“my wife says that you can be hooked on face powder
or whatever you call it”

A Corpus Based Study on *Extender Tags* from a Learner Language Perspective

Marjo Ojanperä
MA Thesis
English, Degree Program for Language Teaching and Learning
School of Languages and Translation Studies
Faculty of Humanities
University of Turku
October 2020

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

UNIVERSITY OF TURKU

School of Languages and Translation Studies / Faculty of Humanities

OJANPERÄ, MARJO: “my wife says that you can be hooked on face powder *or whatever you call it*” – A Corpus Based Study on Extender Tags from a Learner Language Perspective

MA Thesis, 66 p., 27 app.p.
English, Degree Program for Language Learning and Teaching
October 2020

This thesis studies how native speakers of Finnish use *extender tags* in spoken English. This linguistic phenomenon in English has been examined since the 1970’s. The learner corpus perspective, however, is a newly developing research area in the field. It has been made possible by the availability of learner corpora. This study utilizes the oral performances from the YKI-corpus which is comprised of the tests of The National Certificates of Language Proficiency in Finland. The data includes 360 recordings and the background information on the test participants from the YKI-corpus. The research employed a mixed-methods approach to examine the use of extender tags. The study was based on a quantitative research approach, but a qualitative approach was used to examine the environment of extender tags. In addition, the results are compared with two studies conducted abroad. The effect of social variables on the use of extender tags was studied by using the Orange toolkit.

According to the quantitative analysis, the three most popular extender tags are *and so on*, *or something like that* and *or something.* The analysis of social variables suggest that a higher level of education and youth promote the usage of extender tags. In terms of pedagogical implications of the study, it can be stated that the use of extender tags should be taken into consideration in language teaching. Students should learn to recognize different variants of extender tags and gradually also learn how to use them. Future research could concentrate on the functional properties of extender tags by conducting a deeper qualitative analysis.

Key words: extender tag, learner language, Finnish EFL learners, language variation, corpus study, social linguistics

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CA	Contrastive Analysis
CA	Classification accuracy (in Section 4.5.5)
CALS	the Centre for Applied Language Studies
DCPSE	Diachronic Corpus of Present-Day Spoken English
EA	Error Analysis
EFL	English as a Foreign Language
ELF	English as a lingua franca
ELFA	the English as a Lingua Franca in Academic Settings
F1	F-1 is a weighted harmonic mean of precision and recall
IPCU	income per consumption unit
LINDSEI	the Louvain International Database of Spoken English Interlanguage
LOCNEC	the Louvain Corpus of Native English Conversation
MWUs	Multi word Units
NNSs	Non-native speakers
NSs	Native speakers
SLA	Second language acquisition

1 Introduction

> I mean . nowadays even even having your address on . a public on a on a portal or webpage *or whatever* . might mean something to someone yo= you never know and and fake ids *and everything* that are are real life (YKI Corpus)

The words written in italics represent a linguistic phenomenon that at first glance might seem to be some arbitrary extra element that speakers use to fill in gaps in speech. While speaking the speaker might not remember a particular word, or the speaker might imply that there are other items, but these items are not mentioned. In these kinds of utterances, the speaker might add an and/or coordinator with a noun phrase at the end of the sentence. This way, the utterance is extended with "recognizable chunks of language" as O'Keeffe calls them (2006, 130). It is important to bear in mind that extenders vary in shape, length and terms of their position in a clause. Extenders have been labeled with a wide variety of names and thus, it is not surprising to find that the field lacks consistent terminology. I will be using the term *extender tag* throughout the thesis when referring to the linguistic phenomenon that is under investigation in this study. When citing previous studies, I will be using the terms the linguists use in their respective research.

Extender tags are not a novelty. In fact, they are as old as the English language itself (Carroll 2008, 7). According to Tagliamone & Denis (2010, 340), Shakespeare utilized *and things* already circa 1596. However, the first mention of the extender tag *and things* is recorded in the Oxford English Dictionary, in a text written by poet and playwright Benjamin Jonson in 1601: "And with-all calles me at his pleasure; I knowe not how many Cocatrices, *and things*" (OED, s.v. "thing," n.1). In addition to the English language, the linguistic phenomenon in question can also be found in other languages and has been studied, for example, in Montréal French (Dubois 1992), Swedish (Winter and Norrby 1999), German (Overstreet 2005), Spanish (Cortés Rodríguez 2006), Japanese (Watanabe 2014) and Persian (Parvaresh et al. 2012). A discussion of other languages in regard to extender tags falls outside the scope of this thesis and therefore will be excluded.

Although extender tags have been used throughout the history of the English language, the earliest attempts to analyze extenders did not emerge until the 1970's by Cristal & Davy (1975) and Ball & Ariel (1978). A sociolinguistic analysis of Australian women was conducted by Dines (1980) and terminal tags used by Scottish coal miners was investigated by Macaulay (1985). An exhaustive analysis on extenders within

discourse approach is illustrated by Overstreet (1999). Other significant studies have been conducted by Meyerhoff (1992), (Ward & Birner (1993), Channell (1994), Winter and Norrby (1999), Cheshire (2007), Cucchi (2007), Carroll (2008), Terraschke (2010), Martinez (2011), Pichler & Levey (2011) and Wagner et al. (2015). Aforementioned linguistic research has shown that there are indeed rules, that govern the usage of these lingual components, though the rules are very flexible by their nature.

As mentioned above researchers have investigated extender tags from various perspectives. The learner corpus perspective, however, can be considered to be a newly developing research area in the field (Buysse 2014, 4). Aijmer (2015, 212) underlines that it is the availability of learner corpora that has made it possible to investigate the use of general extenders by non-native speakers and to compare it with native speakers' use of general extenders. Buysse (2014, 3) points out that it takes time for the language research community to expand the research on a specific phenomenon to embrace a new domain such as learner language. Learner language is considered to be a variety of language that generates a separate research field within the study of language variation.

The object of this thesis is to widen the current knowledge of the use of extender tags in learner language particularly with regard to Finnish speaking EFL (English as a Foreign Language) learners among whom few studies have been conducted. Oral performances from the YKI-Corpus which is comprised of the tests of The National Certificates of Language Proficiency in Finland, are used as the basis for the data. No studies on extender tags in the YKI-Corpus have been made so far which offers a unique opportunity to investigate how extender tags are utilized by Finnish EFL learners. This is a corpus-based study and will aim to answer the following questions:

1. What is the distribution of extender tags in oral performances of the YKI-Corpus?
2. Which structures of extender tags are used by Finnish EFL learners?
3. What is the possible influence of social variables such as age, education level and test level in the use of extender tags by Finnish EFL learners?

The purpose of the first research question is to examine the occurrence of extender tags in the speech of the Finnish EFL learners. For counting the distribution of extender tags, only the oral part of the examinations of The National Certificates of Language Proficiency will be used. The oral performances were selected as the object of the

investigation of the present thesis since previous research indicates that extender tags are most frequently found in spoken discourse. As mentioned above, the access to ready-made learner corpora makes it possible to study learner language. That is why I chose the YKI-Corpus, which is intended for research purposes.

Through the second research question, my aim is to discover which structures of extender tags are used in the dataset. Analyzing the second research question, besides the extender tag itself, the scope of the extender tags will be included in the investigation. Thus, the structure in my analysis covers both the different variants of extender tags and the preceding context of the extender tags. Considering the scope of the thesis, I selected *and stuff/things (like that)*, *and so on* and disjunctives for a more detailed analysis of the preceding context. The distinct variants only used by the Finnish EFL learners are presented in Section 4.1. The results of the more detailed analysis of the above-mentioned extender tags are presented in Sections 4.2, 4.3, and 4.4.

The purpose of the third research question is to deploy the background information given by the test participants and to analyze the data by using the Orange software. In addition, the results of the first and the second questions will be compared to the findings of Buysse's (2014) and Aijmer's (2015) studies in order to obtain valuable knowledge about how Finnish EFL learners rank internationally. The comparison of the results of the first and second research questions is made by comparing the percentual shares and the number of tokens of extenders, the number of extender tag variants, and the ranking of the most frequently occurring extender tags. In addition, the percentual shares and tokens of short and long variants of two adjunctives and disjunctives will be compared.

Following the introduction, I will explain the structure of extender tags and consider their position and the ability to refer back in a sentence in Section 2. In addition, I will review most of the terms related to the phenomenon. Furthermore, previous research and studies with a learner language point of view will be briefly discussed. The source of the data of the study will be presented in Section 3 as well as the Orange toolkit which is used to analyze social variables in terms of the usage of extender tags. I will introduce the findings of my analysis in Section 4 followed by discussion in Section 5. Some possible future developments in the research regarding extender tags will be discussed in this section. Section 6 is the conclusion to the thesis.

2 Theoretical Background

Since the terms and approaches applied to this field of research vary with the interests of scholars investigating the discourse phenomenon in question, this thesis will focus on the key concepts related to this study. After explaining the key concepts, I will introduce previous research and will conclude the section by presenting studies with a learner language viewpoint.

2.1 Key Concepts

In this section I will describe how extender tags are formed, what they refer to and their placement in a sentence. Furthermore, I will be presenting the most significant labels on extender tags. Additionally, the multifunctionality of extender tags is discussed. The section ends with a brief overview of learner language, which is a constituent part of this study.

