UNIVERSITY OF TARTU

DEPARTMENT OF ENGLISH STUDIES

THE COMPILATION OF THE SPOKEN SUB-CORPUS FOR THE TARTU CORPUS OF ESTONIAN LEARNER ENGLISH

MA thesis

ANNE RAHUSAAR

SUPERVISOR: *Lect.* JANE KLAVAN, *PhD*

TARTU

2019

ABSTRACT

Computer learner corpora (CLC) are electronic stored collections of either written or spoken texts which are produced by learners of a language, as a foreign language (Granger 2004: 124). Computer Learner Corpus research is a fairly new and growing discipline. The problem with most corpora which have been compiled so far is that they are not publicly available, thus they are not accessible for researchers outside of the specific corpus team. In Estonia, studying learner language and using corpora has become more and more popular during the recent years, yet there is still much to learn about the Estonian learners of English. There have been some studies about written learner corpora but studying and compiling a spoken learner corpus is not a common practice yet, mainly because compiling a spoken corpus is a more time-consuming process.

The main purpose of this thesis is to describe the process of compiling the spoken sub-corpus for the Tartu Corpus of Estonian Learner English (TCELE). The aim was to collect the data so that it can be included in the international LINDSEI corpus. Yet, the thesis also sets out to examine why learner corpora can be beneficial for teachers and students, and finally puts this knowledge into practice. A short empirical analysis is done to analyse how Estonian learners of English use the word *well* as a pragmatic marker, comparing the results to Swedish learners as well as native speakers. To further illustrate what can be done with a corpus inside the classroom, one example exercise for the students is created.

TABLE OF CONTENTS

LIST OF ABREVIATIONS

BNC – British National Corpus

CECL – Centre for English Corpus Linguistics

CLC – Computer Learner Corpora

DDL – Data-driven Learning

EFL – English as a Foreign Language

ELT – English Language Teaching

ICLE – International Corpus of Learner English

KWIC – Key Word in Context

LINDSEI – Louvain International Database of Spoken English Interlanguage

LOCNEC – Louvain Corpus of Native English Conversation

OIEC – The Oslo Interactive English Corpus

POS – Part-of-Speech

TCELE – Tartu Corpus of Estonian Learner English

INTRODUCTION

Over the years, a great amount of research has been done to observe how languages work, how teachers should teach languages and how students learn, however the observation of learner language, the actual output of the learner has quite often received less attention (Granger 2002: 3). Luckily, more and more research is done in the field of learner output and one of the main ways for this type of research is the compilation of computerised learner corpora. A corpus is defined as a large collection of written and spoken text, "it refers to a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of it" (McEnery and Wilson 1996: 215). However, Dash (2005) has argued that a corpus does not necessarily have to be large in size as long as it represents the common and specific features, uses the authentic language that was gathered (meaning the text has not been modified to fit the standards) and it has to be available in electronic form. Computer learner corpus (CLC) is an electronic stored collection of written or spoken texts which are produced by learners of a language, as a foreign language (Granger 2004: 124). The decision whether the corpus that will be compiled is a written or a spoken one, depends on the aims of the compiler as well as available resources. Compiling a written corpus is less time-consuming and requires less effort, especially in case the written text is already in an electronic form. Compilation of a spoken corpus requires more time, since speech needs to be transcribed into a text and this process needs to be done with care and in detail, also, the recording process itself requires time and equipment. Although compiling a spoken corpus requires more work, it provides the researcher with valuable material because language is a tool of communication and being able to analyse spoken language thoroughly is necessary for a more comprehensive view on natural language use.

As the compilation of a corpus requires much time and work, it makes sense to think through all of its aims and further uses as there is little point in creating a large-scale corpus for only one study. Tartu Corpus of Estonian Learner English (TCELE) is a corpus being compiled at the Department of English Studies at the University of Tartu. It collects written as well as spoken data from Estonian learners of English. Currently there are more written texts in the corpus than there are spoken ones. The present study therefore provides valuable material for TCELE.

The thesis has two main aims. The first aim is to compile a spoken learner corpus of the texts provided by Estonian learners of English and to describe the whole process of compiling a corpus from the beginning to the end. The second aim has to do with pedagogical implications. The thesis aims to give an overview and illustrate how, in addition to researchers, teachers and students can benefit from corpora. As it is pointed out by many researchers (see, for example, Sinclair 1996, Granger 2008, Gilquin 2015, Timmis 2015), a (learner) corpus can be of great use for a teacher in explaining certain aspects of the language to the students and for providing students with more hands-on tasks.

The thesis is structurally divided into two main chapters. The first chapter gives an overview of the essence and types of learner corpora, mainly focusing on the spoken corpus and the difficulties that commonly arise with respect to compiling a spoken corpus. Finally, the chapter offers knowledge about how corpora have already been engaged in the classroom and what other possible ways there are for teachers and educators to benefit from the use of corpora. The second chapter describes the whole process of compiling a spoken learner corpus, starting from the necessary tools and ending with some of the ways of doing corpus research. The sections of chapter two describe the participants and the methodology, as well as the transcription process and its tools. Included in the second

chapter is an empirical study on the use of the word *well* as a pragmatic marker which analyses the usage of Estonian learners of English in comparison to Swedish learners of English and draws conclusions based on it. Finally, the thesis provides one example exercise created on the basis of the gathered data to illustrate one possible way of making use of corpus and concordance in the classroom. In the Discussion section, the whole process of compiling a corpus and its benefits are discussed. Furthermore, the section discusses the limitations of this study and offers some ideas for future research. The Discussion is followed by Conclusion which summarises the research findings. At the end of the thesis, there is one appendix which includes the example exercise designed on the basis of the compiled corpus.

1. LEARNER LANGUAGE AND LEARNER CORPUS RESEARCH

Computer learner corpora (CLC) are electronic stored collections of either written or spoken texts which are produced by learners of a language, as a foreign language (Granger 2004: 124). The definition of a language learner is a problematic one as people tend to have different understandings about its meaning. Yet, one definition that is accepted by the majority is that a language learner whose data is suitable for a learner corpus is a foreign language learner, a speaker who learns a language which is not their first language nor the second official institutionalised language of the country they live in (Granger 2008: 259). Granger (2008) also notes that another way to explain is to say that only the data from the varieties of English in the Kachru's (1985) expanding circle belongs to a learner corpus. Sinclair (1996, cited in Granger 2008: 260) has defined authentic data as "All the material is gathered from the genuine communications of people going about their normal business". Therefore, in a way, learner data cannot really be authentic, and it is always considered somewhat unnatural because most of the learning in EFL happens inside the classroom and therefore, it is always in a sense artificial. Yet, if students are allowed to write freely (in the case of written corpora), not focusing on the aspect the researcher is interested in, it is considered to be an authentic classroom activity, in terms of learner corpora (Granger 2002: 5). In the case of spoken learner corpora informal interviews, picture descriptions and some types of summaries (re-telling a story, for example) are considered suitable (Granger 2008: 260).

Computer learner corpus is a great step from the way data was previously stored. Now it is much more convenient to analyse and investigate the different aspects of learner language by using certain computer programs. CLC has made it possible to look at different linguistic patterns with the help of various types of annotation, for example error

tagging, part-of-speech tagging (POS), semantic tagging, which all help to describe and analyse learner language (Pravec 2002: 81). Error tagging means annotating the errors in the corpus with a "standardized system of error tags" (Granger 2003: 1). POS tagging is done by assigning each word in a chosen sentence its corresponding part of speech, e.g. verb, adjective, noun, etc. (Brill 1992: 152). Semantic tagging, which is thought to be a step higher on the difficulty level compared to POS tagging, is a type of tagging where words are assigned semantic categories. Semantic tagging therefore gives semantic information about the words, compared to POS tagging, which focuses on the grammar of the word and not on its meaning (Abdou et al 2018: 4881).

According to Granger (2009: 14), learner corpora are more complex type of linguistic corpora as they contain data from language learners and thus require more from the research and the researchers. Granger (2009: 14) lists the following core components of learner corpus research: corpus linguistics, linguistic theory, second language acquisition, foreign language teaching. Therefore, the researcher must have knowledge about linguistics as well as corpus linguistics, about acquiring a second language and some background knowledge about foreign language teaching.

1.1 TYPES OF LEARNER CORPORA

There are different types of learner corpora, most commonly a distinction is made between general or specific, written or spoken, cross-sectional (synchronic) or longitudinal, mono-L1 or multi-L1 data (Gilquin & Granger 2015: 419). In the case of general learner corpora, data can be collected to show the language in all its contexts of use

but specific learner corpora focus on a specific context or specific users (Gabrielatos 2005). In addition to cross-sectional, which comprises data from different types of learners at a single period of time, and longitudinal learner corpora, which means that data is collected from the same learners over a longer period of time, there is also a quasi-longitudinal learner corpus which "gathers data at a single point in time but from learners of different proficiency level" (Granger 2008: 261).

A learner corpus is different from the classic error analysis as it is more contextualised and thorough, since the focus of learner corpus is not on errors or mistakes but on the usage of language in general. Ellis (1994: 64) has stated that to analyse the learner or the learner language we have to be aware of what the learner does correctly as well as what is wrong, not only either one or another. Seeing the used words in context is also one of the most beneficial aspects of a learner corpus, as this is the only way to analyse and make generalisations about how some aspects of the language are actually used.

According to Wichmann (2009: 188), spoken language corpora are mostly compiled for or by linguists who want to understand the essence of human language and communication. There is a difference between two types of learner corpora: the one that provides a written transcription of the spoken text and the other which also provides access to the actual sound file/ the recording (Gilquin 2015). The recordings of the conversations for corpora are often not available, therefore the transcription of the recording is treated as data for the corpus and it is analysed the same way as it would be in the case of data which was written in its origin (Wichmann 2009: 189). Ballier and Martin (2013: 35) distinguish between three types of spoken learner corpora: mute spoken corpora (recording-based transcripts), truly speaking corpora (with access to the recording) and phonetic corpora.

The corpus created for this thesis is currently a mute spoken corpus, meaning that only the created transcriptions are used in the corpus compilation and analysis.

