foreign language, but also allowed performing experiments on automatic approaches for classifying learner essays into proficiency levels (CEFR Level Prediction) that have been conducted by Vajjala and Lõo (2014).

Despite the wealth of corpora that can now be found or is being compiled worldwide, there have yet not been assembled any corpus of Estonian learners from the perspective of English as a foreign language and the current thesis aims at filling the gap. With acquired knowledge about assembling the corpus, the current thesis compiles and provides the Estonian–English Interlanguage Corpus that is utilised to gain greater insight into the use of conjunctive adjuncts in the written essays by Estonian EFL learners.

According to Estonian National Curriculum for Secondary School (2011: Appendix 2, section 3), one of the main aims of foreign language learning is that the students are able to communicate purposefully in the target language, both orally and in writing. By the end

of the secondary school education, students are required to reach the level of B1–B2 according to Common European Framework of Reference for Languages[3]. From the perspective of the writing skill, this level demands students to be able to write coherent argumentative essays, where viewpoints and arguments are overtly explained.

Conjunctive adjuncts are necessary for building connections between ideas in text and are used to link the text semantically and logically (Muddhi and Hussein 2014: 18). Conjunctive adjuncts are most commonly used in academic writing, where the main objective is to present and support explanations and arguments for a wide readership (Biber et al. 2002: 392). Analysing variability and frequency of conjunctive adjuncts in the target learners' writing is highly beneficial to English language teachers for facilitating the teaching methods or improving study materials for EFL learners.

The goal of the current thesis is two-fold – to give an overview of corpus design criteria for assembling the Estonian–English Interlanguage Corpus and to report the results of an experimental quantitative investigation on the use of the conjunctive adjuncts among Estonian EFL learners. The current thesis designs and compiles learner corpus from the written part of the entrance examination (2014) of the Department of English Studies at the University of Tartu and analyses the conjunctive adjuncts that are found in it. Computer-aided analysis (concordance tool) is applied in the study to examine the frequency and variability of the found conjunctive patterns. The current thesis will also use the results adopted from the reference corpus MICUSP (the Michigan Corpus of Upper–level Student Papers) for the comparative aspect in quantitative analysis.

There are three research questions in this study:

1. What kind of conjunctive adjuncts are used in Estonian ESL students' essays?

[3] Available at http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf (20.04.2015)

2. What are the most frequent conjunctive adjuncts in Estonian ESL students' essays?
3. To what extent do results differ from native speakers' (reference) corpus?

The thesis is divided into four sections – describing corpus design criteria, providing the theoretical framework about stages in corpus research, the theoretical background on the notions of coherence and cohesion, quantitative analysis of the frequency and variability of the conjunctive adjuncts in the compiled Estonian–English Interlanguage Corpus as well as comparison of the results with the reference corpus.

2. CORPUS COMPILATION AND DESIGN CRITERIA

The current chapter is devoted to corpus compilation principles and addresses the main issues and decisions to be made when compiling a corpus. The chapter aims to provide the theoretical perspective for building and using a learner corpus for linguistic research and presents the basic corpus typology. The current chapter presents the four principal criteria that are necessary to be applied to corpus design – simplicity, representativeness, quantity and the quality aspects and explains why it important to design corpus carefully. The possible contradictions in the four default corpus design criteria are brought out. In addition, the chapter discusses the compilation of the Estonian–English Interlanguage Corpus that has been used for quantitative analysis in the current thesis.

For the reason that researchers are interested in different aspects of learner language, the design of the learner corpus will naturally vary from project to project (Tono 2003: 800). It is therefore necessary to design a corpus that would gather and present findings that could, apart from the initial study or analysis, be further subjected to scrutiny of other researches (Tono 2003: 801). Leech (1998: 17) has stressed the importance of careful and practical design criteria for corpora and stated: “the creation of corpora demands a great deal of spadework to be done before any research results can be harvested.” Similarly, Tono (2003: 801) has claimed that corpus will be unlikely to be of much use if data is gathered in an opportunistic way without proper control and documentation of learner and task variables. It is therefore of the utmost priority to present and apply clear and exhaustive criteria to corpus design and ensure that each component illustrates the homogeneity of the material before any legitimate conclusions or comparisons are drawn from the study.

2.1. Corpus typology

The fundamental distinction of corpus typology is one of the most open-ended theoretical questions concerning the corpus research. As Bowker and Pearson (2002: 11) have noted, "language is so diverse and dynamic that it would be hard to imagine a single corpus that could be used as a representative sample of all language" and thus, "there are almost as many types of corpora as there are types of investigations." It is difficult to create an exhaustive list of different types of corpora, however, it is possible to identify some broad categories of corpora that can be compiled on the basis of different criteria in order to meet different aims (Bowker and Pearson 2002: 11). For example, Bowker and Pearson (2002:11) and Hunston (2006: 14) have brought out such types as general vs. specialised corpora, parallel (monolingual vs. multilingual) corpora, synchronic vs. diachronic corpora, written vs. spoken corpora. The thesis will only contemplate the differences between general and specialized corpora, because these are directly connected to the compilation of learner corpora in the current thesis and thus deserve a more detailed explanation.

The notion of corpus being *general* denotes a corpus that requires a large amount of texts of many types (written as well as spoken language) that may be produced in one or many countries (Hunston 2006: 14). In addition, general reference corpus represents a given language as a whole and is used to make observations about a language for general purposes (the language used by ordinary people in everyday situations) (Bowker and Pearson 2002: 11–12). One of the well-known general corpus is BYU–BNC (the British National Corpus) that consists of 100 million written/spoken words of the late 20th century British English[4]. BNC includes a broad cross-section of text types and genres, such as newspapers, research journals or periodicals from various academic fields, fiction and

[4] British National Corpus. Available at http://corpus.byu.edu/bnc/ (9.04.2015)

academic books, letters, essays written by students of different academic levels, brochures and many other types of texts. For the reason that general corpus includes as wide a range of data as possible, it can be used as a reference for making comparisons with more specialised corpora (Hunston 2006: 15).

Sinclair (1996: 7) has stated in his seminal work on corpus typology: "anything which involves the linguist beyond the minimum disruption required to acquire the data is a reason for declaring a special corpus." Essentially, the specialised corpus differs from the general corpus in the data type it contains. The specialised corpus focuses on the particular type of language to represent and investigate (Bowker and Pearson 2002: 12). The degree of specialisation involved is not restricted and depends on the research interest (Hunston 2006: 14). In view of the fact that in the current thesis the learner corpus was compiled and observed, we refer to it as a specialised corpus, because learner corpus is always designated for a particular purpose – collecting and analysing the texts produced by learners of a language (thus consists of particular kind of texts and subjects).

Corpus compilation is a complex and time-consuming undertaking and it is the objective of the researcher to elaborate such a methodology that would allow extracting reliable evidence and ensure the accountability of the research. As stated by Tognini–Bonelli (2001: 49), "The corpus provides all the evidence and demands adequate explanations." However, in order for corpora to provide meaningful evidence, it is important to secure that the default characteristics of corpus compilation design would be present. The default characteristics are listed and discussed in the following paragraphs.

2.2. The default criteria in corpus design

Sinclair (1996: 6–8) named four default characteristics that any corpus needs to include – quantity, quality, simplicity and the requirement of being documented (Sinclair 1996: 5–8). Over the years, these qualities have remained the same, except for variation in terminology. For example, Bowker and Pearson (2002: 9) have listed such default values for corpora as 'authentic', 'electronic', 'large', and set the requirement to have 'specific criteria'. Similarly, Hunston (2006: 25) sets the four design criteria for corpora as 'size', 'content', 'representativeness' and 'permanence'. All of these criteria are interchangeably linked and balance each other as presented in Table 1. An overview of these basic notions is provided in the following sections. In addition, debates and contradictions in the default criteria are presented and discussed.

Sinclair (1996)	**Bowker and Pearson (2002)**	**Hunston (2006)**
1. Quantity	Large	Size
2. Quality	Authentic	Content
3. Simplicity	Electronic	Representativeness
4. Documented	Specific design criteria	Permanence

Table 1. The four default criteria in corpus design.

2.3. Simplicity

The default values of simplicity, representativeness and the requirement of being 'electronic' do not pose any significant contradictions. According to Sinclair (1996: 8) the default value of simplicity is plain text, which means that a researcher can expect an unbroken, linear string of ASCII characters, with any mark-up clearly identified, separable as well as retrievable from the text. The term *text* denotes a file of machine-readable data and thus refers directly to the default value of being 'electronic' (Flowerdew 2012: 3).

Thus, to ensure the representativeness and simplicity, it is necessary to gather data in electronic, machine-readable format from which the data is extractable.

According to Granger et al. (2002: 10) learner corpus can be produced in a variety of formats, for example in the form of a raw corpus, where only plain texts with no extra features are presented, or in the form of an annotated corpus which already includes linguistic/textual information. The Estonian–English Interlanguage Corpus was compiled in a raw format that was optimal for acquainting readers with the basic functionalities that linguistic software can provide and allow to perform the quantitative analysis of conjunctive adjuncts.

2.4. Quantity

As stated by Hardy and McEnery (2012: 28), it would be unimaginable to explore and revise research questions regarding the frequencies of word forms, phrases, or errors without the evidence produced by machine-readable corpora. The feature of frequency has been noted as the most reliable source of evidence that corpora present (McEnery and Hardy 2012: 28). According to Granger et al. (2002: 4), frequency is an aspect that indicates not only what is possible in language production, but also what is likely to occur. The strength of conducting quantitative analyses with the help of corpora is that it is possible to extract distinctive linguistic patterns automatically and results can be then classified and counted as well as compared with other corpora (McEnery and Wilson 2001: 76).

It has been claimed (Nomura 2012: 281; McEnery and Wilson 2001: 75–76; Granger et al. 2002: 4) that quantitative analyses that are based solely on frequency information do not reveal sufficient level of understanding of learner language and suggested that for a comprehensive analysis it is essential to include qualitative analysis.

As McEnery and Wilson (2001: 76) state, in quantitative research we classify and count features as well as construct statistical models, whereas in qualitative research the data is used as a basis for identifying and describing aspects of the language use. Therefore, in order to provide meaningful research, it is necessary to employ quantitative and qualitative analyses interchangeably – quantitative method for gathering objective and representative data, qualitative analysis for identifying, classifying and describing the linguistic instances.

Similarly to Sinclair (1996: 6), Bowker and Pearson (2002: 9) set the adjective 'large' for the value of quantity, because the principal aim of assembling a corpus is to gather data in quantity. Although quantitative information, such as frequency lists, are useful for identifying possible differences between the corpora and can be further studied in more detail (for example, to establish norms of frequency, draw comparisons between learner groups), the criterion of the exact size for corpus remains unspecified (Hunston 2006: 5, 25–26; also see Flowerdew 2012: 4). The optimum corpus size depends most importantly on the specific linguistic investigation and the type of the corpus, however, factors such as the availability of the data, financial investment in software along with the capacity of computer (speed and efficiency to access software) as well as the amount of time that a researcher is able to devote to a comprehensive (in terms data amount) analysis should also be considered before compiling a corpus (Hunston 2006: 25–26, also Bowker and Pearson 2002: 45–48).

Granger (2004: 124) has claimed that for the reason that corpus data is stored electronically, it is possible to collect a large amount of data fairly quickly and, as a result, "learner corpora are now counted in the millions of rather than the hundreds of thousands of words." Nevertheless, Hunston (2006: 25), Bowker and Pearson (2002: 45) as well as Flowerdew (2012: 4) emphasise that it is not accurate to assume that bigger corpus is always better, because the sheer quantity of information can become overwhelming for the

observer. In support of this argument, Granger (2012: 4) has claimed that the wealth of occurrences that learner corpora provides often leads to the point when researchers cannot study the whole set of evidence and ultimately have to select a representative sample for conducting the research.