2.1.1 The structure of extender tags

This section unravels the basic structure of extender tags. It is important to note that extender tags are characterized by the diversity of the form which the subsequent examples will clearly exhibit. On the other hand, there are extender tags that appear in a fixed form, for example, *and so on* and *et cetera*. Consequently, the illustrations cannot include all possible forms. The aim is to provide a description of the most frequent variants.

Essentially, extender tags are comprised of the following elements in their canonical form: the conjunctions *and/or*, a modifier or multiple modifiers and a head noun (see Overstreet 1999; Carroll 2008, Wagner et al. 2015). This basic formula is illustrated in Figures 1 and 2 after which Figures 3 to 6 illustrate the most common variants of extender tags. In Figure 2 (example 1) *thing* functions as a semantically empty head noun and *other* as "a modifier that extends the denotation of the noun" Carroll (2008, 8). In addition to *thing*, there are other head nouns such as *one, people, where, shit, crap* (Tagliamonte & Denis 2010, 337). Typically, modifiers are quantifying determiners such as *every* or *any* (Wagner et al. 2015, 708).

Figure 1 The canonical formula of extender tags

The following example of an extender tag demonstrates the breakdown of the formula in Figure 2.

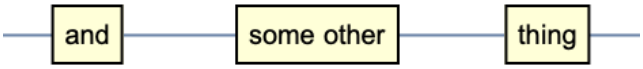


Figure 2 Extender tag example 1

The head noun can be, for instance, *stuff* as illustrated in Figure 3.

Figure 3 Extender tag example 2

In the following example 3 the head noun is replaced with a pronoun *that*.

Figure 4 Extender tag example 3

Next, an adverbial phrase *so on* is placed instead of a head noun.

Figure 5 Extender tag example 4

The most common extender tag variants are exemplified in Figure 6. It can be noted that a modifier can occur after the head noun or it can be omitted. Furthermore, a conjunction can be dropped out.

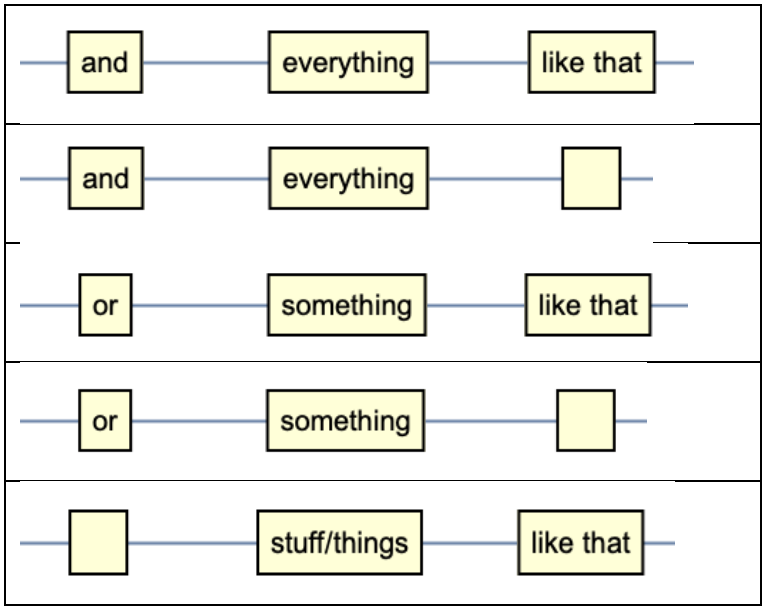


Figure 6 Extender tag examples 5 to 9

As seen above the extenders typically begin with the coordinators *and* or *or*. Overstreet (1999,4) was the first researcher to categorize extender tags into two groups according to conjunctions. The first group of tags begins with an *and*-conjunction and are called *adjunctive extenders*. The second group beginning with an *or*-conjunction is called *disjunctive extenders*. Overstreet adds the adjective *general* to the concept referring to their feature being nonspecific. With the term *extender* Overstreet refers to the extenders' ability to lengthen clauses that are already grammatically completed (1999, 3).

Besides *general extenders*, Overstreet also names *specific extenders* that can be identified according to the specific lexical items they contain (1999, 12). A lexical item can be an adjective that modifies a head noun of the extender or a relative clause that restricts the scope of the extender (1999, 51). As an example of the latter, Overstreet (1999, 52) provides a Bible verse:

(1) Neither shall you covet your neighbor's house, or field, or male or female slave, or ox, or donkey, **or anything that belongs to your neighbor**. Deuteronomy 5:6-21.

In this thesis, I have adopted the broad definition of *extender tags* provided by Ruth Carroll (2008, 12). Therefore, both *general* and *specific extenders* are embraced by the term *extender tag*.

2.1.2 Anaphoric reference

One of the fundamental characteristics of extender tags is their ability to refer anaphorically. In other words, there are a number of constituent types to which extender tags are appended. These constituent types aka exemplars (Channell 1994, 132) can be noun phrases, verb phrases, prepositional phrases or even clauses (Aijmer 2015, 220). It is also possible to find adjective and adverb phrases performing as exemplars, but they are rare (Channel 1994, 132). In the prototypical form the head noun is *stuff* as illustrated in example (2) by Cheshire (2007, 168). The head noun *stuff* points to a preceding mass noun Indie music.

(2) they're probably more into like Indie music *and stuff like that*

In example (3) from Aijmer (2015, 221), the extender tag *stuff* refers to a clause and example (4) from Pichler & Levey (2011, 450) illustrates the anaphoric reference to a verb phrase getting murdered.

(3) and she's moved now so she's okay they were really demanding and like we went over to stay . and she would they wouldn't let her have time off *and stuff* so that was a bit unfair

(4) You never hear of people getting murdered *and things like that* in Berwick, you know.

In addition to the above noted, extender tags can also refer to quoted speech or numerical expressions. Example (5) from Aijmer (2015, 221) portrays the numerical expression.

(5) eh it was because my: boyfriend he studied there in Auckland so after two-and-a-half months *or something like that* I: visited him

Furthermore, example (6) illustrates the instance of an adjective exemplar. Adjectives inquisitive and enquiring are followed by the adjunctive extender *and things like that*. Two adverb exemplars, in a different way and more efficiently are followed by the disjunctive *or something* in example (7) (Channel 1994, 136).

(6) I think the purpose of school is to make people inquisitive and enquiring and broader minded *and things like that*

(7) Don't you remember cos didn't you do something–tweak it up there to make it work in a different way or more efficiently *or something*

In addition to the above presented exemplars, the head noun *stuff* can even refer to states and events particularly when spoken by native English speakers (Aijmer 2015, 220). The odd one among extender tags is *and that* since its reference can be any in the preceding context (Chesire 2007, 168).

It is worth noting that there are also borderline cases for which it is difficult to define the reference. The next example by Cheshire (2007, 170) illustrates the ambiguity of the reference well.

(8) Sally: I might go over and teach languages *or something* in France *or something* . I'd quite like to go to Canada

Sally might mean that she wants to teach something other than languages or she might intend to have another type of job than being a language teacher (ibid.). In the former case the extender tag refers to the noun languages and in the latter case the referent is the predicate with its object, teach languages (ibid.). As Cheshire continues that there is not ambiguity with the second extender tag *or something* which refers clearly to France since in the next sentence another country is mentioned (ibid.).

In addition, the scope of the extender tag might be unclear as to how far the scope of the extender tag reaches out in the preceding context, particularly, if there is an *and*-conjunction between the nouns in the preceding context. In that case the categorization of the referents may be more complicated. I found this type of case in Aijmer's (2015, 219) study from which the following example is included here.

(9) I like sort of writing short stories and poetry *and stuff*

Aijmer presents this excerpt as an example of the canonical form of the extender tag *and stuff* where the head noun *stuff* has the same properties as the referent which is the case with the uncountable noun poetry. As I see it, *and stuff* also refers to the noun *short stories* that is appended with an *and*-conjunction to the noun poetry and thus both of the nouns belong to the scope of the extender tag. In that case, there is also a countable noun in plural as a referent which can no longer be considered as a canonical form. However, Aijmer leaves the matter unaddressed.

2.1.3 The position of extender tags in a sentence

Next, I will discuss the position of extender tags in a sentence. Generally, *tags* in a broader meaning are utilized to refer to clause-final elements, for example, questions tags (Carroll 2008, 12). Extender tags are observed to be used mostly in a clause-final position (Wagner et al. 2015, 708). This usage is exemplified in example (2). However, they are also discovered in clause-internal position which can be seen in example (10) provided by Overstreet (2014, 109). The extender tag *or something like that* is placed in the middle of adverbial phrases prior to verbs.

(10) I thought some poison in his drink *or something like that* a few weeks down the road would kill him.

In addition to the fact that extender tags are characterized by a flexibility in their form as can be seen in Figures 2-6, they also seem to be flexible in respect to their position. Overstreet (2005, 1849) points out that the distribution of tags can be truly person or group related. In her data, the extender tag *an'stuff* was used almost at the beginning of an utterance. This usage is exemplified in (11).

