UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

A CORPUS-BASED STUDY OF THE FILLER *SO* IN THE SPEECH OF ESTONIAN EFL LEARNERS

BA thesis

ROMI VARULA
SUPERVISOR: *Assoc. Prof.* JANE KLAVAN

TARTU
2024

ABSTRACT

Discourse markers (DMs) in speech have been extensively researched, particularly since the late twentieth century, with growing focus on DMs in learner language in recent decades. However, filler words in learner language have not been as vastly researched and thus, this thesis at hand aims to address this gap by contributing to a few studies made in Estonia so far. The thesis focuses primarily on the use of the filler *so* in the speech of Estonian learners of English as a Foreign Language (EFL) and native English speakers utilizing the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI-EST) and the Louvain Corpus of Native English Conversation (LOCNEC). Additionally, it aims to analyse whether *so* is used more as a filler, a conjunction or an adverb between the two groups.

The introduction part of the thesis introduces the topic and corpora at hand, states the two research questions and gives an overview of all the sections of the thesis. The first part is concerned with previous studies conducted on both discourse markers and fillers in speech and how *so* functions as a filler, a conjunction and an adverb. The empirical section of the thesis gives an overview of the methodology, the corpora used in the study and analyses the use of *so*. The last part of the thesis discusses the findings based on the results gathered from the data. The conclusion answers the two research questions presented in the introduction and suggests possibilities for further research in similar field of study.

TABLE OF CONTENTS

ABSTRACT 2

LIST OF ABBREVIATIONS 4

INTRODUCTION 5

1. DISCOURSE MARKERS AND FILLER WORDS 7

1.1 Discourse markers in speech 7

1.2 Filler words in speech 9

1.3 *So* as a filler, a conjunction and an adverb 12

2. EMPIRICAL ANALYSIS OF *SO* 14

2.1 Methodology 14

2.2 Corpora used in the study 15

2.3 Analysis of *so* 17

2.4 Discussion 25

CONCLUSION 29

REFERENCES 31

RESÜMEE 33

LIST OF ABBREVIATIONS

DM – Discourse marker
EFL – English as a Foreign Language
KWIC – Keyword in context
LINDSEI-EST – The Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage
LOCNEC – Louvain Corpus of Native English Conversation
PM – Pragmatic marker

INTRODUCTION

Discourse markers (DMs) and filler words are commonly used words or phrases in daily spontaneous speech. Despite the considerable number of scholarly articles written over in recent decades on the topic of DMs and fillers, it continues to hold a central position in this specific field of research (Beeching et al 2022: 1). Authors such as Schiffrin (1987) were among the first to define DMs and explain their purpose. However, significantly less information is available concerning fillers, particularly in the context of learner language. Similarly, there is limited research done on fillers, particularly the filler *so*, among Estonian learners of English as a Foreign Language (EFL) and native speakers.

Previous research in a related field, utilizing the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI-EST), has predominantly focused on pragmatic markers (PMs) such as *well* and *like*, as indicated in the studies by Rahusaar (2019) and Konso (2021) respectively. However, no research has been conducted on the filler s*o* utilizing LINDSEI-EST and the Louvain Corpus of Native English Conversation (LOCNEC) corpora. The author's primary focus on *so* in this research stems from its frequent occurrence in the author's own English speech.

The thesis is concerned with the following research questions: how many times is *so* used as a filler compared to other parts of speech, such as an adverb or a conjunction, among Estonian learners of English and native speakers and what are some of the primary differences in the usage of the filler between these two groups of speakers. In order to answer these research questions, an empirical corpus-based study was carried out.

The first part of this paper focuses on previously conducted research on both DMs and filler words. In this section, both of these phenomena are defined and explained and the definitions used in the present research highlighted. The following two subsections discuss DMs and fillers in speech in other learner languages that have been compared with English.

The third subsection discusses previously conducted studies on *so* as a filler, a conjunction and an adverb.

The second part of this paper focuses on the empirical analysis of *so*. The first subsection gives an overview of the methodology, followed by overviews of the LOCNEC and LINDSEI corpora. The last two sections analyse the use of *so* and discuss the findings from both corpora.

1. DISCOURSE MARKERS AND FILLER WORDS

This part of the thesis aims to give an overview of previous research conducted on discourse markers and fillers in speech. The first part introduces discourse markers, their meaning and function. The subsequent part gives an overview of filler words, and their meaning and main functions. The last part is concerned with previous research conducted on s*o* as a filler, a conjunction and an adverb.

1.1 Discourse markers in speech

In the literature it is recognized that DMs can be labelled in multiple different ways due to their diversity, and thus, there is not one single and correct term in use to describe them. In her article, Alami (2015: 1) lists some of the ways in which researchers have labelled DMs in their studies: for example, discourse markers, pragmatic markers, discourse connectives, cue phrases, discourse particles, pragmatic particles and so on. Despite the lack of general agreement on what to call these linguistic elements, researchers have attributed several of their definitions for these elements. According to Levinson (1983: 87, 88), DMs are words or phrases that denote a relation between an utterance and the preceding discourse. In Zwicky's (1985: 303, 304) view, DMs are words that are separated from other function words as they often appear at the beginning of the sentence to continue the conversation. He views them inherently independent from other context that appears in a sentence. However, one of the foremost researchers in the field of DMs is Schiffrin (1987: 31), who has defined them as "sequentially dependent elements which bracket units of talk". On the basis of Schiffrin's study, these elements are *oh, well, and, but, or, so, because, now, then, I mean* and *you know* (Schiffrin 1987: 31).