1.2 DIFFICULTIES CONCERNING SPOKEN CORPORA

Spoken corpora begin as a sound not a text, thus, the first step of compiling a spoken corpus is to record the output and afterwards the oral text needs to be transformed into a written text. Yet, the process of transcribing any speech, especially learner speech is tricky because, firstly, the transcriber can hear things differently, secondly, the learner might make some grammatical or phonetical errors (Gilquin 2015: 9). And thus, a problem arises: in case the transcriber decides to use the faulty form of the word, it will not be shown when the standard form of this word is searched, yet, when the transcriber corrects or normalises the word, valuable information might be lost (for example, the error might be specific to learners with a certain mother tongue); so the transcriber is responsible for the transcription that is created, compared to written corpus where the learner is the one responsible for the end result (Gilquin 2015: 10).

Furthermore, in case of a large-scale corpus, there might be several people who work on the transcripts and although they might follow the same set of guidelines, individual differences might still occur in the transcriptions. This will later on result in inconsistency within and across the corpus as well (Creer et al 2019).

In addition to the actual speech the speaker produces, spoken language involves many other factors, such as pauses, laughter and other sounds (e.g. sighing, throat-clearing). All of these factors add value to the produced text and should be marked in the transcriptions, which is a time-consuming process that requires the transcriber to pay

attention to very small details and make notes of those. In the case of learners, long pauses and sighing often indicate some kind of emotional attitude, often hesitance or insecurity (Wharton 2003).

The recordings that were recorded for this study were transcribed as a plain text, extralinguistic features were omitted for these transcriptions because for the sake of this thesis and study, the author needed only the actual speech of the participants. Yet, the transcripts will be later on edited and made suitable for LINDSEI corpus and the extralinguistic features will be added according to the transcription guidelines of LINDSEI. Transcribing the recordings in detail would have beyond the scope of this thesis. As the transcriptions were not phonetical, the author of this thesis did not have to decide whether to write the version of the speaker or the correct one. Even if the speaker pronounced something in a non-traditional way, the word itself was still written using the correct grammar. Speakers made repetitions of some words quite often in the sentences and these repetitions were included in the transcriptions.

1.3 THE IMPORTANCE OF LEARNER CORPUS FOR TEACHERS, FOR TEACHING AND FOR LEARNING

Over the years, a great amount of research has been done to observe how languages work, how teachers should teach languages and how students learn, however the observation of learner language, the actual output of the learner has quite often received less attention (Granger 2002: 3). Yet, researching and analysing learner output gives us much more detailed information about the problematic areas in foreign language learning and teaching. Analysing learner language draws attention to the aspects and elements of the language which could be explained more or differently by teachers as well as what

could be done differently in designing materials and curricula. However, in recent years, more and more research is done in the field of learner output and the main tool for this kind of research is compiling learner corpora, as Granger (2002: 4) states "it is to be hoped that learner corpora will contribute to rehabilitating learner output by providing researchers with substantial sources of tightly controlled computerised data".

According to McEnery and Xiao (2010: 365), a distinction can be made between two different approaches to the collection and use of corpora: the direct use and the indirect use. The direct use of corpora to teaching means that the teachers themselves gather the data from their students, analyse the data, draw conclusions and based on it the same students can benefit from it. This approach also suggests that students should be involved in it and they should be taught about corpora. However, this approach requires more effort and resources because the age of the student has to be taken into consideration, as well as different equipment and programs are required, and more importantly, it is time-consuming and most probably it is quite difficult for the teachers to incorporate this into their curriculum. The indirect use of corpora is usually done by researchers, such as linguists and lexicographers. The data is gathered and analysed by the researchers and as this is a long process, the subjects who provided the data are not usually the ones who finally benefit from the results (Granger 2015: 488). The drawn conclusions and results are mainly used for syllabus and material design. Most language textbooks, curricula, tests and examinations are compiled according to what the creators of these reseources believe are relevant to learners in the target language, but there is no reference to learner language (Mark 2002).

According to Granger (2015: 487), the data from a learner corpus can provide help with three components of material design: the selection, description and sequencing. In terms of selection, learner data from a corpus provides useful information about areas and

aspects of language which are overused, underused, not used at all or misused, this information can, therefore, be taken into consideration when selecting which areas of the language should be taught in which amount (Granger 2015: 487). Computer Learner Corpus makes it possible for material writers do consider the context as well as the background of the learner to draw attention to certain aspects and rules, making the descriptions clear and understandable for learners with specific mother tongues (Granger 2015: 488). For example, the Estonian language does not have different pronouns for males and females, thus mixing *he* and *she* can be a common mistake the Estonian learners of English make. With the help of CLC analysis, the textbook writers could write special notes about it in the textbooks, giving examples of real misuses of the pronouns in context.

Sequencing the order in which different linguistic aspects should be taught is one of the most difficult stages, especially as CLC cannot offer here only one correct answer. However, using CLC definitely helps to make it easier. One way is to use cross-sectional data which is stratified for proficiency (Granger 2015: 488). Here, the researchers need to look at longitudinal or quasi-longitudinal studies in accordance with the proficiency and materials used by those specific learners, then based on the overuses, underuses and misuses, the sequence of teaching linguistic aspects can be adapted.

Ute Römer (2004), for example, did a comparative analysis of modal auxiliaries by comparing the usage of modal verbs in two textbook series (*Learning English Green Line* and *Learning English Grundgrammatik)* to the usage in British National Corpus (BNC). Based on the analysis, Römer drew several conclusions and made several suggestions, including changing the order in which those textbooks teach modal auxiliaries, as she believes that the "more frequent verbs (i.e. more important verbs, at least from the communicative point of view) should be introduced at an earlier stage in the learning process" (Römer 2004: 195).

In addition to the aforementioned benefits of learner corpora for designing study materials and study aids, learner corpora are also used for creating dictionaries and pedagogical grammars for learners. De Cock and Granger (2005) state that monolingual learners' dictionaries and learner corpora make an ideal match, as corpora give information about the problems learners have when learning the language and the dictionaries aim to help the learners with those problems. Learner corpora are also used for pedagogical grammars, here Cambridge University Press is one of the first ones that started to use learner corpora. Learner corpora provides an opportunity to give examples and warnings about common problematic aspects and areas (Granger 2015: 491).

So far, I have discussed how the students can benefit from learner corpora in a rather indirect way, meaning that the researchers or teachers have done all the work with the corpora and students get the end-product. Yet, there are ways to incorporate learner corpora into the study process and into lessons, so that students themselves can analyse, make conclusions and learn by using corpora. However, this practice is not that common yet.

There are several reasons as to why so far using a corpus in an EFL classroom is not that common, for example, time-management issue – lessons are short and there is not enough time for these types of extra activities; syllabus – teachers have to follow specific guidelines as to what they teach and when they do it; proficiency – it is important to consider the language level of the students for corpus-driven tasks. Yet, as pointed out by Granath (2009: 47), most teachers themselves have not had the necessary training as a part of teacher training courses, therefore they do not have the necessary skills to use nor teach using learner corpora.

Granath (2009) also provides three example exercises how students can get familiar with the essence of a learner corpus, find examples to given grammar rules and also realise that rules have exceptions and in different contexts there can be alternative rules. For example, the following exercise (Figure 1) aimed to illustrate the grammar section about British English and American English verb forms with collective nouns.

MicroConcord search SW: couple (The New York Times 1996)

1	for older people. There's an elderly	couple	next door to us who haven't used
2	Soraya Wiriadinata, an Indonesian	couple	who appear on seating charts at a
3	35 freelance writers busy, but the	couple	still run the operation from the
4	e labor would be cheaper. But the	couple	were committed to the mill and to
5	later, they were married. The	couple	now live in Brooline, Mr. Feuerst
6	at Oxford Brookes University. The	couple	were introduced by friends who two years
7	products manager. The	couple	were introduced at the wedding of
8	Rhyne, of the founding family. The	couple	have two children, the oldest, 12-
9	stormed out of the apartment. The	couple	were not married and did not live
10	arpeted platform, upon which the	couple	has placed a chaise longue. A deep

Figure 1. Granath (2009) "Investigating variation in the verb form used with collective nouns."

The provided exercise gives students a selected set of sentences from a corpus (see Figure 1) but depending on the age and level of students, they could also search the corpus themselves, although this would most probably give them too many results which makes it more difficult and time-consuming for the students, yet, they will have a better understanding about the essence of a corpus.

The University of Oslo has created a website (Oslo Interactive English) for their undergraduates where the students can do multiple exercises in which they need to use the Oslo Interactive English Corpus.

> The Oslo Interactive English Corpus (OIEC) is a 7-million-word monolingual English language corpus comprising texts mainly from the 20th century (with a few texts from the 21st century). The corpus includes fictional and non-fictional texts in addition to political speeches and film scripts from the 1980s and 1990s. The major contribution comes from UK and US sources, but texts from other English-speaking countries are also included. (Oslo Interactive English 2015)

OIEC sets a great example as to how a corpus can be used as well as what type of exercises students could be provided with. The corpus itself can only be accessed by the

students of the University of Oslo but the webpage of OIEC gives instructions on how to use a corpus and how to understand the corpus search results and concordance. Although, the corpus is not publicly available, the different exercises are. The exercises are divided into seven main categories. For example, there are exercises dedicated to adjectives and adverbs, function words, differences between English and Norwegian, etc. Overall, the webpage offers great ideas for teachers who would like to make use of corpora in their lessons.

Another situation where a learner corpus can come in handy is when a student asks a question from the teacher, but it is about something the teacher either has never thought about or has taken for granted (Granath 2009: 54). Then the teachers can either do some research with the help of a (learner) corpus on their own and present the students with the results later or the teacher could turn it into an exercise and let students themselves (alone or as a groupwork) find answers to the question.