(11) Donna: You mean quarter to four.
Karen: Yeah, *an'stuff*—after I took care—after I took care of the body, so. I just figured I didn't wanna leave that hanging till the evening shift.

It is worth pointing out that the above example represents a rather rare usage of extender tags since it has already been stated in Section 2.1.2 that extender tags typically refer back to exemplars in the preceding context. The example (11) truly substantiates the flexible nature of extender tags.

2.1.4 Variety in terminology

Extender tags have been labeled with various terms by researchers over the years, starting from the 1970s. In this section I will highlight the most significant of them. As has been stated previously, extender tags typically occur in the clause-final position which is equally reflected in terminology. Most of the terms depict the tag feature.

Among the first to study extender tags were Ball and Ariel (1978). They focused on the interaction of *or-something*, which they classified as a member of *or-tags.* Their

data contained a collection of 190 sentences from overheard conversations and published presidential transcripts in the USA. One of the conclusions of the study was that *or-tags* are used to mitigate the directness of a speech act (1978, 41). This observation already points to the usage of politeness strategies that were introduced in the following decade by Brown and Levinson.

Dines (1980) studied extender tags from the sociolinguistic point of view and examined the speech of Australian women. She discovered that working-class women use *set-marking tags*, as she calls them, more frequently than middle-class women. She argues that *set-marking tags* serve as cues for the listener whereupon the listener is able to infer the more general underlying category or notion (1980, 22). The point of departure for her investigation was Labov's model of variation analysis which she extended to cover the discourse level (Dines 1980, 15). Variants of set-marking tags such as *and things*, *and stuff like that*, are considered as discourse variables which are "differentially distributed" (Dines 1980, 216). She concludes that although a stereotype suggests that set-marking tags serve as markers of vague and inexplicit speech, it cannot be stated on account of her analysis (Dines, 1980, 30). It is important to note that Dines excludes, for example, extender tags *and so on* and *et cetera* from her study. Furthermore, the range of disjunctive extenders was rather limited in the study.

Channell (1994) agrees with Dines' definition on *set-marking tags* but claims that the exemplar + tag-construction that she calls extender tags, function as *vague category identifiers*. According to her, by using *vague category identifiers,* the speaker refers vaguely to entities of which s/he has given an exemplar in the preceding context. Extender tags' role as a category identifier is significant, but it is only one of the multiple functions they serve in discourse.

Meyerhoff (1992) took the approach of hedging on nouns in the study of the extenders *sort of* and *or something* and named these extenders *post-noun hedges*. She studied whether there are differences in hedging strategies between genders. The term describes the structure of extender tags well, but as noted in section 2.1.1, there are also other word classes serving as constituent types.

Macaulay (1985) has studied extender tags from the point of view of social and regional variation in language. He introduced the term *terminal tags* in his analysis of the speech of a particular Scottish coal miner who used terminal tags more frequently than other interviewed community members. The most used terminal tag in his speech was *and that* and the second most used terminal tag was *et cetera* which is rather surprising

since it is not a part of the everyday vernacular of the Scottish mining community (Macaulay 1985, 112). According to Macaulay, the use of *and that* in everyday speech of the Scottish people is common but the frequency of *and that* used by this miner was unusually high (ibid.). Furthermore, there are some extender tags which seem to be rare and which did not occur in the learner data of YKI corpus either, for example, *and/or that there, and the likes of that*. Macaulay (1985, 113) suggests that *and that* is in some instances "a reduced form of *and things like that*" and serves a set-marking function in the utterance. The case of the Scottish coal miner offers an illustrative example of how the distribution of extender tags can be unique to an individual as explained in the previous section.

Jefferson (1990) studied extender tags from the list construction point of view. The focus of the study was on lists in conversation. In general, the list construction raises a question of the number of the parts involved and whether there is a list in progress with only one item. According to Jefferson (1990) that is the case. However, Lerner (1994) argues that there must be two items involved. Jefferson's claim has been corroborated by Bertrand & Priego-Valverde (2017) who studied list construction in French conversation. The rationale behind their conclusion was demonstrated by prosodic features. The prosodic dimension is not examined in this thesis, but otherwise the construction of one item plus the extender tag *and so on* is considered to be a list-in-progress in my analysis.

Jefferson's (1990, 74) analysis revealed the speakers' preference for formulating lists that comprise of three parts in which *generalized list completers* serve two main functions. By using *generalized list completers,* the speaker signals that her/his list construction process containing two parts is completed, and that it is the listener's turn to talk. It might not be a perfect list of items per se since adjunctive extenders imply that there are more in the group, but an extender tag closes the list construction process. This is exemplified in excerpt (12) offered by Jefferson (1990, 67).

(12) We were building, <u>camps</u>, and <u>airfields</u>, and, uh, *everything like that*.

As seen in the example above the first unit in the list is <u>camps</u>, the second unit in the list is <u>airfields</u> after which the extender tag occupies the third slot which signals transition to the listener. In her analysis Jefferson makes an interesting observation related to the phonetic environment prior to extender tags. The data of the study showed that speakers construct the list containing a list completer in such a way that a list completer

is acoustically consonant with the prior list items (Jefferson 1990, 709). The following example (13) from her data illustrates acoustic congruence (vowel/L) marked with bold font.

(13) Samuel jus' takes things casu**all**y en natur**all**y en, – ***all*** *that*

Before Jefferson, Crystal and Davy (1975) had taken somewhat a similar approach. They referred to extender tags as *summarizing phrases* when describing their functions as phrases that the speaker utilizes in summarizing the content of the utterance. They identified *summarizing phrases* as starting with the *and*-conjunction (1975, 113). Furthermore, they categorized *summarizing phrases* as vague language which occurs "at the end of a sequence of lexical items (such as a list)" (ibid.). However, it is important to recognize that their object of study was conversational English in general not just extender tags.

2.1.5 Multifunctionality of extender tags

There are diverse factors that affect the use of extenders in conversation. Among the factors can be found, for example, the familiarity of speakers, the formality of speech and type of discourse (Aijmer 2015, 212). Furthermore, the different roles that extenders have in a discourse influence which extenders speakers choose to deploy. Hence, it can be stated that they are characterized by multifunctionality (ibid). Some of their functions have already been mentioned briefly in the previous section such as the textual functions, including turn-exchange.

In conversation, speakers can use extender tags to intensify or to emphasize the content of their speech (Aijmer 2015, 226). Particularly, the extender tags *and everything* and *and all* are used for this purpose. The next excerpt from Buysse's (2014, 14) data illustrates the intensifying function. By using the extender tag *and everything* the interviewee highlights the beauty of the scenery.

(14) Interviewee: that was very beautiful that was really impressive in terms in terms of the beauty .. and when I went it was a black beautiful summer's evening all the light reflecting on the lake
Interviewer: oh yeah
Interviewee: *and everything*

Overstreet (1999) derives her functional categories of general extenders from Halliday's multifunctional principle, which I will be discussing next, since it lays down the foundation to better understand the approach taken by Overstreet.

According to Halliday, language use simultaneously embraces three different functions: ideational, interpersonal and textual. All three are related to how the meaning of clauses are processed (Halliday et al. 2003, 18). Searle (1965) had earlier introduced the notion of *communicative functions.* Halliday, in turn, termed these functions as *metafunctions* (Halliday et al. 2003, 18). The ideational metafunction refers to how speakers construe the surrounding reality including one's own consciousness. The speaker constructs the presentation of the natural world through language resulting in clauses. The ideational level comprises of participants, circumstances and processes. According to Halliday, there are distinct processes that take place in language production on this level. To give an idea of these processes, one example is given. The so-called *material processes* can be recognized by action verbs that occur in a clause (Halliday et al. 2003, 22). As in the clause "David closed the door" a concrete action is materialized and represented by the verb close. The interpersonal metafunction relates to the social world and to an interaction in which speakers and hearers utilize clauses as means of exchange. On the third level a clause is analyzed based on the textual metafunction and describes how the flow of information is organized in a clause. This includes *theme* and *rheme* (Halliday et al. 2003, 278).

There are two main categories in Overstreet's study. The first category covers representational function of language and the second refers to interpersonal function. Representational function, which is speaker-based, covers the roles general extenders have as vague language markers, list completers or category markers. Furthermore, the interpersonal function, according to Overstreet, contains the social relationships of interactants to which the function of general extenders is tied (1999, 18). Overstreet goes even further in claiming that the interpersonal function outweighs the importance of representational function (1999, 11). By using extender tags speakers show their stance towards matters discussed or towards the hearer (ibid.). According to Overstreet the usage of extender tags is also related to shared knowledge (ibid.).

In this connection it is relevant to briefly discuss the concept of linguistic politeness that is related to the functions of extender tags. The field of linguistic politeness dates back to the 1970's when Robin T. Lakoff (1973) initiated the modern study of politeness. In 1987 Brown and Levinson developed the model for politeness strategies.

Behind the theory lies the scale of Face-Threatening Acts. These are, as the term indicates, speech acts that imply a possible threat of losing face on the listener's part, on a scale of losing one's face entirely to not losing one's face at all (1987, 60). The speaker needs to decide which strategy s/he uses in order to mitigate the possible threat that his/her directive speech act causes to the hearer's face, depending on the circumstances.

