UNIVERSITY OF TARTU

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

THE USE OF ADJECTIVES AND ADVERBS IN ESTONIAN AND BRITISH STUDENT WRITING: A CORPUS COMPARISON

MA thesis

ANNA DANIEL

SUPERVISOR: ÜLLE TÜRK (MA)

TARTU

2015

Abstract

Corpus analysis of learner language is currently being conducted in various parts of the world and the field is gaining momentum (Granger 2004: 123; 2015: 9). Since the interlanguage of Estonian learners of English has not yet been studied on a larger scale, this thesis aims to be part of filling this gap. It examines the use of adjectives and adverbs in Estonian and British student essays. The aim is to compare Estonian learners' usage to native-speaker usage from the perspectives of lexical variation and sophistication, the proportion of academic words and the types of adjectives and adverbs used.

The two corpora under scrutiny are Estonian-English Interlanguage Corpus and Louvain Corpus of Native English Essays. The methodology used is corpus comparison, which was conducted with the aid of the tool AntConc (Anthony 2014) and the lexical profilers available at the Lextutor website (Cobb 2002).

The thesis is structured into two main chapters. The Literature Review of Corpus Research on Advanced Interlanguage covers research on late interlanguage at the end of the 1990s and in the 2000s with a focus on vocabulary studies; the uses of word frequencies in analysing learner language together with the pitfalls that they entail; and aspects to consider when compiling and comparing corpora. The second chapter reports on the empirical study and begins by explaining the methodology and tools, describing the data and their limitations, and detailing the measures that were taken to make the two corpora more comparable. This is followed by the presentation and interpretation of the findings. The results show that learners use less varied, sophisticated and academic adjectives and adverbs than native speakers.

Table of Contents

Abstract 2

Introduction 4

1. Literature Review of Corpus Research on Advanced Interlanguage 10

1.1. Late Interlanguage Studies 10

1.2. The Use and Caveats of Word Frequencies 15

1.3. Comparable Corpora 17

2. Corpus Comparison of Adjectives and Adverbs used in EEIC and LOCNESS 22

2.1. Methodology and Tools 22

2.2. Data and Limitations 23

2.3. Exclusions from the Lists of Adjectives and Adverbs 28

2.4. Results 32

2.4.1. Lexical Variety and Sophistication in the Use of Adjectives 32

2.4.2. Lexical Variety and Sophistication in the Use of Adverbs 35

2.4.3. Twenty Most Common Adjectives and Adverbs 37

2.4.4. Types of Adjectives and Adverbs 40

2.5. Discussion 42

Conclusion 47

References 49

Appendix 1. Entrance Examination 2014: Task Description and Source Text 54

Appendix 2. Misspelt Adjectives and Adverbs 57

Resümee 58

Introduction

Corpus research can be said to have begun with the completion and distribution of the Brown Corpus in 1964 (Leech 2011: 10). This first computerised corpus contained native-English-speaker texts and gave rise to the first word frequency lists and corpus-based studies of language. To complement this one-million-word corpus of American English, the Lancaster-Oslo/Bergen (LOB) Corpus of British English was published in 1976 (Leech 2011: 10), permitting comparison of American and British English. At the time, one million words were considered abundant for research purposes though by today's standards and in comparison with the British National Corpus (BNC) of 100 million words, for instance, the early corpora are minuscule (Granger 1998: 4).

The advent of corpora led to the development of a field of study called 'corpus linguistics', which is a methodology founded on the use of electronic collections of naturally occurring texts, i.e. corpora (Granger 2002: 3–4). Prior to that, in the 1940s and 1950s, linguists interested in language teaching and learning mostly employed contrastive analysis and slightly later error analysis. Contrastive analysis entailed comparing two languages, native and foreign, with the aim of identifying interlingual differences that were likely to cause difficulty for learners (MacDonald 2013: 37; Hasselgård and Johansson 2011: 34). The problem with contrastive analysis was that it was restricted to interlingual transfer as the only cause for difficulty and it relied on behaviourist habit-formation theory as a means of eliminating errors from learners' production (MacDonald 2013: 37). As an alternative method, error analysis, whose heyday was in the 1960s and early 1970s, entered the scene. Error analysts sought to identify and classify learners' errors and by so doing infer what the learner has already acquired and what he still needs to learn (Hasselgård and Johansson 2011: 34–35). Though undoubtedly insightful and still in use, error analysis has

been criticised for its narrow focus on errors rather than the whole of learner production or performance, including learners' achievements (MacDonald 2013: 37).

Both of these approaches aimed at explaining features and mechanisms of second language acquisition (SLA) but fell short on several accounts. In retrospect, it may be surprising that although corpus research on native-speaker language usage had already begun to contribute to language description, research on learner language moved along a separate path for quite some time. It was not until the late 1980s and early 1990s that "academics, EFL [English as a Foreign Language] specialists and publishing houses alike began to recognize the theoretical and practical potential of computer learner corpora" (Granger 1998: 4). In broad terms, as Granger (2004: 129) states, computer learner corpora can be divided into two main categories: commercial and academic corpora. The two major commercial learner corpora are the Longman Learners' Corpus and the Cambridge Learner Corpus, which are both very big (ten million words for the Longman corpus and sixteen million for the Cambridge corpus). The academic corpora, though far more numerous, are highly variable in size, and Sylviane Granger (2004: 129–130) finds it paradoxical that despite the abundance of learner corpora, hardly any of it is available for academic research. This might be the reason why Granger's own project, the International Corpus of Learner English ("ICLE"), which contains over two million words from over ten mother-tongue backgrounds[1], has become one of the most cited learner corpora today (Hasselgård and Johansson 2011: 38).

While most corpora contain written language, corpus research has not restricted itself to the written medium. Only five years after the start of the ICLE project, the Centre for English Corpus Linguistics launched a new project called Louvain International Database of Spoken English Interlanguage ("LINDSEI"). Presenting a list of all prominent

[1] The first edition of ICLE contained 2.5 million words from eleven mother-tongue backgrounds and was published on CD-ROM in 2002. The second version (ICLEv2) was released in 2009 and contains 3.7 million words from sixteen mother-tongue backgrounds ("ICLE").

learner corpora compiled to date would be nearly impossible and for that reason suffice it to say that they are numerous, they come from various L1 backgrounds all over the world, and they continue to be analysed by individual researchers and research groups.

When discussing learner corpus research, one cannot look past the two neighbouring fields of study, namely SLA and ELT or FLT (English / Foreign Language Teaching). Rod Ellis and Gary Barkhuizen (2005: 15–23) list three types of data that can be collected from learners:

1) non-linguistic performance data (e.g. measuring learners' reaction time and comprehension, intuition about grammaticality);
2) samples of learner language;
3) reports from learners about their own learning (e.g. questionnaires, think-aloud-protocol)

They agree that the primary data for investigating L2 acquisition should be written and spoken samples of learner language. This category can, in turn, be divided into three sub-categories: naturally-occurring samples, clinically elicited samples, and experimentally-elicited samples. Granger (2002: 5) argues that much of current SLA research favours introspective and experimental data and tends to dismiss natural language use data. She points to the constraints that using such data entails, noting that "SLA specialists regularly rely on a very narrow empirical base, often no more than a handful of informants" (Granger 1998: 5), which undermines the generalisability of the results. It is in this respect that learner corpus researchers feel that they have a great deal to offer to SLA research. Granger (2002: 4) writes that computer corpus methodology is particularly suitable for conducting quantitative analyses, which was previously unfeasible or at least very time-consuming.

Given the increasing number of people who speak English as a foreign language, the link between learner corpus research and FLT is equally as important as the connection with SLA. In her overview of learner corpus data and ELT, Granger (1998: 6–7) relates that native English corpora began to inform ELT materials design in the 1990s, with Collins Cobuild's pioneering dictionary project. This approach was driven by the understanding that better descriptions of authentic native English lead to better EFL materials. While this is certainly true, native corpora can provide no information about which structures and concepts are difficult for learners. Therefore, Granger advocates for the parallel use of authentic native and authentic learner data in materials and curriculum design. In this way, the native corpus would highlight what is typical in English, and the learner corpus would help clarify what poses problems for learners.

Much of the work in the English-speaking community has naturally focused on learner English. In Estonia, however, it is the interlanguage[2] of Russian speakers learning Estonian that has been studied most widely so far. The Estonian Interlanguage Corpus (EIC) comprises written Estonian language texts and contains around 500,000 tokens (Eslon and Metslang 2007: 106). However, the interlanguage used by Estonian learners of English has not yet been studied on a larger scale and to the best of the author's knowledge, this thesis is one of the first attempts in this direction.

The author's interest in students' use of adjectives and adverbs arose from the personal perception (and that of several fellow EFL teachers) that secondary school students tend to use quite a limited range of adjectives and adverbs in English classes. Teachers of English complain that students often resort to using simple adjectives, such as *good*, *bad* and *interesting*, which, though widely applicable, lack precision and detail. Corpus research has indicated (Granger 1998; Cobb 2003) that there is a strong case for an

[2] The terms 'interlanguage' and 'learner language' are used interchangeably throughout the thesis as such is also the practice in this field of study.

overuse hypothesis in learner language in the sense that learners of a language tend to use a narrow set of rather common words more frequently than native speakers.

Considering the lack of Estonian-English interlanguage corpora and the perceived shortcomings of learners in using adjectives and adverbs, the aim of this thesis is to characterise and analyse the use of adjectives and adverbs in Estonian student writing with the aid of a learner corpus and a native-speaker (NS) corpus, and to compare NS and non-native-speaker (NNS) usage of adjectives and adverbs. The two corpora employed are Estonian-English Interlanguage Corpus (EEIC) and Louvain Corpus of Native English Essays (LOCNESS). The research questions are as follows:

- Which corpus uses a more varied selection of adjectives and adverbs?
- Which corpus uses more sophisticated adjectives and adverbs?
- Which corpus uses more academic adjectives and adverbs?
- What types of adjectives and adverbs are used and how do they differ in the two corpora?
- Which adjectives and adverbs tend to be overused or underused in EEIC?

Longman Grammar of Spoken and Written English (LGSWE, Biber et al. 1999: 64–65) distinguishes between adverbs, which are most typically in the role of modifiers, and adverbials, which function as elements of the clause. In the present thesis such distinction is not made partly due to technical reasons discussed in detail in section 2.3. and partly because it is not considered essential in the light of the main focus of the thesis

The thesis is structured into two main chapters. Following the Introduction, the Literature Review of Corpus Research on Advanced Interlanguage is divided into three sections (1.1.–1.3.). The first gives an overview of research on late interlanguage at the end of the 1990s and in the 2000s with a focus on vocabulary studies. The second section

discusses the uses of word frequencies in analysing learner language, together with the pitfalls that they entail. The third section focuses on aspects, such as size and design, to consider when compiling and comparing corpora. After the theoretical discussion, the second chapter reports on the empirical study conducted as part of the thesis. The first three sections (2.1.–2.3.) explain the methodology and tools, describe the data and their limitations, and finally, explain what kinds of measures were taken to make the two corpora more comparable. The fourth section (2.4.) presents the findings of the study in four sub-sections dealing with the lexical variety and sophistication of adjectives and adverbs, the twenty most common adjectives and adverbs in both corpora, and the types of adjectives and adverbs used. The final section (2.5.) in Chapter 2 relates the findings of the study to theoretical considerations and previous research in the field.