The different functions of DMs have been fairly thoroughly analysed and researched. However, research demonstrates that these functions vary depending on different types of approaches, particularly as observed in speech. These differences in functions are present in DMs *and* and *but* in both Schiffrin's (1987) and Cuenca and Crible's (2017) studies. For example, Schiffrin (1987: 152) finds that *and* and *but* are both important units in speech because of their structural and cohesive role. According to her research, *and* and *but* are structural due to the fact that they link other units, like clauses and phrases, and cohesive due to the fact that the understanding and interpretation of the whole conjunctive sentence depends on both of those conjuncts (Schiffrin 1987: 153). However, based on their study, Cuenca and Crible (2017: 153) claim that when an argument following *but* is not verbally expressed at the end of the sentence, it leaves the reader or listener to automatically infer what is implied. On the basis of their analysis, the authors' point seems to be that it is possible to understand the sentence after *but*, even if it is not completed. In this case, a sentence is not completed when the speaker uses *but* as the last item of the sentence and another speaker starts talking, such as in the example (1).

(1) Speaker 1: I was hoping it'd be a week so I could get off the ward for a bit **but**
Speaker 2: oh already? (Cuenca and Crible 2017: 153)

In addition, it is possible to infer speakers' intended meaning of the sentence in the analysis of this current study even when it is not completed, although this is not always the case.

DMs have also received more attention in learners' speech in various non-English speaking countries. The extensive research has been carried out with both native English speakers and learners of English worldwide. In this section, Chinese speakers of English as a Foreign Language (EFL) and native English speakers, and Pakistani EFL learners and native English speakers will be compared. Both studies draw similar conclusions on the use

of some of the DMs that they mention. According to Jabeen et al (2011: 69), the DMs they look at are *I mean, you know, I think, kind of, sort of, well, you see* and *so* in the case of Pakistani and native speakers of English. In the case of Chinese and native speakers of English, Bu (2012: 36) looks at DMs such as *like, yeah, oh, you know, well, I mean, right* and *ok* due to their extensive use in the corpus, in academic settings and in the speech of native Chinese speakers. She notes that in the speech of native English speakers, the use of *ok* and *right* is more frequent than the use of *like* and *you know*, especially when the former are used as transition markers (Bu 2012: 37). She goes on to add that DMs such as *you know, I mean* and w*ell* are used more frequently in native speech than in the speech of Chinese speakers (Bu 2012: 37).

Similarly to the study about the use of *you know* in the Chinese and native English speakers, the corpus-based study comparing Pakistani and native English speakers' speech shows that *you know* is used five times more in the speech of native speakers than Pakistani EFL learners (Jabeen et al 2011: 79). *I mean* is also used more frequently in native speakers' speech (Jabeen et al 2011: 79). While it appears in almost all cases in the speech of native speakers, it is entirely absent in the speech of Pakistani English learners (Jabeen et al 2011: 80). However, while *well* is one of the most frequently used DMs in the speech of native English speakers, the discourse marker is only used by one out of thirty Chinese students in this study (Bu 2012: 39).

1.2 Filler words in speech

Further research indicates that DMs can be seen as hesitation elements that show how the speaker constructs utterances and what they choose to say while they talk (Beeching et al 2022: 1). Therefore, when taking this into consideration, it seems plausible that DMs also

emerge in other classifications, for example, in the form of verbal fillers and filled pauses (Beeching et al 2022: 2). Verbal fillers can be lexicalised or unlexicalised, which means that they appear in a sentence as words such as *well* and *so* or as vocalisations such as *um* or *em* (Rose 1998: 7, 8). Filled pauses are voiced pauses, while unfilled pauses are silent pauses (Rose 1998: 7). In linguistics, researchers acknowledge various definitions for fillers. According to Carter and McCarthy, the authors of the *Cambridge Grammar of English* (2006), fillers

> "refer to vocalisations (*er, erm, um, mm*) or words that are used to fill gaps in conversations. A filler can either mark hesitation, a shift in topic, or indicate speaker's online process of thinking (***Well**, of course, **erm**, I think we should take our time before taking any action*)" (Carter and McCarthy 2006: 903).

However, in her article, Hirschman (1993: 431) defines fillers as words or phrases that can appear anywhere in speech and that could be deleted from the sentence without any change in meaning such as in the example (2).

(2) I would **uh** agree with you **like** in from from my past experience (Hirschman 1993: 431)

In her view, fillers can also be words such as *um* and *you know* (Hirschman 1993: 431). Furthermore, Fitriati et al (2021: 30) partly agree with Carter and McCarthy since they define fillers as lexically empty items that do not have a certain function within discourse, except to fill a gap in speech. However, in their view, speakers are most likely to use fillers such as *well, so, I mean, actually, let me think* to show hesitation in order for the conversation to go smoothly (Fitriati et al 2021: 30).

Fitriati et al (2021: 31) also note that using fillers in speech and communication in another language helps the speaker appear more natural and fluent, especially in case of spontaneous speech. Fillers are also one of the most frequent conversation elements used by

learners (Fitriati et al 2021: 30). As well as DMs, these fillers and their functions tend to vary in use in speech. According to Clark and Fox Tree (2002: 97) fillers can be used to convey various meanings, for example, to "hold the floor" in conversation, which means that when a speaker uses a filler, he or she lets their listener know that they will continue speaking. Similarly, according to Corley and Stewart (2008: 592), while fillers convey uncertainty or hesitation in the speaker's discourse, it does not necessarily imply that the speaker indicates a forthcoming delay in their speech solely due to hesitation. The definition used in this paper is based on the one provided by Carter and McCarthy (2006: 903): fillers can indicate a speaker's hesitation, thought process or a shift in the topic that they are currently talking about. The choice of this definition is primarily due to the author's belief that fillers are mainly used for these purposes, together with "holding the floor" as mentioned by Clark and Fox Tree (2002: 97), in her speech and in the speech of Estonian EFL learners and native speakers.

Research indicates that filler words in learners' speech have not been as vastly researched as DMs, although a few comparisons can still be made. For example, Okazawa (2014: 54) investigates and compares the speech of Japanese EFL learners with an aim to find whether there are any differences between the number and function of fillers between native English speakers and Japanese EFL learners. Okazawa (2014: 56) argues that while Japanese fillers are prominent throughout all the students' speech while speaking English, the only English filler words in use are *well* and *or how can I say*. However, other research on the topic of filler use in Japanese EFL learners has shown that these fillers are not the only ones used among Japanese English learners. In Shimada and Miura's (2019: 200) view, they also include fillers such as *anyway, kind of, you know, so, right, really, oh great* and so on.