According to Brown & Levinson (1987), the concept of face consists of two specific kinds of wants. The first one, a positive face, is one's desire to be approved of and liked. The second face is a negative one, which embraces the hearer's want for freedom from imposition. For example, when the speaker asks for some kind of help from the hearer, it threatens the hearer's negative face, i.e. his want not to be imposed on by anybody. According to the model, the speaker estimates the weightiness of his/her request and chooses the most appropriate strategy to soften his/her request. Brown & Levinson' model offers 15 positive and 10 negative politeness strategies that are available to the speaker. When estimating the risk of losing face and calculating the weightiness of the Face Threat Act, the speaker must consider three elements that exist in the speech situation: distance, power and the degree of imposition (Brown & Levinson 1987, 76). The first element distance denotes social distance. In other words, it conveys the closeness of interlocutors whether they are close friends or mere strangers to each other. The second element in a speech situation, power, refers to the relative power that exists between speaker and hearer, for example, a boss who has power over his/her employee. Thirdly, the degree of imposition equals the risk the speaker takes in uttering a specific message such as requests or orders (Brown & Levinson 1987, 77)

An example of such a directive speech act could be a request. The speaker asks the hearer, who could be a very close friend, to perform some favor in the most direct manner "Take me to the airport!" However, if the hearer is a person in a superior position, the request would impose the hearer and the speaker would need to choose a strategy which reduces the force of the imposition. Hence, the speaker might instead ask: "I'd like to ask you a big favor: Could you possibly take me to the airport?"

Overstreet states that by using adjunctive extenders speakers used positive politeness strategies especially when it comes to presupposing common ground (1999,99). In particular, the adjunctive extender *and stuff* serves this function (ibid.). Furthermore, disjunctive extenders such as *or whatever, or something* are use as negative politeness strategies and thus mitigate the imposition on the hearer's face (Overstreet 1999, 104). Speech acts where the speaker needs to weigh the possible threat on the

hearer's face are, for example, requests, proposals and offers (Overstreet 1999, 107). These events also pose a threat to the speaker's face. By using extender tags the speaker offers options to the hearer to choose from and thus the preferred answer will possibly emerge (Overstreet 1999, 108).

2.1.6 Learner language

Since this is a corpus-based study with a learner language perspective, it is relevant to discuss learner language. Language skills take a long time to develop and normally learning a new language takes place in progressive steps. Second language acquisition (SLA) theorists have studied foreign language learning from various angles resulting in various frameworks. The focus has been on the errors language learners make. Thus, linguists have attempted to resolve the underlying reasons by comparing a native language to a target language according to which the approach was named *Contrastive Analysis* (CA) (Carl 1990, 205). Since then researchers examined learner language by using *Error Analysis* (EA) in order to discover how learner language develops and what kind of development phases there are (Brunni & Jantunen 2015, 383). Studies have shown that there are universal similarities in learning phases and they are not reliant on the learner or their native language (Brunni & Jantunen 2015, 384).

However, it was noticed that neither CA nor EA could exhaustively explain the language learner's errors. The next turn in the field of SLA occurred when the concept of *interlanguage* was introduced by Selinker in 1972 (Sayer 2008, 405). According to this approach, a language learner develops an independent linguistic system with its own grammar (ibid.). An interlanguage is characterized by particular processes: "language transfer, transfer of training, strategies of L2 learning, communication strategies, and overgeneralization of L2 rules" (ibid.). The interlanguage theory has received much criticism. Particularly, the tip of the criticism has been directed to the goal of a language learner to be like a native speaker in all its aspects (Cook 2007, 18).

Subsequently, a more holistic view of learner language has been adopted. In 1991 Cook coined the term *multi-competence* to better describe the two-language system in the learner's mind (Cook 2007, 17). I am convinced that this broader view is opening the way to better understand leaner language and liberating language learners to consider themselves more as foreign language users than merely language learners for the rest of their lives. As Cook states "With multi-competence, the competence of a monolingual

native speaker became in a sense irrelevant; it was the competence of the successful L2 user that mattered." (Cook 2007, 18). The new holistic view brings many improvements. One of them is the notion that language transfer or interference goes both ways, from L1 to L2 and backward (Cook 2007, 20). As I see it, when there is more than one language in the learner's mind, transfer takes place as a network-like process in the brains.

A language learner receives new vocabulary and new grammar and gradually learns native-like phrases. It is worth pointing out that a language learner perceives words within a text either spoken or written whereupon words are associated with other words (Vetschinnikova 2019, 1). It seems to be obvious, but it has profound effects on language learning. Consequently, a language learner is compelled to learn multi word units (MWUs) that contain more than one word.

MWUs have also been described as phraseological units (ibid.). It must be noted that the concept of MWUs is extensive and concerns, for example, phrasal verbs, idioms, lexical bundles and extender tags. It has been proved that 50 % of the running text contains multi word units (Vetschinnikova 2019, 57). However, the occurrence of specific multi word units is far less (ibid.). The focus has been more on language learning processes within different linguistic theories and thus the same phenomenon has been termed *chunking* or *holistic processing* (Vetschinnikova 2019, 3). Researchers have discussed phraseology in relation to learner language since phraseological units of the target language are laborious to learn (Brunni & Jantunen 2015, 392). Even after spending some time in a target-language country, learners have difficulties in mastering them (ibid.). The mechanism of human memory lies behind all this and prefers meaning to surface structure (Vetschinnikova 2019, 63). Much exposure of a specific form is needed in order an accurate representation can be built in the brains (ibid.). In addition, other linguistic factors such as frequency and collocations occurring with MWUs should be taken into consideration. Thus, having a command of MWUs in alignment with other linguistic factors reveals the level of phraseological competence (Fernandez & Yuldashev 2011, 2625).

The reported findings of research on multi word unit processing by native speakers (NSs) and non-native speakers (NNSs) are conflicting. Some studies (Wray 2002, Arnon & Christiansen 2017) show that L2 learners have difficulties in acquiring MWUs. On the other hand, other studies present evidence on similar kind of learning process of NSs and NNSs, for example, on verb-argument constructions (Ellis, Römer & O'Donell 2016). Research has shown that nonnative speakers use MWUs with less

accuracy when compared to NSs (Vetschinnikova 2019, 58). It is suggested that the language processing of NNSs in terms of MWUs deviates from that of NSs resulting in inaccuracy (ibid.).

Researchers have given different explanations on divergence and one of the latest has stemmed from the studies on English as a lingua franca (ELF). When speakers with different native language communicate in English, the main goal is to get a message through not to focus on the production of the perfect formal English language (Vetschinnikova 2019, 58). This results in variability in the use of ELF and also in the use of MWUs. Mauranen (2009, 230) calls those deviations *approximations*, which are minor replacement of a word or structure (Mauranen 2011), such as a non-standard use of a preposition, which is found in a preposition phrase like *in/on my point of view* in which the prepositions *in* and *on* are used instead of the preposition *from* (ibid.). In addition, the findings of previous research show that the use of these approximations are not random, but systematic irrespective of a speaker's mother tongue or whether a speaker is a native English speaker or not (Vetschinnikova 2019, 64).

In addition to accuracy, there are other distinct development areas that are important parts of learner language. These are fluency and complexity (Brunni & Jantunen 2015, 387). Fluency is characterized by the speed of writing and speaking and possible breaks and corrections. In addition, other parts of the fluent language use are, for example, how automatized the language production is or how idiomatic expressions are (ibid.). Since non-native speakers use extender tags to improve their fluency, it is an excellent way for language learners to obtain more time for language production (Aijmer 2015, 229). For example, possible breaks can be avoided by adding an extender tag at the end of the sentence in speech on-line when an appropriate word is not found.

Complexity denotes how complex linguistic structures language learners are able to command and how non-automatized and complicated language learners are capable of producing (ibid.). Complexity can be measured, for example, by counting subordinate clauses within sentences (ibid.).

All three aforementioned features are affected by language teaching (ibid.). In this respect, teachers need to reflect their teaching with respect to what expressions they use, what task types students are given and whether students are given the possibility to plan their language production (ibid.).

One significant factor that affects the development of learner language is crosslinguistic influence. Previously, the influence of a learner's mother tongue was

considered to be a negative one, but nowadays researchers have revealed positive effects too (ibid.). Furthermore, the crosslinguistic influence affects the usage of extender tags. There are languages, for example the Finnish language, in which extender tags are not used in the same way as in the English language.

There are some traits that are characteristic to Finnish EFL learners. Research on speech rate and pauses in the English of Finns showed, that Finnish EFL learners use more silent pauses and their rate of speech was lower when compared to Swedish-speaking Finns or Swedish EFL learners (Lehtonen 1979, 49). According to Lehtonen, one explanation could be that it is customary to leave longer pauses in speech than in various other languages in Europe (ibid.). The speech rate of the spoken English by Finns is influenced by the characteristics of the Finnish language since all of the syllables of words in Finnish are pronounced (Lehtonen 1979, 45). Therefore, reduction in unstressed syllables is hardly produced by the Finnish EFL learners (Lehtonen 1979, 46). Consequently, it is difficult for Finnish EFL learners to acquire a native-like rhythm in spoken English.