The author of the thesis would like to thank her supervisor for her comments and guidance throughout the process and all the lecturers and professors at the department of English for their support on the way to becoming a better teacher of English.

1. Literature Review of Corpus Research on Advanced Interlanguage

The aim of this chapter is to bring together theoretical discussions and reports on empirical research in the field of advanced interlanguage. Firstly, an overview of what has been done in late interlanguage studies so far is given. Secondly, the uses of word frequencies and frequency lists are discussed. The final section deals with corpus building and aspects to consider when choosing comparable corpora.

1.1. Late Interlanguage Studies

The term 'interlanguage' originates from Larry Selinker (1972) and entails the understanding that "learner language displays systematicity and opportunity for intelligent intervention rather than random error" (cited in Cobb 2003: 394). This applies to language learners on all levels of proficiency, including advanced learners. Though this idea was expressed forty years ago, researchers are still investigating what this systematicity means for late interlanguage. In 2003, Cobb (394, 396) argued that intermediate and advanced interlanguage remained relatively uncharted and that, until then, the main advice given to advanced learners was simply to get lots of practice without any specific focus. Cobb (2003: 396) justly posits that "[i]f, instead, advanced learners are seen as learners nonetheless, moving systematically through acquisition sequences and overcoming shared misconceptions about the L2, then instruction can be focused more effectively". What, then, has been written on the subject of late interlanguage?

In 1998 an influential volume on computer-aided learner language analysis was published. Several contributors to this collection of articles describe advanced interlanguage as "vague and stereotyped"; "dull, repetitive and unimaginative" (Ringbom 1998: 49–50); generating "an impression of 'non-nativeness' or 'lack of idiomaticity'"

(Lorenz 1998: 53). In fact, the terms ‘late interlanguage’ and ‘advancedness’ have not yet been clearly defined. A research group at the University of Bremen lists the task of defining terms such as ‘advanced learner’ and ‘near-native competence’ as one of the issues the field is still struggling with (“Lexico-grammatical variation in advanced learner varieties”). They use the term ‘advanced learner varieties’, which hints at the plurality of forms late interlanguage can take.

It is commonly agreed that advanced learners are advanced by virtue of having mastered the basic rules of syntax and morphology, but they still have difficulty with finer points of lexico-grammar and style. A definition of advanced interlanguage is provided by Granger (2004: 135), who writes that it is “the result of a very complex interplay of factors: developmental, teaching-induced and transfer-related, some shared by several learner populations, others more specific”. The typical characteristics of late interlanguage are succinctly presented by Marcus Callies (2010) from the above-mentioned German research group:

1. overuse of high-frequency vocabulary;
2. overuse of a limited number of prefabricated patterns (prefabs);
3. a much higher degree of personal involvement;
4. stylistic deficiencies (overly spoken style, mixture of formal and informal markers).

These characteristics have been formulated on the basis of several studies on late interlanguage, some of which are considered in the following. As already mentioned, the compilation of articles edited by Granger proved to be insightful not only because of the results achieved by the contributing researchers but also because of the further investigation it has encouraged. In his 2003 article on late interlanguage, Cobb reports on his replications (and expansions) of three European learner corpus studies. These are presented below.

Firstly, Håkan Ringbom (1998: 41–52) compared the 100 most frequent words in the ICLE and the LOCNESS corpora of argumentative essays and was able to demonstrate that advanced learners across seven L1 backgrounds consistently use these 100 very high frequency words about 4–5% more than NS writers. In his replication, Cobb (2003: 398–407) confirmed this hypothesis and showed that almost 90% of vocabulary items used in writing by the Québec advanced learners are common words from the 0–1000 (or K-1) frequency range, which indicates that non-native speakers' vocabulary is less varied than that of NSs. Hasselgren (cited in Hasselgård and Johansson 2011: 40) gives an apt metaphor of this phenomenon by observing that learners cling to their "lexical teddy bears", i.e. "words they feel safe with" at the expense of more precise synonyms.

The second study expanded by Cobb (2003: 407–415) also supports the initial findings by Sylvie De Cock et al. (1998: 67–79) that although advanced learners do use prefabricated expressions, these are not necessarily the same as those used by NSs, and they might have different syntactic and pragmatic functions. Cobb found that learners have at their disposal a limited number of prefabs, which they repeat more frequently than NSs. Originally conducted by Stephanie Petch-Tyson (1998: 107–118), the third study was on writer-reader visibility. In his replication, Cobb (2003: 415–418) confirms the view that advanced learner writing resembles informal spoken language written down as it contains a much higher degree of involvement in the form of personal pronouns, references to the writer's mental processes via phrases such as 'I think', and conversational monitoring of information flow.

With respect to advanced learners' use of adjectives and adverbs in particular, it must be noted that studies with such a focus are hard to come by. However, there are three research reports also in Granger's 1998 publication worth mentioning here. The first focuses on adjective intensification, the second on the use of adverbial connectors by

advanced Swedish learners, and the third on automatic profiling of learner texts by word class.

Firstly, Gunter Lorenz (1998: 53–66) examines advanced German students' practices of adjective intensification. After he disproves three hypotheses as to why German students tend to over-intensify, he arrives at a significant conclusion that one of the reasons for over-intensification of adjectives lies in the way the students structure information within the clause. He notes that much of the over-intensification happens at the beginning of a clause rather than the end, which defies the theme-rheme principle of the English language.

Secondly, Bengt Altenberg and Marie Tapper (1998: 80–93) conclude that Swedish learners' use of adverbial connectors compares fairly well to that of English students, adding that they tend to underuse conjuncts, unlike many EFL learners of other L1 backgrounds.

Thirdly, Sylviane Granger and Paul Rayson's (1998: 119–131) study builds word category profiles of LOCNESS and ICLE essays (namely the French sub-corpus) which give evidence that it is mainly short adverbs of native origin that are overused, especially those expressing place and time, and mainly *-ly* adverbs that are underused: amplifiers (greatly, truly, widely, readily, highly), disjuncts (importantly, traditionally, effectively), modal adverbs (possibly, supposedly), time adverbs ending in *-ly* (newly, currently, previously, ultimately). Yet this is the category of adverbs that academic texts use the most (Granger 1998: 128; Biber et al. 1999: 540).

Some additional fragments of research results relevant to this thesis are presented below. In a study focusing on error types and patterns on the intermediate (B-) and advanced (C-) levels, Jennifer Thewissen (2013: 88–89) points to an interesting finding concerning adverb order errors. She shows that not only do C-level learners place adverbs

more accurately in a sentence, but they also use significantly more adverbs overall than B-level learners. This shows that adverb placement develops comparatively late.

In a study on English for Academic Purposes (EAP), Gaëtanelle Gilquin et al. (2007: 328) mention register problems in learner writing, noting that learners tend to use expressions typical of speech rather than of writing. This is evidenced by their overuse of adverbs expressing a high degree of certainty, such as *really*, *of course* or *absolutely* and underuse of hedging adverbs, such as *apparently*, *possibly*, *presumably*, which are common in academic writing.

To conclude this section, there is still a great deal of ground to cover in late interlanguage research. At the beginning of his article from 2003, Cobb (2003: 394) names two reasons why late interlanguage has been studied less extensively than beginner and intermediate interlanguage: firstly, lack of data, and secondly, lack of theory. As far as data is concerned, it can be said that this problem is being tackled on a yearly basis since new learner corpora are emerging in various parts of the world and with various L1 backgrounds. As to the lack of theory, or a systematised approach to late interlanguage development, when at the beginning of the 2000s Thomas Cobb (2003: 395) expressed the hope that perhaps, as advanced learner data and studies accumulate, theories and hypotheses will start to emerge, in the middle of the 2010s, Granger (2015: 11) writes that "[a]s a result of the many CIA [contrastive interlanguage analysis] studies carried out over the years, we now have a much clearer picture of the complex interplay of lexico-grammatical and discourse features that characterizes advanced interlanguage." The present thesis seeks to be part of this mosaic of describing late interlanguage development. In order to test the overuse hypothesis on Estonian learners' language production, the concept of frequency is essential.

1.2. The Use and Caveats of Word Frequencies

Geoffrey Leech (2011: 7), one of the authors of the BNC frequency lists, writes that, "[i]f asked what is the one benefit that corpora can provide and that cannot be provided by other means, I would reply 'information about frequency'". Leech (2011: 7–8) differentiates between three usages of the term 'frequency':

a. *Raw frequency* shows how many times a linguistic phenomenon occurs in a corpus, text or collection of texts.
b. *Normalised frequency* (sometimes also called *relative frequency*) expresses frequency relative to a standard yardstick (e.g. tokens per million words).
c. *Ordinal frequency* shows how the frequency of one item is compared with the frequency of another item, thus yielding a rank frequency list, in which words are listed in the order of frequency.

Leech argues that in terms of language learning, the most useful measure of these is ordinal frequency. He explains that it is of little use for the language teacher to know that *shall* occurs 175 times per million words in a corpus, but to know that *will* is much (15 times) more frequent than *shall* is more likely to be pedagogically useful. In order to pinpoint areas where further instruction is necessary, comparing frequency data can be insightful both within a single corpus and between a learner and a reference corpus.

The key terms in discussing word frequencies are overuse, underuse and misuse or deviant patterns. It is important to note that the terms 'overuse' and 'underuse' are descriptive rather than prescriptive and simply mean that a linguistic feature is found more or less often in the learner corpus than in the reference corpus (Paquot and Granger 2012: 143). These terms have been criticised by SLA theorists and corpus linguists (cited in Granger 2015: 18–19) as being condemning and overly obsessive of the target language norm. Guy Aston (cited in Granger 2015: 19) warns against treating "all quantitative

differences from reference group behaviour as undesirable". Hasselgård and Johansson (2011: 55) also maintain that "the concepts of overuse and underuse are not straightforward, and quantitative findings need to be carefully considered and cross-checked with qualitative analyses before exposing learners to them". These are valid considerations, which will certainly benefit learner corpus research. As to terminology, Granger (2015: 19) concludes that since the terms 'overuse' and 'underuse' are now well established in the field, "they can continue to be used in their technical meanings of 'containing more or less than'" and such is also the practice in the present thesis.