1.3 *So* as a filler, a conjunction and an adverb

The filler *so* alone has received limited attention in research in the past, often being grouped together with other fillers to provide context regarding their use in sentences. To understand the function of *so* in sentences, a deeper study of its role within a sentence is necessary. As noted earlier, the functions of fillers are often closely associated with hesitation in speakers' spontaneous speech. Rose (1998: 5) lists some classifications of hesitation phenomena that help to understand how and when fillers can be used furthermore. Hesitation phenomena occur in a variety of features in spontaneous speech to slow the process of conveying lexicalized information to the other person (Rose 1998: 4). The six classifications are as follows: false starts, which happen when a speaker starts speaking but then stops mid-sentence; repeats, which happen when a speaker repeats a lexical item mid-sentence; restarts, which happen when a speaker utters some words and returns to say the same words again; self-corrections, which happen when a speaker goes back to correct themselves mid-sentence; lengthenings, which happen when a speaker draws out the pronunciation of a word; and pauses, which happen when a speaker pauses in order to breathe, to think about what to say next or to fill the pause with either an unlexicalized or lexicalized filler (Rose 1998: 4-6). The phenomena are illustrated with examples from this study in the analysis section of the paper.

So has received more attention in research as a conjunction, although it is often paired and studied with other conjunctions. The types of conjunctions that *so* most often appears in are: cumulative or copulative conjunctions, which join statements or add one statement to another such as in the example (3); conjunctions as a result and consequence, which means that *so* and *that* are often used together to express cause and reason such as in the example (4); and correlative conjunctions, which means that word-pairs (such as *either …or*, *so …that*) are used together to emphasize the combinations of two structures such as in the

example (5) (Unubi 2016: 204, 205, 206, 208). The second and last types could essentially convey the same meaning.

(3) He is my father **so** I respect him (Unubi 2016: 205)

(4) I will help him now **so that** he can help me tomorrow (Unubi 2016: 206)

(5) She was **so** late **that** I had already entered the classroom without waiting for her (Unubi 2016: 208)

So as an adverb has mainly one function. In the classification of adverbs by function, it typically belongs to the category of adverbs of degree, which means that it mainly modifies adjectives or other adverbs (see example 6) (Sarifuddin 2023: 86).

(6) He talked **so** fast (Unubi 2016: 206)

Other words that fall into that category are words such as *almost, nearly, quite, just, too, enough, extremely* and so on (Sarifuddin 2023: 86). Both the classifications of conjunctions and adverbs are further illustrated with examples from this study in the analysis section of the paper.

2. EMPIRICAL ANALYSIS OF *SO*

This part of the thesis aims to give an overview of the methodology that was used to gather the data. The subsequent section of the paper introduces the two corpora utilized for analysis. The last section analyses the filler *so* between Estonian EFL learners and native speakers and ultimately, results are drawn and conclusions made.

2.1 Methodology

As mentioned, this thesis is mainly concerned with two research questions: to investigate how many times *so* is used as a filler in the speech of Estonian learners of English and native speakers of English in relation to how much it is used as an adverb or conjunction and what are the main differences between how the students use the filler. In order to answer these questions, a corpus-based analysis was conducted and both quantitative and qualitative approaches were used. The former was used to analyse how many times *so* appears in both the LINDSEI-EST and LOCNEC corpora using a free corpus analysis toolkit called AntConc (Anthony 2024), and the latter was used to analyse how many times it appears in each sentence as a filler and the differences in the use of *so*. However, the main body of the research was conducted using the qualitative method, primarily involving sentence analysis.

In order to determine the total number of *so* used in both corpora, all the interviews were uploaded to AntConc, which is used for concordancing and text analysis (Anthony 2024). After that, *so* was searched as the keyword in context (KWIC). The KWIC method helped discern the context preceding and following the target word.

All of the occurrences of *so* were then copied from AntConc to Microsoft Excel. After that, each occurrence of *so* could be assigned the corresponding function: filler,

conjunction or adverb. As both corpora are incomparable because of the difference in the size of total words, the frequencies had to be normalised. All the uses could then be compared and further inferences of the differences in the use of *so* made.

2.2 Corpora used in the study

The LINDSEI corpus, short for the Louvain International Database of Spoken English Interlanguage, is a corpus consisting of informal interviews with English learners ranging from higher intermediate to advanced levels. It contains over 1 million words of transcribed text, with nearly 800,000 words of transcribed text from learners representing diverse mother tongue backgrounds, such as Chinese, Japanese, Dutch, German, Bulgarian, German, Greek, Italian, Polish, Spanish, Swedish, Estonian and so on. (Gilquin et al 2010: para. 1)

The LINDSEI-EST corpus, short for the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage, serves as the primary source for analysis in this thesis and comprises, at its current status, 38 informal interviews (82,239 words of transcribed text recorded in 2024) with Estonian university students whose native language is Estonian. All interviewees are either enrolled or were previously enrolled as third-year BA and first-year MA students of English at the University of Tartu. The interviews are structured into three distinct sections: a warm-up exercise, in which learners can freely choose and discuss one of three set topics, an informal discussion between the interviewee and interviewer as the main part of the interview, and finally, a picture description exercise. The LOCNEC corpus, which will be discussed in the following section, is used as a counterpart of the LINDSEI corpus, making the comparisons between the two possible.