The pronunciation skills play a significant part in Finnish EFL learning, since speaking a foreign language exposes the language learner to criticism. Furthermore, if the language learner feels unsuccessful in speaking the English language, it generates negative emotions. The shame of speaking English was studied among Finnish students in academic settings (Immonen 2020). The findings of the study showed that the pronunciation skills were the most influential factor in bringing shame (Immonen 2020, 62). Particularly, if the interlocutors were Finnish, Finnish students compared their pronunciation skills to their peers and, in order to keep a positive self-image, English-speaking situations and even speaking English in general were avoided (ibid.).

2.2 Previous research

Some significant studies have thus far been discussed in this thesis in the context of the terminology, Section 2.1.4. Furthermore, two studies will be presented in this section as well as a few studies with a learner language approach.

2.2.1 Recent studies

The following studies are presented in this thesis since both of them are based on ready-made corpora and investigate the usage of the same extender tags that are under

investigation in the present thesis: *and stuff, and things, and so on, et cetera* and *or something* to be specific.

Spoken discourse of British teenagers aged 13 to 17 was studied by Martínez (2011) and it was compared with the language of adults from Diachronic Corpus of Present-Day Spoken English (DCPSE). The study focused on investigating three extenders tags: *and stuff*, *and things* and *and everything* since they are typically utilized by teenagers (Martínez 2011, 2452). The findings of the study were in keeping with previous research and showed that in speech general extenders are used widely whereas in written discourse extenders are mostly found in informal writing such as emails or in texts composed by authors who reproduce conversations (Martínez 2011, 2459). Furthermore, the set of variants of general extenders utilized by adults is wider than the one utilized by adolescents (ibid.). However, the teenagers in the study utilized the general extenders *and that* and *and stuff* almost three times as often as adults (ibid.). The striking result of the study is the total absence of the general extenders *and so forth* and *so on and so forth* in the adolescents' speech (ibid.). Mention should also be made of the preference to use adjunctive extenders which was evident in the data of both the teenagers and the adults. Summing up, the results of the study showed that general extenders served interpersonal functions in discourse.

A similar kind of result is supported by the investigation conducted by Mari Metsä-Ketelä (2016, 325). Metsä-Ketelä examined the use of general extenders *and so on*, *et cetera* and *or something (like that)* in intercultural communication in academic settings in the dataset from the English as a Lingua Franca in Academic Settings (ELFA) corpus. It is comprised of a million-word database of English as lingua franca interactions which are recorded at four Finnish universities (Metsä-Ketelä 2016, 330). The speakers studied in the project were from several linguistic backgrounds, only 5 % of them were native English speakers (ibid.). The aim of the study was to investigate the pragmatic functions of general extenders (Metsä-Ketelä 2016, 326). The findings of the study show, for example, that the use of general extenders facilitated spoken interaction between international students and academic staff (Metsä-Ketelä 2016, 325). Instances in the data showed that extender tags triggered a quick response from the hearer, confirming that they had understood the message (Metsä-Ketelä 2016, 345).

2.2.2 Studies with a learner language viewpoint

Researchers have recognized that there are differences in the use of extender tags between native speakers and non-native speakers. In recent years the development in communication technology has offered new material for research on extender tags. Fernandez & Yuldashev (2011, 2615) deployed different instant messaging platforms such as iChat, Google Talk, Gmail, and Jabber in collecting data in order to examine the variation in the use of general extenders between native and non-native English language users. For the study, 524 instant messages were collected from the researchers' social network comprising of in total 53 participants at major universities in the United States (ibid.). There were 22 advanced level non-native English speakers among the participants. The focus of the study was on vague language particularly in the sense of how general extenders convey vagueness. The findings of the study showed that non-native speakers utilized more adjunctive extenders as opposed to native speakers (Fernandez and Yuldashev 2011, 2623). Researchers pointed out that the instant messaging medium might affect the usage of extender tags since the message sometimes had split into different lines and consequently the extender tags were dropped alone on a separate line (ibid.).

Moreover, Buysse and Aijmer have conducted comparative corpus-based studies on extender tags. Next, I will proceed to look at their studies in more detail. Buysse (2014, 4) examined the learner data of native speakers of Dutch in the Louvain International Database of Spoken English Interlanguage (LINDSEI) corpus and compared the data with the Louvain Corpus of Native English Conversation (LOCNEC) corpus of native speakers of English. The data was derived from 50 interviews of university students majoring in English. Each of the interviews lasted some fifteen minutes (Buysse 2014, 4). First, the students were asked to describe a travel experience they had had or to talk about a film or a book for two minutes (Buysse 2014, 5).

The findings of the study showed that the mother tongue appears to affect the learners' choice of certain general extenders, particularly, when the forms bear close resemblance to each other (Buysse 2014, 30). Furthermore, the results of the analysis confirm the results of previous research in terms of interpersonal interaction. It is worth pointing out that the command of language skills is not only about mastering grammar, containing phonology, syntax and semantics (Leech 1983, 12), but also requires the perception of language use in context, i.e. the speech situation. Leech underlines that pragmatics is concerned with the meaning of an utterance that the speaker and hearer

construe in cooperation in a speaking situation (Leech 1980, 80). In other words, one sentence has a different meaning depending on the context. This can be illustrated by Leech's (1980, 83) example sentence "Can you play the piano?". There can be two interpretations for this question. For example, two people, A and B, are at a music shop where there are pianos for sale. Person A asks this question with a view of getting to know whether person B is able to play the piano. Thus, this can be interpreted as a direct question. We can also picture a birthday party where person A asks person B the same question with the intention of asking the person B to play the piano. In this case the question can be interpreted as an indirect request.

In addition to the content meaning of a sentence, language users add lexical expressions to their utterances in order to, for example, comment on the basic message. (Fraser 1996, 67). These linguistic expressions are called pragmatic markers (Fraser 1996, 68). Fraser does not, however, include extender tags in his definition.

According to Buysse, (2014, 30) problems in language production cause the Dutch learners to use disjunctive extenders, particularly *or something (like that),* which exhibit approximation. However, the range of these pragmatic devices used by language learners is limited and they settle for few linguistic expressions (ibid.). Moreover, the analysis highlighted that the learners largely use general extenders when problems in oral performance occur (ibid.).

Aijmer also utilized the LINDSEI corpus in her research. She conducted a study on *General extenders in learner language* utilizing the Swedish component of the LINDSEI corpus, which contains Swedish learners' oral performances in English. Aijmer compared them to the data from the LOCNEC corpus (2015, 214). She adopted the term *general extender* and justified her decision by claiming that the term had been more frequently employed in the field in the recent years (2015, 212). The data from the learner corpus was compared with the data from the corpus of native speakers of English (Aijmer 2015, 214). Aijmer's corpus contained 50 interviews of university students majoring in English in their third or fourth year (ibid.), whereas the student interviewees in Buysse's (2014, 4) study were in their second or third year at university.

Aijmer's findings show that, on the whole, the Swedish learners use general extenders in different ways than native speakers (2015, 211). In particular, the use of and-extenders was less frequent among learners than among native speakers while the results regarding or-extenders showed the opposite (Aijmer 2015, 217). These findings are in line with Buysse's results (2014, 30). As mentioned previously in section 2.1.1 there are

short and long extender variants at speakers' disposal. There were no significant differences to be found between the learners and the native speakers in the use of short and long general extenders (Aijmer 2015, 222). The study showed that both learners and native speakers mix short and long variants in their speech, but short variants are preferred by both non-native and native speakers (Aijmer 2015, 219).

According to Aijmer's findings, language learners used general extenders either excessively or scantily (2015, 230). In addition, there is less variation in the structure of general extenders (ibid.), which is in keeping with Buysse (2014, 30). There was some variation in the use of general extenders among the Swedish learners. For example, 10 participants out of 50 used no general extenders during the interview (Aijmer 2015, 222). It is significant that the highest occurrence of general extenders was found in the performance of the participant who had lived in an English-speaking country for a long time (ibid.).

In Buysse's study the five most frequently used general extenders are respectively: or *something (like that), and stuff (like that), and so on, or so, or anything (like that)* (Buysse 2014, 6-7). Correspondingly, the equivalent list of Aijmer's study displays as follows: *or something, and stuff (like that), and so on, or anything, and things (like that)* (Aijmer 2015, 216-217).

The Functional aspects related to the findings from Aijmer's study should also be mentioned. The study revealed the multifunctionality of general extenders. Firstly, general extenders were used to foster interpersonal relations by sharing knowledge. Secondly, they were used to express hesitation. The third function was related to intensification (Aijmer 2015, 230).

3 Data and Methods

In this section I will present the data and the methods used to conduct this study. First, the source of the data, the YKI Corpus will be introduced. Then, I will discuss the methods used to examine the data.

3.1 YKI-Corpus

The source of my data is the examination of The National Certificates of Language Proficiency aka the YKI-corpus which is available as an online web application. The corpus consists of the tests of The National Certificates of Language Proficiency in Finland and the data are intended for research purposes only. The tests are available for adults and they are not a part of any curriculum. After each test round, which are normally organized twice a year, new data is added to corpus. There are nine test languages, English being one of them, and three test levels: Basic, Intermediate and Advanced.