1.4 LOUVAIN INTERNATIONAL DATABASE OF SPOKEN ENGLISH INTERLANGUAGE

The Louvain International Database of Spoken English Interlanguage (LINDSEI) contains oral data from learners of English with different mother tongues. All the data for the database is gathered following the guidelines set by LINDSEI. The data is gathered from interviews which have a set structure: three conversation topics to choose from, free discussion and a picture description. All the interviews are transcribed and marked using the given conventions. Each interview is linked to a profile which provides information about the interviewer and the interviewee. The database can be used for various research purposes, the studies that have been done using the data from LINDSEI have studied

different aspects of learner English, including lexis, syntax, phraseology, pragmatics and discourse. (UCLouvain 2019) For example, Karin Aijmer has conducted various studies about pragmatic markers, e.g. the use of *well, I don't know* and *dunno;* the research paper written on the use of pragmatic marker *well* (2011) will be further discussed in this thesis. Sylvie De Cock has done numerous studies where she compares spoken and written language as well as use of language by native and non-native speakers of English.

LINDSEI is a project by the International Corpus of Learner English (ICLE), which provides a large-scale comparative study of the learner language of advanced EFL learners who have different language backgrounds. The main purpose of the corpus is to compare the data of learners cross-linguistically to investigate whether certain errors are language specific or universal (Pravec 2002: 83). In addition, Sylviane Granger, the founder and project director of Centre for English Corpus Linguistics (CECL) and ICLE, is determined that ICLE has to benefit the learners, therefore, the CECL has created different types of study aids, for example, an error tagging tool (*Error Editor)*, which allows researchers to tag learner errors and in such way compile lists of typical errors (Pravec 2002: 84).

The data collected for this thesis followed the guidelines of LINDSEI because the aim is to gather the material from Estonian learners of English for the Estonian learners component of LINDSEI. This in turn gives the researchers of TCELE access to the database of LINDSEI, which provides great opportunities for future research in the field, since, it is then possible to compare Estonian learners of English with other non-native learners.

1.5 EXISTING RESEARCH ON LEARNER CORPUS IN ESTONIA

Computer Learner Corpus research is a fairly new and growing discipline, which is said to have its roots in the 1980s (Granger 2004: 123). In Estonia, studying learner language and using corpora has become more and more popular during the recent years, yet there is still much to learn about the Estonian learners of English. In 2016, Lenne Tammiste studied how Estonian learners of English form different types of collocations in written language, based on the learner corpus compiled by Anna Daniel and Elina Merilaine in 2015. This learner corpus comprised 127 essays written as a part of the entrance examination to English Language and Literature BA programme. Using this corpus, Anna Daniel (2015) analysed the use of adjectives and adverbs of Estonian learners and compared it to the usage of native speakers. In her Master's thesis, Merilaine (2015) gave a thorough overview about the compilation of a learner corpus and conducted an empirical study to investigate the frequency and variability of conjunctive adjuncts. The same written learner corpus was also used by Eliza Podburtnaja (2018) for her Bachelor's thesis, as she focused on the correct and incorrect usage of the words *it* and *this.* The previously mentioned studies all investigated the written learner language which seems to be more accessible and less time-consuming. In 2017, Merle Kirsimäe compiled an Estonian spoken mini-corpus of English as a lingua franca and conducted a lexico-grammatical analysis.

In addition to compiling corpora and analysing their content, Aare Undo (2018) calculated the error rate of an automated part-of-speech tagger used for the written sub-component of the Tartu Corpus of Estonian Learner Language and also compared it to the error rates of native English corpora.

2. THE COMPILATION OF THE SPOKEN SUB-CORPUS FOR TCELE

The compiler of the corpus needs to decide upon many important aspects, including what type of text is suitable for the corpus and how it will be added to the corpus; how the files are named and in which format should the files be (Reppen 2012: 32).

The following sections will discuss the stages in the compilation of learner corpus and provide information about the current thesis in connection with those stages. According to Timmis (2015: 15-17), there are following aspects to consider when compiling a corpus and these can be seen as the main broad stages in the compilation of a (learner) corpus:

1. Choosing the participants and the context: deciding what kind of language the corpus is going to represent and whether it is spoken or written.

2. Collecting the texts and storing them electronically. Also deciding upon the size of the corpus.

3. Using mark-up format for labels.

4. Tagging.

5. Analysing corpus data: word frequency count, concordance and collocation.

Stages 1 and 2 are the main and essential stages for corpus compilation because as a result of these stages, the compiler has the collection of text which is considered a corpus (Timmis 2015: 17). The use of mark-ups (codes that keep the labels or extra information about the text separate from the corpus data) depends on the purpose of the corpus. Tagging, which is giving each word in the corpus a ‘label’ is also an optional stage,

depending on the aims of the corpus. Finally, as most probably compiling a corpus has a more meaningful purpose than just the process of creating it, the corpus can be analysed with the help of software. These five stages in the compilation of a corpus are rather broad in the sense that each stage requires that the compiler considers different smaller stages.

The purpose of this chapter is to describe the process of compiling the spoken sub-corpus for the Tartu Corpus of Estonian Learner Language (TCELE). The aim was to collect the data so that it can be included in the international LINDSEI corpus. Therefore, the given guidelines for LINDSEI corpus were used, firstly, for choosing the participants, secondly, for gathering the data and thirdly, for naming the files. LINDSEI has also provided their guidelines for the transcription process, but as this aspect did not fit within the workload of the present project, the transcription guidelines are not yet implemented. LINDSEI collects orthographic transcriptions, as opposed to phonological transcriptions, thus the plain texts that were created for this thesis can later easily be edited and modified to be suitable for LINDSEI.

2.1 DATA COLLECTION

According to Hunston (2008), the compiler of a spoken language corpus has to decide upon three important aspects: selection of the speakers and social contexts; management of data collection; the choice of transcription system.

The mini-corpus of this study is compiled of 10 interviews and about 133 minutes of speech. The interviews were conducted in the period of October to December 2018. The interviews followed the format outlined by the LINDSEI corpus. The average length of the interviews is 12 minutes, the shortest interview lasted for 8 minutes and the longest lasted for 17.5 minutes. The interviewer was the author of this thesis who is a native speaker of Estonian.

The interviews were semi-structured, the interviewees were limited to choosing one

topic to talk about but afterwards the interviewer asked additional questions about topics that are related to the one chosen by the interviewee as well as questions about general topics, such as hobbies, languages, life at university, etc.

The interview consisted of two major parts, which both can be divided into some smaller segments. First, the interviewee was asked to choose one topic out of the given three, to talk about for about three to five minutes as a monologue. The topics also suggested some ideas about what the interviewees should mention in their monologue. The three topics were: an experience which has taught an important lesson, a visit to an impressive country, description and opinion about one film or play. The interviewees had a moment to choose a topic and they were expected to start talking about it instantly, they were not allowed to take any notes, as it was important that their speech was spontaneous and natural. When the interviewee had spoken for a while, the interviewer started to ask additional questions related to what the interviewee had talked about. As the last segment in this part, the interviewer asked the interviewee questions about some general topics, mainly about hobbies and university life. The second part of the interview, which was meant to be shorter than the first one, was a picture description task. The interviewee was asked to look at four pictures which made up a story and retell the story by describing the pictures and analysing the situation. As the last part of the interview, the interviewer asked some follow up questions about the story.

The most popular topic discussed was the review of a movie or play. Two of the participants decided to talk about a country they have visited and only one person spoke about an important life lesson they have learned.

All the participants of the study had to fill a learner profile questionnaire and a consent form. As these interviews and all the documents connected to those were primarily done for the LINDSEI corpus, the documents cannot be included in the appendices of this

thesis for privacy and copyright reasons.

2.2 PARTICIPANTS

All the participants interviewed for the study were native speakers of Estonian. They were third or fourth year students of English philology at the University of Tartu, all together 10 interviewees. 7 of them were students on the bachelor's level, 3 of them were students on the master's level. There were 4 male and 6 female interviewees. They were not asked to assess their English skills but they all had to fill in a learner profile which gave a general overview about their English language background. The learner profile included a question about the mother tongue of the interviewees as well as their parents. For the purpose of gathering data for the LINDSEI corpus, it was important that all the chosen participants were native speakers of Estonian.

They were also asked about the medium of instructions on different levels of education (primary school, secondary school, university). All the participants of this study had Estonian as their main medium of instruction in primary school and secondary school. In university, the main medium of instruction for them was English but as there are also courses in their native language, Estonian, they also added this as their medium of instruction in university. The participants were also asked to write other foreign languages, other than English, they have studied. Most common foreign languages were Russian, German, French, Spanish, other languages that were written by some students were Latin, Ancient Greek, Korean and Japanese. The average number of foreign languages each participant could speak was two, the minimum was one and maximum was four.

2.3 TOOLS

The next steps, after deciding on the aim and context of the corpus and choosing

the participants, is choosing a recording device and the location for the interviews. When choosing a device for recording, it is important to do some research in order to choose the one that can provide quality sound files, as the recordings form the basis of transcriptions (Gilquin 2005: 19).

For the present thesis, the interviews were recorded using Tascam DR-05 which is a 24-bit/96kHz Digital Recorder. It was easy to use - transferring multiple files from the recorder to computer took only about a minute and the files can be accessed in MP3 or WAV format. During the recording, it was necessary to adjust the input levels manually at the beginning of each recording to ensure that the levels match with the volume level of each speaker.

All the interviews were recorded in the same small seminar room with little furniture and a glass wall, in addition to the wall with windows. Therefore, there might have been a slight echo in the room, but after listening the recordings, the author did not notice echo. One of the problems with the room was that at certain times, other people would enter or exit the rooms next to the interview room and as the walls were not really soundproof, this type of noise can also be heard from the recordings. Yet, fortunately, the text of the speakers was not lost due to this and was still understandable. Also, during one interview the chimes of the town hall were playing for several minutes and again, although, this can be heard from the recording, it did not interfere with the transcribing process.

Although the recorder used for the interviews created high-quality recordings, there were some recordings in which the quality was not very good, mainly the volume was too low. This was mostly because the interviewer did not adjust the input level, or the interviewee changed their volume during the course of the interview or somehow sat further away from the recorder. Therefore, an additional program, Audacity (2000) was used to amplify the voice of some recordings. Audacity is an open source, cross-platform

editing software, which can be downloaded and used for free of charge. The author first tried to manage without using this program, but the automatic transcription program did not detect most parts of the interview if the volume was too low, which is understandable, as it was even hard for the author to hear what was said. After amplifying the sound, the author was able to hear the recording better and make necessary changes in the automatic transcription.