In discussing corpus size criterion, such term as type-token ratio (TTR) needs to be introduced. In corpus linguistics, token denotes a word. However, as words reoccur (verbs, articles) in the text, the number of token types in corpus is always smaller than the total number of the tokens. The type-token is expressed in percentage terms and it can be calculated by dividing the number of token types by the number of the total tokens, (Flowerdew 2012: 324). Type-token ratio demonstrates the lexical diversity (range of vocabulary) in the corpus. For example, a low type-token ratio indicates that there is a lot of repetition, whereas a high type-token ratio suggests a greater degree of lexical diversity in the corpus (Flowerdew 2012: 324). The type-token ratio of the Estonian–English Interlanguage Corpus will be calculated and discussed in the succeeding chapter.

The principal concern with the notion of quantity lies in the fact that the narrow empirical base can rarely allow making any definitive statements about learner language. In reference to longitudinal SLA studies, Gass and Selinker (2008: 55) have stated that it is difficult to claim with any degree of certainty whether the results obtained from a small corpus are applicable only to the one or two learners studied, or whether they can characterize a wide range of learners. This argument is valid especially in relation to the research conducted in ELT framework, where the goal is to improve ELT tools (dictionaries or grammar reference books). In such cases the quantity is a major consideration, because the results must be representative, meaningful as well as beneficial for the whole learner population (Granger 1998: 11).

According to Granger et al. (2004: 125; also see Granger: 1998 10) the factor that has a direct influence on the size of learner corpora is the degree of control exerted on the variables, which in turn depends on the objectives of the researcher. As Granger (2012: 5) states: “However large it [the corpus] might be, a learner corpus will only be useful if it has been compiled on the basis of strict design criteria.” It is therefore possible to conclude that if the research question is specific and corpus design criteria are presented explicably, the data reduced to a manageable amount retains the advantages of coverage of a large corpus and allows making conclusions from the research. Granger (1998) has closely observed the level of detail on learner attributes within the International Corpus of Learner English (ICLE) database compilation. The learner attributes refer to the notion of authenticity and are also closely connected to the notion of documentation that are both worthy of further comment in the following paragraphs.

2.5. Quality

According to Sinclair (1996: 7) and Tognini–Bonelli (2001: 55), authenticity is a core value for corpus, meaning that a special restriction on the choice and collection of texts needs to be done by the researcher and this should be clearly stated in the documentation. For the reason that foreign language teaching context usually involves some degree of artificiality, the notion of authenticity is difficult to ensure (Sinclair 1996: 7; also see Granger et al. 2002: 8 and Granger 2012: 3). Sinclair (1996: 7) suggests that if corpus data does not contribute to a description of ordinary language (contains a high proportion of unusual features), it should be regarded as a special corpus, belonging to the general category of experimental corpora. According to Sinclair (1996: 7), this is necessary

in order to avoid any statements made about language collected in experimental conditions or artificial circumstances of various kinds.

If applied to EFL field, this means that learner corpus research is inevitably built on experimental data, because most of the learner language samples result purely from elicitation techniques or are gathered in artificial conditions or circumstances (Granger et al. 2002: 8). As Granger et al. (2002: 8; also see Granger 2012: 3) states, even the most authentic data from non-native speakers is rarely as authentic as native speaker data, especially in the case of EFL learners, who learn English in the classroom and have few opportunities to use the target language in everyday situations. It is thus important to note that the concept of quality covers a different degree of authenticity in learner corpora (Granger 2002: 8–9).

According to Sinclair (1996: 7) and Granger et al. (2002: 8), although it is difficult to ensure authenticity in learner corpora, there are nevertheless opportunities for corpora to be designed within the limits of reasonable expectation, ensuring that corpora contributes to a description of ordinary language. Granger (2012: 3) states that between natural and fully experimental data there is a wide range of data types situated at various points on the scale of naturalness. In the case of the written corpus, essay writing can be considered authentic written data, because it represents ‘free writing’ (Granger et al. 2002: 8; also see Granger 2012: 3). Although it is necessary to set the task variables, such as time limit or topic of the essay, this particular data presents the researcher precisely what students are able to produce independently. In the Estonian–English Interlanguage Corpus, essays were collected and thus, it accounts for representative and accountable data, where students had an opportunity to independently produce an essay in equal and strictly limited conditions.

2.6. Documentation

In order to produce legitimate and reliable results, corpus needs to present careful and inclusive documentation of its design criteria. In case of interlanguage, which is highly variable in linguistic, situational and sociolinguistic factors, it is important to be as explicit as possible. Therefore, documentation of a learner corpus needs to include two domains – one with the features that concern the learner and the other that concerns the language. Granger et al. (2002: 9) states: "The usefulness of a learner corpus is directly proportional to the care that has been exerted on controlling and encoding the variables." It is thus important to include specific information about the learners as well as the language situation, because any additional information about the learners and task variables may be necessary and beneficial for further research. Documentation will enable researchers to make comparisons between corpora – for example, compare the results between the same learner populations (over the period of time) or compare learners with different mother tongues (Granger et al. 2002: 10).

Yukio Tono (2003: 800) has illustrated the three main design considerations that assist the researcher in compiling the corpus documentation (Table 2). These categories include a) language-related criteria, b) task-related criteria and c) learner-related criteria. The list of features singled out by Tono (2003) is exhaustive in language, task and learner-related detail and should be consulted in order to include as much learner-oriented detail as possible.

language-related	**task-related**	**learner-related**
mode	data collection	internal-cognitive
[written/spoken]	[cross-sectional/longitudinal]	[age/cognitive style]
genre	elicitation	internal-affective
[letter/diary/fiction/essay]	[spontaneous/prepared]	[motivation/attitude]
style	use of references	L1 background
[narration/argumentation]	[dictionary/source text]	L2 environment
topic	time limitation	[ESL/EFL, level of school]
[general/leisure etc.]	[fixed/free/homework]	L2 proficiency
		[proficiency standard test score]

Table 2. Design considerations for learner corpus. Adopted from Tono (2008: 800).

According to Granger (2012: 5), even if learner corpus has been carefully designed, not all variables can be recorded, because it is rarely possible to include information on such specific aspects as the teaching methods, the course materials or the L1 or L2 status of the teachers, which all comprise crucial factors in the setting of foreign language learning. Admittedly, it is doubtful that any corpora can guarantee total coverage of every possible criterion, however researcher should include as detailed account of variables as possible in order to contribute for future research and avoid any subjectivity. The documentation of the Estonian–English Interlanguage Corpus will be introduced in the Chapter 5 that is devoted to describing the research method of quantitative analysis in the current thesis.

In order to ensure and foster accountable and representative linguistic research, the four default criteria are necessary to be applied to corpus design. The relevant constituents of corpus design that should be addressed already during the initial stages of corpus compilation are quantity, quality, simplicity and documentation. In case of compiling a learner corpus, which ultimately is a specialised corpus that collects and analyses the language from a particular group of people (interlanguage), the level of detail should not

be restricted and researchers should strive to make their corpus documentation as representative as possible. The explicit design criteria in corpus construction will facilitate the projects by anticipated users of the corpus as well as other specialists from different research fields.

3. STAGES IN LEARNER CORPUS RESEARCH

In spite of the growing interest towards corpus linguistics and its usefulness in learner language study, it is difficult to find a universal instruction for conducting corpus research that would be suitable for every study, as research interests and aims vary greatly. The general typologies of different corpora and default characteristics that should be attributed to corpus design have been discussed in previous chapter. The present chapter seeks to take a step towards illustrating the learner corpus research technique, that is, the planning of the research project by naming and discussing the main stages in learner corpus research and demonstrating the corpus tool (*AntConc* software) that was used in linguistic analysis in the Estonian–English Interlanguage Corpus.

3.1. Main stages in learner corpus research

In order to examine a corpus and retrieve useful information from it, it is first important to look into the central steps in corpus analysis. According to Granger (2012: 9) there are seven main stages in learner corpus research:

1. **Choice of methodological approach**
2. **Selection and/or compilation of learner corpus**
3. Data annotation
4. **Data extraction**
5. **Data analysis**
6. **Data interpretation**
7. Pedagogical implementation

Granger states (2012: 9) that five of the stages are mandatory – the choice of methodological approach, the selection and/or compilation of learner corpus, the data extraction, the data analysis and the data interpretation. The stages of data annotation and

pedagogical implementation are not necessarily required, however are regularly met in the learner corpus research. The five mandatory stages are present in the current thesis and each of them will be described in the following sections.

3.2. Choice of methodological approach

Corpus analysis is an empirical approach, because it is derived from observing and describing authentic data, more precisely the analysis and the description of language use as realised in text(s) (Tognini–Bonelli 2001: 2). The observation of linguistic instances may either lead to the formulation of a hypothesis or be studied according to a pre-existing research question, therefore it is possible to make a distinction between corpus-driven and corpus-based language studies (Tognini–Bonelli 2001: 2). These binary terms were originally introduced by Tognini–Bonelli (2001) and determine the corpus research into one or the other group according to their method.

As stated by Tognini–Bonelli (2001: 65), corpus-based studies typically use corpus data in order to explore a theory or test the hypothesis about the language to then either exemplify, refine, validate or reject it. The primary interest for corpus-based linguists is to test how well their theories account for the data (Granger 2012: 10). Such a relationship between theory and data is common in linguistics, because corpus-based linguists carry out the analysis according to the pre-existing hypothesis or research question (Tognini–Bonelli 2001: 65–66).

In contrast to corpus-based approach, which always works within accepted theoretical frameworks, corpus-driven method sees the corpus itself as the sole source of hypotheses about language (McEnery and Hardy 2012: 6). As Tognini–Bonelli (2001: 85) claims, corpus-driven approach requires the commitment to inspect and analyse the corpus

data as a whole, because the aim is to look for possible extensions in the pre-existing theories. It is thus claimed that the corpus itself embodies its own theory of language and should not be adjusted in any way to fit the predefined theoretical categories of the analyst (Tognini–Bonelli 2001: 84–85; also see Granger 2012: 10). Tognini–Bonelli (2001: 86–88) states that although the corpus-driven approach potentially leads the scholar to uncover new grounds and posit new hypotheses about language, it remains largely unexplored, as it needs an exhaustive account of evidence (in terms of quantity) to be representative for making any conclusions about language.

The learner corpus in current thesis was compiled for investigating Estonian–English interlanguage and the current study aims to explore the variability and frequency of conjunctive adjuncts by the Estonian EFL learners. The research questions about the conjunctive adjuncts determine the theoretical framework (in current thesis, the categories of conjunctive adjuncts) according to which the corpus data is analysed. The results are inextricably linked to the specific corpus as well as specific student group. Therefore, the current analysis can justly be defined as corpus-based, not corpus-driven.

3.3. Selection and/or compilation of learner corpus

The learner corpus compilation is time-consuming and complex undertaking that consists of two phases – selecting or collecting suitable texts for corpus and including or transferring these into machine-readable, electronic format. Granger (2012: 11) has suggested that it is advisable to first survey the field to find out whether there have been already compiled any suitable and available corpora for the research. However, if there have yet not been assembled (or available) any learner corpus that could meet the purposes and requirements for conducting required research, it is necessary to compile a suitable

corpus. Regardless of whether the corpus represents spoken or written language, researcher needs to decide upon several questions already during the corpus design phase – what constitutes a suitable text and how it will be inserted in the corpus, how can the files be named and which format should be used for data storing (Reppen 2012: 32).

In case of specialized subject fields that learner corpus belongs to, it is necessary to convert suitable material in electronic form by either typing in the texts or using technology that would do it automatically, for example, optical character recognition software or dictation software (Bowker and Pearson 2002: 59). The main disadvantage of manually typing in the texts as well as using assisting technologies lies in the difficulty of ensuring the notion of accountability, therefore the converted texts must be carefully proof-read and edited by a researcher so that the authentic learner spelling and grammar structures would be preserved (Bowker and Pearson 2002: 59; also see Reppen 2012: 34). The solution to this problem is making sure that every text is read and edited at least twice and text is inserted and proof-read by different researchers.