There are four different subtests that are assessed in the test: reading comprehension, writing, listening comprehension and speaking. The reading and listening comprehension tests are comprised of three different types of tasks: multiple choices tasks, true-false tasks and open-end questions. The writing test contains three different writing tasks: a letter, e-mail and a reply to the letter to the editor. The Speaking test includes four different types of tasks: an account, conversations situations, situation tasks and presenting and justifying your opinion on given topics, for example, “Carpooling need to be promoted!” or “Would strong marriages be better for society?”.

The intermediate level speaking test form fall 2008 is on display in Appendix 2. Since the fall 2016 examination, the English intermediate level examination has been monolingual, with all task descriptions and questions in English. Previously, the participant was allowed to choose a booklet in either Finnish or Swedish. Regarding to the fourth subtest in the speaking test, a test participant is given a two-minute preparation period before the speaking time of two minutes. It is these recordings of the fourth subtest that are the object of the present study. It is also worth noting that task types are altered periodically.

The YKI-corpus is divided into older and newer parts the dividing line being the year 2010 at the time of retrieving the data. The lists of variables, which consist of the background information, are available in pdf format. The lists are attached to this thesis

as Appendix 3. It is voluntary for participants to fill in the background information and the data is not therefore available for every participant. Every examinee is represented by id numbers. The testing dates are not available in the background information. Recordings of oral performances are available in .mp3 format. Since extender tags are mostly found in spoken discourse, I decided to concentrate only on spoken language. Therefore, the written performances of participants were excluded.

It is worth mentioning that the interview part of Buysse's (2014) study is nearly identical to the oral performances of the YKI Corpus in the current study, though the topics discussed in the language tests deviate from each other.

In its entirety the YKI-corpus contained 898 recordings at the time of the accessing of the corpus. The YKI-corpus is broken down into the three sections: There are 48 recordings at the basic level, 694 at the intermediate level and 156 at the advanced level recordings. In total 360 recordings from the YKI-corpus were included in the data (Table 1). The duration of the collected data is 12 hours. All of 48 recordings from the basic level were included in the data. The basic level breaks down to 43 recordings from the old material and 5 recordings from the new part. 156 recordings from both the intermediate and advanced level were collected to be analyzed so that the size of the corpora on the advanced level and the intermediate level would be in balance.

Table 1 The distribution of recordings in the dataset

	≤ 2010	> 2011
basic	43	5
intermediate	126	30
advanced	126	30
	295	65

Since the Centre for Applied Language Studies (CALS) does not provide transcripts of oral test performances, the total word count for the oral performances is not available. The transcription conventions used in this thesis are based on the ones used in Buysses's (2014) study. However, they are simplified and modified to facilitate transcribing. The list can be seen in Appendix 1.

3.2 Data

Next, I will describe the election and processing of the data. Since there are three test levels in the YKI-corpus, the recordings must be collected separately from each test category. In addition, the collection of the recordings must be conducted from the old and new material separately. There were six rounds of collected audio recordings in total. The recordings were collected in sequential order, for example, beginning from the first until the amount of 126 from the old material at the intermediate level. There were a few recordings that had to be left out due to distinct problems such as technical reasons.

The first step in processing the data was to listen to the oral performances, to identify extender tags, to write them down and to count them. Extender tags were collected by listening to the multi word units starting with *and*- and *or*-connectives. In addition, extenders occurring without connectives were included in the collection.

The second stage was to enter all of the occurrences of the extender tags with participant ids into Microsoft Excel, where the total number of occurrences could be counted. Furthermore, I created an xlsx-file with worksheets for different categories and test levels: old and new material, adjunctives, disjunctives, extenders without connectives and the extender tag *and so on*. In addition, I named different columns for participant ID's, the preceding clause(s), extender tags and in some cases clauses subsequent to extender tags after which I transcribed the sentences containing extender tags. The transcriptions were added to worksheet columns. Furthermore, all worksheets were printed in order to facilitate the qualitative analysis and comparison between different categories.

3.3 Methods

As mentioned in the introduction, I have combined quantitative and qualitative research methodology in order to gain a deeper knowledge of the use of extender tags among Finnish EFL learners. Consequently, three research questions were formulated for the present study:

1. What is the distribution of extender tags in oral performances of the YKI-Corpus?
2. Which structures of extender tags are used by Finnish EFL learners?
3. What is the possible influence of social variables such as age, education level and test level in the use of extender tags by Finnish EFL learners?

In order to discover how the Finnish EFL learners rank internationally, the findings of the first and second research questions will be contrasted with the results of the native speakers of Dutch in Buysse's (2014) study and the results of the native speakers of Swedish in Aijmer's (2015) study.

The purpose of the first research question concerning the distribution of extender tags is to obtain quantitative results on which specific types of extender tags are used most frequently in the oral performances of the language test participants. In addition, it includes the division of extender tags to short and long variants and the distribution of both types. This is done by first listening to the selected recordings, then writing down the extender tags used by the test participants and categorizing them as explained in the section above.

To analyze the variants of adjunctives with a head noun such as *stuff* or *thing* in the data, I will be using the categorization utilized by Aijmer (2015). Aijmer's categorization is more detailed than those used by Cheshire (2007) and Tagliamonte & Denis (2010). The categories are named after the phrases which the head noun refers to. Aijmer's five categories are: Expected noun phrase (NP), other noun phrase (NP), verb phrase (VP), other class, which contains adverbial (AdvP) and prepositional phrase (PP), and the last category is labeled as clause. I made some additions to the labels of categories changing, for example, "correct NP" to NP count in relation to *and things* or NP non-count in relation to *and stuff* in order to better describe the properties of noun phrases.

Bearing in mind the scope of this thesis I decided to select only a few extender tags for closer investigation in terms of list construction. One of them is *and so on*, which according to Buyesse's (2014, 8) findings was over-presented in the learner data. In addition, its equivalent counterparts *and so forth* and *et cetera* are also included in the analysis.

Furthermore, a collocation analysis is made by examining whether the following collocates from Aijmer's (2015, 228) list occur with disjunctives. The list contains the following collocates: "*maybe, some, I don't know (dunno), sort of/kind of, probably, I don't know what it's called, like, I think (I suppose) ….or something, I can't remember"*. When disjunctives collocate with above mentioned hesitation markers, they signal uncertainty or hesitation (ibid.).

My third research question concerns the social variables that might affect the use of extender tags in oral performances. By using the data mining application Orange my

aim is to investigate whether the social variables such as age, test level, education level and socio-economic status correlate with the use of extender tags by the test participants. I have presumed the relationship between two variables are normally distributed and more or less linear. My aim has been to use as many ordinal variables as possible in order to use Pearson's correlation, which deals with linear dependency (Farrús et. al. 2012, 6-7). It is the simplest form of a relationship of dependence and I wanted to concentrate my research on the simplest model.

Logistic regression is used to predict probabilities. "Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables" (StatisticsSolutions 2020). In this study extender usage which has a "yes" or "no" interpretation equals binary variable and other information gathered from test participants such as gender is an example of nominal variables. Regression analysis has been used to predict, for example, age groups of Twitter users (Morgan-Lopez 2017) and author age from text (Nguyen et al. 2011).

In addition to the Orange toolkit, the computer software Mathematica will be used to draw illustrative figures which present the basic structures of extender tags (Figures 1 to 6 in Section 2.1.1), pie charts with percentage figures (Figure 8 in Section 4.5.1), the line plot (Figure 11 in Section 4.5.4). The Orange toolkit will be briefly presented in the next subsection.

3.4 Orange toolkit

There are several data mining tools available for use in this type of research, both open source and commercial. The Orange toolkit is a type of software that combines machine learning, data mining, data visualization and data analysis. The Orange toolkit is written in several programming languages, primarily Python and C++. It was originally released in 1996 at the University of Ljubljana and its latest release 3.25.0 (May 2020) is used in this study (Orange.biolab.si n.d.). I chose to utilize the Orange toolkit because it is an open source program and also user-friendly in the sense that its graphical user interface does not require any coding in order for the user to perform analyses. Instead the user chooses the widgets, connects them and loads datasets for analysis (See Figure 7).

In corpus linguistic the Orange toolkit has been used as a platform for automatic linguistic annotation, for example, for Slovene and English (Pollak et al. 2012). Another

study related to Machine Translation also used the Orange toolkit for feature selection and machine learning algorithms. Orange toolkit also offers the possibility to evaluate the precision of automatic translation since the significance of the quality of machine translation is increasing (Avramidis 2012, 84). The language pair used in the study was English-Spanish. The Orange toolkit was used in a pilot study in which language practices of adolescents in Austrian Carinthia were investigated (Zorčič 2019).

The Orange toolkit consists of graphical units that are called widgets which in turn include numerous machine learning, pre-processing and data visualization algorithms. Widgets serve various functions such as reading the data, training predictors and visualizing data elements. The distinct widgets are linked together using channels. To illustrate the function of the Orange toolkit, workflow contains a widget set $\{W_1, W_2, \ldots, W_n\}$ and a channel set $\{C_1, C_2, \ldots, C_m\}$ so that each $C_i = (W_j, W_k)$ where data output (W_j) equals data input (W_k). Thus, in most cases a widget has both input and output. Normally, the first input to a widget is a File widget. Workflows are constructed by dragging and dropping widgets onto a workflow canvas.