Gathering data for a spoken corpus means that the oral components need to be transferred into a written form in order to analyse them. The corpus compiler has to decide whether orthographic or phonological transcriptions are needed for the corpus and then transcribe the recordings accordingly. Transcribing is the first step that has to be taken in order to further work with the text. It is possible to have the spoken texts transcribed automatically, instead of listening to the recordings and manually transcribing them. Finally, when the spoken texts are in written form, the corpus compiler can use programs to start analysing the data, either manually or using different programs, depending on the aims of the study. According to Timmis (2015), there are three most commonly used analytical operations: word frequency counts, concordance and collocation. If necessary for the study, the texts can also be annotated and tagged, using different tools.

For the purpose of this thesis, the author decided to try different programs available for automatic speech to text transcription and for corpus analysis to demonstrate some of the things that can be done with the spoken sub-corpus. Furthermore, the following paragraphs will discuss the benefits and problems of different programs as well as the different stages and actions necessary for transcribing and analysing the data further. The sections will discuss transcribing, annotation, word frequency counts (statistics about the data), concordance, collocations and the tools used for those procedures.

2.4 AUTOMATIC TRANSCRIBING: TEMI

For this MA thesis the author decided to test automatic speech to text programs that can be found online, instead of transcribing all the interviews manually. De Cock (2010) differentiates two types of transcriptions: broad and narrow transcriptions. Broad transcriptions provide just the text, what was said, yet, in addition to plain text, narrow transcriptions also provide details about the intonation, voice quality and stress (De Cock 2010: 125). The study for the present thesis did not require specific details about the way the speaker spoke, thus, the transcriptions are broad.

After testing some programs and websites, temi.com was chosen for the purpose of this thesis. Temi (2019) is an online speech to text software which allows registered users to upload their audio files and transcribes them in about five minutes. Although using the program is very convenient and user-friendly, it still cannot completely do the entire transcribing process for the user. Gilquin (2015: 11) explained that it will probably take years or decades before "spontaneous learner speech can be transcribed accurately in a fully automatic manner". Depending on the quality (background noises, echo, voice and sound level, etc) of the audio file, it gives as accurate transcription as possible, but the user definitely has to make a few (see Figure 2 and Figure 3) or in some cases quite many changes and adjustments (see Figure 4 and Figure 5).

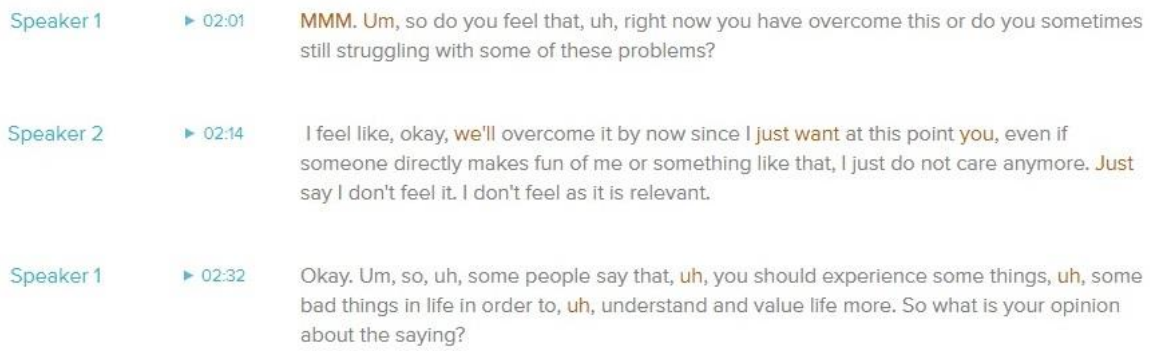

Figure 2. Segment of an automatic transcription by Temi.

The program marks the words and phrases of the conversation which were more difficult to

understand with a different colour so that the user can notice those places easily. Here, in the first example (Figure 2), the program has only marked eight parts of the conversation and most of them are fillers (*MMM, Um, uh).* It is understandable that an automatic program is not able to detect for certain whether the speaker used fillers or maybe just the recording quality was not good in that part for some reasons (the speaker turned away from the microphone or there might have been a loud background noise).

Speaker 1	▸ 02:01	Mm. Um so do you feel that, uh, right now you have overcome this or do you sometimes still struggle with some of these problems?
Speaker 2	▸ 02:14	I feel like I've overcome it by now since I just um well at this point even, even if someone directly makes fun of me or something like that, I just do not care anymore. So to say I don't feel it. I don't feel as it is relevant.
Speaker 1	▸ 02:32	Okay. Um, so, uh, some people say that, uh, you should experience some things, uh, some bad things in life in order to, uh, understand and value life more. So what is your opinion about this saying?

Figure 3. Edited version of Figure 2.

The user can then easily listen to the exact parts of the text and make necessary changes in the same program on the website. In addition to paying attention to the parts with different colour, the user should still listen and edit other parts too. For example, in these provided segments (see Figure 2 and Figure 3), some other words had to be edited (instead of *struggling*, the speaker said *struggle;* instead of *the*, the speaker actually said *this).* Another aspect that can be seen from this example, is that in some parts the automated program avoided repetition of words, yet in reality, the speaker sometimes started to say a sentence, then paused and started the sentence again by repeating the words they had already said.

In some cases, the automated transcription had a rather poor quality and quite a big part of the given text had to be edited. There can be many possible explanations as to why some transcriptions have better quality and some have worse. The author of the thesis noticed some reasons why the quality of the automatic transcriptions can be poor: the speed of speaking is too fast, the speaker spoke too quietly, the speaker said something

while laughing, the speaker did not properly finish one word before saying the next one.

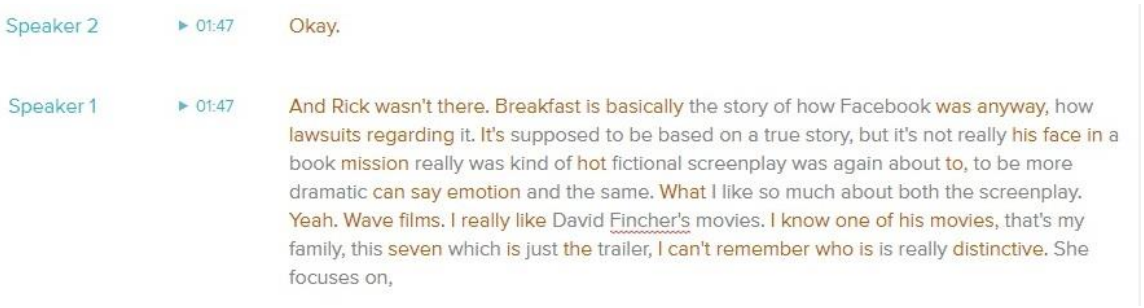

Figure 4. Segment of an automatic transcription by Temi.

Here is an example of a segment from the transcription (see Figure 4 and Figure 5) which had noticeably more parts in which the program was not sure about.

Figure 5. Edited version of Figure 4.

In the case of this interview, the speaker talked rather fast and quite often said some things under breath. It was quite hard to distinguish some words even after listening the interview several times, therefore, it is understandable why the program struggled that much too. The speaker referred to three famous people, the program was able to automatically detect one out of three (*David Fincher*).

Also, one aspect that can be noticed when looking at the raw version of the transcript (Figure 4) and the edited one (Figure 5), is the difference in the way the speakers are labelled. Temi automatically assigns labels to the speakers, the one who starts the conversation is Speaker 1 and the speaker who starts talking next is Speaker 2, yet quite often the program did not distinguish the two voices and mixed up the speakers. Therefore,

in addition to correcting the spoken text, the user also has to pay attention to assigning the right text to the right speaker. If the user wants to use other names for the speakers, they can be easily changed according to the needs.

Overall, the author of the thesis concludes that Temi is a good program to use for transcribing, especially if the recordings have good and clear quality. It takes about 5 minutes, after uploading the recording file, for the program to provide the automatic transcription which can then be edited by the user. It also sends an email to the user when the transcription is ready. It is important to note that Temi is not a free program. The user can first test the program for free with one recording and transcription. Therefore, it may be quite a good deal for someone who only has one recording. After that, the user has to pay for each recording separately, the price depends on the length of the recording. It is also worth mentioning that currently TEMI supports only English audio and also video files.

Temi also allows the user to download the final transcription in different file formats. Considering the aims of this thesis, all the transcripts were downloaded in *plain text* format, which makes it suitable for different programs, for example AntConc.

2.5 DATA ANNOTATION

Data annotation is one of the optional stages in learner corpus research. The use of linguistic annotations depends on the aims of the research. In case the aim is to concentrate on analysing one word/term, the "raw corpus" would be enough, yet to analyse, for example, one group of words or a category, manually searching through a raw corpus can be time-consuming and more prone to errors (Granger 2012: 18). One of the most common corpus tools are Part-of-Speech (POS) taggers, programs which assign "contextually appropriate grammatical descriptors to words in texts" (Voutilainen 2005). Yet, although

using programs is less time-consuming and in some cases might be more accurate, the researcher should still remain critical and take the time to double-check the results, as the program does not always provide accurate tags (Granger 2012: 18). According to the study done by Undo (2018), where he calculated the error percentage of an automated POS tagger, the results showed that while the error rates for TCELE written text corpus were rather low and acceptable (1.06% in average), the error rate for the spoken sub-corpus of TCELE was remarkably higher, 23.14%. It is therefore concluded that for spoken data, additional manual tagging should be applied. POS-tagging has not been undertaken within the present thesis due to time constraints; this remains an important task for future research.

2.6 CONCORDANCE: ANTCONC

After collecting all the texts and considering it a corpus, the corpus researcher needs to start analysing the data from the corpus. Depending on the aims of the research, the researcher has to choose and use programs to apply linguistic techniques. One of the most common ways to display search results in a corpus is using a concordance, which "is a list of examples of a word as they occur in a corpus, presented so that the linguist can read them in the context in which they occur in the text" (Wynne 2010: 706).