In much of the research on learner corpora, various frequency lists have been employed as yardsticks against which learner data are measured. For a long time it was the General Service List (GSL) compiled by Michael West in 1953 that was used. In his article from 2010, Cobb relates (rather humorously) how in 2000, under Paul Nation's supervision, Averil Coxhead,

> [c]apitalizing on some accidents in the development of English (the Norman conquest and bifurcation of the language) /…/ showed in a corpus study that a smallish set of 570 mainly Greco-Latin word families, of medium (post-2,000 level) frequency in English as a whole but [of] much higher frequency in the discourse of scientific texts, when added to the 2,000 families of the [GSL] will normally give academic learners about 90% coverage in the texts they are studying (or a little more since they will also know some technical items in their subjects). (Cobb 2010: 190–191)

By using a newer frequency list based on the BNC and showing that the first 2,000 words in BNC provide as much coverage as was previously done by the GSL and Academic Word List (AWL) together, Cobb (2010: 190–195) questions the relevance of the AWL. Still, while he achieved similar results with many text types, Cobb concedes that for texts heavy in academic and scientific vocabulary, there might still be room for an AWL. Conveniently, for the 60th anniversary of the initial GSL, in 2013 a New GSL saw the light of day ("A New General Service List"), and to complement it, a New AWL was also published ("A New Academic Word List"). Both the previous and the new set of lists are

available on the Lextutor website[3] for lexical profiling of texts. In view of the criticism of and revisions to the old lists, in this thesis the new lists will be used in profiling the use of adjectives and adverbs in NS and NNS writing.

Besides these frequency lists there are other quantitative measures that can provide useful information. Hasselgård and Johansson (2011: 35–36) make reference to a study by Linnarud (1986), who used measures such as "lexical individuality (lexical words unique to the writer), lexical sophistication (the number of less frequent words), lexical variation (type-token ratio), and lexical density (the proportion of lexical or content words in relation to the total number of words)". Alternatively, the term 'diversity' can be used for variation, and the term 'richness', which shows the proportion of low-frequency words in a piece of writing, can be used instead of lexical sophistication (Laufer, 1994; Laufer & Nation 1995 cited in Tami Levitzky-Aviad and Batia Laufer 2013: 129).

As was mentioned above, the frequency of a linguistic feature *per se* does not prove half as insightful as when it is compared with data from a reference corpus, which raises the question of what the aspects that should be borne in mind when selecting corpora for comparison are.

1.3. Comparable Corpora

Corpora are indispensable in tracing differences between language varieties. One can compare different varieties of the same language: "spoken vs. written, general vs. domain-specific, current-day vs. earlier varieties, standard vs. other regional or social varieties"; but also different languages (Granger 2015: 8). In learner language research, there are two main types of comparison: firstly, a comparison with native language (L1 vs.

[3] Most Vocabprofilers on that website are based on Laufer and Nation's Lexical Frequency Profiler and have been adapted for the Web by Thomas Cobb, according to whom "Vocabulary Profilers break texts down by word frequencies in the language at large, as opposed to in the text itself."

L2), and secondly, a comparison of different varieties of learner language (L2 vs. L2), especially from different mother tongue backgrounds (Granger 2015: 8). In the present thesis, the discussion is restricted to comparisons between native and learner language (L1 vs. L2).

Any corpus building is subject to strict design criteria and this equally applies to learner corpus building (Granger 2004: 125). Yukio Tono (2003: 800–802) gives an overview of the aspects influencing learner production which must be taken into account when building a corpus. He divides the variables into three categories: language-, task- and learner-related criteria. Granger (2004: 126) also presents a chart of the general and L2-specific variables contained in the ICLE database but Tono's presentation seems more comprehensive and useful to present here. The only two aspects that could be added to Tono's table from Granger's variable list are other foreign languages (learner-related) and length of the piece of writing (task-related).

Table 1. Design considerations for learner corpora (Tono 2003: 800)

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

As can be expected, in order to make fruitful and legitimate comparisons, the data contained in both or all of the corpora must match in terms of most of these aspects. Depending on the aims of comparison, some variation in the characteristics of corpora may

be necessary, for instance, when contrasting the language of different genres or age groups or different modes. An additional rule concerning corpora intended for comparison states that they should be of similar size and produced under similar circumstances (Cobb 2003: 396–297).

Even assuming that most of these requirements for comparing corpora are sufficiently met, there is still a range of issues that can arise. The most important of these is the question of the target linguistic behaviour or norm (Leech 2011: 25–26), i.e. which variety should be the target in ESL/EFL research and teaching? Should all learners have the same target? Can general-purpose corpora be used for all kinds of comparison or should English for Specific Purposes or EAP corpora be employed (Meunier 2011: xv)? On the one hand, some researchers, such as Lorenz (1999) (cited in Gilquin 2007: 326), criticise the practice of using expert writing as a norm against which to compare learner writing, arguing that learner language should be compared to native-speaker *student* texts, which would therefore likewise be novice writing. On the other hand, student language may not be high enough an ideal for language teaching and learning (Gilquin 2007: 326–327). Another option would be to use a corpus of non-native speakers using English as a lingua franca, such as the VOICE corpus. This relates to the broader question of whether native-speaker language should still be regarded as the ideal standard at all (Leech 2011: 26). Granger (2015: 15–16) in her reappraisal of CIA also discusses the issue of norm with respect to Lingua Franca Englishes and World Englishes, and notes that "[t]he conclusion is not to abandon the terms native and non-native altogether but to avoid using them as de facto generic terms". From the point of view of pedagogy, Leech (2011: 26) states that for "the normal EFL educational curriculum, the ideal corpus should be longitudinal, representing competent target language use appropriate to the age cohort of

the learners." The only problem is that longitudinal data are far more difficult to gather and most corpora are cross-sectional.

Despite the numerous issues to consider, the comparison of learner and native language can be very useful. The compilers of ICLE solved many of the issues by creating a similar NS corpus, namely LOCNESS, to match ICLE (Hasselgård and Johansson (2011: 38). Indeed, a substantial proportion of learner language research has been conducted by comparing one or several sub-corpora from ICLE and LOCNESS. LOCNESS consists of argumentative and literary essays written by British and American university students and British A-level pupils. Hasselgård and Johansson (2011: 38) note that though ICLE and LOCNESS are relatively closely matched for text type, writer age and experience, there is less information available on contributors in LOCNESS and their texts are more heterogeneous in terms of essay topics. For this reason, Hasselgård and Johansson (2011: 38) admit, many researchers have decided to use only a sample of LOCNESS. Nonetheless, LOCNESS "remains the best available comparable corpus to match ICLE and continues to be widely used" (Hasselgård and Johansson 2011: 38).

Due to technical and practical reasons, the reference corpus used in the empirical research of this thesis is LOCNESS. It has the advantage of being readily available on the Internet and a history of having been used in numerous other studies in learner corpus research. More details about the suitability of LOCNESS for comparison with EEIC will be provided in the next chapter.

In conclusion, this chapter provided a short overview of what has been discovered about late interlanguage development with a few notes on learners' use of adjectives and adverbs in particular. As a result of numerous studies in the field, some of which have been presented above, there is now a much clearer understanding of what characterises late interlanguage. In the middle section, the benefits and pitfalls of word frequencies were

under scrutiny with the general conclusion that they can be useful as long as the data is carefully interpreted. The last section dealt with aspects of corpus building and corpus comparisons and concluded that despite numerous aspects to be borne in mind, it is possible to find suitable comparable corpora for research purposes.

2. Corpus Comparison of Adjectives and Adverbs used in EEIC and LOCNESS

The aim of this chapter is to compare the use of adjectives and adverbs in Estonian-English Interlanguage Corpus, a learner corpus, and Louvain Corpus of Native English Essays, a native corpus. Before reporting on the findings, the methodology and tools are described, followed by an account of the data and the limitations that arise from the ways the data are treated (both by the tools and the researcher), after which the exclusions from the adjective and adverb lists are explained. In the analysis, the use of adjectives and adverbs is quantitatively described from the perspectives of lexical variety and sophistication, and the types of adjectives and adverbs used.

2.1. Methodology and Tools

The broader methodological framework employed in this study is corpus analysis. The more specific data extraction and treatment methods, however, will be discussed step-by-step in the following sections. The data were mostly processed using two tools: AntConc (Anthony 2014) and online VocabProfilers (Cobb 2002). AntConc was used to create wordlists and make concordance searches in the corpora. Two VocabProfilers were used to create lexical profiles: VP-Compleat, the Neo-Classic sorter (with the New GSL and New AWL) and VP-Compleat, the BNC sorter. VP Neo-Classic is a four-way sorter, which divides words into the first and second thousand levels according to the NGSL, the NAWL, and the remainder or 'off-list' (which also contains proper names, numbers and misspelt words). In the case of this sorter it was mainly the proportion of academic vocabulary in the corpora that was of interest. The second sorter, VP BNC, stratifies words

into 20 one-thousand-word frequency bands plus ‘off-list’ and provides the basis for describing lexical variety and sophistication.

2.2. Data and Limitations

The essays comprising EEIC were written as part of the entrance examination to the English Language and Literature BA programme at the University of Tartu in July 2014. The essays were typed in manually and checked by two people. The prerequisite to taking the entrance exam was proof of secondary education; however, no data were gathered as to prior higher education or length of study of the English language. The corpus contains altogether 127 essays of (ideally) around 200 words written as a response to an academic text on the topic of the future of the English language. The task together with the source text is provided in Appendix 1.

The corpus can be characterised by the following aspects:

- the number of words in the corpus is 24,457 tokens[4];
- the gender division is 88 females and 39 males;
- the age range is 18–35, with the average age of approximately 19 years;
- the length of the essays varies from 60 to 320 words, with the average length of 193 words;
- all participants hold Estonian citizenship, but their mother tongue is unknown;
- no reference tools were allowed.

[4] The number of tokens varies in different programmes due to the way they treat raw data. The figure given above is taken from the MS Office Word file word count tool, which counts contracted forms (don’t, let’s) as single words. The online VocabProfiler counts 24,590 tokens and deletes single letters (except “I” and “a”), which yields forms like “don” and “won”, which are categorised under the off-list. The AntConc programme counts 24,610 word tokens and treats single letters (‘t and ‘s) as separate tokens, which is logical for ‘t, which stands for “not”, but ambiguous for ‘s, which can mark either the possessive or the contraction of “is”.