In order to further explain the functions of the Orange toolkit, the following example is given. Figure 7 portrays an example of the workflow of topic modelling of tweets. A main window can be seen on the upper left corner where there are widgets on the left side bar and canvas on the right side. The workflow starts from the Twitter widget. The following widget to the right is the Preprocessed Text widget which is connected to the Twitter widget by a channel. The Twitter window is a dialog window where the user sets the filters based on which the tweet will be searched. It is possible to define the time period and the number of tweet output. Consequently, the Twitter widget contains the tweets downloaded from Twitter which form the corpus of the study, in other words, the output of the Twitter widget is the input of the Preprocessed Text widget.

There are several procedures available to process the text, for example, filtering, transformation and tokenization. Next, the Preprocess Text widget is connected through a channel to the Top Modelling widget in which a particular algorithm is used to search ten topics in the tweets. The Top Modelling window organizes words filtered by the Preprocessed text widget according to topics. This is done by latent Dirichlet allocation (Blei et al. 2003, 993).

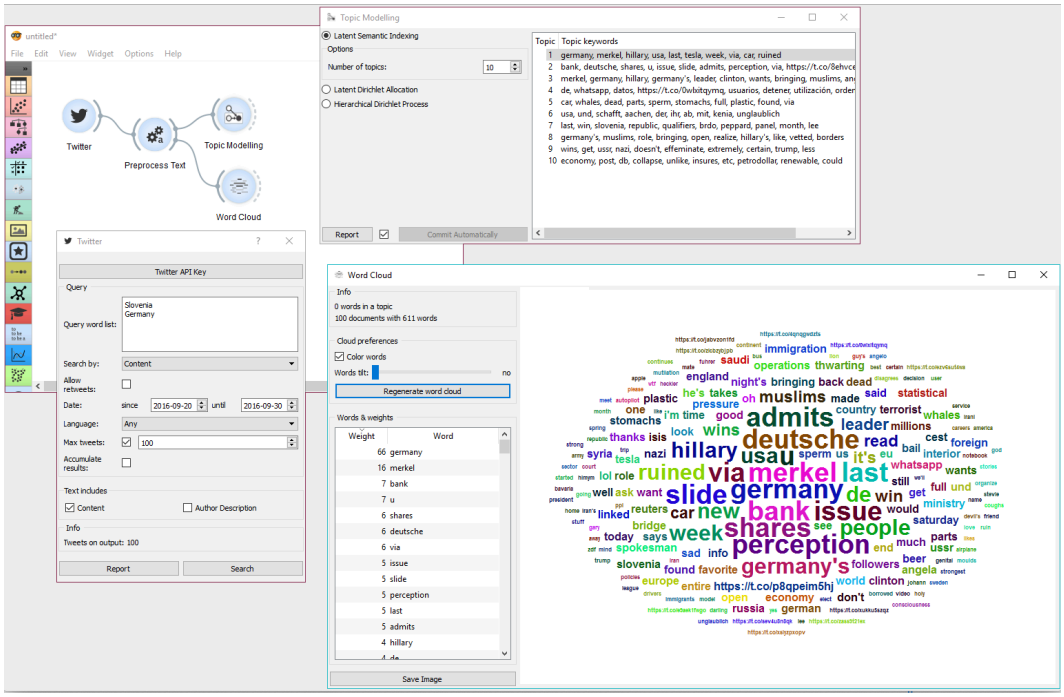


Figure 7 An example of topic modelling of tweets in the Orange toolkit

The most frequent words are displayed in a word cloud where it is possible to choose one word which will be shown in the separate Corpus Viewer window (not seen in the picture). All of the documents containing the word in question are listed in the Corpus Viewer Window for the user.

4 Analysis

I will begin this section by presenting the results of the analysis on the distribution of variants of extender tags after which the findings of the analysis on *and stuff (like that), and things (like that), and so on* are presented. Next, the analysis on collocations will be introduced. The Orange analysis on social variables completes the section. Samples from the data will be presented together with discussion.

4.1 The distribution of variants of extender tags

As mentioned earlier extender tags were collected by listening and dividing them into three different categories according to the connectives they occur or do not occur with. Therefore, counting the percentual share of each category provides the general overview of the use of extenders in the learner corpus. In addition, the figures are compared with the results in Buysse's (2014) study. The percentage figures and tokens of the three categories of the present study and Buysse's (2014, 6) study are combined in Table 2.

Table 2 The percentual shares and the number of tokens of extenders among the YKI, Dutch and Native Speakers (NS) corpora

			Buysse 2014					
	YKI		DUTCH			NS		
extenders	pct.	tokens	pct.	tokens	rel.	pct.	tokens	rel.
adjunctive	67 %	109	37 %	76	9,54	61 %	199	15,84
disjunctive	28 %	45	61 %	125	15,69	37 %	122	9,71
others	5 %	8	2 %	5	0,63	2 %	5	0,40
	100 %	162	100 %	206	25,86	100 %	326	25,95

The adjunctives form 67 % of all extender tags in the YKI-corpus, the total number of tokens being 162. The number of disjunctives is composed of 45 tokens covering 28 % of all extenders. The third category in which extenders are not preceded by connectives is comprised of the minority share of 5 % with 8 tokens. The percentage shares of three extender categories are more or less the same as in Buysse's (2014) study on the native corpus. However, the striking difference between the Finnish and Dutch learners is that the Dutch learners utilize approximately twice as many disjunctives as the Finnish learners. In addition, it is important to note that the Finnish EFL learners use

slightly more adjunctives 67% as opposed to the native speakers in Buysse's study, in which adjunctives used by native speakers take up 61 % of all extenders. The high occurrence of adjunctives in the Finnish learner data deviates from the results of both Aijmer's (2015) and Buysse's (2014) studies in that the Finnish EFL learners used substantially more adjunctives (67 %) than Dutch (37 %) and Swedish (47,7 %) EFL learners.

The high usage rate of adjunctives is partially explained by the substantially large amount of *and so on* extender that occurs in the Finnish learner corpus which can be seen in Table 3. The overuse of *and so on* extenders by language learners is observed also by Buysse (2014) and Aijmer (2015). In addition, Buysse (2014, 9) reported on two heavy users of *and so on* in the Dutch data. They are heavy users in a sense that one of them used the extender tag in 7 instances and the other interviewee used it in as many as 9 instances during the interview (ibid.). This type of heavy users was not found in the Finnish learner data. However, there were 3 participants who were each responsible for 3 tokens of *and so on* in their speech. Furthermore, 7 participants were found to utilize 2 tokens of *and so on* each.

The high occurrence of the extender tag *and so on* can be considered as overuse since in native speaker corpora it has been observed nearly exclusively in written (Biber et al. 1999, 117) and formal speech (Overstreet 1999, 7). One possible explanation in terms of Finnish EFL learners might be the L1 inference. The resemblance of the meaning and the form of *and so on* to the equivalent expression in the Finnish language "ja niin edelleen" is evident. It is also used similarly both in English and in Finnish, as a means to complete the constructed list of items. Furthermore, it is easy and quick to add in clause-final position.

Next, the distribution of different variants of extender tags and the number of tokens occurred in the YKI-corpus are shown in Tables 3, 4, and 5. As mentioned above, the most remarkable finding is the use of the extender tag *and so on,* which will be discussed in more detail in Subsection 4.3.

It is interesting to notice that the extender tag *et cetera* with its variants occurs several times in the YKI corpus. It is significant that it is totally absent from the non-native corpora in Buysse's (2014) and Aijmer's (2015) studies. Moreover, Aijmer reports that *et cetera* was not used by the German learners (2015, 224). However, the Finnish EFL learners have used *et cetera* and its two distinct variants. The three tokens of *et cetera* extenders are explained by the fact that on the intermediate level, there was a test

participant who used it three times during the oral performance. Another test participant on the intermediate level used a double *et cetera* once in his speech. The third variant *and et cetera* was also used by a test participant on the intermediate level (See example (15)). It is noteworthy that there are two (*and, et*) conjunctions in this variant. The extender *et cetera* originates from Latin denoting in English and (*et*) the rest (*cetera*) (OED, s.v. "et cetera"). It may be that the speaker is not aware of it or that there is language transfer from the native Finnish to the target language English. No combination of two conjunctions was found in Buysse's (2014) and Aijmer's (2015) data whereupon the findings of the YKI corpus in terms of *et cetera* and its variants can be seen exceptional.

(15) healthy minerals you need . every day . like vitamins *and et cetera*

The following excerpt (16) embodies the double construction.

(16) if every other employees would . let them to do exercises . in mid= in the middle of the working day . go swimming . running . *et cetera et cetera* hmm too much exercise could be a problem . for injuries .