The LOCNEC corpus, short for the Louvain Corpus of Native English Conversation, was compiled in 1995 by its creator, Sylvie De Cock, affiliated with the Catholic University of Louvain in Belgium. It is a corpus constructed on a similar basis to the LINDSEI corpus and comprises, at its current status, 55 informal interviews primarily conducted for research purposes at the Centre for English Corpus Linguistics. The interviews feature British university students, all of whom are native speakers of English at Lancaster University. (De Cock 2004)

All interviews consist of the same three parts as the LINDSEI-EST corpus: a warm-up, during which learners are encouraged to discuss one of three set topics, an informal discussion between the interviewee and the interviewer, and a picture description exercise. The corpus comprises approximately 170,000 words of transcribed text, with around 120,000 words of transcribed text produced by the interviewees themselves. (De Cock 2004)

All 38 interviews from LINDSEI-EST and 55 interviews from LOCNEC were used to make comparisons. Since this thesis is concerned with learner language, the speeches of both the interviewers (A) and interviewees (B) were used as all of them are university students. The interviewers' parts could not be deleted as the context for only interviewees would not have been enough to investigate each sentence containing *so*.

As previously mentioned, given the disparity in the size of both corpora, normalisation of all occurrences was necessary before the comparisons between the two corpora could be made. In order to calculate the relative frequency, all absolute frequencies in both corpora had to be divided by the corpus total word count and multiplied by 10,000.

2.3 Analysis of *so*

In general, *so* appeared 1,060 (relative frequency 128.9 occurrences per 10,000 words) times in all interviews in the LINDSEI-EST corpus and 2,472 (relative frequency 145.4 occurrences per 10,000 words) times in all interviews in the LOCNEC corpus. Additional words mentioned in the literature review that are often categorized as fillers, such as *well*, *you know*, *kind of*, *I mean*, *like* and *actually* were also present in both corpora but occurred less frequently in almost all cases (see Table 1). However, it should be noted that this is a general overview of the words previously mentioned. Given that *well*, *you know*, *kind of*, *I mean*, *like* and *actually* are not the primary focus of investigation in this thesis and are not subjected to further analysis regarding their function as fillers or other parts of speech, as is the case with *so*, it remains uncertain whether they indeed all appeared less frequently as fillers compared to *so*. However, inspecting the overall frequency of these different fillers was the necessary first step before making a decision to zoom in on the use of *so*. Relative frequencies are presented in Table 1, with absolute frequencies presented in brackets.

Table 1. The occurrence of *so* and the total number of words other than *so* often categorized as fillers.

Words	LOCNEC	LINDSEI-EST
so	145.4 (2472)	128.9 (1060)
well	68.2 (1159)	42.2 (347)
you know	41.1 (699)	15.0 (123)
kind of	10.9 (186)	29.9 (246)
I mean	48.9 (832)	6.1 (50)
like	106.7 (1814)	154.9 (1274)
actually	15.1 (256)	16.5 (136)

Total	**436.4 (7418)**	**393.5 (3236)**

To determine the number of occurrences of *so* that functioned as fillers out of 1,060 instances in the LINDSEI-EST corpus and out of 2,472 instances in the LOCNEC corpus, each sentence containing *so* had to be checked manually. As the context of some sentences in both corpora was indiscernible, making it difficult to determine whether *so* was used as a filler or as another part of speech, these instances were removed. Specifically, 10 out of 1,060 occurrences in the LINDSEI-EST corpus and 23 out of 2,472 occurrences in the LOCNEC corpus were removed. The difficulty in understanding the context that included *so* in the deleted sentences stemmed from three main reasons: (1) limited context, which means that the absence of sufficient context in a sentence surrounding *so* hindered determining its function (see example 7), (2) missing words, which means that some words were missing due to poor audio file and therefore the sentence could not be analysed further (see example 8), and (3) unintelligible speech, which means that the speech was unclear at times, making it difficult to understand whole sentences (see example 9). The speech of either the interviewer (A) or interviewee (B) is written between the brackets of said letters.

(7) <A> of teaching <A> <B> mm oh it was <B> <A> **so** <A> <B> yeah it was really hectic about six (LOCNEC_12)

(8) <A> English that century and <A> <B> oh right <B> <A> **so** .. had to know well then you had that (LOCNEC_21)

(9) <A> good experience <laughs> <A> <B> <overlap> <laughs> <B> <A> <overlap> **so** <overlap> <laughs> recommend . whatever you do (LINDSEI-EST_002)

After examining each sentence containing *so* in both the LINDSEI-EST and the LOCNEC corpora, it was found that there were 767 instances of fillers out of 1,050 in the

former, and 997 instances of fillers out of 2,449 in the latter. After the normalisation, the relative frequency was 93.3 and 58.6 respectively. It is already possible to draw initial conclusions based on this information. It is evident that Estonian EFL speakers use the filler *so* nearly two times more frequently than native speakers of English. Due to the substantial disparity in filler usage, it can be said that the removed sentences do not significantly impact the comparison of *so* between the two corpora. Additionally, it is important to highlight that the comparisons drawn here are not indicative of all Estonian EFL speakers using more fillers more frequently than native speakers of English. This conclusion is solely based on the findings of this particular study. Table 2 indicates the different functions of *so* used in both corpora based on the functions highlighted by the author in the 1.2 subsection of this thesis. Table 2 provides the normalized frequency counts with the absolute number of occurrences given in the brackets.

Table 2. Functions of the filler *so*.

Functions	**LOCNEC**	**LINDSEI-EST**
Hesitation phenomena	12.2 (208)	17.8 (146)
Shift in topic	29.5 (501)	39.4 (324)
Holding the floor	5.2 (88)	2.9 (24)
Other	11.8 (200)	33.2 (273)
Total	**58.7 (997)**	**93.3 (767)**

When looking further into how the filler *so* is used in sentences across both corpora, numerous distinctions and some similarities occur. One way to recognize *so* as a filler is by observing the words it commonly appears in a sentence with. It is evident that the words could differ between the Estonian EFL speakers and native speakers of English because of

the way they use the English language. However, as mentioned, there were some similarities between the two as well. *So* is most frequently used before words such as *I think* (example 10) *it was* (example 11) and the vocalisation *eh* (example 12) in the LINDSEI-EST corpus. As mentioned by Corley and Stewart (2008: 592), fillers are most often used to express hesitation or uncertainty. As shown by the examples below, the first and last combinations are used by the interviewees to express uncertainty. *So* used before *it was* could be used to convey a story and can express both hesitation and uncertainty.