It must be taken into account that the *crème de la crème* of the applicants for the programme was exempt from taking the entrance examination. Namely, students who had scored at least 95 points out of 100 in the National Examination in English or had received a certificate[5] of English did not have to take the entrance examination. On a more technical note, the corpus currently contains only raw information, meaning it is not tagged for part of speech nor is it graded (the essays have not been assigned levels according to the Common European Framework of Reference for Languages or other frameworks). Fortunately there is data available on the participants' scores for the National Examination in English. Since the examination format was changed recently, Table 2 presents the distribution of scores for both the previous and the new examination. The 95-points exemption applies in both cases. According to the new examination, B2 level was given to students who scored at least 75 points out of 100. It must be noted here that these students may in fact already have a higher level of proficiency, such as C1, but since the examination was not designed to test above B2 level, such claims cannot be made. As to the previous examination, it is known that only 4 participants scored below 75 point, although it must be taken into account that these two examinations are not straightforwardly comparable. Nevertheless, it can be concluded that the majority of the candidates form a rather homogeneous sample on the upper-intermediate level, since there are very few essays that are markedly weaker than the rest and the best candidates did not write any essays at all.

[5] The acceptable certificates are:
- Certificate in Advanced English (CAE) CEFR level C1 or above
- Certificate of Proficiency in English (CPE) CEFR levels C2 or C1
- The International English Language Testing System (IELTS) overall score 7 or above
- Test of English as a Foreign Language (TOEFL) overall score 100 or above

Table 2. National Examination scores of the 127 participants of EEIC

Previous Examination		**New Examination**	
Score	Participants	Level	Participants
90–94	13	B2	89
80–89	10		
70–79	5	B1	8
below 70	2		
Total:	**30**		**97**

As discussed in the section on comparable corpora, certain design criteria have to be met for the comparison to be fruitful. Unfortunately there is noticeably less information available on the contributors in LOCNESS; for instance, the gender division, length of the essays and use of reference material are unknown. The accessible data are presented below.

LOCNESS contains in total 324,304 words of argumentative and literary essays written by American (168,400 words) and British (95,695 words) university students and British A-level pupils (60,209 words). The essay topics range from French literature and philosophy to parliamentary system and fox hunting in the UK. In order to better match the size of the Estonian corpus, a selection, which would correspond to the total length of around 24,000 or 25,000 words, had to be made from LOCNESS. This selection containing 29 essays was guided by a number of considerations. It should first be noted that some of the criteria were easier to match while others required more tailoring. The criteria of age and experience of contributors, and the style of writing were easier to satisfy than the criteria of theme and length of the essays. Since the Estonian contributors' average age is 19 years, it seemed logical to choose the British A-level pupils' (usually aged 16–18) essays over the university students' texts, thus making LOCNESS match EEIC in terms of age and experience. As to the style of writing, the description of LOCNESS simply states that the texts are "A-level argumentative essays" while the task (see Appendix 1) set for the Estonian students specifies that their "answers should be logically structured and use

appropriately academic and grammatically correct English." The task also sets a question before them and asks for their "opinion", which means that although the style of writing is expected to be academic, it can still contain some elements of slightly less formal language (such as, perhaps, the use of the first person). Notwithstanding this minor difference, it can be concluded that both the Estonian students' and the British pupils' essays are (supposed to have been) written in argumentative and more or less academic style.

As mentioned above, some of the requirements for comparison were more difficult to meet. In the light of the aims of the present thesis, the most problematic issues are the theme and length of the texts. EEIC contains essays of the average length of 193 words written on a single topic while LOCNESS essays vary both in length and theme. The number of words in LOCNESS A-level essays seems to fall roughly between 200 and 600 words and the topics include transport, parliamentary system, monarchy, fox hunting, boxing, national lottery, the effects of technology on modern life, genetic engineering and beef consumption. Out of this plethora of topics, five were selected for analysis. Although the topics in LOCNESS vary greatly (five vs. one in EEIC), there is still some thematic uniformity as over 85% of the essays (20,786 words) are related to technology and/or medicine; the remainder of the essays are about political systems. Both length and topic can, however, influence language production. A longer piece of writing generally gives the student more opportunity to use language, including lexical words such as adjectives and adverbs. In a similar fashion, a set topic determines the choice of vocabulary. This is, perhaps, more relevant in the case of adjectives than of adverbs as the choice of adjectives tends to depend more on the subject matter than the use of adverbs. In order to remedy this situation, some precautions were taken, which are detailed in section 2.3.

Among other difficult choices were those concerned with the situation of data elicitation, use of reference material and the existence of a source text to work with.

Contributors to EEIC wrote their essays in a potentially stressful examination situation where no reference material was allowed. As for LOCNESS, the context of producing the essays is unspecified, as is the use of reference material. It is possible to hypothesise that the use of reference sources would be more influential for learners of a language than for native speakers since learners have less linguistic knowledge at their disposal and could, thus, make more use of dictionaries and grammars. The final point of divergence between the two corpora is that the Estonian students had a source text (see Appendix 1) accompanying the task, which they were expected to comment on and consider in their responses. Their language production, therefore, cannot be viewed as completely neutral in this respect because they are discussing someone else's ideas together with their own. LOCNESS description file does not specify whether or not the British students were working with a source text.

Table 3 below highlights the important similarities and differences between EEIC and LOCNESS. Despite all these discrepancies, it is hoped that the selection from LOCNESS is compatible enough with EEIC to be insightful.

Table 3. Comparison of EEIC and LOCNESS

	EEIC	**LOCNESS**[6]
similarities		
- size	24,457 words vs. 24,089 words	
- age	roughly the same age	
- style	academic style, expository writing	
- experience	novice, not expert writers	
differences		
- situation	examination	not specified
- topics	future of the English language	genetic engineering (8,452 words) computer vs. the human brain (6,280 words) in vitro fertilisation (6,054 words) monarchy (1,815 words) parliamentary system (1,488 words)
- source text	yes	not specified
- length	ca. 200 words	ca. 200–600 words
- reference	not allowed	not specified
- variety	N/A	British

[6] "LOCNESS" is hereafter used to mean the selection from the corpus, not LOCNESS in its entirety.

The lists of adjectives and adverbs were extracted manually from the wordlists generated by AntConc. Where part of speech was dependent on the context (e.g. in the case of participial adjectives), the decision was made based on the concordance results of the word (i.e. sentence context) and a new ranking of the frequency of adjectives and adverbs was created, as opposed to the ranking of all words present in the corpus. Early on in the process of extraction, it was decided that two groups of words would not be included in the adjective and adverb lists: (1) misspelt words and (2) adjectives relating to nationalities and cultures (such as 'English', 'Spanish', 'Nigerian') and places ('Oxfordian', 'Western'). The second category was regarded as atypical examples of adjectives. These words were not checked for part of speech (where it might have been ambiguous) and they do not figure in the initial adjective and adverb lists ("initial total" in Table 4). Misspelt words are grouped together and presented in Appendix 2. Table 4 presents the numbers of misspelt words in both corpora.

Table 4. Misspelt adjectives and adverbs in EEIC and LOCNESS

	EEIC (all words 24,457)		**LOCNESS** (all words 24,089)	
	adjectives	adverbs	adjectives	adverbs
initial total:	**380**	**174**	**593**	**253**
misspelt:	+42	+24	+33	+19
total misspellings:	66		52	
all attempts:	422	198	629	272

2.3. Exclusions from the Lists of Adjectives and Adverbs

All of the issues considered in the previous section can influence the results of the study. Since one of the research questions was whether, and to which extent, native speakers employ a more varied selection of adjectives and adverbs, comparing the vocabulary of a thematically unified and a thematically diverse corpus would inevitably yield results that are swayed in favour of the more varied corpus. For this reason and in

order to neutralise the effect of the theme on the choice of vocabulary, topic-specific adjectives were removed from both the LOCNESS and the EEIC adjective lists. In the case of EEIC, 18 adjectives which were semantically clearly connected to the topic of English as an international language and which appeared in the task description or the source text were removed. From LOCNESS, 70 adjectives were excluded. Table 5 presents these adjectives in the order of frequency in both corpora. The categorisation into two topic areas in LOCNESS is not very strict but it was useful in making the selection. It must be conceded that in this kind of elimination of words, there is subjectivity involved on the part of the researcher. Nevertheless, it is hoped that this procedure will make the two adjective lists more comparable and eventually yield more objective results. A similar approach to reducing topic sensitivity was taken by Ringbom (1998: 48), who admits that topic-sensitivity will still "to some extent be present whenever word frequency patterns are established for texts with different content."

Table 5. Topic-specific adjectives excluded from the adjective lists

EEIC	**LOCNESS**		
English as an international language:	technology and/or medicine:		
1. international	1. human	19. homosexual	37. fertilised
2. native	2. genetic	20. immoral	38. fractal
3. standard	3. moral	21. unborn	39. infected
4. foreign	4. post-menopausal	22. atomic	40. inherited
5. official	5. scientific	23. heterosexual	41. manipulated
6. national	6. nuclear	24. infectious	42. moralistic
7. non-native	7. medical	25. male	43. muscular
8. expanding	8. infertile	26. married	44. neural
9. extinct	9. aborted	27. pre-menopausal	45. post-fertility
10. global	10. biological	28. psychological	46. pre-born
11. outer	11. mental	29. unethical	47. pregnant
12. linguistic	12. physical	30. antibiotic	48. radioactive
13. standardised	13. technological	31. anti-cancer	49. sensory
14. grammatical	14. electronic	32. bacterial	50. sexual
15. non-English	15. engineered	33. chemical	51. sterile
16. emerging	16. hereditary	34. computerised	52. superhuman
17. globalised	17. ethical	35. congenital	53. toxic
18. inner	18. female	36. environmental	54. virtual

	political systems:		
	1. political	7. conservative	12. civil
	2. royal	8. diplomatic	13. constitutional
	3. electoral	9. liberal	14. democratic
	4. economic	10. non-elected	15. elected
	5. undemocratic	11. apolitical	16. monetary
	6. parliamentary		
Total: 18 words	**Total: 70 words**		

In addition to this, there were two other groups of adjectives that were excluded from the lists and these were compound adjectives (some of them are also topic-specific) and unrecognised words. Multiword adjectives had to be removed from the lists because both the online VocabProfiler and AntConc only recognise single words. These programs separate a hyphenated word into its constituent words (so 'round-the-clock' becomes round, the, clock), which would give irrelevant frequency data about adjectives and adverbs. The excluded compound adjectives are listed in frequency order in Table 6. Additionally, there is only one multiword adverb in LOCNESS, 'in vitro', which is not listed in the table. It is evident that NSs use compounding five times more than NNSs.