The program AntConc (Anthony 2014) was used for creating concordances for the corpus. AntConc is a program which can be freely downloaded without any additional costs or regulations. It is used for concordancing, creating word lists and keyword lists, analysing collocates and clusters.

Currently one of the most common corpus-linguistic tools in use is the key word in context (KWIC) concordance, which enables to see "the word of interest in its immediate

context" (Gries and Newman 2013: 277). This tool makes it very easy to analyse the use of one word in its context.

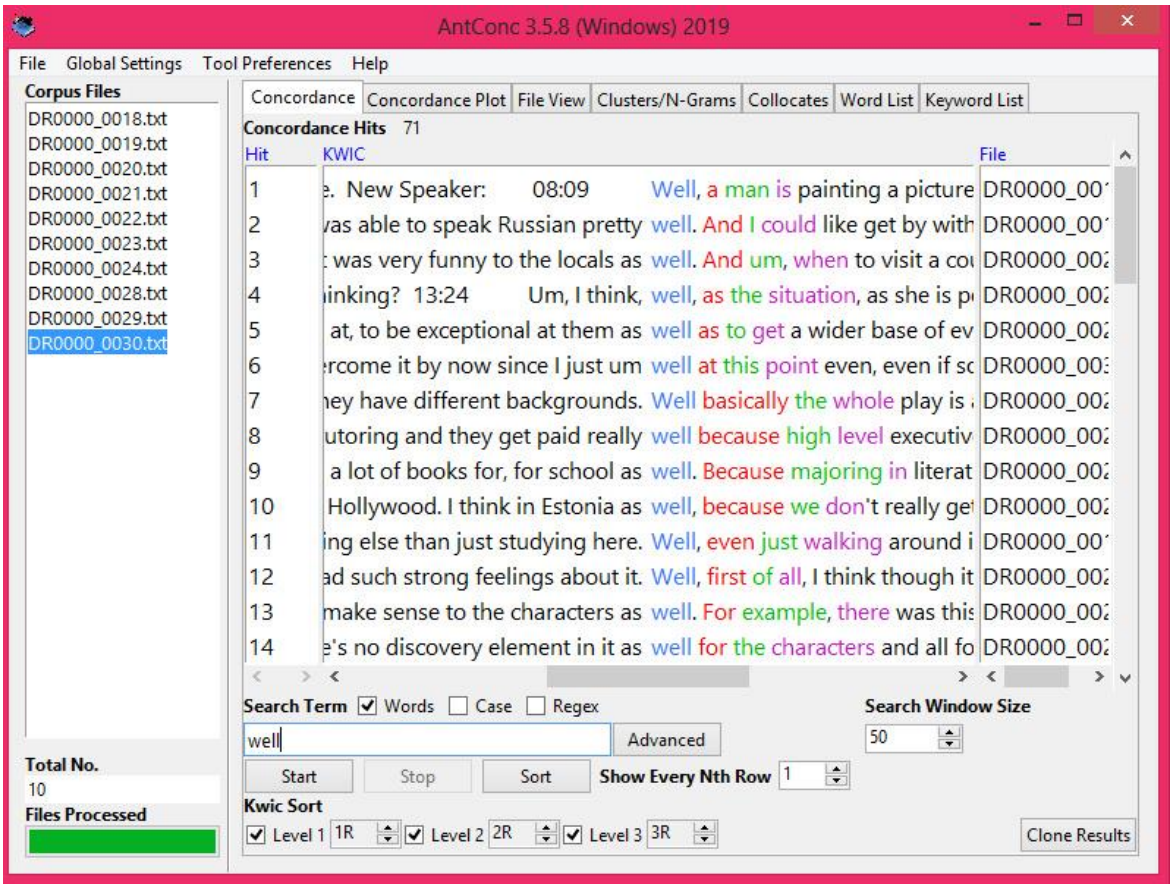

Figure 6. AntConc concordance search with the word *well*.

Figure 6 shows the concordance search results of the word *well*, the first 14 results to be more exact. In total, there were 66 concordance hits. Here, the author used all 10 interview transcriptions. The program makes it really convenient to analyse the usage of a word (here *well)* in its context, as well as make notes about which text the lines originate from.

2.7 FREQUENCY COUNT

Although it is possible to manually search and create lists according to the frequency rates of one word or term, it is much more convenient to use a program for it. There are frequency lists for words, collocations and chunks (Timmis 2015: 42). For

example, word frequency lists are especially beneficial for teachers who would like to see or demonstrate to students how they tend to overuse some words. Frequency lists for collocations can show what words are most often used together with the searched word.

All together the corpus created for this thesis consists of 14,640 words (including the text produced by the interviewer as well as the interviewee). There are 1727 different words used. The most popular word used is the personal pronoun *I*, which has been used 554 times (see Table 1). The second most frequent word is the preposition *to*, followed by the definite article *the.*

Table 1. Ten most frequent words in the compiled corpus.

Rank	Frequency	Word
1	554	I
2	485	to
3	471	the
4	401	and
5	353	you
6	342	it
7	311	a
8	282	of
9	281	that
10	252	so

The ten most frequently used words in this corpus are pronouns, prepositions, conjunctions and articles, therefore this list would not actually help a teacher in showing how students tend to overuse some words because all the words are essential. Although, the word *so* (ranked number 10 in Table 1) can be an important component in a sentence, it is often also used as a pragmatic marker. The program allows the user to click on a certain word and explore the sentences in which the words have been used. If looking at the contexts of the word *so*, it appears that it is mostly used in phrases such as *or so* and *and so on* but in most cases it is used as a word that starts a question or a new sentence.

When comparing the frequency list of Estonian speakers of English to the frequency list provided by British National Corpus (BNC), almost all the words are the same in both tables, only there is the preposition *in* instead of *so* in the BNC list (see Table 2). The most frequent word for British speakers, based on the list, is the definite article *the,* followed by the personal pronoun *I,* which was number one in the list based on Estonian speakers of English. Generally, the order of frequency of the words is mostly the same.

Table 2. 10 most frequently used words in spoken language based on BNC (Leech *et al* 2001)

Rank	Word
1	the
2	I
3	you
4	and
5	it
6	a
7	to
8	of
9	that
10	in

2.8 COLLOCATIONS

Collocations are language-specific, meaning that an English collocation, for example, cannot mostly be directly translated into Estonian using the same words (Loigu 2006: 9). *Oxford Collocations Dictionary for Students of English* defines collocation as the way words combine within a language to produce "natural-sounding speech and writing" (2003: vii). Yet, for a beginner learner of a language the notion "natural-sounding" is difficult to comprehend because learners tend to depend on their mother tongue when creating their phrases and sentences in a foreign language. This is completely normal and part of the learning process, but teachers have to be aware of this problem and be ready to talk about the differences in languages as well as teach and compare collocations. One way to teach students collocations and show them in context is by the use of a corpus.

Nesselhauf (1996: 41) states that the use of learner corpora has many advantages compared to "elicitation tests" (basic exercise types such as fill-the-gap or judgement tasks). It provides the students with the opportunity to analyse their own text, the collocations they used and which they were aiming to use (Nesselhauf 1996: 41).

To learn and teach about collocations, the corpus should be compiled of a number of texts. The corpus designed for the current study did not provide any significant search results for collocations. Yet, the 'collocates' function in AntConc program can be useful in other ways as well. For example, the choice of prepositions in different phrases or phrasal verbs can be observed. It is common that students switch between saying "in the picture" and "on the picture". The correct way to describe a picture or a photo is to say "There is a … IN the picture", unless something is physically on top of the picture, then ON could be used. When using the 'collocate' function in the created corpus to search the word *picture*, there are various words connected to it, including *in* and *on.* There are 7 instances of using *on the picture* and only 5 instances of *in the picture.* It is very easy to analyse and observe this kind of use of prepositions with the help of the 'collocate' function, in addition, this could be shown to the students when discussing common mistakes. Or the teacher could have the students work out the rule on their own, using the corpus research results as well as the material that can be found online.

2.9 AN EMPRICAL CASE STUDY: THE USE OF *WELL* BY ESTONIAN LEARNERS OF ENGLISH

The main purpose of this thesis is to focus on the compilation of a learner corpus and its stages and also draw attention to the ways a learner corpus (and the findings of its analysis) can be used. It aims to show how, in addition to syllabus and material designers

and creators, teachers and students can benefit from the use of learner corpora as well. Thus, to further illustrate the theoretical part about the useful aspects of the corpus, the author of the thesis will also analyse the use of the word *well* by the interviewees, draw conclusions and create an example exercise for students based on the analysis. The exercise would show how the teacher can make use of a (learner) corpus in a lesson.

The way Estonian learners of English use the word *well* is compared to Swedish learners of English. Karin Aijmer (2011) conducted a study using the data from the Swedish component of the LINDSEI corpus and LOCNEC (*Louvain Corpus of Native English Conversation)* to analyse and compare the use of pragmatic marker *well*. The results of Aijmer's (2011) study showed that Swedish learners tend to overuse *well*, and learners mostly use *well* "as a fluency device to cope with speech management problems but underuse it for attitudinal purpose". She determined the overuse and underuse by comparing the frequency of the use of *well* among Swedish learners to native speakers' usage, considering that the frequency of native speakers is the norm.

Since Aijmer used the interviews done with Swedish participants for LINDSEI corpus for her analysis and LINDSEI has the same guidelines and structure for all the interviews in the corpus, then the results of the study done by Aijmer are nicely comparable with the current study. In her study, Aijmer (2011: 251) suggested that for further studies about *well,* it would be important to find the similarities and differences between learners with different mother tongues; to study whether all the learners have the same problems. She also pointed out that the native language of the learners might also affect the use of pragmatic markers, depending on how often and which type of pragmatic markers are used in their native languages.

The word *well* is mainly taught at school as either an adverb or an adjective, not as a pragmatic marker. Pragmatic markers are connected to, usually informal, spoken language and they "express pragmatic aspects of communication" (Andersen 1998: 147). These kinds of pragmatic markers are typical for native language speakers or speakers of higher proficiency. Since they are not taught at school in this way, learners might pick up the use of pragmatic markers from either listening to native speakers (e.g. from media) or also from their teachers.