The notion of text type is often predetermined by the research interest (Reppen 2012: 32). For example, if the research aim is to analyse the written production of learners, the written learner data should be collected. It is important to note that it is preferable to collect the full texts rather than the extracts, because examining the location of the pattern or the structure of the text may become relevant in the future research (Bowker and Pearson 2002: 49). In addition, it is necessary to determine the file format in which the data will be stored, as saving files in a format that will not be compatible with the tools that will be used for analysis may result in many extra hours of work (Reppen 2012: 33). It is therefore advisable to store the data in the format of *plain text* that functions well with most of the corpus analysis software (Reppen 2012: 34).

The conventions in file naming (labels, tags, codes) are important to be established clearly from the beginning of the corpus compilation procedure, because files need to be directly relatable to the information about the learners and the text itself (date or place when it was produced) (Reppen 2012: 33). The issues of corpus storage will not be discussed in much detail here, however, it is worth mentioning that multiple locations of the corpus as well as using backup software (keeping several copies of the corpus) are strongly advisable in order to avoid any threat of losing the corpus through computer malfunction (Reppen 2012: 33).

The Estonian–English Interlanguage Corpus that was compiled for the current research was manually transcribed into electronic form by a group of postgraduate students that enabled the texts to be typed in and later be edited by different people. Each text was proof-read twice in order to preserve the original layout (in terms of paragraph structure) and authentic learner language. The essays were first transcribed in the form of a *Word* file that could later be easily transformed into plain *txt* format, depending on the software chosen by a researcher for the analysis. The names of the subjects and any other personal information about students were excluded from the texts and instead given an anonymous code. The corpus was stored in multiple copies as well as computers.

3.4. Data annotation

The linguistic annotation of data is not listed as a compulsory step by Granger (2012) for the reason that it primarily depends on the objectives of the study and in many cases the data can be successfully analysed in the format of raw (unannotated) corpus. There are various levels of annotation, one of the most common approaches in annotation is part of speech (POS) tagging, where each word in the corpus is assigned to a

grammatical tag corresponding to the word class it belongs to (Bowker and Pearson 2002: 83; Reppen 2012: 35; also Flowerdew 2012: 178–179).

Nowadays, the improved accessibility of computers and corpus annotation applications enables researchers to manipulate with corpus data mostly automatically, whereas the initial research on learner language was conducted manually. Automated approaches are absolutely essential in the case of large datasets when it becomes impossible to search for some instance manually (McEnery and Hardy 2012: 2). The benefit that a researcher can derive today from using already standardized automation/mark-up software is invaluable, as the input of new technologies affect the methodological frame of the linguistic enquiry by speeding it up, systematising it, and making it possible to analyse large amounts of data in short terms (Tognini–Bonelli 2001: 5). However, it is still a matter of financial funding as well as the investment of time and effort from the side of the researcher that is necessary for becoming familiar with automated methods and tools (Granger 2004: 126).

The type of annotation that is noted as particularly relevant in terms of the learner corpora is error tagging, where the corpus needs to be preliminarily annotated with the help of comprehensive error classification (Granger 2012: 13; also Dagneaux et al. 1998: 163). However, as the corpus analysis tools have for the most part been developed on the native corpus data, there are several drawbacks in learner corpus error analysis (CEA) (Granger 2012: 12). For the reason that the learner language (interlanguage) errors differ from the native speaker errors and the variability in learner performance is prominent, it is difficult to construct a universal error checker. In fact, even if one error checker is created for the research, it is essential to start the analysis with a pilot study to check the accuracy of the tagging mechanism (Granger 2012: 12).

The annotating process is largely manual and time-consuming procedure that needs to be designed and produced by a group of researchers. For example, the corpus error analysis developed at Louvain University (ICLE corpus) requires the collaboration of at least two researchers – ideally native and non-native, because bilingual team heightens the quality of error correction (Dagneaux et al. 1998: 165). Before the error-tagged files are analysed, the learner data must be corrected manually by a native speaker of English, who inserts correct forms in the text; after that the analyst assigns to each error an appropriate error tag, documents it in the error tagging manual (creating a hierarchical tagging system) and finally inserts the tags into the corpus (Dagneaux et al. 1998: 165–166).

Although the corpus annotation can facilitate linguistic analysis and allow performing sophisticated linguistic investigations, for example detecting interlanguage errors in the learner corpus, it demands a careful instruction of a research group to first apply and check the tag-set as well as decipher it later during the data analysis. Such an investment of time and effort may not be practical and achievable in terms of every research. In current thesis and the Estonian–English Interlanguage Corpus, no annotation has been performed. Nevertheless, it is a worthwhile enterprise that will hopefully be performed in the future research.

3.5. Data extraction: The *AntConc* software

Corpus analysis software enables researchers to automatically extract a wealth of information from the learner corpus and assists the researcher in describing the linguistic phenomena objectively. There are numerous programmes as well as software existing today for linguistic analysis. The most popular commercial programmes are *WordSmith Tools* and *MonoConc*. Majority of the programmes for linguistic analysis include the basic

functionality of concordancer. Concordances capture and highlight a single word or word combination and present it in its immediate contexts (Granger 2012: 14). In the current section the *AntConc* software is presented. This programme was used for conducting the quantitative analysis in the current thesis and proven to be suitable for providing insights into Estonian EFL learners writing, namely the use of conjunctive adjuncts.

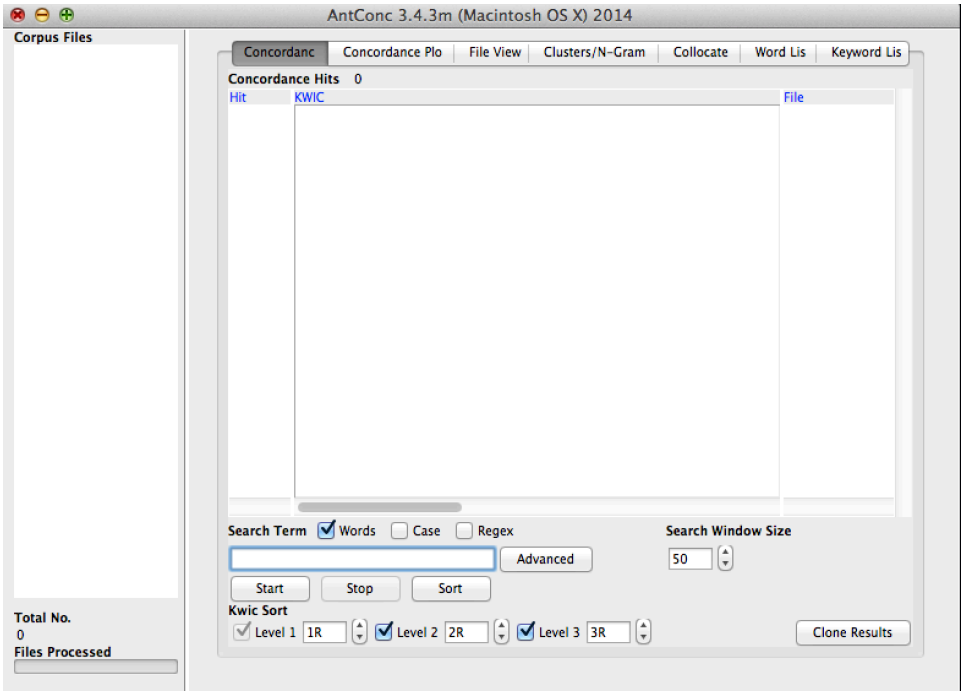


Figure 1. Top screen of *AntConc* 3.4.3m for Macintosh.

The *AntConc* programme was developed by Anthony Laurence in the Waseda University. The *AntConc* software is straightforward in its use and does not require installation on the computer. For the reason that the programme is freely downloadable from the *AntConc* homepage[5], it is ideal for individuals or educational institutions with limited financial resources (Laurence 2005: 729). The first version of *AntConc* was released in 2002. The *AntConc* programme is sporadically updated and the version used for analyses in the current thesis is *AntConc* 3.4.3m (2014) for Macintosh. The *AntConc* programme allows manipulating with plain *txt* files as well as with data in *xml* and *html*

[5] Available at http://www.laurenceanthony.net (5.03.2015)

formats. Among the central concordance technique, the *AntConc* programme also enables to create word lists, examine collocates and clusters that can be found in the corpus. In the succeeding paragraphs an overview and illustrations of the basic functionalities that were used in current analysis are presented and discussed.

The 'Word List' tool in the *Antconc* programme enables to create alphabetical and frequency-sorted lists. Once compiled, the lists can be opened and edited in the spreadsheet format (new file) or any other standard text editor. Word lists are useful in linguistic enquiry mainly because they highlight the most frequent words in the corpus and allow comparing the results cross-linguistically with various corpora (for example with different levels of language proficiency) (Römer and Wulff 2010: 104; also Laurence 2005: 732). Word list also allows calculating the lexical variety (type-token ratio) of the corpus. This aspect can be particularly relevant in the preliminary stages of corpus analysis in order to map and present the basic characteristics of the corpus, such as vocabulary (Flowerdew 2012: 9–10).

For example, the type-token ratio in the Estonian–English Interlanguage Corpus is 8.5% in comparison to the reference corpus (the Michigan Corpus of Upper–level Student Papers), where type-token ratio is 11.8% (Muddhi and Hussein 2014). The fact that the type-token ratio in the Estonian–English Interlanguage Corpus is lower than in the reference corpus can be explained by the fact that the topic for essays in the Estonian–English Interlanguage Corpus was the same for all of the learners and the number of words for the essays was strictly limited (altogether 127 essays consisting of 24,796 tokens). In the Michigan Corpus of Upper–level Student Papers (altogether 25 essays consisting of 95,538 tokens) students have written longer essays on various topics.

From Figure 2 it is possible to see the most frequently occurring words in the Estonian–English Interlanguage Corpus (EEIC). The most frequently occurring word in

EEIC is the definite article *the*. Indefinite article *a* also falls into the list of the most common words. There are two frequently occurring words that are connected to the topic of the essay – *English* and *language*. This is a noteworthy fact that explains the lower type-token ratio in EEIC, as students had a specific topic to write about and thus, the choice of vocabulary was inevitably limited. Among the other frequently occurring words in EEIC are content words (*would, will, be*) or function words (*and, it, or*) that have a little lexical meaning, but are nevertheless necessary for building grammatical relationships between words and sentences.

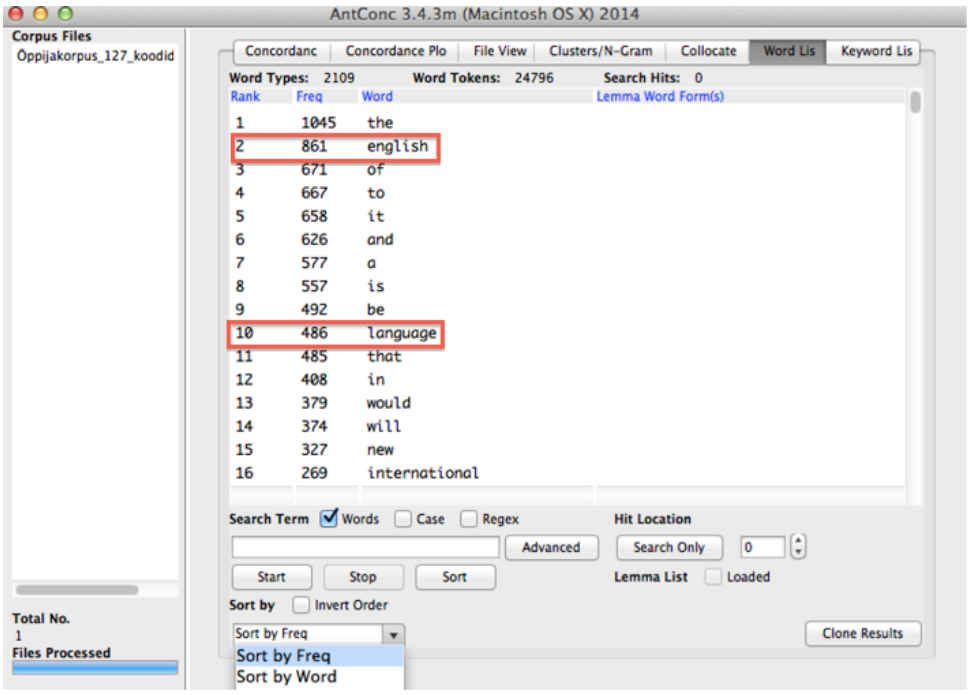


Figure 2. List of the most frequent words in the Estonian–English Interlanguage Corpus.