Table 6. Compound adjectives excluded from the adjective lists

Compound adjectives*			
EEIC	**LOCNESS**		
1. English-speaking	1. in vitro	15. computer-orientated	27. much needed
2. made-up	2. third-world	16. computer-simulated	28. oil-based
3. multi-language-speaking	3. first-past-the-post	17. ever-increasing	29. present-day
4. native-speaking	4. long-term	18. far-fetched	30. round-the-clock
5. non-English-speaking	5. cancer-causing	19. full-time	31. sex-linked
6. open-minded	6. ever-expanding	20. hand-printed	32. short-term
7. third-world	7. free-thinking	21. job-motivated	33. tailless
8. well-known	8. money-motivated	22. labour-saving	34. three-pronged
	9. old-fashioned	23. large-scale	35. three-way
	10. time-saving	24. life-saving	36. well-built
	11. well-documented	25. man-made	37. well-meaning
	12. business-orientated	26. mind-numbing	38. wheel-based
	13. computer-controlled		39. wheelchair-bound
	14. computer-generated		
Total: 8 words	**Total: 39 words**		

*Not all compounds were correctly hyphenated in the corpora as they are in this list, but as hyphenation in itself is not under scrutiny in this thesis, they are spelt correctly here for the sake of clarity.

Finally, there were words that were not topic-sensitive nor compounds but that the programs did not recognise. Since they were classified under the off-list and did not, therefore, contribute to rank frequency lists, they were also excluded. This remaining category is a small one, as can be seen from Table 7.

Table 7. Other unrecognised adjectives excluded from the adjective lists

EEIC	**LOCNESS**		
1. non-business	1. non-malicious	3. overpopulated	5. over-reliant
2. non-popular	2. non-skilled	4. overprotective	6. uninventive
Total: 2 words	**Total: 6 words**		

Table 8 shows all three types of exclusions from the adjective and adverb lists in numbers. Through the process of exclusion, more adjectives were removed from LOCNESS adjective list (19% vs. 7% in EEIC), but still LOCNESS contains more adjectives and adverbs than EEIC.

Table 8. Exclusions from adjective and adverb lists

	EEIC (all words 24,457)		**LOCNESS** (all words 24,089)	
	adjectives	adverbs	adjectives	adverbs
total initial:	**380**	**174**	**593**	**253**
topic-specific:	18		70	
compounds:	8		39	1
unrecognised:	2		6	
total excluded:	**28 (7%)**		**115 (19%)**	**1**
included for analysis:	**352**	**174**	**478**	**252**

After such extensive truncation, the lists are finally suitable for obtaining data from the computer programs and making comparisons between the corpora, which is the focus of the next sections.

2.4. Results

Sub-sections 2.4.1. to 2.4.4. present the findings of the study. The first two sub-sections analyse lexical variety and sophistication in the use of adjectives and adverbs separately. The third sub-section studies the twenty most common adjectives and adverbs in both corpora, and the fourth sub-section provides information on the types of adjectives and adverbs used based on their form.

2.4.1. Lexical Variety and Sophistication in the Use of Adjectives

In this section, quantitative measures are used to describe lexical variety and sophistication in EEIC and LOCNESS. Since it is not all the words in the two corpora but only the lists of adjectives, where each word appears once, that are under scrutiny in this section, variety is not understood in its usual sense of type-token ratio (as explained on in section 1.2.) but rather as the number of *different* adjectives used. The feature of sophistication is employed in the same sense as was explained earlier (in section 1.2.), that is, as a measure of showing the proportion of less frequent words.

Tables 9 and 10 below show the distribution of adjectives across the 20 BNC frequency bands in the NS and NNS corpora. At first glance, it seems that the use of adjectives is equally sophisticated in both corpora, since both lists are exhausted at K-12 level. There is one exception to this in LOCNESS, which contains one word from the K-20 frequency band and this word is 'jurassic'. With the help of AntConc, it was possible to ascertain that this word was used twice and both times in the title of a film, namely *Jurassic Park*. On closer examination, however, the cumulative token percentages reveal that EEIC uses consistently more adjectives from the high-frequency bands though the differences with LOCNESS are not remarkable. Cumulatively 53.69% of the adjectives in

EEIC are from amongst the 1000 most frequent words whereas slightly fewer, 50.84% of adjectives in LOCNESS come from the same frequency band. However, when analysing the proportion of each thousandth level (either the types or tokens percentages) separately from the cumulative percentages, EEIC can boast higher percentages in half of the frequency bands from K-3 to K-12. These are K-3, K-7, K-9, K-10 and K-12. This leads to the conclusion that although Estonian learners tend to use more adjectives from the top frequency bands, they still manage to cover frequency bands beyond the first 2000 words relatively well compared to the British students. Still, LOCNESS uses proportionally more words from K-3 to K-20 frequency bands than EEIC since 76.7% of adjectives in EEIC are from amongst the 2000 most common words, as opposed to 74.69% of adjectives in LOCNESS, which shows that NS use more infrequent words and this makes their usage of adjectives marginally more sophisticated.

Table 9. EEIC adjectives across BNC frequency bands

EEIC				
Freq. Level	**Families (%)**	**Types (%)**	**Tokens (%)**	**Cumul. token %**
K-1 Words :	153 (49.04)	186 (53.76)	189 (53.69)	53.69
K-2 Words :	78 (25.00)	81 (23.41)	81 (23.01)	76.7
K-3 Words :	34 (10.90)	35 (10.12)	35 (9.94)	86.64
K-4 Words :	14 (4.49)	14 (4.05)	14 (3.98)	90.62
K-5 Words :	11 (3.53)	11 (3.18)	11 (3.12)	93.74
K-6 Words :	3 (0.96)	3 (0.87)	3 (0.85)	94.59
K-7 Words :	7 (2.24)	7 (2.02)	7 (1.99)	96.58
K-8 Words :	2 (0.64)	2 (0.58)	2 (0.57)	97.15
K-9 Words :	5 (1.60)	5 (1.45)	5 (1.42)	98.57
K-10 Words :	2 (0.64)	2 (0.58)	2 (0.57)	99.14
K-11 Words :	1 (0.32)	1 (0.29)	1 (0.28)	99.42
K-12 Words :	2 (0.64)	2 (0.58)	2 (0.57)	99.99
K-13 Words :				
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :				
K-18 Words :				

K-19 Words :				
K-20 Words :				
Off-List:	??	0 (0.00)	0 (0.00)	
Total (unrounded)	312+?	346 (100)	352 (100)	100

Table 10. LOCNESS adjectives across BNC frequency bands

LOCNESS				
Freq. Level	**Families (%)**	**Types (%)**	**Tokens (%)**	**Cumul. token %**
K-1 Words :	190 (46.00)	237 (50.75)	243 (50.84)	50.84
K-2 Words :	104 (25.18)	114 (24.41)	114 (23.85)	74.69
K-3 Words :	35 (8.47)	37 (7.92)	37 (7.74)	82.43
K-4 Words :	27 (6.54)	27 (5.78)	27 (5.65)	88.08
K-5 Words :	17 (4.12)	17 (3.64)	17 (3.56)	91.64
K-6 Words :	10 (2.42)	10 (2.14)	10 (2.09)	93.73
K-7 Words :	8 (1.94)	8 (1.71)	8 (1.67)	95.4
K-8 Words :	6 (1.45)	6 (1.28)	6 (1.26)	96.66
K-9 Words :	3 (0.73)	3 (0.64)	3 (0.63)	97.29
K-10 Words :	2 (0.48)	2 (0.43)	2 (0.42)	97.71
K-11 Words :	9 (2.18)	9 (1.93)	9 (1.88)	99.59
K-12 Words :	1 (0.24)	1 (0.21)	1 (0.21)	99.8
K-13 Words :				
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :				
K-18 Words :				
K-19 Words :				
K-20 Words :	1 (0.24)	1 (0.21)	1 (0.21)	100
Off-List:	??	0 (0.00)	0 (0.00)	
Total (unrounded)	413+?	467 (100)	478 (100)	100

The second characteristic of interest is the proportion of academic vocabulary as in both corpora the required style of writing was (more or less) academic. Since the NGSL provides quite similar (only less detailed) information on lexical variety, it is only the data according to NAWL that is presented here. Tables 11 and 12 show that in LOCNESS the proportion of academic adjectives is higher by approximately one percentage point (4.55% vs. 5.65%) in the case of tokens.

Table 11. Academic adjectives in EEIC

Freq. Level	Lemmas (%)	Types (%)	Tokens (%)	Cumul. token %
NAWL [963 lemmas]	16 (6.45)	16 (4.62)	16 (4.55)	78.41

Table 12. Academic adjectives in LOCNESS

Freq. Level	Lemmas (%)	Types (%)	Tokens (%)	Cumul. token %
NAWL [963 lemmas]	27 (9.00)	27 (5.78)	27 (5.65)	69.04

The results confirm the tendency of learners using more high-frequency and less academic adjectives than native speakers. What is delightful to see is that in the case of the present sample, Estonian learners of English do not compare poorly to the sample of the British A-level students.

2.4.2. Lexical Variety and Sophistication in the Use of Adverbs

As was the case with adjectives, so does EEIC contain more adverbs from the top 1000-word frequency band (76.44% vs. 72.22% in LOCNESS). Tables 13 and 14 demonstrate that in both corpora the top seven frequency bands are represented and that, not surprisingly, starting from the K-2 level towards the less frequent word zones, LOCNESS contains proportionally more adverbs than EEIC. This continues until the K-7 level. Beyond that, however, the scene changes interestingly. By the K-7 level, the adverb list from LOCNESS has been exhausted (cumulative token % is at 100) while in EEIC, the NNS corpus, there is an adverb used from the K-9 list and another from the K-17. These adverbs are ‘someday’ and ‘retroactively’, respectively. This shows that although NS in the selected sample use a larger number of different adverbs, i.e. the variety is greater, the Estonian students are capable of showing knowledge of some more sophisticated items than the British students.