Aijmer (2011) grouped the use of *well* into two categories: speech management and attitudinal (see Figure 9). In order to make further comparisons with the use of *well* between Estonian and Swedish learners of English, this study will also follow the same way of categorising. In the speech management category, the use of *well* functions as a way of breaking the utterance, thus, it is used when planning or searching for a word, clearing and reformulating what has been said (Aijmer 2011: 236). *Well* also functions as an expression of attitude "to the hearer or the preceding discourse", it can be used in disagreeing contexts (e.g. correcting what has been said, denying something, also when refusing to give a direct answer), for providing an opinion or when confirming something that has been said (ibid.).

Speech management
- Choice
- Change
- Prospective (introducing new turn)
- Marking stages in a narrative
- Quotative

Attitudinal
- Opinion
- Disagreement

Figure 9. "The functional typology of *well* used to compare native and non-native speakers." (Aijmer 2011: 236)

Biber et al. (1999: 1096) found that *well* is a more frequent pragmatic marker than, for example, two other frequent ones, *I mean* or *you know*. Aijmer (2011) also found that *well* was used more often than *I mean* or *you know* among both, the Swedish speakers as well as the native speakers. In the corpus created for this thesis, there were 10 instances of *I mean* and 16 instances of *you know*. Therefore, *well* was used 6.6 times more often than *I mean* and 3.9 times more than *I know*, which concludes that well is definitely a more frequent choice. Swedish learners had 282 instances per 100,000 words for *you know*, Estonian learners had even less, 109 instances. Also, Estonian learners had used *I mean* noticeable less than Swedish learners (229 instances per 100,000 words in the Swedish corpus, 68 instances for Estonian learners).

2.9.1 DATA

To understand whether Estonian learners of English tend to overuse or underuse *well* as a pragmatic marker and whether they use it more for speech management or attitudinal functions, the texts had to be looked through manually, divided into those two categories (Speech management and attitudinal) and the percentages had to be calculated. Altogether, the program detected 71 times the word *well* was used (see Table 3), yet after going manually through them, there were 57 instances of *well* being used as a pragmatic marker, thus 14 instances were not considered for this analysis. These were instances where *well* was used for comparison (*as well as*) or as an adverb (*well, as well*).

Table 3. Distribution of the two functions of *well* among Estonian and Swedish learners.

	Estonian learners		Swedish learners
	n	**%**	**%**
Speech management	51	89	79.6
Attitudinal	6	11	20.4
Total	57	100	100

Figure 10 shows that both, the Estonian as well as the Swedish learners use *well*

considerably more for speech management. In most cases *well* was used at the beginning of the sentence, to signal that the interviewee understands that it is now their turn to talk and the answer is coming. This notion is called the prospective *well* (Aijmer 2011: 241-242). In her study Aijmer found that Swedish learners overuse well for speech management when comparing her results to the results of LOCNEC (considering the percentage of native speakers as an average), in which the speech management percentage was 65% (Aijmer 2011: 248). The same applies then to Estonian learners of English as well, since the difference between the percentages, which show the percentage of function use from the total, is 24%. Although, Aijmer (2011) also noted that this type of interview situation definitely has an influence on the way how and how often pragmatic markers are used by the learners. In everyday speech or other kind of natural conversation, a person does not feel the need to signal their turn nor pays that much attention to the choice of words. Therefore, these results give us an insight and an overview of how *well* is used as a pragmatic marker by learners but only in a rather artificial environment.

Table 4. Use of speech management and attitudinal functions by each Participant

	Speech management	**Attitudinal**	**Total use of well**
Speaker 1	7	2	9
Speaker 2	0	0	0
Speaker 3	5	0	5
Speaker 4	6	0	6
Speaker 5	5	0	5
Speaker 6	2	1	3
Speaker 7	10	2	12
Speaker 8	8	0	8
Speaker 9	0	0	0
Speaker 10	8	1	9
	51	**6**	**57**

Table 4 shows the distribution of *well* as a pragmatic marker in the chosen two categories among all the participants of this study. As it can be seen, there were two interviews (number 2 and 9) in which no instances of *well* as a pragmatic marker occurred.

It is important to note that interview number 2 was also the shortest interview, only 8 minutes long. Both of those participants did not really use any other frequent pragmatic markers (*I mean* and *you know)* as well, only the speaker in interview number 2 used *I mean* once. It can also be seen that one of the participants (Participant 7) used *well* particularly more than others, 12 times in total. This participant gave a lengthy overview about a theatre play and Aijmer stated that *well* is very often used for marking stages in a narrative, which is what Participant 7 mostly did (Aijmer 2011: 243). For example, *Well basically the whole play is about family trauma, because every person in the house, is different in the sense of personality.*

2.9.2 EXAMPLE MATERIAL

As has been mentioned before, not only researchers and teachers can benefit from using corpora, students can learn from it as well. The teacher can provide the students with tasks and exercises to focus on a certain learning aspect while making use of a corpus. This kind of method is called data-driven learning (DDL), a term that was coined by Tim Johns in 1990. In DDL, the teacher might give the students a printed out version of a concordance with a number of instances of a certain word or phrase and the students are then asked to make some observations on the use, meaning and grammatical properties of the word (Timmis 2015: 10).

There are several factors to consider when creating a corpus-based exercise for students: language level, age, time amount, exercise type, topic, source. Depending on the age and capabilities of the students, the teacher needs to decide whether the students have to search the corpus for answers themselves or they are provided with excerpts or chosen sentences. If the teacher decides to let the students search the corpus on their own, it is essential that they get educated on the essence of corpus prior the research, because

otherwise it will take too much time and effort for the students to figure it out. Of course, the teacher could try the inductive approach and let the students figure out how to use a corpus and a concordance on their own, yet, in most cases teachers have time-constraints and it just is not reasonable. Also, in case the students have to search the corpus themselves, it is necessary that they all have the technical possibility to access the corpus.

The teacher can decide to compile a corpus and provide students with exercises based on that corpus and their own language or the teacher can use an existing corpus. It all depends on what the aim of the exercise is and how much time the teacher can spend on the whole activity. Another way of making use of a corpus is asking the students themselves to create small exercises for their classmates by using a concordance, however, this is possible only when the students are well-acquainted with using concordance and corpus (Loigu 2007: 37).

To give one example how a learner corpus can be used as a study material, I created an exercise (see Appendix 1) using the current study about the word *well*. For the exercises other instances of well, beside *well* as a pragmatic marker, were also included. This exercise does not ask the students to search the corpus themselves but provides them with 10 selected sentences from the corpus. The students are asked to find instances of *well* in each sentence (there are actually two in sentence number 9) and decide whether the sentence would also work without it. This exercise can be used to teach the students about pragmatic markers and to draw their attention to the fact that *well* has other functions beside the most frequent (*well* as an adjective or adverb). Then the teacher could also talk about other pragmatic markers and generally about why these kinds of markers are used. The teacher can explain that pragmatic markers can make sentences more fluent or give the speakers some additional time to think about what they are saying next.

The teacher could also talk about corpora and show the students the concordances from which the sentences of the exercise originate from. They could search some keywords together and look at the most frequent words – there are many aspects that could be discussed in the classroom.

This exercise is not suitable for beginners or elementary level but could be modified for pre-intermediate and everything above it. With this exercise, it is important that the students have previously studied the comparative forms with *well* and know the function of *well* as an adverb and an adjective. Of course, the teacher could make the necessary changes and make the exercise suitable for elementary level as well, but the approach and the aim of this exercise should then be different.

2.10 DISCUSSION

The aim of this study was to compile a spoken learner corpus, describe the process and give an overview why learner corpora can and should be used by the teachers as well as students. Compiling a corpus is a multilayer process, it begins with choosing the participants, thus, choosing what language is going to be the source material for the corpus. The compiler also has to decide whether it is going to be a written corpus or a spoken corpus.

Compiling a spoken corpus is a lengthy process and takes much time and effort. Nevertheless, spoken language of learners is a valuable material, providing very many different aspects for research. A spoken corpus should definitely not be compiled just for one study and then forgotten, there are so many more possibilities to make use of it. It is especially fascinating to find a study which is done with non-native learners of English with another mother tongue and analyse the same aspects about your mother tongue for comparison. Here, it is very important that the way the material was gathered and the background (level of proficiency) are the same, because only then it is possible to make authentic and reasonable comparisons.

The process of transcribing spoken language is tricky, because there are so many things to pay attention to. For example, in addition to the speech, the non-vocalised aspects (e.g. yawning, sighing, laughing) which add a certain meaning to the speech and have to be marked in the transcription as well. It is then important that a fixed guideline for transcribing is followed, to ensure consistency, although even then there could be individual differences, especially if the guideline itself has some gaps (Andersen 2016). For this study only the sound was recorded, but it could also be possible to record a video which would then add more meaning to the text, as body language and gestures also form a

crucial part of the communication. Diemer *et al*. (2016) discussed the ways how *Skype* could also be a way of collecting informal spoken data. I believe that another type of data collection should definitely be experimented, in case of gathering spoken language material produced by learners, because this type of interview with guidelines is still a quite artificial form of communication. Thus, it does not always provide accurate and authentic language actually used by learners.

For an ordinary teacher, it is definitely easier to gather material for a written corpus, for example, the students could write essays. If the students write the essays electronically and submit them to the teacher, the teacher will already basically have the corpus. Then the teacher needs to decide upon ways of analysing it, and reasons behind it: whether the teacher is going to analyse the written language and aspects of it or the students are going to receive something out of it as well. Using the AntConc program or any other similar program for concordancing could definitely benefit teachers. The program is easy to use and fast, yet, advantages for a teacher are enourmous. The teacher can analyse the frequency of words and decide whether there are words that are overused or underused.