The core functionality of the *AntConc* programme is the concordance tool that is very straightforward and allows compiling concordance lists instantly. Once the necessary word or word combination is entered into the search pattern, it is possible to see the total number of instances that can be found in the corpus where each of them is easily visible and traceable to its original context (file view). By looking at the concordance line it is possible to access the phraseological patterns and the meanings of the given instance in

diverse natural contexts (Laurence 2005: 730). In the current thesis, the concordance tool was used for making quantitative analysis – each conjunctive adjunct was searched and the corresponding number of the found occurrences was written down, resulting in lists that could be compared with the findings adopted from the reference corpus.

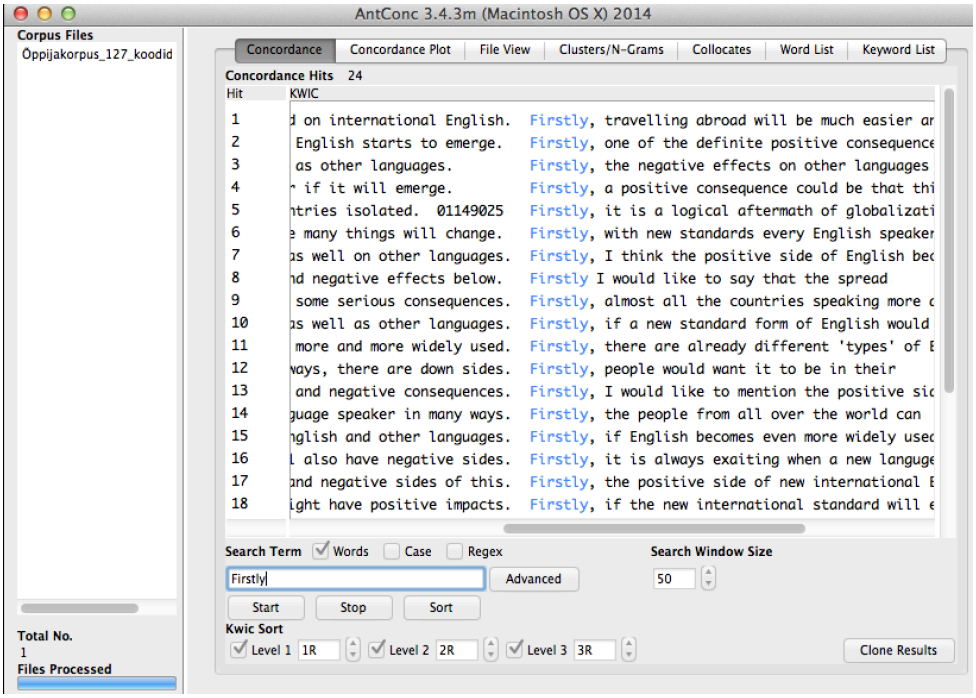


Figure 3. Concordance of the conjunctive adjunct *firstly* in the Estonian–English Interlanguage Corpus.

Although the *AntConc* software is multiplatform, non-profit-oriented programme, easy to use and effective in its basic tools, there are numerous limitations to it. First of all, there are no possibilities for performing qualitative analyses within the programme – the results can only be copied to new file and sorted and counted manually later (Laurence 2005: 735). In addition, the *AntConc* programme only enables performing the analyses in raw (unannotated) files, which greatly influences the flexibility of the performance (Laurence 2005: 735). For the reason that the Estonian–English Interlanguage Corpus is relatively small (total of 127 essays) and not annotated, the *AntConc* programme was suitable and efficient. However, if the Estonian–English Interlanguage Corpus will

continue to grow (essays will be added every year) and annotation will be performed, much more powerful software will be necessary for conducting analyses.

3.6. Data analysis

The data analysis depends on the research interest and on the software that is chosen for particular study. The range of linguistic phenomena that can be investigated with the help of learner corpus is diverse – some of the analyses may focus exclusively on interlanguage errors, whereas the others may compare differences in vocabulary between various corpora. According to Granger (2012: 17) and Flowerdew (2012: 172–173), contrastive analyses where two or more learner groups (ideally native and non-native) are examined, are particularly effective, as they allow to uncover typical features of learners' interlanguage – not only errors, but also instances of under– or overuse of words, phrases and grammatical structures. In the current research, quantitative analysis is performed, where frequency and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus is investigated and compared with the reference corpus (the Michigan Corpus of Upper–level Student Papers).

3.7. Data interpretation

The results obtained from the corpus research need to be described (analysed) as well as interpreted. Granger (2012: 20–21) states that "LCR has so far been stronger on description than interpretation" because "the majority of the studies focused on varieties of interlanguage that were badly in need of description, viz. the upper intermediate and advanced stages of proficiency". For the reason that the current thesis only provides an overview of the corpus compilation theory and compiles the first Estonian–English

interlanguage corpus in Estonian educational arena, it has no substantial base to compare the results with. Therefore, for comparative analysis, the results from the reference corpus were adopted for the current thesis.

Nevertheless, it is hoped that future research will enable to carry out comparative (possibly longitudinal) interlanguage studies in the same field and allow making theoretical conclusions within Estonian–English interlanguage perspective. The current research will serve as a descriptive base on what has been found during the initial analysis of the Estonian–English Interlanguage Corpus in reference to conjunctive adjuncts.

3.8. Pedagogical implementation

The use of corpus-based methods contributes directly to language pedagogy by influencing and improving pedagogical tools, methods as well as teaching and learning materials. According to Braun (2006: 1) corpus-based observations have helped to uncover and remove discrepancies between what is taught in schoolbooks and what kind of language is actually used by learners, which allows making corresponding changes in teaching methods and place greater emphasis on suitable teaching materials.

In addition to the identification of difficult areas in interlanguage at various stages of the language learning process, the direct use of corpora in the classroom (data-driven learning) has recently become increasingly popular (Braun 2006: 1). The corpus-consulting activities (for example concordance-based exercises) can result in entirely learner-centred corpus-browsing projects that foster autonomous learning (Mukherjee 2006: 12). One of the pedagogical approaches could be the compilation of a local learner corpus by the students of the same class, so that the teachers and students could notice and observe their progress over time and compare it to the reference corpora (Mukherjee 2006: 19).

The corpus-based analysis of the frequency and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus contributes directly to pedagogical aims, as it provides a description of results obtained from the authentic Estonian–English interlanguage. The quantitative investigation of conjunctive adjuncts in students' writing allows to determine exactly what kind of conjunctive adjuncts are used by Estonian ESL learners and discover whether they pose any significant difficulties for them (for example overuse or underuse of particular conjunctive patterns). The current research provides quantitative overview of the results that may become useful for improving Estonian EFL teaching methods and materials in the future.

4. THE CONCEPTS OF COHESION AND COHERENCE

Cohesion and coherence are central properties of text that ensure the comprehensibility for the reader. It is first necessary to make a distinction between coherence and cohesion, because both terms are used in discourse analysis and text linguistics. According to Rummel (2010: 46), scholars have not yet fully agreed on what cohesion and coherence denote, because both terms are partly overlapping in meaning. However, as Rummel (2010: 46) further states: "Although coherence and cohesion are both attained by the means used to order parts of a text, generate causal links, maintain topic continuity, determine relations among discrete units of discourse and establish connectivity between distinct parts of discourse, these two notions denote clearly distinct properties of text and discourse".

The term coherence refers directly to the property of text, which is, according to Halliday and Hasan (1997: 1): "any passage, spoken or written, of whatever length, that does form a unified whole." Blanpain (2012: 25) has stated that coherence depends to a large extent on readers' familiarity with the text schemata – the expectations shaped by the knowledge of discourse patterns of how the text will further develop. Coherence can be realized in the structure of the text that is the division of text into chapters, sections or paragraphs (Blanpain 2012: 26, 28–29). According to Rummel (2010: 46– 47), coherence is achieved if the content of the discourse has a logical progression and organisation and thus, coherence is reflected not only in cohesive ties but also to a large extent on readers' expectation of generally accepted way of organising ideas.

While coherence operates on the unit of the text, which, according to Halliday and Hasan (1997: 2) is realized by sentences, the notion that concerns the grammatical and lexical features within and between the sentences is cohesion. In order to carry the meaning, sentences are in cohesive relation with each-other. Blanplain (2012: 25) claims

that cohesion is a surface phenomenon that concerns the grammatical and lexical features that create ties between sentences. According to Halliday and Hasan (1997: 5), cohesion forms an integral part of the language system and is primarily realized in cohesive devices. Halliday and Hasan presented cohesive devices in their seminal piece *Cohesion in English* in 1976 that investigates the relationship between cohesive devices and writing quality and serves as theoretical base for the current thesis.

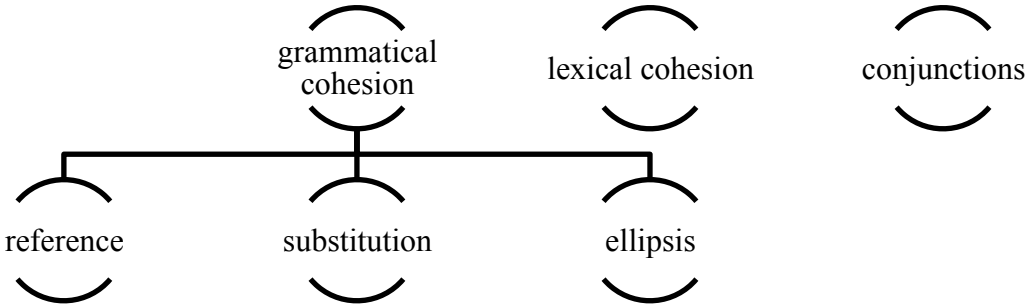


Figure 5. Cohesive devices discussed by Halliday and Hasan (1997).

Halliday and Hasan's (1976) taxonomy of cohesive devices (Figure 5) consists of two larger categories – grammatical cohesion and lexical cohesion. The class of conjunctions has been placed separately, because Halliday and Hasan (1997: 6) state: "conjunction is on the borderline of the two; mainly grammatical, but with a lexical component in it" and therefore this category is also treated individually. The elaboration of each class is beyond the scope of the present research and thus the emphasis is only put on presenting the categories of conjunctions, because this class is directly connected to current research.

4.1. Definition of conjunctions and conjunctive adjuncts

According to Halliday and Hasan (1997: 226), conjunctions are different in nature from the other cohesive relations, because they are cohesive not in themselves but indirectly, by virtue of their specific meanings. Conjunctions presuppose the presence of other components in the discourse and allow them to occur in succession, although they may not be semantically or structurally connected (Halliday and Hasan 1997: 226–227). The current paper adopts the general term of *conjunctive adjunct* proposed by Halliday and Hasan (1997: 228) for the entity of conjunctive expressions (adverbs, such as *but, so;* compound adverbs, such as *furthermore, besides;* prepositional phrases, such as *on the contrary, as a result;* prepositional expressions such as *in spite of that, instead of that*).

Conjunctive adjuncts are used in writing to link the text semantically and logically (Muddhi and Hussein 2014: 18). The main function of conjunctive adjuncts is to clarify the connection between the units of discourse and serve as a link between the passages of a text (Biber et al. 2002: 389). Halliday and Hasan (1997: 238) state that there is no single inventory of the types of conjunctive relation and different classifications are possible. The current thesis adopts the initial scheme of Halliday and Hasan (1997: 242–243) of four basic conjunctive categories – additives, adversatives, causals and temporals.

Additive conjunctives allow connecting additional statement(s) to the pre-existing information. The additives allow introducing discourse units further – emphasise the key point or add new relevant information to the previously mentioned statements (Suswati et al. 2014: 16). Additive conjunctions are *and, nor, neither, either, or, futher(more), besides that, in addition, allternatively, moreover* etc.