Table 13. EEIC adverbs across BNC frequency band

EEIC				
Freq. Level	**Families (%)**	**Types (%)**	**Tokens (%)**	**Cumul. token %**
K-1 Words :	114 (73.55)	123 (75.00)	133 (76.44)	76.44
K-2 Words :	22 (14.19)	22 (13.41)	22 (12.64)	89.08
K-3 Words :	9 (5.81)	9 (5.49)	9 (5.17)	94.25
K-4 Words :	2 (1.29)	2 (1.22)	2 (1.15)	95.4
K-5 Words :	3 (1.94)	3 (1.83)	3 (1.72)	97.12
K-6 Words :	1 (0.65)	1 (0.61)	1 (0.57)	97.69
K-7 Words :	2 (1.29)	2 (1.22)	2 (1.15)	98.84
K-8 Words :				
K-9 Words :	1 (0.65)	1 (0.61)	1 (0.57)	99.41
K-10 Words :				
K-11 Words :				
K-12 Words :				
K-13 Words :				
K-14 Words :				
K-15 Words :				
K-16 Words :				
K-17 Words :	1 (0.65)	1 (0.61)	1 (0.57)	99.98
K-18 Words :				
K-19 Words :				
K-20 Words :				
Off-List:	??	0 (0.00)	0 (0.00)	
Total (unrounded)	155+?	164 (100)	174 (100)	100

Table 14. LOCNESS adverbs across BNC frequency bands

LOCNESS				
Freq. Level	**Families (%)**	**Types (%)**	**Tokens (%)**	**Cumul. token %**
K-1 Words :	157 (69.16)	167 (70.76)	182 (72.22)	72.22
K-2 Words :	43 (18.94)	43 (18.22)	43 (17.06)	89.28
K-3 Words :	14 (6.17)	14 (5.93)	14 (5.56)	94.84
K-4 Words :	4 (1.76)	4 (1.69)	4 (1.59)	96.43
K-5 Words :	5 (2.20)	5 (2.12)	5 (1.98)	98.41
K-6 Words :	3 (1.32)	3 (1.27)	3 (1.19)	99.6
K-7 Words :	1 (0.44)	1 (0.42)	1 (0.40)	100
K-8 Words :				
K-9 Words :				
K-10 Words :				
K-11 Words :				
K-12 Words :				
K-13 Words :				
K-14 Words :				
K-15 Words :				

K-16 Words :				
K-17 Words :				
K-18 Words :				
K-19 Words :				
K-20 Words :				
Off-List:	??	0 (0.00)	0 (0.00)	
Total (unrounded)	227+?	236 (100)	252 (100)	100

As regards the proportion of academic adverbs, the variance between the two corpora is greater than in the case of adjectives. While the proportion of academic adjectives differs only by one percentage point, adverbs differ by 3 percentage points (3.45% in EEIC vs. 6.72% in LOCNESS) as shown in Tables 15 and 16.

Table 15. Academic adverbs in EEIC

Freq. Level	Lemmas (%)	Types (%)	Tokens (%)	Cumul. token %
NAWL [963 lemmas]	6 (4.17)	6 (3.66)	6 (3.45)	86.22

Table 16. Academic adverbs in LOCNESS

Freq. Level	Lemmas (%)	Types (%)	Tokens (%)	Cumul. token %
NAWL [963 lemmas]	17 (9.24)	17 (7.17)	17 (6.72)	78.26

Overall, the Estonian students' use of adverbs is less academic. With regard to sophistication, the results are somewhat conflicting, as EEIC uses more adverbs from the top frequency band while also using some more sophisticated adverbs than LOCNESS. Generally, though, LOCNESS shows higher percentages for most frequency bands beyond the first level.

2.4.3. Twenty Most Common Adjectives and Adverbs

In order to detect possible lexical overuse and underuse, twenty most frequent adjectives and adverbs from both corpora are juxtaposed in Table 17 and Table 18. Words highlighted in grey are those present in both corpora. The frequency counts for adjectives

in EEIC are noticeably higher than for LOCNESS. Considering the higher frequencies and the fact that EEIC contained fewer adjectives than LOCNESS to begin with, adjectives at the top of the list can be viewed as being overused. It is important to note here that the words *new*, *positive* and *negative* were mentioned in the task description of the entrance examination (see Appendix 1) and could, thus, be expected to appear more often, though various synonyms exist (e.g. advantageous, beneficial, fortunate, opportune, disadvantageous, adverse, harmful, unfavourable). Even so, if these three words were removed and the list extended, the next three words from the rank list would be *due* (freq. 17), *original* (freq. 17) and *bad* (freq. 14), which are still more frequently used than the last word on the LOCNESS list, which is repeated 11 times in the corpus.

Table 17. Twenty most common adjectives

EEIC			**LOCNESS**		
rank	**freq.**	**ADJECTIVE**	**rank**	**freq.**	**ADJECTIVE**
1	327	new	1	43	other
2	207	other	2	41	able
3	156	positive	3	31	possible
4	148	negative	4	27	new
5	126	different	5	25	good
6	86	easier	6	20	responsible
7	60	own	7	19	redundant
8	55	same	8	18	due
9	49	good	9	17	fair
10	28	better	10	16	great
11	28	likely	11	15	certain
12	26	used	12	15	general
13	24	smaller	13	15	natural
14	23	important	14	14	normal
15	22	small	15	14	own
16	21	big	16	13	modern
17	19	possible	17	12	old
18	18	able	18	11	beneficial
19	18	difficult	19	11	high
20	18	main	20	11	important

A similar tendency is evident from the adverbs list. The top adverbs in EEIC are repeated much more frequently than the top adverbs in LOCNESS. In addition, EEIC contains some very common linking adverbs used for enumeration, such as *firstly* and *secondly*, while LOCNESS contains such linking adverbs as *however* (used to introduce contrast or concession) and *therefore* (used to present result of inference). While the first of them is also represented in EEIC, it is used much less frequently than in LOCNESS (19 times vs. 68 times).

Table 18. Twenty most common adverbs

EEIC			**LOCNESS**		
rank	**freq.**	**ADVERB**	**rank**	**freq.**	**ADVERB**
1	240	more	1	173	not
2	209	there	2	110	there
3	161	not	3	78	also
4	126	also	4	68	however
5	61	so	5	55	more
6	60	even	6	53	then
7	47	only	7	49	even
8	46	well	8	49	only
9	39	very	9	36	very
10	38	already	10	27	therefore
11	33	where	11	26	how
12	32	now	12	24	now
13	29	widely	13	21	so
14	28	much	14	22	often
15	26	as	15	18	where
16	24	most	16	17	just
17	24	firstly	17	17	still
18	20	secondly	18	16	as
19	19	however	19	16	never
20	19	probably	20	15	much

By contrasting figures in this way it cannot be inferred immediately that learners in this sample are overusing these most frequent words on the list because this would have to be confirmed by further analysis. Nonetheless, remarkable differences in frequencies do hint at potential overuse.

2.4.4. Types of Adjectives and Adverbs

In the following discussion of the types of adjectives and adverbs used in the two corpora, the emphasis is only on morphological and not syntactic or semantic features. While a more thorough investigation would undoubtedly be useful, it would be beyond the scope of this thesis. In this section the aim is to compare the formation of adjectives and adverbs in EEIC and LOCNESS with the large-scale findings presented in LGSWE.

According to LGSWE (Biber et al. 1999: 65), which categorises findings into four registers (conversation, fiction, news, academic texts), adjectives as a word class are most common in "registers with the highest frequency of nouns, i.e. news reportage and academic prose" and least common in conversation. Adverbs together with verbs, on the other hand, are most common in conversation and fiction.

Based on their formation, LGSWE (Biber et al. 1999: 520–536) distinguishes between three categories of adjectives: participial (present and past participles), derived and compound adjectives. Among these, derived adjectives are by far most common in academic prose. Derived adjectives formed with *-al* are markedly more common than those formed with any other derivational suffix. Adjectives ending in *-ent*, *-ive*, and *-ous* are moderately common while adjectives formed with *-ate*, *-ful*, *-less*, *-like*, and *-type* are relatively rare in all registers. Table 19 below shows that LOCNESS uses a more varied selection of prefixes and suffixes than EEIC (25 vs. 18). However, the most commonly used suffixes are roughly the same for both corpora.

Table 19. Most common prefixes and suffixes in adjectives

	EEIC		**LOCNESS**		
different affixes	18		25		
rank		**freq.**	**rank**		**freq.**
1	al	30	**1**	al	56
2	y / able	11	**2**	able	25
3	ive / ful / ant	10	**3**	y / ous	16
4	ous / ic	8	**4**	ic / un	15

A fourth category, namely simple adjectives, could be added to the three types of adjectives presented in LGSWE, which then allows for a distribution into simple, participial, derived and compound adjectives. However, as the hyphenated compounds were removed from the corpora, the remaining compounds, such as *worldwide*, *widespread* and *homemade*, are the ones written as one word, which the programs do recognise. The figures in Table 20 demonstrate once again that EEIC uses more simple adjectives and fewer derived adjectives than LOCNESS, which confirms the earlier finding that learners use less academic vocabulary.

Table 20. Distribution of types of adjectives

	simple	**participial**	**derived**	**compound**
EEIC	40%*	21%	38%	4 words
LOCNESS	32%	18%	49%	6 words

*The percentages are rounded and, thus, do not add up to exactly 100%.

With regard to the formation of adverbs, LGSWE (Biber et al. 1999: 539–542) divides adverbs into five categories: simple, derived by suffixing -*ly*, derived using other suffixes, compound adverbs and fixed phrases. Fixed phrases were excluded from the lists and the compound adverbs featured here are written as one word (e.g. *anyway*, *beforehand*, *nevertheless*). The majority of adverbs are either simple forms or derived using the -*ly* suffix. “Conversation and academic prose represent opposite extremes of use” (Biber et al. 1999: 540) as in conversation, over 60% of adverbs are simple and only 20% are -*ly* forms, whereas in academic prose, around 55% are -*ly* forms and a little over 30% are simple. Table 21 shows this to be precisely the case for LOCNESS while in EEIC, simple and -*ly* adverbs are distributed more evenly, though -*ly* adverbs are slightly more numerous.

Table 21. Distribution of types of adverbs

	simple	*-ly*	compound	other suffix
EEIC	41%*	47%	11%	0 words
LOCNESS	36%	55%	8%	2 words

*The percentages are rounded and, thus, do not add up to exactly 100%.

This analysis of the types of adjectives and adverbs used in both EEIC and LOCNESS has confirmed that NS usage is more varied and more appropriate in terms of register. Despite these observations, NNS usage does not show overly limited use of vocabulary nor can it be described by ignorance of style considerations (there are still more *-ly* forms than simple adverbs, for instance).

2.5. Discussion

The aim of this section is to analyse the findings presented in sections 2.4.1. to 2.4.4. in more detail, to consider possible reasons for these results, and to compare them to previous studies and issues discussed in Chapter 1.

Based on the results from the two corpora, it can be concluded that both in the case of adjectives and adverbs, LOCNESS uses more varied and sophisticated vocabulary than EEIC, an outcome which is more or less predictable. What is surprising is that for both parts of speech the two corpora shared the extent of coverage across the BNC frequency bands: the first 12 bands for adjectives and the first 7 bands for adverbs (with the two exceptions in EEIC beyond K-7). Therefore, even though LOCNESS does use more varied adjectives and adverbs, EEIC is not too far behind. Of course, the scene would be different if the adjective lists had not been shortened on account of topic-specificity. Had all the adjectives been included for analysis, there would have been only 5 bands out of 20 left empty in LOCNESS whereas in EEIC, the coverage would have remained roughly the same (8 unrepresented frequency bands).