(10) country I visited .. should I just talk <overlap> ok (eh) **so I think** I'm gonna go with Norway (LINDSEI-EST_020)

(11) <B> (eh) just a little bit yes <B> <A> okay. **so it was** enough to manage <A> <B> enough to (LINDSEI-EST_014)

(12) <B> in cultural studies <B> <A> <overlap> ah . (mhm) . okay and **so (eh)** area-wise . London and <A> <B> yeah London (LINDSEI-EST_006)

However, *so* is most frequently used before words such as *I mean* (example 13), *you know* (example 14) and *you are going to* (example 15) in the LOCNEC corpus. These instances are almost all expressed due to uncertainty or hesitation. Several of the instances that were more frequent in one corpus, could, in most cases, also be present in the other one (see Table 3).

(13) <B> yeah I wouldn't mind going <B> <A> **so I mean** .. don't don't be scared I (LOCNEC_02)

(14) <B> sixteen <B> <A> oh yeah <A> <B> years old **so you know** I I don't think they quite (LOCNEC_47)

(15) <A> I guess <laughs> <A> <B> <laughs> yeah <B> <A> erm .. **so you're going to**: to live off campus next (LOCNEC_28)

A notable similarity between both corpora is the frequent use of the phrase *you can* before *so* (see example 16), which is a phrase commonly used by the interviewers to guide the conversation or change the topic. An example from the LINDSEI-EST corpus illustrates this below.

(16) is then: have four pictures and they tell a story **so you** can study the pictures and then make up (LINDSEI-EST_035)

Table 3 displays all of the relative and absolute frequencies together with words used after *so* according to how often they appear, from most frequent to least.

Table 3. Most frequently used words after *so*.

Words	LOCNEC	LINDSEI-EST
you can	8.2 (139)	11.2 (92)
I think	1.2 (21)	10.8 (89)
you are going to	6.8 (115)	3.6 (30)
I mean	5.4 (91)	3.6 (30)
you know	6.1 (103)	2.3 (19)
it was	1.1 (19)	6.8 (56)
eh	0.0 (0)	7.7 (63)
Total	**28.7 (488)**	**46.1 (379)**

All the words frequently used with *so* among native speakers are also found in the LINDSEI-EST corpus. The same cannot be said about the LOCNEC corpus since the vocalisation *eh* does not appear in it.

When looking at *so* as a filler in view of several classifications of hesitation phenomena from Rose's (1998: 4–6) study, interesting information emerges. The study at

present includes several phenomena, yet not all of them are represented. However, several of the same phenomena are used in both corpora. The classifications that are present in the LINDSEI-EST corpus include repeats, which happen when a speaker repeats a lexical item mid-sentence. An example of this is given in (17) where the speaker says *so* again right after having said it. Other classifications that appear in the corpus are lengthenings, which happen when a speaker draws out the pronunciation of a word. An example of this is given in (18) where *so* is drawn out in speech. And lastly pauses, which happen when a speaker pauses in order to breathe, get more time in the conversation to think or to fill the pause with either an unlexicalized or lexicalized filler such as in example (19) where the speaker pauses right after *so* (Rose 1998: 4, 5, 6).

(17) a sort of a . (mm) (er) a communal sort of **so so** like a social event . in its tiny cinemas (LINDSEI-EST_011)

(18) was lucky enough to (eh) spend time with two families **so:** with the first family we even got to go (LINDSEI-EST_028)

(19) <A> want to . you don't have to <A> <B> ok . **so ..** (em) .. Maria . asked . Charles to paint her .. and (eh) (LINDSEI-EST_025)

In the LOCNEC corpus, hesitation phenomena include false starts, which happen when a speaker starts speaking but then stops mid-sentence such as in (20) where the speaker stops right after *so*; repeats such as example (21) where the speaker uses *so* twice; lengthenings such as in example (22) where *so* is drawn out; and pauses such as in example (23) where the speaker pauses right after *so*. Table 4 shows the relative and absolute frequency of each hesitation phenomena.

(20) know . I'm <X> actually a drummer in a band **so** <B> <A> oh yes [uhu <A> <B> the thing (LOCNEC_39)

(21) far as as far as I know <B> <A> oh **so so** you have to: .. to learn to startlearning (LOCNEC_26)

(22) can't really say you like cos it's like **so:** so different <B> <A> it's so different (LOCNEC_26)

(23) girlfriend back home <B> <A> uhu <A> <B> <X> erm .. **so ..** so that er I couldn't really start anything (LOCNEC_05)

Table 4. Most frequently used hesitation phenomena.

Hesitation phenomena	LOCNEC	LINDSEI-EST	Both corpora
False starts	1.0 (17)	0.0 (0)	1.0 (17)
Repeats	2.5 (42)	8.4 (69)	10.9 (111)
Lengthenings	2.7 (46)	2.4 (20)	5.1 (66)
Pauses	6.1 (103)	6.9 (57)	13.0 (160)
Total	**12.2 (208)**	**17.8 (146)**	**30.0 (354)**

The examples suggest that both Estonian EFL learners and native English speakers use hesitation phenomena similarly, however this data suggests that there is a slight difference in their frequencies. From the frequencies it can be seen that learners use hesitation phenomena more frequently than native speakers. Whereas native speakers use more false starts and lengthenings, learners use more repeats and pauses. There is a higher frequency of false starts among native speakers since they are absent in the speech of Estonian EFL learners. In total, both corpora include 354 of hesitation phenomena. Given that this section of the thesis focuses solely on hesitation phenomena, other potential uses of *so* that make up whole corpora (except for previous functions mentioned such as a shift in topic and holding the floor) are not further explored.