Concordancer can be used for searhing and analysing the typical errors of those students – quite often teachers create lists of most common mistakes in essays or tests to show the students afterwards, the use of a concordancer could make this process of searching less time-consuming. Thanks to a program like AntConc the teacher can notice aspects of the language that need more work and additional tasks can be created. As Granger (2015: 487) said, the data from a learner corpus can have an effect on three stages of material design: the selection, description and sequencing. The teacher might realise after analysing the corpus that the current method or materials are not suitable or something could be improved. Overall, the use of a corpus is beneficial for the teacher and

consequently for the students. Therefore, I suggest that the teachers should be introduced to corpora and learner corpora already during the teacher training years. This way the teacher is aware what the advantages of using or compiling a corpus are and will be less reluctant to make use of it during the actual teaching years.

As this study collected and analysed the language of proficient non-native learners of English, it made sense to look at a more complex language aspect for analysis and empirical study. Pragmatic markers are more common in spoken language than they are in written language and the reasons as to why they are used vary greatly. Estonian learners of English use the pragmatic marker *well* mainly at the beginning of the sentence as there were 32 instances of *well* as the first word in a sentence (out of the total 57 instances of *well* as a pragmatic marker). Aijmer (2011) compared the use of *well* by Swedish learners of English to other non-native language learners and suggested that it could then be possible to analyse the use from a mother tongue specific point of view. Although, *well* was used for speech management more by both the Estonian and Swedish learners than it was by the native speakers, it was clearly underused for attitudinal function. Aijmer (2011) stated that this could be because Swedish does not have a pragmatic marker which completely or partial corresponds to *well* and thus, the learners are not familiar with using *well* in its attitudinal function (to disagree with the speaker or to give their opinion).

The Estonian equivalent to *well* could be *noh* but this also mostly occurs at the beginning of the sentence or when clarifying something, it is not commonly used for disagreeing. Another reason why *well* might have been used less for disagreeing or giving an opinion, is the form of the conversation which was an interview. The interviewer mainly asked questions and did not say anything provocative nor gave the interviewee statements to react on. Therefore, the study of attitudinal *well* could be replicated when the

environment and context are more natural, for example, a conversation between two learners.

The example exercise (Appendix 1) that was created for this study is just one example how the results of a concordancer can be used for creating an exercise. As mentioned in section 1.4, there are webpages (e.g. OIEC web page) and books dedicated to different types of corpus-based or data-driven learning exercises. In his book *Corpus Linguistics for ELT*, Timmis (2015) provides theoretical as well as practical knowledge for teachers of English about the use of corpus in a classroom environment. The exercise created in the context of the present thesis is connected to the empirical study of the pragmatic marker *well* and enables the students to later discuss other topics in addition to the regular grammar and vocabulary. The instructions of the exercise can be modified by the teacher to be suitable for certain students with a certain level of proficiency. The sentences could also be used for something completely other than the exploration of the word *well.* Also, the sentences provide other pragmatic markers (*so* and *like)* and phrases that in school environment are taught to be informal (*and stuff)*. Students are generally encouraged to not use such colloquial words and phrases and they need to find alternatives which are more formal or academic. Therefore, this exercise could lead to an interesting discussion in the classroom about the register and different situations where different registers could be used.

To sum up, the compilation of a spoken corpus is a lengthy process. Considering there were 10 interviewees for this study, the average time spent with each interviewer was about 18 minutes (in addition to the time of the interview, the description of the interview process as well as the time it took for the participants to fill out the paperwork.), meaning that in total the interviewing process took about 180 minutes, which is 3 hours. In addition, the interviewer had to sometimes wait for the interviewees when they arrived late or when

some participants did not show up at all. Furthermore, transcribing one interview, even with the help of the automatic program, took about 2 hours, in case the speech was clear and understandable and TEMI was able to detect most of the speech. Yet, with interviews where the participants talked fast or they had a specific accent, it took considerably more time because certain parts of the interview had to be listened to several times. Therefore, the time spent on transcribing was at least 20 hours in total. Baring in mind, that for the purpose of this thesis, the author did not make notes of non-vocalised aspects which would make the process even more time-consuming. Nevertheless, now that the material is collected and transcribed and it can be analysed, the effort is overweighed by the benefits. Especially since the goal was to collect material for the Estonian component in the LINDSEI corpus, in order to gain access to their database which provides researchers of TCELE numerous opportunities for future studies.

CONCLUSION

The compilation of any kind of a corpus is a work that requires time and effort and a clear goal. Yet, when the material is collected and the "pre-work" is done, the corpus can offer unlimited possibilities for linguistic as well as educational research. The compilation and analysis of learner corpora is beneficial for researchers of language, material and syllabus designers, teachers and lecturers and for students themselves as well. McEnery and Xiao (2010: 365), have written about two different approaches to the collection and use of corpora: the direct use and the indirect use. The first means that the teachers themselves gather the data from their students, analyse the data, draw conclusions and based on it the same students can benefit from it. The indirect use of corpora is usually done by researchers, such as linguists and lexicographers. The data is gathered and analysed by the researchers, since this is a long process, the subjects who provided the data are not usually the ones who finally benefit from the results (Granger 2015: 488). Usually the drawn conclusions and results are used for syllabus and material design and improvement. Most language study materials are compiled according to what the creators of these believe are relevant to learners in the target language, but there is often no reference to learner language (Mark 2002).

The thesis had two main aims. The first aim was to compile a spoken learner corpus of Estonian learners of English and to describe the whole process of compiling a corpus from the selection of participants and tools to the concordancing process. The second aim of the thesis was to provide an overview and illustrate how, in addition to researchers, teachers and students could benefit from corpora.

The corpus designed for this thesis is a spoken sub-corpus of TCELE and it consists of 10 interviews with Estonian learners of English. The participants were all English

philologists who had studied English at the university for 3 or 4 years. The reason why the author decided to have these parametres for the study and for this corpus, is that the study also collected material for the Estonian component of the LINDSEI corpus. LINDSEI provided guidelines for the compiler about the form and topics of the interview, requirements for the participants as well as guidelines for transcriptions.

It is very important that a corpus has guidelines that ensure consistency within the corpus because only then comparative studies can be done with the data (Creer *et al.* 2004). The author of the thesis decided to try automatic speech to text program TEMI for the transcription process. Although the program made the process of transcribing less time-consuming, it did not provide perfect transcriptions. All the transcripts had to be looked through and corrected, some more than others. The quality of the transcript depended on the quality of the audio and the speed and accent of the speaker. Overall, the author of the thesis thinks the program is very helpful and would recommend it for others.

For concordancing and frequency counts, this study used the freely accessible and rather easy-to-use program AntConc. The program can be downloaded from the web, it is compatible for different operating systems and can also be used offline. The webpage as well as the internet offer different manuals as to how to make the most of this program but the basic things like concordancing and frequency counts are so easy to access that reading a manual might not even be necessary.

The author conducted an empirical study on the use of the word *well* as a pragmatic marker among the Estonian learners of English. The study was based on the study done by Karin Aijmer in 2011 where she used the Swedish component of LINDSEI to analyse the use of *well.* The study revealed that the way Estonian learners use the pragmatic marker *well* is quite similar when compared to the Swedish learners. *Well* is mostly used for

speech management, for example, at the beginning of the sentence to mark turns or to indicate that an answer is on the way. It is not that much used for showing disagreement or for giving an opinion, which is a function that native speakers tend to use more.

Finally, the thesis provided an example exercise created on the basis of the gathered data to illustrate one possible way of making use of the corpus and concordance in the classroom. The thesis provides a selection of ideas as to what can be done with a corpus to engage the students and it talkes about a website created for the students of University of Oslo, which offers good examples of different corpus-based exercises. The example exercise done by the author makes use of the data collected for this study and allows the teacher to have a discussion about pragmatic markers with the students.

REFERENCES

Abdou, Mustafa, Artur Kulmizev, Vinit Ravishankar, Lasha Abzianidze and Johan Bos. 2018. What can we learn from Semantic Tagging? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, October 2018, Brussels, Belgium. Available at https://www.aclweb.org/anthology/D18-1526

Andersen, Gisle. 1998. The Pragmatic Marker like from a Relevance-theoretic Perspective. In Andreas H. Jucker and Yael Ziv (eds.). *Discourse Markers: Descriptions and theory*, 57: 147-171.

Andersen, Gisle. 2016. Semi-lexical features in corpus transcription. In John M. Kirk and Gisle Andersen (eds.). *Compilation, transcription, markup and annotation of spoken corpora*, 21: 3, 323-347.

Anthony, L. 2019. AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Aijmer, Karin. 2011. Well I'm not sure I think… The use of well by non-native speakers. In *International Journal of Corpus Linguistics*, 16: 2, 231-254.

Audacity 2019. Available at https://www.audacityteam.org/.

Ballier, Nicolas and Philippe Martin. 2013. Developing Corpus Interoperability for Phonetic Investigation of Learner Corpora. In Ana Diaz-Negrillo, Nicolas Ballier and Paul Thompson (eds.). *Automatic Treatment and Analysis of Learner Corpus Data*, 33-65. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Biber, D., S. Johansson, G. Leech, S. Conrad and F. Finegan. 1999. *The Longman Grammar of Spoken and Written English.* London: Longman.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *proceedings of the third conference on Applied natural language processing*, Trento, Italy. Available at https://dl.acm.org/citation.cfm?id=974526

Creer, Sarah and Paul Thompson. 2004. Processing spoken language data: The BASE experience. In: *Workshop on Compiling and Processing Spoken Language Corpora. 24th May.* Available at

https://www.researchgate.net/publication/228930756_Processing_spoken_language_data_The_BASE_experience

Daniel, Anna. 2015. *The Use of Adjectives and Adverbs in Estonian and British Student Writing: A Corpus Comparison.* Master's thesis. University of Tartu. Available at http://hdl.handle.net/10062/47055.

Dash, Niladri Sekhar. 2008. *Corpus Linguistics: An Introduction.* India: Pearson Education India.

De Cock, Sylvie. 2010. Spoken Learner Corpora and EFL Teaching. In Wolfang Teubert and Michaela Mahlberg (eds.). *Corpus Based Approaches to English Language Teaching*, 123-138. London/New York: Continuum.

De Cock, Sylvie and Sylviane Granger. 2005. Computer Learner Corpora and Monolingual Learners' Dictionaries: the Perfect Match. *Lexicographica*, 20: 72-86.