Adversative conjunctives allow presenting new, contrasting viewpoints or arguments within the same topic, for example *yet, though, but, nevertheless, all the same,*

despite this, but, however, as a matter of fact, to tell the truth, actually, as against that, at the same, on the one/other hand etc.

Causal conjunctives are used to build connections between ideas in text, explain and verify the reasons or viewpoints. Causal conjunctives are *so, thus, therefore, consequently, for this reason, as a result, for this purpose, for, because* etc.

Temporal conjunctives allow building relations between the successive sentences regarding the time, for example *then, next, afterwards, after that, subsequently, at the same time, previously, meanwhile, until then, at this point/moment* etc.

4.2. Previous research on conjunctive adjuncts in learner writing

Corpus-based research provides an opportunity to investigate large samples of learner writing and compare learner-created texts with those written by native speakers. In Estonian educational arena, the research on conjunctive adjuncts from the perspective of the Estonian EFL learners has not yet been performed and the thesis seeks to provide the first corpus-based account on the current matter. Crewe (1990) discussed the main difficulties that EFL learners encounter in regard to the conjunctives and suggested three aspects – the erroneous use of connectors (for example *on the contrary* is used for *on the other hand/however*), overuse of certain connectives as well as underuse (avoidance) of connectives.

According to Crewe (1990), EFL learners often tend to use connectives as stylistic enhancers to give their text more ‘academic’ or ‘educated’ look. However, in case of misuse, the argumentation is not only difficult to process but also appears illogical to the reader (Crewe 1990: 316). Crewe (1990: 316) suggests that students should be offered such a sub-set of connectives that would be comprehensible for them. Crewe (1990: 316–

318) recommends teaching connectives separately, because these should be seen as higher-level discourse units that allow making learners' logical links in text more apparent and thus deserve greater attention.

The previous comparative research regarding the use of conjunctive adjuncts (between native and non-native learner groups) has been quantitative. The practical significance of the quantitative research concerning the over– or underuse of conjunctive adjuncts in learner writing is that it indicates precisely what kind of conjunctive patterns are used by learners and demonstrates the degree of deviation of frequency/variety of conjunctions from native speakers' writing. The comparative analysis reveals possible overuse or underuse (avoidance) of certain conjunctive patterns.

The study conducted by Granger and Tyson (1996) revealed that advanced French, Dutch and Chinese EFL students use certain individual connectors more frequently in comparison to native speakers (such as *actually, indeed, of course, moreover, namely, on the contrary)*. Correspondingly, certain connectors were found to be relatively unpopular in comparison to native speakers' writing (such as *however, instead, though, yet, hence* and *then)*. The study suggests that similarities in three non-native learner populations may have resulted from interlangauge influence (Granger 2002: 9–10).

Tapper (2005) revealed that in comparison to native speakers' writing, advanced Swedish EFL learners overuse the category of adverbial connectives (additive conjunctions) and noted that Swedish learners used slightly more types of connectives than the American students. Although this study also shows certain differences in the use of conjunctions between native and non-native learner groups, no consistent pattern of over– or underuse was revealed. The use of conjunctive adjuncts has also been investigated in the writing of Arab EFL learners. For example, Fakhra (2009) revealed that Syrian EFL learners use most commonly the category of additive conjunctions and such conjunctives

as *so, and, but* and *also* are used most repeatedly in their writing. In the Palestinian EFL learners' writing (beginners and intermediate level), Sharkh (2012) revealed that additive adjuncts (especially conjunctive *and*) were overused in comparison to native speakers' writing.

The current thesis has adopted the model of the study conducted by Saud K. Muddhi and Riyad F. Hussein (2014). This particular study was chosen because the data was analysed according to the original classification of conjunctive adjuncts (proposed by Halliday and Hasan in 1976). The study by Muddhi and Hussein (2014) investigated and compared the frequency and variability of conjunctive adjuncts in two learner corpora – the Kuwaiti Leaner Corpus and the Michigan Corpus of Upper–level Student Papers.

The results of the study conducted by Muddhi and Hussein (2014) indicated that variability of conjunctive adjuncts in Kuwaiti Learner Corpus is smaller than in the reference corpus. In addition, the results showed that Kuwaiti EFL learners overused some conjunctive adjuncts, namely *in addition, for example, so* and *but* whereas certain conjunctives, such as *however, though* and *thus* were significantly underused in comparison to native speakers' writing.

The study by Hussein and Muddhi (2014) concluded that conjunctive adjuncts and their variability should be given more attention in the EFL education. In addition, two recommendations were given – the conjunctives should be investigated further, in different writing types (genres) and investigation should be continued among different student groups, in order to trace the use/development of the conjunctive adjuncts in different learning stages.

5. METHOD

5.1. Participants

The participants in this study were adult Estonian secondary school graduates (who were learning English a foreign language. The age of the students varied from 18 to 35. The total number of participants was 132, among whom 88 were women and 39 were men. Only 127 of the essays were included into the corpus, as 3 of the students did not provide the written task that was necessary for current analysis and 2 of the students did not have Estonian citizenship. The requirement for the participants was the certificate of secondary school education. Information about the previous educational background (possible higher education in other speciality), the ethnicity or L1 of the participants was not specified.

The English proficiency level of the participants was not indicated beforehand, however, the participants who had fulfilled any of the succeeding requirements were exempt from the examination:

- Scored at least 95 points at the State Examination of English language
- Have a Certificate in Advanced English (CAE) level C1 or higher
- Have a Certificate of Proficiency in English (CPE)
- Scored at least 7 points in The International English Language Testing System (IELTS)
- Scored the maximum (that is 100 points) in the Test of English as a Foreign Language (TOEFL)

5.2. Corpus data

The entrance examination was held in July 2014. The primary goal of the examination was to test the candidates' proficiency in English language upon entering the Department of English Studies in the University of Tartu. The examination measured two

constituent language skills – reading and writing. Duration of the examination was 2 hours and participants could administer their time as they wished (choose the order for fulfilling the tasks). The essays (up to 200 words) from the third part of the examination were elicited for corpus data. The third part of the examination also included two shorter writing tasks that were not included in the corpus.

For completing the task, participants had to read a text and provide the answer to the corresponding (essay-type) question. The reading passage concerned the future of English language and was adopted from Guy Cook's *Applied Linguistics* (2008). The topic of the essay required participants to provide arguments in favour and against the main statement as well as explain their personal opinion. The ability to provide correct, logically structured and linguistically appropriate answers was evaluated.

5.3. Corpus data collection and data analysis

The collected hand-written manuscripts were manually transcribed into electronic, machine-readable form. Once the texts were transcribed, every text was additionally proof-read by two postgraduate students. The total number of essays was 127 that made up 24,457 tokens. The data was used solely for research purposes. Each text was assigned a code. The corpus has a documentation file that deciphers the codes and can be used to relate the text to its author. The documentation is stored separately and is not directly attached to the data and thus, no additional information is visible during the corpus analysis.

The compiled raw corpus was investigated with the help of the *AntConc* programme. In order to perform quantitative analysis, concordance tool and word list tool were used. Concordance lines and word list tool were utilised to bring together instances of

each conjunction and allowed the researcher to observe regularities in use and sort the conjunctive adjuncts into corresponding categories for making comparisons to the reference corpus.

5.4. Reference corpus

For the reference corpus, the Michigan Corpus of Upper–level Student Papers (NSC) was used. The results (percentages of the use of conjunctive adjuncts) from the native speakers' corpus were adopted from the study conducted by Hussein and Muddhi (2014) and juxtaposed with the results obtained from the Estonian–English Interlanguage Corpus (EEIC). The study by Hussein and Muddhi (2014) was chosen because it used the initial classification of conjunctive adjuncts proposed by Halliday and Hasan (1976).

The reference corpus consisted of 25 essays that were graded as 'excellent' and were written by native speakers of English on different topics concerning linguistics and English in general. The number of tokens in the reference corpus was 95,538. The current analysis uses only the percentages of conjunctive adjuncts across corpora and thus, the lexical density (discussed in section 3.5) do not hinder the results.

6. RESULTS – DATA ANALYSIS AND INTERPRETATION

In the current chapter, the results from the Estonian–English Interlanguage Corpus (EEIC) are presented. The results were obtained with the help of the *AntConc* programme that displayed the number of occurrences of each conjunction. The results were transformed into percentage terms and juxtaposed with the results from native speakers' corpus (NSC) that were adopted from the study conducted by Hussein and Muddhi (2014). Each category of conjunctive adjuncts is presented and analysed separately.

6.1. Frequencies of additives in EEIC

Additives	EEIC	EEIC%	NSC%
and	626	64.7%	6.8%
also	126	13%	52.3%
or	86	8.8%	1%
for example	38	3.9%	9.9%
on the one/other hand	29	2.9%	2.9%
that is	17	1.7%	1.5%
furthermore	9	0.9%	3.6%
thus	9	0.9%	11.2%
for instance	7	0.7%	1.3%
moreover	7	0.7%	0.7%
in addition	6	0.6%	3.1%
nor	3	0.3%	1.5%
i mean	2	0.2%	
either	1	0.1%	1.2%
in the same way	1	0.1%	0.7%
alternatively	0		0.7%
besides	0		0.2%
likewise	0		2%
not only that	0		0.52%
neither	0		0.26%
similarly	0		2%
to put it another way	0		2%
TOTAL	967	100%	100%

Table 3. Frequencies of additives in the Estonian interlanguage corpus.

Table 3 illustrates that the native speakers' corpus consists of more additive conjunctions (altogether 21 variants) than the Estonian–English Interlanguage Corpus (altogether 15 variants). In EEIC the most popular additive conjunctions are *and* with 64%, *also* with 13% and *or* with 8.8%. These are followed by additive conjunctions *for example* and *on the one/other hand* that both account for less than 4% in EEIC. In NSC the most popular additive conjunctions are *also* with 52.3%, *thus* with 11.2% and *for example* with 9.9%. Less popular additive conjunctions in NSC are *and* with 6.8% and *furthermore* with 3.6%.

It is possible to conclude from the results that Estonian EFL learners and native speakers of English use different additive conjunctions, because the contrast in the frequencies is significant. Estonian ESL learners tend to use coordinating conjunction *and* frequently (64.7%), native speakers mostly use *also* (52.3%). Significant contrast in the frequencies concerning additive conjunctions can also be seen in the use of *thus,* which is 11.2% in NSC and only 0.9% in EEIC as well as in the use of additive conjunction *or,* which is 8.8% in EEIC and only 1% in NSC.

The overall variety of additive conjunctions is larger in the reference corpus and there are plenty of variants that account for at least 2% in NSC, but are absent in EEIC, namely, *likewise, similarly* and *to put it another way.* This may partly explain the relatively high occurrence of *and* (64.7%) in EEIC, as Estonian EFL learners use a smaller variety of additive conjunctions overall.

There are numerous additive conjunctions that are relatively unpopular in both corpora, although their frequency is notably higher in NSC rather than in EEIC. For example, *furthermore* with 3.6% in NSC is only 0.9% in EEIC; similarly, *in addition* with 3.1% in NSC is only 0.6% in EEIC. This also explains why the use of *and* is so high in EEIC – the possible alternatives are relatively unpopular.

There was only one additive conjunction that was present in EEIC but absent from NSC – *i mean* (0.2%). However, as this conjunction belongs to the informal register, its absence in NSC is explainable. In EEIC were not detected the following additive conjunctions *alternatively, besides, neither, not only that,* that were relatively unpopular but nevertheless present in NSC.

Based on the results drawn from the reference corpus and the Estonian–English Interlanguage Corpus, it is possible to conclude that Estonian EFL learners should be given more instruction on additive conjunctions, as *and* in comparison to other additive conjunctions is overused (64.8%) and overall variety of additive conjunctions in EEIC is smaller from NSC.