Greater variation (and sophistication as well, to some extent) in LOCNESS can be partly attributed to the two most influential differences between the corpora, namely, length and topic of the essays. As discussed on in section 2.2., longer assignments give the writer more opportunity to use words and to demonstrate their skill. Considering that EEIC contained 127 essays and LOCNESS only 29, there are a great deal more introductory and concluding sentences and paragraphs in EEIC and many more body paragraphs in LOCNESS. Since the main discussion of an issue takes place in the body paragraphs of a piece of writing, the British students were able to dedicate more space to the development of ideas. It does not follow that introductions and conclusions contain fewer adjectives and adverbs than body paragraphs (at least no such evidence is currently known to the author of the thesis), but since adjectives and adverbs are the third and fourth word classes after nouns and verbs in terms of overall frequency, a shorter text normally also contains fewer adjectives and adverbs. Though depending on the idiosyncratic style of the writer, producing a longer text does not always mean using more adjectives and adverbs[7].

As for the other variable, the subject matter, excluding topic-sensitive adjectives was deemed the best course of action to neutralise the effect of topic-specificity, but the fact that LOCNESS essays were written on five subjects instead of one may still have influenced the results in favour of the NS corpus in terms of both variety and sophistication. Collectively and cumulatively in the NS texts, this diversity of topics amounts to more varied and sophisticated vocabulary. Despite all of the above, part of the difference in variation and sophistication can still be ascribed to differences in language proficiency in NS and NNS.

Another parameter related to sophistication is the use of academic words. As with the previous two features, LOCNESS boasts once again a higher proportion of academic

[7] One of the most famous examples of this in literature is Hemingway's intentionally unadorned style.

adjectives and adverbs than EEIC. In the case of adverbs the difference is more marked (3.45% in EEIC vs. 6.72% in LOCNESS), which can be explained by the types of adverbs used. While in both corpora the largest category is *-ly* adverbs, in the case of LOCNESS the proportions of simple and derived adverbs correspond exactly to the general profile of adverb usage in academic texts presented in LGSWE (Biber et al. 1999: 540). In EEIC the proportions of simple and *-ly* adverbs are more equal and the percentages are closer together. These findings correspond to Granger and Rayson's (1998: 119–131) results, according to which learners overuse mainly short adverbs of native origin and underuse mainly *-ly* adverbs, such as amplifiers, disjuncts, modal adverbs etc. This seems to be a general and strong tendency on the part of learners since it has become evident from such a small sample as EEIC (24,000 words) and is in line with findings from a study on a much larger scale with a different L1 background[8]. Therefore, this feature of late interlanguage is not specific to Estonian learners.

Among the four characteristics of late interlanguage summarised by Callies (2010) only one was under scrutiny in this thesis, namely overuse of high-frequency vocabulary. Although a much larger corpus would be necessary in capturing patterns of overuse in learner language, some tentative conclusions can be drawn. The juxtaposition of the twenty most frequent adjectives and adverbs in EEIC and LOCNESS points to potential overuse of the most frequent adjectives and adverbs. For some words the difference between the number of repetitions in both corpora is over 100, for instance, the word 'other' has 164 more occurrences and the word 'more' has 185 more occurrences in EEIC than in LOCNESS. Although many researchers (e.g. Hasselgård and Johansson 2011: 55) have highlighted the need for careful qualitative analysis of the initial quantitative findings, this

[8] Granger and Rayson used a NNS corpus of 280,000 words and a reference corpus of 230,000 words (1998: 131). The mother tongue of the learners was French.

would have been beyond the scope of this thesis and will have to remain an area for further research.

As mentioned in the Introduction, part of the value of this thesis lies in the fact that it is the first attempt to describe Estonian-English interlanguage with the aid of a corpus. With that in mind, it would be pertinent if some statements could be made about L1 influence as well. Unfortunately, the small size of the corpus does not allow for broad generalisations. What is more, in order to determine the relatedness of a linguistic feature to the learners' mother tongue, interlanguage varieties with several L1 backgrounds would have to be contrasted. There is still one aspect, which though not considered in the main analysis of the data, deserves a mention in this context. It is the number of multi-word adjectives excluded from the adjective lists due to technical reasons. The number of compounds removed from LOCNESS is approximately five times larger (8 vs. 39 compounds) than the number of compounds removed from EEIC. According to LGSWE (Biber et al. 1999: 533–535), compounds "lend themselves to a compact and integrated expression of information", and adjectival compounds are common in the written registers, especially news. Why then do learners use them so sparingly? It might be argued that compounding as a technique is complex and requires thorough knowledge of the language and thus learners might feel unsure about using such constructions. While this may be true, it may also be that seeing as writing words as one word or separate words is often viewed as one of the most difficult aspects to master in the Estonian grammar (Habicht "Määrsõnade kokku- ja lahkukirjutamisest"), learners assume that it is equally as difficult in the foreign language and refrain from any attempts. This is simply a hypothesis that needs further investigation. Students may also be confused as to what constitutes a compound in English since in Estonian a compound is always written as one word or is hyphenated, whereas in English there is also the third option of writing a compound as

separate words. As noted in Table 6 in section 2.3., not even NSs were certain about the rules concerning the format and orthography of compounds.

One of the more philosophical issues discussed in Chapter 1 is the question of the target norm. From the pedagogical perspective, it seems logically sound to agree with Leech (2011: 26), who states that "the ideal [control] corpus should [represent] competent target language use appropriate to the age cohort of the learners". Having extracted misspelt words (see Appendix 2) from both EEIC and LOCNESS and briefly analysed innumerable sentences to determine the instances of adjective and adverb use in both corpora, the general impression of the author of the thesis is that the target for advanced learners should be higher than novice writing but at the same time, novice writing serves as a good midway benchmark for comparison.

In conclusion, it is clear that in corpus studies, the larger the corpus the more generalisable the results. With such a small sample of only one type of text written on a single topic, broad generalisations are not possible. Therefore, the findings of this thesis are representative, first and foremost, of the interlanguage of the Estonian learners who took the entrance examination at the University of Tartu under the conditions specified above. Still, in the hope that EEIC will grow in the future, these early conclusions (and many more hunches and inklings about learner language) can be tested later on a larger sample. The pedagogical implications arising from the present results are that learners could benefit from more instruction on *-ly* suffixation and perhaps derivation in general, and that more attention should be given to developing the skill of compounding. In the future, when EEIC has become larger and more representative, other aspects, such as multi-word units, various types of collocations, and recurrent word patterns can be studied.

Conclusion

With the arrival of computer learner corpora, research on learner language changed gear. While research on learner English is currently being conducted in various parts of the world and with numerous L1 backgrounds, the interlanguage of Estonian learners of English has until very recently not been studied on a larger scale. The aim of this thesis is to contribute to filling this gap in interlanguage research by examining the use of adjectives and adverbs in Estonian and British student writing.

Chapter 1 gave an overview of three topic areas: firstly, late interlanguage research, secondly, using frequency data in learner language analysis, and thirdly, aspects to consider when choosing corpora for comparison. While in the 1990s, learner corpus research was starting to gain popularity, by the 2000s and 2010s, a great deal of knowledge has accumulated on late interlanguage development. Countless studies have focused on lexis and word frequencies in learner and native language, with the goal of establishing linguistic (and pragmatic) aspects where learners might need further instruction. Native-speaker reference corpora must, however, be selected carefully and consciously in order to ensure fruitful results.

Chapter 2 presented the empirical study conducted as part of this thesis. The two sources of data employed for comparison are Estonian-English Interlanguage Corpus and Louvain Corpus of Native English Essays. The methodology used is corpus comparison. The data from the two corpora were extracted with the aid of the tool AntConc (Anthony 2014) and characterised using the lexical profilers available at the Lextutor website (Cobb 2002). The results show that the Estonian learners' use of adjectives and adverbs is less varied and sophisticated and they opt for less academic words. The comparison of the types of adjectives and adverbs used reveals that learners could benefit from more instruction on derivation of adjectives and adverbs, especially *-ly* suffixation. The

juxtaposition of the twenty most common adjectives and adverbs also showed that learners repeat the same words much more frequently than native speakers. Even so, based on the selected samples, Estonian learners' language usage compares relatively well to native-speaker usage.

In the discussion of the results it was concluded that in order to make inferences about L1 transfer, in this case the influence of the Estonian language on learners' English, having a much larger corpus is crucial. Nonetheless, some tentative conclusions were made about learners' hesitant use of derived adverbs and adjectival compounding. There is a myriad of aspects that can be studied via EEIC in the future. The free program AntConc has multiple options for data retrieval, enabling the study of multiword units, prefabricated patterns and various kinds of collocations among others. Provided that EEIC grows larger year by year, the results will become more and more representative of Estonian-English interlanguage and its varieties. As there is presently very little longitudinal data available for interlanguage research, another suggestion would be to collect further samples from those students who were admitted to the English Language and Literature BA programme, and to track their development. It seems intuitive to believe that some development takes place during undergraduate studies, but the question is how it manifests itself and whether there is a way to enhance the process. Yet another angle can be provided by comparing advanced learners' language use to expert, not novice usage, as was done in the present thesis.

On the whole, while the results obtained in the present thesis might be somewhat predictable, the extent and the nature of the differences between native-speaker and non-native-speaker usage can point to some specific and unexpected issues. It is hoped that in the hands of meticulous and creative researchers, this corpus project will be useful for many learners of English in the future.

References

Corpora

Estonian-English Interlanguage Corpus (EEIC). Available at the Department of English Philology of The University of Tartu.

Louvain Corpus of Native English Essays (LOCNESS). Available at http://www.uclouvain.be/en-cecl-locness.html , accessed February 22, 2015.

Tools

Anthony, Laurence. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/ , downloaded February 25, 2015.

Cobb, Thomas. 2002. *Web Vocabprofile*, an adaptation of Heatley, Nation & Coxhead's (2002) *Range*. Available at http://www.lextutor.ca/vp/ , accessed March 7, 2015.

Literature

Altenberg, Bengt. 2011. Preface. In Fanny Meunier et al. (eds). *A Taste for Corpora: In Honour of Sylviane Granger*, xiii–xv. Amsterdam, Philadelphia: John Benjamins.

Altenberg, Bengt and Marie Tapper. 1998. The use of adverbial connectors in advanced Swedish learners' written English. In Granger (ed). *Learner English on Computer*, 80–93. London and New York: Longman.

A New Academic Word List. Available at http://www.newacademicwordlist.org/ , accessed May 2, 2015.

A New General Service List (1.01). Available at http://www.newgeneralservicelist.org/ , accessed May 2, 2015.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*, Harlow: Pearson Education.

Callies, Marcus. 2010. Learner Varieties and Interlanguage Variation. Available at http://www.staff.uni-mainz.de/mcallies/Ringvorlesung/interlanguage%20variation.pdf , accessed April 30, 2015.