Diemer, Stefan, Marie-Louise Brunner and Selina Schmidt. 2016. Compiling computer-mediated spoken language corpora. In John M. Kirk and Gisle Andersen (eds.). *Compilation, transcription, markup and annotation of spoken corpora*, 21: 3, 348-371.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a fling or wedding bells? *The Electronic Journal for English as a Second Language*, 8: 4.

Gilquin, Gaëtanelle. 2015. From Design to Collection of Learner Corpora. Sylviane Granger (eds.). *The Cambridge Handbook of Learner Language Research*, 9-34. Cambridge: Cambridge University Press.

Gilquin, Gaëtanelle and Sylviane Granger. 2015. Learner Language. In D. Biber and R.Reppen (eds.). *The Cambridge Handbook of English Corpus Linguistics*, 418-436. Cambridge: Cambridge University Press.

Granath, Solveig. 2009. Who Benefits from Learning How to Use Corpora? In Karin Aijimer (ed.). *Corpora and Language Teaching*, 47-65. Amsterdam/ Philadelphia:

John Benjamins Publishing Company.

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

Granger, Sylviane. 2003. Error-tagged learner corpora and CALL: A promising synergy. In *CALICO Journal*, 20: 3, 465-480.

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

Granger, Sylviane. 2008. Learner Corpora. In A Lüdeling and M. Kytö (eds.). *Corpus Linguistics. An International Handbook.* 1: 259-275.

Granger, Syviane. 2012. How to use foreign and second language learner corpora. 62 In Mackey, A. and Gass, S. (eds.) *Research Methods in Second Language Acquisition: A Practical Guide*.

Hunston, Susan. 2008. Collection Strategies and Design Decisions. Anke Lüdeling and Merja Kytö (eds.). *Corpus Linguistics Handbook*, 154-168. Berlin: Walter de Gruyter GmbH & Co.

Johns, Tim. 1990. From printout to handout: Grammar and vocabulary teaching in the context of datadriven learning. *CALL Austria*, 10, 14-34.

Kirsimäe, Merli. 2017. *The Compilation and Lexicogrammatical Analysis of an Estonian Spoken Mini-Corpus of English as a Lingua Franca*. Master's thesis. University of Tartu. Available at http://hdl.handle.net/10062/57562.

LINDSEI, available at https://uclouvain.be/en/researchinstitutes/ilc/cecl/lindsei.html

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

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

Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Loigu, Lembi. 2006. Teaching Collocations in English Using Corpora (I). *The EATE Journal*, 30: November, 9-13.

Loigu, Lembi. 2007. Teaching Collocations in English Using Corpora (II). *The EATE Journal*, 31: May, 33-37.

Mark, Kevin L. 2001. A PARALLEL LEARNER CORPUS Using Computers in a Humanistic Approach to Language Teaching and Research. *The Japanese Journal of Language in Society*, 4: 1, 5-16.

McEnery, T & Xiao, R. 2010. What corpora can offer in language teaching and learning. in E Hinkel (ed.). *Handbook of Research in Second Language Teaching and Learning.* vol. 2, 364-380. London; New York: Routledge.

Merilaine, Elina. 2015. *The frequency and variability of conjunctive adjuncts in the Estonian–English Interlanguage Corpus.* Master's thesis. University of Tartu. Available at http://hdl.handle.net/10062/47065.

Nesselhauf, Nadja. 1996. *Collocations in a Learner Corpus.* Amsterdam/ Philadelphia: John Benjamins Publishing Company.

O'Keeffe, Anne & Michael McCarthy (eds). 2010. *The Routledge Handbook of Corpus Linguistics.* London; New York: Routledge.

Oslo Interactive English. Available at http://folk.uio.no/signeo/OIE/html/Introduction.html

Oxford Collocations Dictionary for Students of English. 2003. Oxford: Oxford University Press.

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

Reppen, Randi. 2012. Building a corpus. What are the key considerations? In

O'Keeffe, Anne and McCarthy, Michael. *The Routledge Handbook of Corpus Linguistics.* London and New York: Routledge. 31–37

Römer, Ute. 2004. A Corpus-driven Approach to Modal Auxiliaries and Their Didactics. In John Mch. Sinclair (ed.). How to Use Corpora in Language Teaching, viii: 185-199. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, John. 1996. Preliminary recommendations on corpus typology. *Eagles*. Available at http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html, accessed March 9, 2019.

Tammiste, Lenne. 2016. *The use of adjective-noun, verb-noun and phrasal-verb-noun collocations in Estonian learner corpus of English.* Master's thesis. University of Tartu. Available at http://hdl.handle.net/10062/53280.

Temi Available at https://www.temi.com/.

Timmis, Ivor. 2015. *Corpus Linguistics for ELT*. London and New York: Routledge.

Undo, Aare. 2018. *Calculating the Error Percentage of an Automated Part-of-Speech Tagger when Analyzing Estonian Learner English – An Empirical Analysis*. Master's thesis. University of Tartu. Available at http://hdl.handle.net/10062/60466.

Voutilainen, Atro. 2005. Part-of-Speech Tagging. In Ruslan Mitkov (ed.). *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press.

Wharton, Tim. 2003. Interjections, language, and the 'showing/saying' continuum. *Pragmatics and Cognition,* 11: 1, 39-91.

Wichmann, Anne. 2008. Speech Corpora and Spoken Corpora. Anke Lüdeling and Merja Kytö (eds.). *Corpus Linguistics Handbook,* 187-207. Berlin: Walter de Gruyter GmbH & Co.

APPENDIX 1: CORPUS-BASED EXERCISE

Please read the following sentences carefully. In each sentence, please underline the word *well* and decide whether it is necessary in the sentence or it could be omitted/removed. In case *well* is needed in the sentence, make notes about its function in the sentence and be ready to explain it. What else do you notice about each sentence (the choice of words, level of formality/register, etc.)?

1. Well, a man is painting a picture of a woman who is sitting on the chair.

2. And, like, when I finished high school I actually was able to speak Russian pretty well.

3. So it was very funny to the locals as well.

4. Well, first of all, I think though it's obvious that techniques have improved in the film-making and stuff and magic.

5. I was good at it in gymnasium and stuff, so I thought, it was my second option, and I thought, well, I'm going to try this.

6. I did well in school in English.

7. And if it means conflict with somebody else, well I can deal with it.

8. So they offer private tutoring and they get paid really well because high level executive, um, workers want to learn how to talk in English more proficiently.

9. Yeah, well, next to school I don't have that much time and I have to read a lot of books for, for school as well.

10. I just um, well at this point, even, even if someone directly makes fun of me or something like that, I just do not care anymore.

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Anne Rahusaar

The Compilation of the Spoken Sub-corpus for the Tartu Corpus of Estonian Learner English/ Eesti keelt emakeelena kõnelevate inglise keele õppijate suulise alakorpuse koostamine

Magistritöö
2019
Lehekülgede arv: 56

Annotatsioon:

Magistritöö peamine eesmärk on koostada suuline õppijakeele korpus ning kirjeldada sellise korpuse loomise protsessi, alustades osalejate ning vahendite valimisest ning lõpetades korpuse analüüsiga. Teine suur eesmärk on anda ülevaade, mil viisil saavad õpetajad rakendada korpuseid, eriti õppijakeele korpuseid, enda igapäeva töös ning mis kasu on korpustest õpetajatele ja õpilastele.

Töö esimene peatükk keskendub teooriale ning selgitab korpuse, õppijakeele ja õppijakeele korpuse olemust, kasutegureid ja puuduseid. Lisaks toob esimene peatükk välja juba varasemalt selles valdkonnas tehtud uurimused nii mujal maailmas kui Eestis.

Töö teine peatükk annab põhjaliku ülevaate suulise korpuse koostamisest ning selle analüüsimisest. Teises peatükis on ka magistritöö empiiriline osa, kus tuuakse üks näide, millist informatsiooni korpus võib pakkuda ning mis uurimusi on võimalik teha.

Magistritöö uurib eesti inglise keelt õppivate tudengite suulist keelekasutust. Viidi läbi 10 intervjuud, kus intervjueeritav pidi rääkima igapäevastel teemadel ning jutustama piltide põhjal loo. Intervjuud transkribeeriti kasutades programmi TEMI, mis viib suulise kõne automaatselt kirjalikku vormi, vähendades seega tunduvalt ajakulu, mis on manuaalse transkribeerimise puhul vältimatu. Kuigi ka antud programmi koostatud transkriptsioonid tuli autoril käsitsi üle vaadata ning korrigeerida.

Empriirilises osas võrreldakse eesti õppijate pragmaatilise markeri *well* kasutust rootsi õppijatega ning tulemused näitavad, et võrreldes inglise keelt emakeelena rääkijatega, kasutavad nii eesti kui ka rootsi õppijad sõna *well* teatud kontekstis liigselt. Lisaks koostas magistritöö autor ka loodud korpusest saadud otsingute tulemusi selleks, et luua üks näiteülesanne õpetajale, tutvustamaks üht võimalust, kuidas korpust ja sealset informatsiooni ka tunnis kasutada.

Töö kokkuvõttes jõuab autor järeldusele, et suulise õppijakeele korpuse loomine nõuab küll palju tööd ja aega, kuid saadud lisainformatsioon õppijate kohta on seda vaeva väärt.

Märksõnad: korpusuuringud, suulise kõne korpus, õppijakorpus, inglise keele õpetamine, pragmaatilised markerid.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Anne Rahusaar,

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

   The Compilation of the Spoken Sub-corpus for the Tartu Corpus of Estonian Learner English,

   mille juhendaja on Jane Klavan,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace´i kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Anne Rahusaar

20.05.2019

Autorsuse kinnitus

Kinnitan, et olen koostanud käesoleva magistritöö ise ning toonud korrektselt välja teiste autorite panuse. Töö on koostatud lähtudes Tartu Ülikooli maailma keelte ja kultuuride kolledži anglistika osakonna magistritöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Anne Rahusaar

20.05.2019

Lõputöö on lubatud kaitsmisele.

Jane Klavan

20.05.2019