6.2. Frequencies of adversatives in EEIC

Adversatives	EEIC	EEIC %	NSC %
but	141	75%	7%
however	19	10%	35.2%
though	4	2.1%	20.7%
actually	4	2.1%	
rather	4	2.1%	4.9%
instead	3	1.5%	2.9%
yet	3	1.5%	9.9%
nevertheless	3	1.5%	2%
at the same time	2	1%	2.9%
at least	2	1%	1.6%
on the contrary	1	0.5%	
in any/either case/event	1	0.5%	1.2%
in any/either way	1	0.5%	
in fact	0		7.4%
at any rate	0		0.4%
despite this/that	0		0.4%
TOTAL	188	100%	100%

Table 4. Frequencies of adversatives in the Estonian–English Interlanguage Corpus.

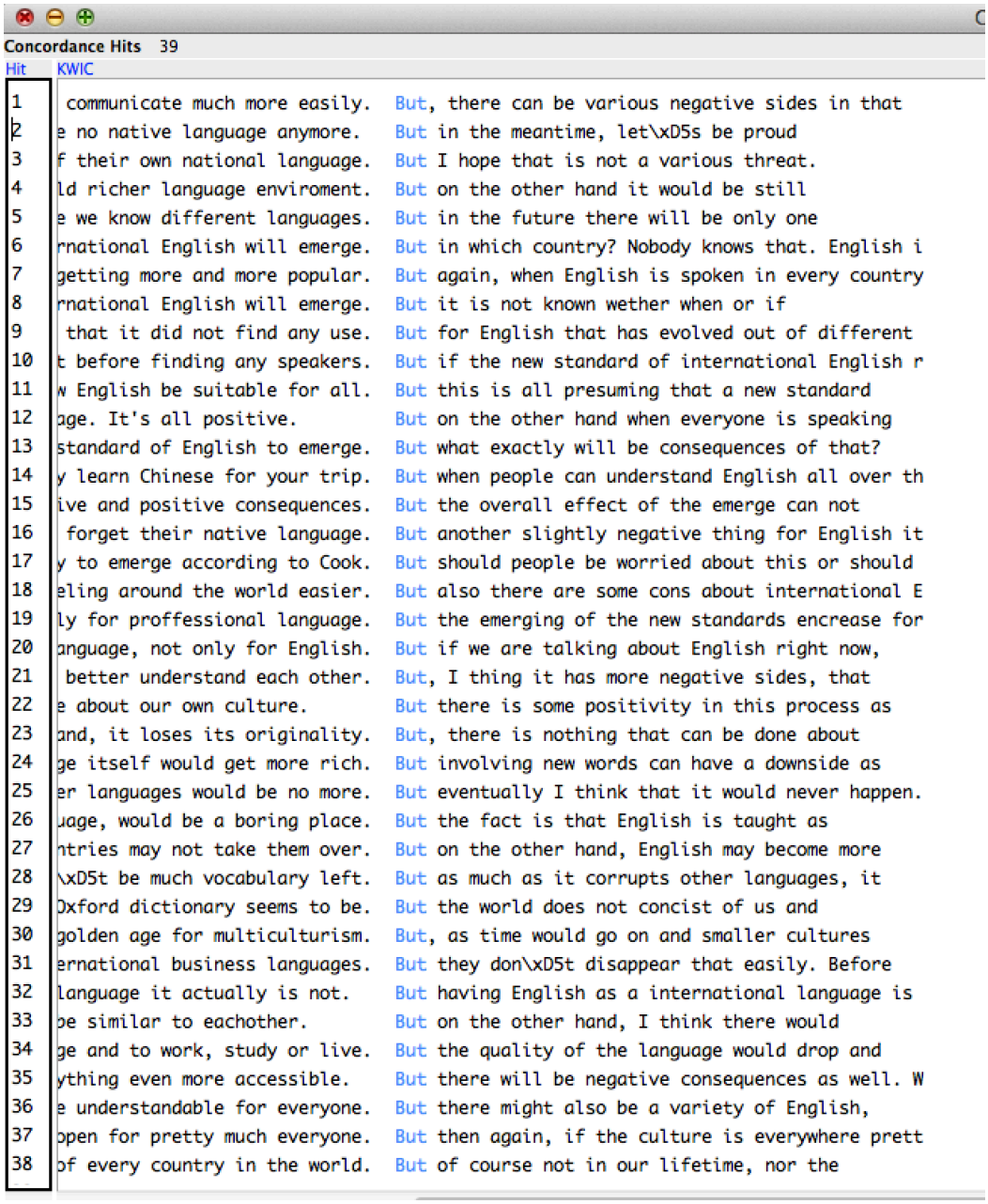
Table 4 illustrates that although EEIC and NSC both include the same amount of adversative conjunctions (in both corpora 13 variants were found), the frequencies in their use are significantly different. The most popular adversative conjunctions in EEIC are *but* with 75% and *however* with 10%. In NSC the most popular adversative conjunctions are *however* with 35%, *though* with 20.7% and *yet* with 9.9%.

Significant contrasts in the frequencies can be found in the use of *but* (75% in EEIC and only 7% in NSC), *though* (20.7% in NSC and only 2.1% in EEIC), *yet* (9.9% in NSC and only 1.5% in EEIC) and *in fact* (7.4% in NSC but absent in EEIC).

The following adversative conjunctions were not found in EEIC – *in fact* (7.4% in NSC), *at any rate* (0.4% in NSC) and *despite this/that* (0.4% in NSC). Correspondingly, in NSC were not found the adversative conjunctions *actually* (2.1% in EEIC), *on the contrary* (0.5% in EEIC) and *in any/either way* (0.5% in EEIC).

It is evident from the results that Estonian EFL students tend to overuse the adversative conjunction *but,* because only one conjunction (*however*) equals 10% and the use of other conjunctions is lower than 3%. It is also evident from the results that native speakers of English use different adversative conjunctions simultaneously, because percentages are distributed evenly in comparison to results drawn from EEIC.

The concordance list (Figure 4) revealed that in 39 sentences (out of 141) the conjunction *but* was used at the beginning of the sentence. For the reason that the conjunction *but* is a coordinating conjunction (used to join words, phrases and clauses that are balanced as logical equals), the use of it in the beginning of the sentence (in academic writing) is considered informal and thus erroneous in the context of argumentative essays.



Concordance Hits 39

Hit	KWIC	
1	communicate much more easily.	But, there can be various negative sides in that
2	e no native language anymore.	But in the meantime, let\xD5s be proud
3	f their own national language.	But I hope that is not a various threat.
4	ld richer language enviroment.	But on the other hand it would be still
5	e we know different languages.	But in the future there will be only one
6	rnational English will emerge.	But in which country? Nobody knows that. English i
7	getting more and more popular.	But again, when English is spoken in every country
8	rnational English will emerge.	But it is not known wether when or if
9	that it did not find any use.	But for English that has evolved out of different
10	t before finding any speakers.	But if the new standard of international English r
11	w English be suitable for all.	But this is all presuming that a new standard
12	age. It's all positive.	But on the other hand when everyone is speaking
13	standard of English to emerge.	But what exactly will be consequences of that?
14	y learn Chinese for your trip.	But when people can understand English all over th
15	ive and positive consequences.	But the overall effect of the emerge can not
16	forget their native language.	But another slightly negative thing for English it
17	y to emerge according to Cook.	But should people be worried about this or should
18	eling around the world easier.	But also there are some cons about international E
19	ly for proffessional language.	But the emerging of the new standards encrease for
20	anguage, not only for English.	But if we are talking about English right now,
21	better understand each other.	But, I thing it has more negative sides, that
22	e about our own culture.	But there is some positivity in this process as
23	and, it loses its originality.	But, there is nothing that can be done about
24	ge itself would get more rich.	But involving new words can have a downside as
25	er languages would be no more.	But eventually I think that it would never happen.
26	uage, would be a boring place.	But the fact is that English is taught as
27	ntries may not take them over.	But on the other hand, English may become more
28	\xD5t be much vocabulary left.	But as much as it corrupts other languages, it
29	Oxford dictionary seems to be.	But the world does not concist of us and
30	golden age for multiculturism.	But, as time would go on and smaller cultures
31	ernational business languages.	But they don\xD5t disappear that easily. Before
32	language it actually is not.	But having English as a international language is
33	be similar to eachother.	But on the other hand, I think there would
34	ge and to work, study or live.	But the quality of the language would drop and
35	ything even more accessible.	But there will be negative consequences as well. W
36	e understandable for everyone.	But there might also be a variety of English,
37	pen for pretty much everyone.	But then again, if the culture is everywhere prett
38	of every country in the world.	But of course not in our lifetime, nor the

Figure 4. Concordance of the adversative conjunction *but* in the beginning of the sentences in EEIC.

It is possible to conclude from the results that Estonian EFL students should be given more instruction on adversative conjunctions in order to avoid the use of *but* (75%). In contrast to the use of *but* (75%) and *however* (10%) the other variants of adversative conjunctions are significantly underused in comparison to reference corpus.

6.3. Frequencies of causals in EEIC

Causals	EEIC	EEIC%	NSC%
because	113	48.7%	6%
so	42	18%	20.4%
then	54	23.2%	26.5%
thus	9	3.8%	7.5%
therefore	7	3%	14.3%
as a result	4	1.7%	12.8%
in that case	2	0.8%	
because of this	1	0.4%	3%
otherwise	0		4.5%
hence	0		0.7%
for this reason	0		3%
consequently	0		0.7%
for that reason	0		0.7%
it follows	0		0.7%
TOTAL	**232**	**100%**	**100%**

Table 5. Frequencies of causals in the Estonian–English Interlanguage Corpus.

Table 5 indicates that the variety of causal conjunctions is larger in NSC (altogether 13 variants) rather than in EEIC (altogether 8 variants). The most popular causal conjunctions in EEIC are *because* with 48.7%, *then* with 23.2% and *so* with 18%. In NSC the most popular causal conjunctions are *then* with 26.5%, *so* with 20.4%, *therefore* with 14.3% and *as a result* with 12.8%.

In NSC can be found numerous causal conjunctions that are absent in EEIC, namely *in that case, otherwise, hence, for this reason, for that reason, consequently, it follows*. The biggest contrasts in frequencies between corpora can be noted in the use of *therefore* (14.3% in NSC and only 3% in EEIC), *as a result* (12.8% in NSC and only 1.7% in EEIC) and *thus* (7.5% in NSC and only 3.8% in EEIC).

It is evident from the results that Estonian EFL learners often use the causal conjunction *because* (48.7%) in their writing. Native speakers use the conjunction *because* less (6%). The distribution of *but* is even across the EEIC (Figure 5 illustrates where the conjunction appears in the corpus). For the reason that the other causal conjunctions, such as *so* and *then* are also relatively popular in EEIC and the contrast is not as high, the overuse of *because* was not detected.

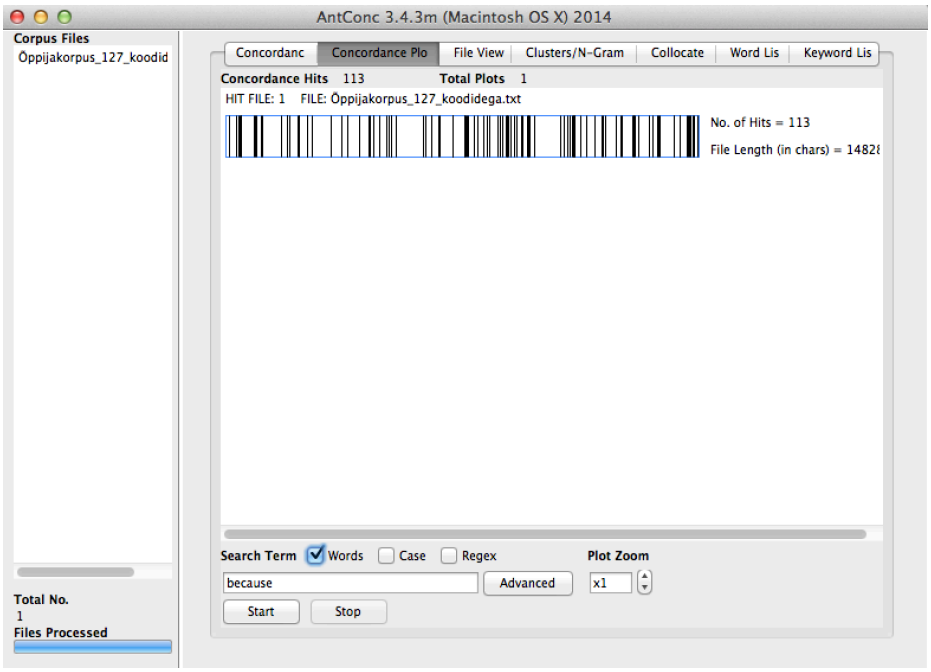


Figure 5. The distribution of causal conjunction *because* in the Estonian–English Interlanguage Corpus.