Cobb, Thomas. 2003. Analyzing late interlanguage with learner corpora: Quebéc replications of three European studies. *The Canadian Modern Language Review*, 59: 3, 395–423.

Cobb, Thomas. 2010. Learning about language and learners from computer programs. *Reading in a Foreign Language*, 22: 1, 181–200.

Coxhead, Averil. 2000. A New Academic Word List. *TESOL Quarterly*, 34: 2, 213–238.

De Cock, Sylvie, Sylviane Granger, Geoffrey Leech and Tony McEnery. 1998. An automated approach to the phrasicon of EFL learners. In Granger (ed). *Learner English on Computer*, 67–79. London and New York: Longman.

Ellis, Rod, and Gary Barkhuizen. 2005. *The Study of Second Language Acquisition, Second Edition*. Oxford, New York: Oxford University Press.

Eslon, Pille and Helena Metslang. 2007. Õppijakeel ja eesti vahekeele korpus. *Eesti Rakenduslingvistika Ühingu aastaraamat*. [Learner language and Estonian Interlanguage Corpus. Estonian Papers in Applied Linguistics], 3, 99–116.

Gilquin, Gaëtanelle, Sylviane Granger, Magali Paquot. 2007. Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6, 319–335.

Granger, Sylviane. 1998. Introduction. In Granger (ed). *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, Sylviane, Joseph Hung and Stephanie Petch-Tyson (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3–33. Amsterdam; Philadelphia: John Benjamins.

Granger, Sylviane. 2004. Computer learner corpus research: current status and future prospects. In Ulla Connor and Thomas A. Upton (eds). *Applied Corpus Linguistics. A Multidimensional Perspective*, 123–145. Amsterdam & Atlanta: Rodopi.

Granger, Sylviane. 2015. Contrastive interlanguage analysis. A reappraisal. *International Journal of Learner Corpus Research* 1:1, 7–24.

Granger, Sylviane and Paul Rayson. 1998. Automatic profiling of learner texts. In Granger (ed). *Learner English on Computer*, 119–131. London and New York: Longman.

Habicht, Külli. Määrsõnade kokku- ja lahkukirjutamisest. Emakeele Seltsi keeletoimkonna otsus 11. juunist 2012. [Writing adverbs as one word or many words. Estonian Language Association, decision from June 11, 2012] Available at http://www.emakeeleselts.ee/otsused/ES-keeletoimkond_Muutumatud-s6nad-kokku-lahkukirj-11.06.12.pdf , accessed May 17, 2015.

Hasselgård, Hilde and Stig Johansson. 2011. Learner corpora and contrastive interlanguage analysis. In Fanny Meunier et al. (eds). *A Taste for Corpora: In Honour of Sylviane Granger*, 33–61. Amsterdam, Philadelphia: John Benjamins.

ICLE. Université catholique de Louvain. Available at https://www.uclouvain.be/en-cecl-icle.html , accessed May 17, 2015.

Leech, Geoffrey. 2011. Frequency, corpora and language learning. In Fanny Meunier et al. (eds). *A Taste for Corpora: In Honour of Sylviane Granger*, 7–31. Amsterdam, Philadelphia: John Benjamins.

Levitzky-Aviad, Tami and Batia Laufer. 2013. Lexical properties in the writing of foreign language learners over eight years of study: single words and collocations. In Bardel, Camilla, Christina Lindqvist and Batia Laufer (eds). *Eurosla Monographs Series 2. L2 Vocabulary Acquisition, Knowledge and Use. New Perspectives on Assessment and Corpus Analysis*, 127–147. European Second Language Association.

Lexico-grammatical variation in advanced learner varieties. Available at http://www-user.uni-bremen.de/~callies/ALV.htm , accessed April 30, 2015.

LINDSEI. Université catholique de Louvain. Available at https://www.uclouvain.be/en-cecl-lindsei.html , accessed May 17, 2015.

Lorenz, Gunter. 1998. Overstatement in advanced learners' writing: stylistic aspects of adjective intensification. In Granger (ed). *Learner English on Computer*, 53–66. London and New York: Longman.

MacDonald, Penny, Amparo García-Carbonell and José Miguel Carot-Sierra. 2013. Computer learner corpora: analysing interlanguage errors in synchronous and asynchronous communication. *Language Learning & Technology*, 17: 2, 36–56.

McEnery, Tony and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In Eli Hinkel (ed). *Handbook of Research in Second Language Teaching and Learning. Volume II*, 364–380. New York, UK: Routledge.

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

Petch-Tyson, Stephanie. 1998. Writer/reader visibility in EFL written discourse. In Granger (ed). *Learner English on Computer*, 107–118. London and New York: Longman.

Ringbom, Håkan. 1998. Vocabulary frequencies in advanced learner English: a

cross-linguistic approach. In Granger (ed). *Learner English on Computer*, 41–52. London and New York: Longman.

Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97: S1, 77–101.

Tono, Yukio. 2003. Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*, University of Lancaster, 800–809.

West, Michael. 1953. *A General Service List of English Words*. London: Longman, Green and Co.

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

Entrance Examination 2014: Section B

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

English and Englishes

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

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

As we observed at the beginning of this chapter, it is usual for speakers of a language, while welcoming the learning of it by others, to feel a sense of ownership towards it. In the case of smaller and less powerful languages, limited to a particular community in a particular place, this is both unexceptional and unremarkable. Once,

however, a language begins to spread beyond its original homeland and the situation changes and conflicts of opinion begin to emerge. Thus, even until surprisingly recently, many British English speakers regarded American English as an ‘impure' deviation, rather as they might have regarded non-standard forms within their own islands. While such feelings of ownership are to be expected, they quickly become untenable when speakers of the ‘offspring’ variant become, as they are in the USA, more numerous and more internationally powerful than speakers of the ‘parent’.

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

The Indian scholar Braj Kachru describes this situation as one in which English exists in three concentric circles: the inner circle of the predominantly English-speaking countries; the outer circle of the former colonies where English is an official language; and the expanding circle where, although English is neither an official nor a former colonial

language, it is increasingly part of many people's daily lives. At issue is the degree to which the English in each of these circles can provide legitimate descriptions and prescriptions. The rights of the outer circle are now reasonably well established. What, though, of the English used in the expanding circle? Could a new standard international English be emerging there, with its own rules and regularities, different from those of any of the 'native Englishes'?

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

Appendix 2. Misspelt Adjectives and Adverbs

EEIC				LOCNESS			
freq.	**ADJ (42)**	**freq.**	**ADV (24)**	**freq.**	**ADJ (33)**	**freq.**	**ADV (19)**
3	singel	4	definetly	2	extreem	2	unfortunatly
2	differend	1	alreary	2	succesful	1	compleatly
1	awalable	1	constanly	2	unsuccesful	1	completly
1	benefitial	1	contraversely	1	appriate	1	definatley
1	collossal	1	culturaly	1	benificial	1	exponentally
1	convienent	1	defenetly	1	avalible	1	extensivly
1	convinient	1	defenitely	1	bioligical	1	genically
1	damageing	1	definatelly	1	collosal	1	immiediately
1	derrifiying	1	definetely	1	concieved	1	increadibly
1	diffrent	1	easely	1	contreversall	1	moraly
1	dominent	1	efficently	1	democatic	1	morrally
1	easir	1	efficially	1	easyer	1	privatly
1	easyer	1	especcially	1	electral	1	reacently
1	emproved	1	inherrently	1	embarassed	1	relitavely
1	enourmous	1	internationaly	1	forseeable	1	stil
1	exaiting	1	necessarely	1	hardcrafted	1	truely
1	exctinct	1	pherhaps	1	harmeless	1	wholy
1	exiting	1	practicly	1	ineffecutual	1	widley
1	extinqt	1	probobly	1	machanical		
1	facinating	1	secondy	1	neglegent		
1	formel	1	thankfuly	1	nessecary		
1	globalasing	1	therefor	1	resonnable		
1	innevitable	1	unfortunatly	1	responsable		
1	instrinctive			1	reticulous		
1	internationl			1	scandalistic		
1	intersting			1	scientic		
1	linguistical			1	starteling		
1	necassary			1	uncreatative		
1	offcial			1	unforseen		
1	orignal			1	usefull		
1	proffessional			1	vareous		
1	substantan			1	wronge		
1	suprising						
1	techical						
1	underestandable						
1	unforseeable						
1	unfumiliar						
1	unnecesary						
1	unnoticable						
1	unregognisable						
1	younge						

Resümee

TARTU ÜLIKOOL
INGLISE FILOLOOGIA OSAKOND

Anna Daniel

The Use of Adjectives and Adverbs in Estonian and British Student Writing: A Corpus Comparison

Omadus- ja määrsõnade kasutus eesti ja briti õpilaste seas: korpuste võrdlus

magistritöö

2015

Lehekülgede arv: 53

Annotatsioon:

Käesoleva töö eesmärk on kirjeldada omadus- ja määrsõnade kasutust eesti-inglise õppijakeeles, võrreldes seda inglise keelt emakeelena kõnelevate gümnaasiumiõpilaste keelekasutusega. Võrdlusaluseks on järgmised parameetrid: leksikaalne variatiivsus ja keerukus, akadeemiliste omadus- ja määrsõnade osakaal ning kasutatud omadus- ja määrsõnade tüübid.

Töö teoreetilises osas antakse ülevaade uurimustest edasijõudnud õppijate vahekeele teemal, sõnasageduste sobivusest ja kitsaskohtadest õppijakeele kirjeldamisel ning olulistest aspektidest korpuste võrdlemisel. Töö empiirilises osas rakendati korpuste võrdlemise metodoloogiat, kasutades andmete saamiseks allalaaditavat programmi AntConc ning andmete kirjeldamiseks veebipõhiseid programme veebisaidil Lextutor, mis võimaldasid näha teksti leksikaalset profiili lähtudes sõnasagedustest inglise keeles Briti Rahvuskorpuse (British National Corpus) põhjal.

Andmete analüüsi tulemusena selgus, et eesti õppijate omadus- ja määrsõnade kasutus on inglise keelt emakeelena kõnelejate keelekasutusega võrreldes väiksema variatiivsuse ja keerukusega. Ühtlasi kasutavad õppijad vähemal määral akadeemilisi sõnu ning kalduvad sagedamini kordama ühtesid ja samu sõnu. Õppijaid võiks aidata suurema tähelepanu pööramine määrsõnade derivatsioonile ning omadussõnade liitmisele.

Märksõnad: vahekeel e. õppijakeel, õppijakorpused, korpuste võrdlus, sõnasagedused, omadussõnad, määrsõnad

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Anna Daniel, (48710270345)

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

The Use of Adjectives and Adverbs in Estonian and British Student Writing: A Corpus Comparison,

mille juhendaja on Ülle Türk,

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 ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 20. mail 2015

_____

(*allkiri*)