When analysing *so* as a conjunction or an adverb as opposed to its role as a filler, a difference in their frequencies emerges between the corpora. As a filler, *so* appears 767 (relative frequency 93.3) times in the LINDSEI-EST corpus and 997 (relative frequency 58.6) times in the LOCNEC corpus. As a conjunction, *so* appears a total of 118 (relative frequency 14.3) times in all interviews in the LINDSEI-EST corpus and 1151 (relative frequency 67.7) times in all interviews in the LOCNEC corpus. As an adverb, *so* appears a total of 125 (relative frequency 15.2) times in all interviews in the LINDSEI-EST corpus and 135 (relative frequency 7.9) times in all interviews in the LOCNEC corpus. All of this is presented in the table below (see Table 5).

Table 5. *So* as a filler, a conjunction and an adverb.

	LOCNEC	**LINDSEI-EST**	**Both corpora**
Filler	58.6 (997)	93.3 (767)	151.9 (1764)
Conjunction	67.7 (1151)	14.3 (118)	82.0 (1269)
Adverb	7.9 (135)	15.2 (125)	23.1 (260)
Other	9.8 (166)	4.9 (40)	14.7 (206)
Total	**144.0 (2449)**	**127.7 (1050)**	**271.7 (3499)**

As indicated earlier, the data reveals that the filler *so* is more prevalent in the speech of Estonian learners compared to that of native speakers. In the case of the two other uses of *so*, the native speakers use *so* more frequently as a conjunction than as an adverb. However, in learners' speech, *so* is used more frequently as an adverb than as a conjunction. Nevertheless, the difference of *so* used as a conjunction and as an adverb is minimal in the LINDSEI-EST corpus (14.3 *vs.* 15.2).

According to Unubi's (2016: 204, 206, 208) study, all types of conjunctions that *so* most frequently appears in are also present in both corpora: cumulative or copulative conjunctions, such as in the example (24), in which *so* joins statements or adds one statement to another; conjunctions as a result and consequence such as in the example (25), in which *so* and *that* are used together to express cause and reason; and correlative conjunctions such as in the example (26), in which word-pairs (such as *so …that*) are used together to emphasize the combinations of two structures (Unubi 2016: 204, 206, 208).

(24) mean I was only what sixteen seventeen at the time **so** I'd never really been drunk before and I (LOCNEC_18)

(25) (em) I'm not exactly from a high-income family . **so that** was really important . for me . to get the (LINDSEI-EST_023)

(26) was cool . (eh) the food as good and cheap **so that** was very nice I got to . eat a (LINDSEI-EST_034)

As stated by Sarifuddin (2023: 86), *so* as an adverb has mainly one function: it modifies adjectives or other adverbs. This can also be seen in both corpora in this study (see examples 27 and 28).

(27) about it and she hated it and found that **so amazing** that we were such opposites [extremes . erm (LOCNEC_35)

(28) <laughs> this makes much more sense I understand you **so much** better now . so I think that they (eh) (LINDSEI-EST_037)

2.4 Discussion

Based on the data collected and analysed from both corpora, the most prevalent difference in using the filler *so* between the learners and native speakers is how *so* is used

with other words. While *so* is most dominant in learners' speech before words such as *I think*, *it was* and *eh*, it is most prevalent in the speech of native speakers before words such as *I mean*, *you know* and *you are going to*. A notable similarity that emerged is the frequent occurrence of the phrase *you can* after *so* in both corpora. It appeared 139 times in the LOCNEC and 92 times in the LINDSEI-EST corpus. Regarding hesitation phenomena, most categories are prevalent in both corpora, with the only distinction that false starts were exclusively in the LOCNEC corpus. Based on the analysis of *so* as a conjunction and an adverb, the results indicate that *so* is more dominant as a conjunction in the speech of British students and more dominant as an adverb in the speech of Estonian students. *So* appears more frequently as a filler than a conjunction or an adverb in the LINDSEI-EST corpora but not more than a conjunction in the LOCNEC corpora.

The results of the current thesis indicate that Estonian EFL learners use the filler *so* almost two times more frequently than native English speakers. Whether the same can be said about other fillers, is outside the scope of this study, although a small overview is given in 2.3 subsection. This could mainly be due to the fact that the LINDSEI-EST corpus is currently much smaller than the LOCNEC counterpart, with only 38 interviews and 82,239 words of transcribed text compared to 55 interviews and 170,000 words of transcribed text in the LOCNEC corpus. Although relative frequencies could be calculated, the total amount of *so* that each speaker produced varied numerously between all interviews in both corpora. Moreover, it can be said that since *so* is closely related to the Estonian words *nagu* or *noh*, often used in Estonian speech as fillers, learners may have tended to use it more frequently for that reason. Since this thesis is not concerned with the Estonian counterparts of *so,* it remains unclear whether Estonian learners in fact used *so* more often due to these fillers in their mother tongue. Additional studies can be conducted on this topic and this analysis can be further expanded.

Based on further results, it can be said that there are not too many differences in the way that Estonian EFL learners and native English use the filler *so* in different functions. Although there are numerous distinctions in frequencies, speakers in both corpora use the functions relatively similarly, meaning that all three groups of functions (hesitation phenomena, shift in topic, holding the floor) as indicated by Rose (1998: 4-6) and Carter and McCarthy (2006: 903) are all featured and used similarly in the context of each sentence. An example of this can be seen in (29) and (30), where *so* is utilized to convey a shift in topic in both corpora.

(29) you like that <A> <B> yeah <laughs> <B> <A> er **so** you're living here on campus do you like (LOCNEC_12)

(30) alright <B> <A> okay . **so** I'd like to interview you informally . in (eh) (LINDSEI-EST_013)

This is mainly due to the fact that *so* in this context is used by the interviewers to guide the interview in both corpora. This relates back to the use of *you can* after *so*, in which the word pair was utilized together to express the same meaning. Likewise, *you can* in this context is utilized more frequently than any other word after *so*. However, when comparing the functions, it appears that "holding the floor" appears only 5.2 per 10,000 words in the LOCNEC and 2.9 per 10,000 words in the LINDSEI-EST corpora, meaning that it is the least frequent function of *so* in this study. In this case, it is likely that both Estonian and British students do not want to appear very assertive in their spontaneous speech in the interview.