Nevertheless, the relatively high number of occurrences of *because* (48.7%) and significantly low variability of other variants of causal conjunctions in EEIC suggests that Estonian EFL learners should be given more instruction regarding the alternative variants of causal conjunctions.

6.4. Frequencies of temporals in EEIC

Temporals	EEIC	EEIC%	NSC%
second(ly)	25	19.6%	
firstly	24	18.8%	7.9%
in conclusion	20	15%	0.7%
soon	13	10.2%	
to sum up	11	8.6%	
eventually	10	7.8%	
finally	6	4.7%	12.9%
in the end	5	3.9%	
here	3	2.3%	28.7%
at first	2	1.5%	0.7%
at the same time	2	1.5%	5%
lastly	2	1.5%	0.7%
anyway	1	0.7%	
at this point	1	0.7%	5%
before that	1	0.7%	
next	1	0.7%	5%
in short	0		0.7%
meanwhile	0		2.1%
previously	0		2.8%
TOTAL	**127**	**100%**	**100%**

Table 6. Frequencies of temporals in the Estonian–English Interlanguage Corpus.

Table 6 shows that EEIC consists of more temporal conjunctions (altogether 17 variants) than the NSC (altogether 13 variants). In EEIC the most popular temporal conjunctions are *secondly* with 19.6%, *firstly* with 18.8% and *in conclusion* with 15%. In NSC, the most popular temporal conjunctions are *here* with 28.7%, *firstly* with 7.9% and *finally* with 12.9%.

The variety of temporal conjunctions differs across corpora greatly. In EEIC were not detected temporal conjunctions such as *in short, meanwhile* and *previously*. Correspondingly, temporal conjunctions such as *secondly, soon, to sum up, eventually, in the end, anyway, before that, from now on* were not detected in NSC.

It is possible to conclude from the results that Estonian EFL learners are more consistent in using temporal conjunctions *firstly* and *secondly* (Figures 6 and 7) in their topic development (sequential relation between paragraphs). Similarly, students are consistent in using conjunctions such as *in conclusion* or *to sum up* in writing conclusions.

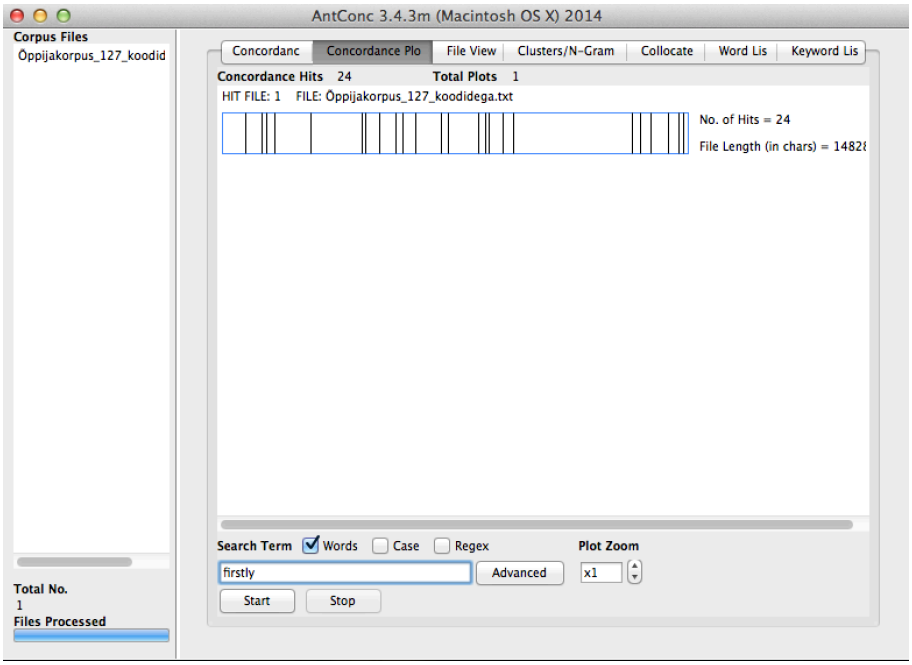


Figure 6. Distribution of temporal conjunction *firstly* in EEIC.

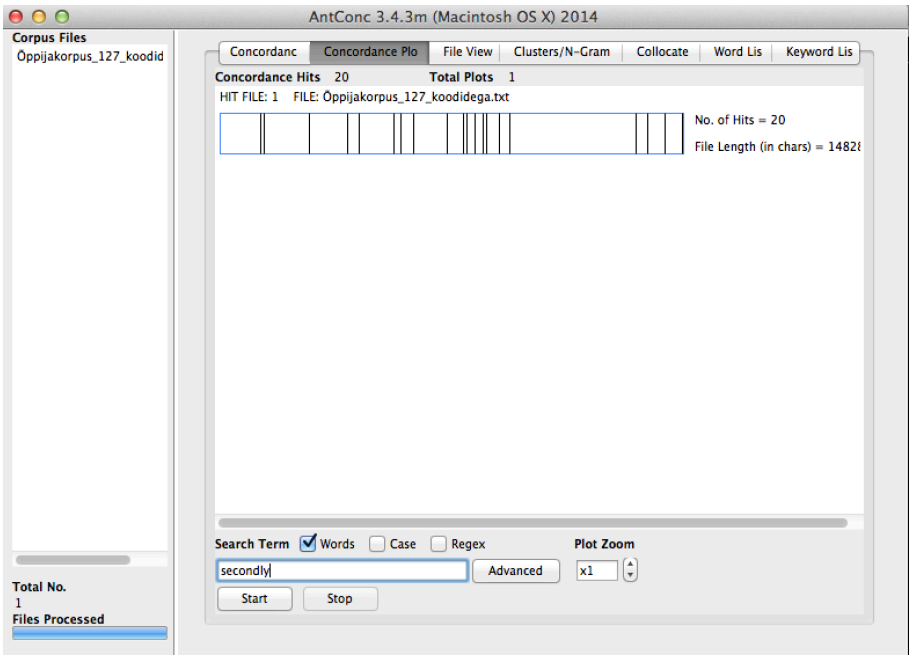


Figure 7. Distribution of temporal conjunction *secondly* in EEIC.

Based on the results adopted from reference corpus, native speakers use temporal conjunction *firstly*, but prefer to proceed with using temporal conjunctions such as *next* and *at the same time* in their topic development and typically use *finally* to express conclusive information in their essays.

It is apparent from the results that Estonian EFL learners' command of temporal conjunctions is systematic, because students are consistent in building logical relations with the certain pattern. It is also important that the category of temporal conjunctions is the only category where Estonian ESL learners' variety of conjunctions was larger from native speakers' list.

6.5. Summary of the most frequent conjunctive adjuncts in EEIC and NSC

The quantitative analysis of conjunctive adjuncts has shown that all four semantic categories of conjunctive adjuncts proposed by Halliday and Hasan (1997) were present in the Estonian–English Interlanguage Corpus. Table 8 presents the most frequent conjunctive adjuncts (that comprise at least 10% in corresponding category) that were found in the Estonian–English Interlanguage Corpus and the Michigan Corpus of Upper–level Student Papers.

As can be concluded from the results, Estonian EFL learners and native speakers of English use different conjunctive adjuncts in written essays, because the conjunctions occur at different frequencies across corpora. The overall variability of conjunctive adjuncts is also different between corpora and reference corpus includes slightly more variants of conjunctive adjuncts than EEIC.

Additives	**EEIC**	**NSC**
	and	also
	also	thus
	or	for example
Adversatives	**EEIC**	**NSC**
	but	however
	however	though
		yet
Causals	**EEIC**	**NSC**
	because	then
	then	so
	so	therefore
		as a result
Temporals	**EEIC**	**NSC**
	secondly	here
	firstly	firstly
	in conclusion	finally

Table 8. The most popular conjunctive adjuncts in EEIC and NSC.

NSC included more variants of conjunctive adjuncts in the categories of additive and causal conjunctions. It is important to note that the variability concerning the causals was almost twice as high in NSC than in EEIC. However, the frequencies of causal conjunctives did not reveal any variant that was particularly overused.

In the category of additive conjunctions, the overuse was revealed in the use of conjunction *and* (64.7%). For the reason that alternative variants of *and* were relatively unpopular (13% or less), the high number of occurrences of this conjunction may refer to avoidance strategy – students are either unaware of other variants or feel insecure about using them.

The category of adversative conjunctions was equal regarding the variability between the corpora. However, in EEIC was detected a strong overuse of adversative *but,* that resulted in total of 75%. The study revealed that 39 occurrences out of 141 with the

adversative *but* were erroneous, because students used the conjunction at the beginning of the sentence. The overuse of *but* thus marks a critical aspect where better instruction should be given to students. The alternative conjunctions for *but* which are common in native speakers' writing are *though, yet, instead* or *in fact*.

The category of temporal conjunctions was the only category where the variability of conjunctions in EEIC was bigger from NSC. In addition, Estonian EFL learners were consistent in using a certain conjunctive pattern in their writing – *firstly* and *secondly* in topic development and *in conclusion* or *to sum up* in writing conclusions. Estonian EFL learners' command of temporal conjunctives can be characterised as systematic.

7. CONCLUSION

With the help of computer and various corpus analysis tools (linguistic software), learner corpus allows to perform quick and efficient manipulation of data through the elicitation and analysis of various linguistic features as well as their diverse manifestations. There are numerous aspects in learner language that can be investigated with the help of a learner corpus. Quick and efficient manipulation of the data via computer constitutes a reliable base for describing learners' authentic language – interlanguage. In Estonia, computer learner corpus research is relatively young and for the current thesis was compiled the Estonian–English Interlanguage Corpus that allowed to provide insights into Estonian EFL learners' writing.

In order to extract objective and meaningful linguistic information and to recognise the theoretical and practical potential of computer learner corpus, it is necessary to be acquainted with the basic corpus design criteria. Chapter 2 was devoted to describing corpus compilation principles that were considered and applied to the design of the Estonian–English Interlanguage Corpus. Although the default corpus design criteria are difficult to be secured, this chapter discussed their importance and took the possible limitations into account. Chapter 3 acquainted the reader with the corpus research method and introduced the *AntConc* software that was utilised in the analysis of the Estonian–English Interlanguage Corpus.

The Estonian–English Interlanguage Corpus consisted of 127 essays (24,796 tokens) that were written by Estonian secondary school graduates upon entering the Department of English Studies in the University of Tartu in 2014. Although the Estonian–English Interlanguage Corpus is obviously very small in comparison to the currently assembled native as well as non-native corpora worldwide, it is nevertheless the first step in the direction of investigating Estonian–English interlanguage.

The current thesis employs corpus-based quantitative analysis and investigates the distribution and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus. Two functionalities were primarily employed in approaching and analysing the Estonian–English Interlanguage Corpus – the concordance tool and the word list tool. The use of the linguistic software (the *AntConc* programme) enabled to observe and compare frequencies and variability of the conjunctive adjuncts found in the Estonian–English Interlanguage Corpus to the reference corpus.

Cohesion is a prerequisite for conveying ideas clearly and it enables writers to convey their knowledge to the intended readership effectively (Rummel 2010: 21). The use of conjunctive adjuncts plays an important role in building connections between the ideas in text and is therefore an essential aspect in successful argumentative writing. For the reason that Estonian secondary school graduates are required to write coherent argumentative essays by the end of their secondary school education[6] it is thus relevant to study and analyse their written essays in regard to the use of conjunctive adjuncts.