Based on the results of hesitation phenomena in this study, it can be seen that the utilization of *so* suggests a contrast between Estonian EFL learners and native English speakers. An example of this is the use of pauses: while in the LINDSEI-EST corpus the relative frequency is 6.9 per 10,000 words, it is 6.1 per 10,000 words in the LOCNEC corpus.

This could be primarily because Estonian EFL learners need more time to think about what they are going to say next, since they are speaking in a foreign language. Another example of this is the use of false starts: there are 17 false starts in total in the LOCNEC corpus, however none in the LINDSEI-EST one. This result could suggest that British students, speaking in their native language, might make more small mistakes because they are less conscious about their speech compared to when speaking a foreign language.

Based on the results of *so* as a filler compared to its function as a conjunction or an adverb, some conclusions can be made. As a filler and an adverb, *so* is more prevalent in learners' speech; as a conjunction, it is more prevalent in native speakers' speech. The contrast between *so* as a filler and a conjunction is quite drastic, with 34.7 occurrences per 10,000 words as the former and 53.4 occurrences per 10,000 words as the latter. The contrast between the two corpora in the use of *so* as an adverb is 7.3 occurrences per 10,000 words. These results indicate that Estonian learners are likely to use *so* more frequently as a filler because they are speaking in a foreign language, and are, therefore, more prone to hesitating in spontaneous speech. *So* as a conjunction and an adverb is used quite similarly. However, the greater contrast in the use of so as a conjunction could suggest that British students use *so* more as a sentence forming unit of speech, whereas the smaller contrast in its use as an adverb indicates that Estonian learners use it more often as a way to indicate the degree of an adjective or an adverb.

CONCLUSION

In learner language, discourse markers and fillers play significant roles in spontaneous speech. Although discourse markers have received considerable attention from researchers, fillers remain relatively unstudied, despite their importance in speech. Fillers have often been studied alongside discourse markers or pragmatic markers. However, it is only in recent decades that researchers have started to explore fillers in various learner languages and conduct comparative studies, such as those involving Japanese EFL learners and Pakistani EFL learners. This current study hopes to become one of many to introduce fillers in the speech of Estonian EFL learners, as there is a noticeable gap in research within this field of study. Previous research in the context of Estonian EFL learners has so far only been concerned with the pragmatic markers *well* and *like*.

The aim of this thesis was to find differences in the use of the filler *so* between Estonian EFL learners and native English speakers. Another aim was to find how often *so* was utilized as a filler, in comparison to its function as a conjunction and adverb. In order to answer these questions, a corpus-based analysis was conducted using the LINDSEI-EST and LOCNEC corpora. It was found that there were 767 occurrences of *so* as a filler in the LINDSEI-EST corpus and 997 occurrences of *so* as a filler in the LOCNEC corpus. Due to the contrast in corpus sizes, all absolute frequences had to be normalised. Subsequently, it was observed that the relative frequency of *so* as a filler was 93.3 per 10,000 words in the LINDSEI-EST, compared to 58.6 per 10,000 words in the LOCNEC corpus. When used as a conjunction, the relative frequency in the LINDSEI-EST corpus was 14.3 per 10,000 words, whereas it was 67.7 per 10,000 words in the LOCNEC corpus. As an adverb, the relative frequency was 15.2 in the former and 7.9 in the latter. These results indicate that learners use *so* as a filler more frequently than native speakers mainly due to hesitation in speaking a foreign language, however specific distinctions between the corpora could not be

drawn from when *so* functioned either as a conjunction or an adverb. The main distinction in using *so* as a filler between students in the two corpora was essentially quite similar: both learners and native speakers employed all functions highlighted in the thesis (hesitation phenomena, shift in topic and holding the floor) in their speech. The first two functions were most prevalent in both corpora, whereas students from both corpora used *so* less frequently as a means to "hold the floor" in their speech.

This paper contributes to the two previous studies utilizing the LINDSEI-EST and LOCNEC corpora. The results found in this study are based on the data of the current corpora, implying that future research on fillers could indicate a different outcome. Nonetheless, future research is very much feasible, given the vast scope for research in fillers in general, including other words such as *you know*, *kind of*, *I mean* and *actually* mentioned in this thesis.

REFERENCES

Alami, Manizheh. 2015. Pragmatic Functions of Discourse Markers: A Review of Related Literature. *International Journal on Studies in English Language and Literature*, 3: 3, 1-10.

Anthony, Laurence. 2024. AntConc (4.2.4). [Software]. Tokyo, Japan: Waseda University. Available at https://www.laurenceanthony.net/software/antconc/, accessed 12 April 2024.

Beeching, Kate, Howie, Grant, Minna Kirjavainen and Anna Piasecki. 2022. *Discourse-pragmatic markers, fillers and filled pauses: Pragmatic, cognitive, multimodal and sociolinguistic perspectives*. Bristol: John Benjamins Publishing Company.

Bu, Jiemin. 2012. A study of the acquisition of discourse markers by Chinese learners of English. *International Journal of English Studies*, 13: 1, 29-50.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide to Spoken and Written English Usage*. Cambridge: Cambridge University Press.

Clark, H. H. and J. E. Fox Tree. 2002. Using *Uh* and *Um* in spontaneous speaking. *Cognition*, 84: 73–111.

Corley, M. and O. W. Stewart. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Language and Linguistics Compass*, 2: 4, 589–602.

Cuenca, Maria-Josep and Ludivine Crible. 2017. Discourse Markers in Speech: Characteristics and Challenges for Corpus Annotation. *Journal of Dialogue and Discourse*, 8: 2, 149-166.

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures*, 2: 225-246.

Fitriati, Sri Wuli, Mujiyanto, Januarius, Endang Susilowati and Perwari Melati Akmilia. 2021. The use of conversation fillers in English by Indonesian EFL Master's students. *Linguistic Research*, 38: 25-52.

Gilquin, Gaetanelle, Sylvie De Cock and Sylviane Granger. 2010. *Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Belgium: Presses universitaires de Louvain.

Hirschman, Lynette. 1993. Female-male Differences in Conversational Interaction. *Language in Society*, 23: 3, 427-442.

Jabeen, Farhat, M. Asim Rai and Sara Arif. 2011. A corpus based study of discourse markers in British and Pakistani speech. *International Journal of Language Studies*, 5: 4, 69-86.

Konso, Johanna. 2021. *A corpus-based study of like in Estonian EFL learners' speech*. BA thesis. Department of English, University of Tartu, Tartu, Estonia.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Okazawa, Sachiyo. 2014. Pauses and Fillers in Second Language Learners' Speech. *Core*, 52-66.

Rahusaar, Anne. 2019. *The compilation of the spoken sub-corpus for the Tartu Corpus of Estonian Learner English*. MA thesis. Department of English, University of Tartu, Tartu, Estonia.

Rose, Ralph Leon. 1998. *The communicative value of filled pauses in spontaneous speech.* MA thesis. Faculty of Arts, University of Birmingham, Birmingham, United Kingdom.

Sarifuddin, Muhamad. 2023. An Analysis Study for the Types of Adverbs Between English and Indonesia. *Journal Transformation of Mandalika*, 4: 7, 84-102.

Schiffrin, Deborah. 1987. *Discourse markers*. London: Cambridge University Press.

Seals, Douglas R. and McKinley E. Coppock. 2022. We, Um, Have, Like, a Problem: Excessive Use of Fillers in Scientific Speech. *Advances in Physiology Education*, October 5.

Shimada, Kazunari and Aika Muira. 2019. Japanese EFL Learners' Use of Discourse Markers in Their Speech and Writing. *The Japan Association for Language Education and Technology*, 56: 187-209.

Unubi, Abraham Sunday. 2016. Conjunctions in English: Meaning, Types and Uses. *International Journal of Social Science and Humanities Research*, 4: 3, 202-213.

Zwicky, Arnold M. 1985. Clitics and Particles. *The Linguistic Society of America*, 61: 2, 283-305.

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Romi Varula

A corpus-based study of the filler *so* in the speech of Estonian EFL learners
Korpuspõhine uurimus Eesti inglise keelt võõrkeelena õppijate täitesõna *so* kasutuse kohta

bakalaureusetöö
2024
Lehekülgede arv: 33

Annotatsioon:

Käesolev bakalaureusetöö võrdleb Eesti inglise keelt võõrkeelena (EFL) õppijate ja inglise keelt emakeelena kõnelejate täitesõna *so* kasutust. Töö esimeseks eesmärgiks on uurida, kui palju esineb *so* Eesti inglise keelt võõrkeelena õppijate ja inglise keelt emakeelena kõnelejate kõnes nii täitesõnana, kui ka sidesõna ja määrsõnana kasutades korpuseid LINDSEI-EST ja LOCNEC. Lisaks annab töö ülevaate täitesõna *so* peamiste kasutuste kohta nii õppijate kui ka emakeelena kõnelejate vahel.

Töö esimese pooles antakse ülevaade peamistest varasematest uuringutest nii diskursuse markerite kui fillerite kohta ning *so* kasutustest nii täitesõna, sidesõna kui määrsõnana. Töö empiirilises pooles tutvustatakse metoodikat ning kahte korpust, mida andmete analüüsiks kasutatakse. Peale seda viiakse läbi analüüs *so* kohta. Viimases osas kirjutatakse lahti peamised tulemused ning arutletakse tulemuste üle.

Täitesõnana kasutatakse *so*'d 767 korda LINDSEI-EST korpuses ja 997 korda LOCNEC korpuses. Suhteline sagedus oli peale normaliseerimist vastavalt 93.3 kasutust 10 000 sõna kohta ja 58.6 kasutust 10 000 sõna kohta. Sidesõnana kasutatakse *so*'d 118 (suhteline sagedus 14.3) korda LINDSEI-EST korpuses ja 1151 (suhteline sagedus 67.7) korda LOCNEC korpuses ja määrsõnana vastavalt 125 (suhteline sagedus 15.2) ja 135 (suhteline sagedus 7.9) korda.

Tulemustest ilmnes, et Eesti inglise keelt võõrkeelena õppijad kasutavad *so*'d täitesõnana peaaegu kaks korda rohkem kui emakeele kõnelejad, mis viitab sellele, et õppijad kalduvad võõrkeelt rääkides rohkem kõhklema kui oma emakeelt kõnelevad õpilased. Sidesõnana kasutatakse *so*'d rohkem emakeelt kõnelevate õppijate kõnes ning määrsõnana õppijate kõnes. Intervjuudest ei ilmne aga erilisi erinevusi selle vahel, kuidas sõnaliike lausetes kasutatakse. *So* kui täitesõna kasutavad nii Eesti EFL õppijad kui emakeele kõnelejad sarnaselt: nii kõhklusnähtused (*hesitation phenomena*) kui teemanihe (*shift in topic*) esinesid rohkelt mõlema kõnes ning jutukorra enda käes hoidmist (*holding the floor*) esines mõlemas korpuses suhteliselt vähe.

Märksõnad: inglise keel, keeleuurimus, korpuspõhine uuring, diskursuse markerid, täitesõnad, *so*

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Romi Varula,

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

   A corpus-based study of the filler *so* in the speech of Estonian EFL learners,

   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 kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, 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 punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

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

Romi Varula

Tartus, 21.05.2024

Autorsuse kinnitus

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

/allkirjastatud digitaalselt/
Romi Varula

Tartus, 21.05.2024

Lõputöö on lubatud kaitsmisele.

Jane Klavan

Tartus, 21.05.2024