The study revealed that Estonian EFL learners use various conjunctive adjuncts in their writing and four basic semantic categories of conjunctions brought out by Halliday and Hasan (1976) were present in the Estonian–English Interlanguage Corpus. In comparison to NSC, the variability of additive and causal conjunctions was smaller in EEIC. Variability was equal in the category of causal conjunctions and only in the category of temporal conjunctions, Estonian EFL learners' variability was larger from native speakers' corpus.

The study revealed that Estonian ESL learners tend to frequently use the additive conjunction *and* (64.7%). In addition, among frequently occurring additive conjunctions were *also* (13%) and *or* (8.8%). The occurrence of other additive conjunctions was

[6] The national curriculum for upper secondary schools. Available at https://www.riigiteataja.ee/en/eli/524092014009/consolide (28.04.2015)

insignificant (3.9% or less). The high number of occurrences of the additive conjunction *and* may refer to avoidance strategy – students are either unaware of other variants or feel insecure about using them. The possible alternative variants for *and* could be *thus, moreover* or *in addition* that were relatively unpopular in EEIC in comparison to NSC. Also, various alternatives could be adopted from NSC that were absent from EEIC, namely *besides, likewise, similarly, alternatively.*

In the category of adversative conjunctions, the variability of conjunctions found in EEIC and NSC was equal. However, an overuse was witnessed in the use of *but* (75%). The study revealed that 27.6% of the total occurrences of this conjunction were erroneous. It is therefore necessary to provide instruction on the category of conjunctive adjuncts, because Estonian ESL learners' should be more accurate in the use of the conjunction *but.* The significant overuse of *but* may also refer to avoidance strategy, when students are either unaware of other variants or feel insecure about using them in their writing. To avoid the overuse of the conjunction *but,* such alternatives could be offered as *though, yet* (commonly used by native speakers). In addition, various alternative conjunctions from NSC could be adopted – *in fact, despite this/that.*

The variability concerning the causals was almost twice as high in NSC rather than in EEIC. The frequency of the causal *because* was found to be relatively high (48.7%). However, for the reason that such causal conjunctions as *then* (23.2%) and *so* (18%) were also relatively popular in EEIC, the causal *because* was not revealed to be overused. Nevertheless, Estonian EFL learners should be offered alternative variants of causal conjunctions, because the variability of causals in EEIC is considerably smaller from NSC.

The study revealed that Estonian EFL learners are systematic in regard to temporal conjunctions. Learners consistently used a certain pattern of such temporal conjunctions as *firstly* and *secondly* in their topic development and *in conclusion* or *to sum up* in writing

conclusions. The category of temporal conjunctions was the only category where total variability of conjunctions was higher in EEIC.

The Estonian–English Interlanguage Corpus that was compiled for current thesis allowed to observe the written essays of Estonian EFL learners and bring out quantitative results regarding the variability and frequency of conjunctive adjuncts. The comparative aspect was produced with the help of the results adopted from the study conducted by Hussein and Muddhi (2014). Such an approach allowed analysing and indicating the differences between two corpora and proposing possible areas of improvement for Estonian EFL learners.

8. REFERENCES

Biber, Douglas, Conrad, Susan and Leech, Geoffrey. 2002. *Student Grammar of Spoken and Written English*. Longman.

Blanpain, Kristin. 2012. *Academic Writing: A Resource for Researchers*. Leuven/Den Haag: Acco.

Bowker, Lynne and Pearson, Jennifer. 2002. *Working with Specialized Language: A practical guide to using corpora*. London and New York: Routledge.

Braun, Sabine. 2006. ELISA: a pedagogically enriched corpus for language learning purposes. *Corpus linguistics and language pedagogy*, Volume 3: 1–5. Europäischer Verlag der Wissenschaften: Peter Lang.

British National Corpus. Available at http://corpus.byu.edu/bnc/, Accessed April 9, 2015.

Center of Estonian Language Resources. Available at http://keeleressursid.ee/en, Accessed February 28, 2015.

Common European Framework of Reference for Languages. Available at http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf, Accessed April 4, 2015.

Cook, Guy W. D. 2008. *Applied Linguistics*. Oxford: Oxford University Press.

Crewe, W.J. 1990. The illogic of logical connectives. *ELT Journal.* 44(4): 316–325.

Dagneaux, Estelle, Denness, Sharon and Granger, Sylviane. 1998. Computer–aided error analysis. *System*, 25: 163–174.

Eslon, Pille and Metslang, Helena. 2007. Õppijakeel ja eesti vahekeele korpus. *Eesti rakenduslingvistika ühingu aastaraamat 3*. 99–116.

Estonian National Curriculum for Secondary School. 2011. Available at https://www.riigiteataja.ee/en/eli/524092014009/consolide, Accessed April 28, 2015.

Fakhra, Amani. 2009. Relative clauses and conjunctive adjuncts in Syrian University student writing in English. University of Warwick.

Flowerdew, Lynne. 2012. *Corpora and Language Education*. London: Palgrave Macmillan.

Gass, Susan M. and Selinker, Larry. 2008. *Second Language Acquisition.* New York and London: Routledge.

Granger, Sylviane. 1998. The computer learner corpus: a versatile new source of data for SLA research. In Granger, Sylviane. *Learner English on Computer.* 3–18. London and New York: Longman.

Granger, Sylviane. 2002. A Bird's–eye view of learner corpus research. In Granger, S., Hung, J. and Petch–Tyson, S. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* 3–33. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Granger, Sylviane. 2004. Computer Learner Corpus Research: Current Status and Future Prospects. In Connor, U., Upton, T. (eds.) *Applied corpus linguistics: a multidimensional perspective.* 123–145. Amsterdam and Atlanta: Rodopi.

Granger, Syviane. 2012. How to use foreign and second language learner corpora.

In Mackey, A. and Gass, S. (eds.) *Research Methods in Second Language Acquisition: A Practical Guide.*

Granger, Sylviane and Tyson, Stephanie. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *Word Englishes.* 15(1): 17–27.

Halliday, M. A. K. and Hasan, Ruqaiya. 1976. *Cohesion in English.* London and New York: Longman.

Hunston, Susan. 2006. *Corpora in Applied Linguistics.* Cambridge University Press.

Lancaster University. Statistics in corpus linguistics. Available at http://corpora.lancs.ac.uk/clmtp/2-stat.php, Accessed April 4, 2015.

Laurence Anthony web–page. Available at http://www.laurenceanthony.net, Accessed March 3, 2015.

Laurence, Anthony. 2005. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. *IEEE International Professional Communication Conference Proceedings*. 729–737.

Leech, Geoffrey. 1992. Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82. In Svartvik, Jan (ed.). *Corpora and theories of linguistic performance*. 105–122 Mouton de Gruyter, Berlin/New York.

Leech, Geoffrey. 1998. Preface. In Granger, Sylviane. *Learner English on Computer.* xiv–xxii. London and New York: Longman.

McEnery, Tony. Wilson, Andrew. 2001. *Corpus Linguistics. An introduction.* Edinburgh University Press.

Mudhhi, Saud K. and Hussein, Riyad F. 2014. A corpus–based study of conjunctive Adjuncts in the Writings of Native and Non–native Speakers of English. *English Linguistics Research,* vol.3, no. 2, 18–30.

Mukherjee, Joybrato. 2006. Corpus linguistics and language pedagogy: The state of the art – and beyond. *Corpus Technology and Language Pedagogy,* Volume 3: 5–24. Europäischer Verlag der Wissenschaften: Peter Lang.

O'Keeffe, Anne, McCarthy, Michael and Carter, Ronald. 2007. *From Corpus To Classroom: Language Use and Language Teaching*. Cambridge press.

Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal*. 26: 81–114.

Reppen, Randi. 2012. Building a corpus. What are the key considerations? In O'Keeffe, Anne and McCarthy, Michael. *The Routledge Handbook of Corpus Linguistics.* London and New York: Routledge. 31–37.

Römer, Ute and Wulff, Stefanie. 2010. Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research.* 2(2): 99–127.

Rummel, Kärt. 2010. *Creating Coherent Texts in English as a Foreign Language: Theory and Practice*. Tartu University Press.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics.* 10: 209–241.

Sinclair, John. 1996. *Preliminary recommendations on Corpus Typology*. Eagles.

Song, Lichao. 2012. On the variability of Interlanguage. *Theory and Practice in Language Studies.* 2(4): 778–783.

Sharkh, Abu B. 2012. Cohesion and coherence in the essay writing of Palestinian college students. Hebron University.

Suswati, Susi, Sujatna, Sari, Eva, Tuckyta and Mahdi, Sutiono. 2014. Additive Conjunction Choice in English Children Short Stories: A Syntactic and Semantic Analysis. *International Journal of Language Learning and Applied Linguistics World (IJLLALW)*. 5(4): 11–21.

Tapper, Marie. 2005. Connectives in advanced Swedish EFL learners' written English: Preliminary results. *The Department of English: Working Papers in English Linguistics.* 5: 116–144.

Tono, Yukio. 2003. Learner corpora: design, development and applications. *UCREL Technical Paper number 16. Special issue.* 800–809.

Tognigni–Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Vajjala, Sowmya and Lõo, Kaidi. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. *Proceedings of the third workshop on NLP for computer–assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings, 107: 113–127.

The list of learner corpora around the world. Available at http://www.uclouvain.be/en-cecl-lcworld.html, Accessed March 5, 201.

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Elina Merilaine

The frequency and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus / Sidesõnade variatiivsus ja sagedus eesti–inglise vahekeele korpuses
2015
62 lk

Magistritöö eesmärgiks oli luua Eesti esimene eesti–inglise vahekeele korpus ning tutvustada selle loomise- ning uurimispõhimõtteid. Kitsamalt uuriti sidesõnade variatiivsust ning sagedust. Tulemusi analüüsiti ning seejärel võrreldi inglise keelt emakeelena kõnelevate õppijate korpusega, milleks oli *Michigan Corpus of Upper–level Student Papers* (Michigani kõrgeima taseme kirjalike tööde õppijakorpus).

Töö koosnes neljast osast. Magistritöö esimene ja teine osa keskendusid korpuse loomise põhimõtetele ning tutvustati ka korpusuurimuse ülesehitust. Arutleti selliste aspektide olulisuse üle nagu kvantiteet, kvaliteet, dokumentatsioon ning lihtsus. Igat aspekti analüüsiti, tuues välja tugevad ja nõrgad küljed ning võimalikud kitsaskohad.

Magistritöö empiirilise osa läbiviimiseks (kolmas ja neljas osa) kasutati vabatarkvara *AntConc*, mis võimaldas luua statistilist andmestikku, mille tulemusi hiljem analüüsiti. Uuringutulemused näitasid, et Eesti õpilased kasutavad erinevaid sidesõnu, mis kuuluvad viide kategooriasse Halliday ja Hasani (1976) jaotuse järgi.

Uurimustulemuste põhjal on näha, et Eesti õpilased on järjekindlad selliste sidesõnade kasutamisel nagu *firstly, secondly, in conclusion* ja *to sum up.* Uuringu käigus tuvastati järgmiste sidesõnade ülekasutus – *but* ja *and.* Sidesõna *but* kasutamist võib hinnata problemaatiliseks, sest õpilased eksisid korduvalt selle kasutamises (asetades sidesõna lause algusesse).

Kokkuvõtteks võib öelda, et sidesõnade variatiivsuse õpetamine Eesti õpilastele aitaks kaasa koherentsuse tagamisel argumentatiivse teksti kirjutamisel. Abiks tuleks emakeelt kõnelevate õppijate korpusest sidesõnade laenamisest, sest seal oli üldine variatiivsus võrreldes eesti-inglise vahekeele korpusega suurem.

Märksõnad: Korpuslingvistika, korpusuuringud, inglise-eesti vahekeele korpus, sidesõnad, variatiivsus ja sagedus, sidesõnade õpetamine

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Elina Merilaine,

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

The frequency and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus**,**

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, **20.05.2015**