Table 3 The distribution of adjunctives in the YKI corpus

	token
and so on	55
and things like that	10
and stuff like that	8
and stuff	4
and something like that	4
and everything	3
and so forth	3
and that kind of things	2
and so on and so on	2
and any other stuff from wood	1
and services like that	1
and that kind of stuff	1
and other such kind of things	1
and all the other stuff of them	1
and healthy stuffs like that	1
and like that	1
and this kind of thing	1
and all that stuff	1
and that kind of thing	1
and all kind of thing	1
and so	1
and et cetera	1
and on and on	1
et cetera	3
et cetera et cetera	1
Total	**109**

As shown in Table 4, the three most common disjunctives used by the Finnish EFL learners are *or something like that, or something and or whatever*. The comparison of the rank order of the three most common extenders between the YKI-corpus, Buysse's LINDSEI-DU and Aijmer's LINDSEI-SW corpora will be presented and discussed in conjunction with Table 8.

Table 4 The distribution of disjunctives in the YKI corpus

	token
or something like that	13
or something	11
or whatever	4
or anything	3
or things like that	3
or anything like that	2
or other places like that	1
or elsewhere	1
or something something else	1
or so on	1
or place like that	1
or whatever you call it	1
or some places like that	1
or something like that kind of diseases	1
or places like that	1
Total	**45**

Example (17) illustrates the use of the most used disjunctive.

(17) I think the Finnish driving culture is good but it could be make . made better by giving more penalties to those who drive . when they (x) (x) when they are drunk or *something like that*

In the next example (18) the disjunctive *or something* is referring to a numerical expression sixteen.

(18) I used to have these pictures . of me . drinking when I was sixteen *or something*

The distribution of the third category containing extender tags without connectives is presented in Table 5. Even though there is no conjunction, it is possible to infer to which category they belong to based on the meaning of the sentence. Thus, all four tokens of *things like that* belong to adjunctive extenders (See example (19)), disjunctive extenders are *any of that kind of stuff*, *anything like that* and *something like that*. The extender tag *places like this* seems not to belong either of the previous categories.

Table 5 The distribution of the extenders without connectives in the YKI corpus

	token
things like that	4
any of that kind of stuff	1
anything like that	1
places like this	1
something like that	1
Total	**8**

An *and* conjunction could be placed in front of the extender tag in the next example.

(19) you should take them to different kinds of basic courses where they could get to know different kind of (er) hob= (er) hobbies . running (er) . orienteering (em) . taekwondo karate . *things like that*

The number of different variants of extender tags used by the Finnish EFL learners and the Swedish EFL learners together with native speakers from the Aijmer's (2014) study are collected in Table 6.

Table 6 The number of different extender tag variants used in the YKI, the LINDSEI and LOCNEC corpora

	YKI	Aijmer 2015 LINDSEI-SW	NS/LOCNEC
adjunctive	25	14	18
disjunctive	15	10	9
others	5	3	3
	45	27	30

Almost half of the different extender tag variants (21 out of 45) used by Finnish EFL learners were ones that were not found in Buysse's (2014) and Aijmer's (2015) studies. However, it must be emphasized that the majority of the invented variants are only used once. They are collected in Table 7.

Table 7 The list of extender tags used only by the Finnish EFL learners

		token
1	and any other stuff from wood	1
2	and services like that	1
3	and other such kind of things	1
4	and all the other stuff of them	1
5	and healthy stuffs like that	1
6	and like that	1
7	and this kind of thing	1
8	and all kind of thing	1
9	and et cetera	1
10	or other places like that	2
11	or elsewhere	1
12	or something something else	1
13	or so on	1
14	or place like that	1
15	or some places like that	1
16	or places like that	1
17	or something like that kind of diseases	1
18	things like that	4
19	any of that kind of stuff	1
20	anything like that	1
21	places like this	1
	Total	**25**

The above list contains instances that deviate from the canonical formula explained in section 2.1.1. It can be noted that there are unexpected variants, for example, *and healthy stuffs like that*. A speaker uses the noun *stuff* in plural although the head noun is uncountable. One cannot say "one stuff" or "five stuffs" but some *stuff*. It is a fact that in the fifteenth century according to Overstreet (2014, 115), the noun stuff has been used in a concrete meaning, for example, to refer "moveable property" but in the present-day English it is used to allude vaguely to things (ibid.).

One of the listed extender tags *or so on* is an exception in a sense that I did not find it in the inventories of other corpus studies (Buysse 2014, Aijmer 2015, Pichler & Levey 2011 and Tagliamonte & Denis 2010). In Buysse's study *or so* is found both in the Dutch learner corpus and native speaker corpus. However, the Dutch learners use it in a wider manner compared to native speakers, which use it when pointing to numerical expressions with approximation (Buysse 2014, 28).

The extender tag *or so on* is used as a counterpart of *and so on* in the sentence which is shown in example (20).

(20) now you can see the same thing in Finland . we are eating too much junk food and fast food and so on . or the hamburger or chains or . pizza restaurants *or so on*

The ranking order of the most popular extenders reveals interesting differences between learners. In the YKI corpus, *and so on* picks up the highest rank followed by *or something like that* and *or something* whereas the ranking order in Buysse's (2014, 6-7) Dutch corpus is: *or something (like that), and stuff (like that), and so on.* Buysse does not break down the difference between the short and long variants for *or something*. The top three general extenders in Aijmer's (2015, 218) study look somewhat different: *or something, and stuff, and stuff like that*. The two highest ranks belong to short variants of extenders in the LINDSEI-SW Corpus. The ranking order of native speakers include the following general extenders: *and things, or something, and everything* all of them being short variants. In addition, Table 8 shows that the short variant *or something* is included in the top three in all four corpora.

Table 8 The ranking of the three most popular extenders in YKI, LINDSEI-DUTCH, LINDSEI-SW and LOCNEC-NS

		Buysse 2014	Aijmer 2015	
Rank order	**YKI**	LINDSEI-DUTCH	LINDSEI-SW	LOCNEC-NS
1	and so on	or something (like that)	or something	and things
2	or something like that	and stuff (like that)	and stuff	or something
3	or something	and so on	and stuff like that	and everything

Mention should also be made of the extender tag *and stuff (like that)* that take up only fifth and sixth position in the ranking of all extender tags in the Finnish learner corpus whereas it is highly common in the Dutch and Swedish learner corpus. Buysse (2014, 11) claims that the vast supply of American television programs and films explains the tendency for the extender tag *and stuff (like that)* in the Dutch-speaking part of Belgium. Even though, we have the same kind of vast supply of American films in Finland, the use of *and stuff (like that)* was not found with high frequency in the Finnish learner corpus in this study.

In terms of length of extender tags there is variation in use. Speakers can choose to use, for example, the short version *and stuff* instead of the long version *and stuff like that*. According to the findings of the present study, the Finnish EFL learners prefer to use long versions over short versions when it comes to the extender tags *and things/stuff like that* since the third and fourth slots in the list of the most popular extender tags are occupied by the long versions of the aforementioned extender tags. The opposite is found in the findings of Aijmer's (2015, 219) study, which showed that both non-native and native speakers prefer to utilize short variants (*and things/stuff*). The short variants are in the majority (66%) of the tokens of *and things/stuff like that/this* in the LINDSEI-SW as indicated in Table 9. The correspondent figures for the LOCNEC corpus are 86 tokens out of 132 equals to 65% (ibid.).

However, when compared the findings of the YKI-Corpus to the DUTCH corpus (Buysse 2014), it was revealed that the Dutch learners of English also prefer to use the long variants. One explanation may be that by using the long variants, language learners obtain more time for language production (Buysse 2014, 13). The comparative percentage figures for short and long versions for *and stuff/things*, *and stuff/things like that* are presented in Table 9.

Table 9 The comparison of short and long variants with two adjunctives in YKI, LINDSEI-SW

	YKI		Buysse 2014 DUTCH		Aijmer 2015 LINDSEI-SW	
	token	pct	token	pct	token	pct
and things/stuff	4	18%	14	38%	62	66%
and things/stuff like that/this	18	82%	23	62%	32	34%
Total	22	100%	37	100%	94	100%

Furthermore, the equivalent numbers for *or something* and *or something like that* are presented in Table 10.

Table 10 The comparison of short and long variants with two disjunctives in YKI, LINDSEI-SW

	YKI		Aijmer 2015 LINDSEI-SW	
	token	percentage	token	percentage
or something	11	46 %	71	83,5 %
(or) something like that/this	13	54 %	14	16,5 %
Total	24	100 %	85	100 %

4.2 And stuff (like that), and things (like that)

In this section I will take a closer look at the extender tags *and stuff (like that)* and *and thing(s) (like that)* using the categorization utilized by Aijmer (2015).

As it was noted in the section 4.1., the Finnish learners of English appear to exhibit a preference for the long variants of extender tags *and stuff* and *and things*. Table 11 shows the use of *and stuff (like that)* and its referents. Only four instances of *and stuff* were used in the data whereas its long variant occurred in eight instances of the total twelve tokens.

Table 11 The use of *and stuff (like that)* and its environments

	short variant	long variant
NP non-count	1	2
NP count	2	3
VP	0	0
Clause	1	2
Other	0	1
Total	4	8

It is worth noting that *and stuff* occurred only once in the dataset in the expected noun phrase. It can be seen as a parallel trend which is also reported by Cheshire (2007) and Tagliamonte & Denis (2010) in their studies.

The extender tags *and stuff* and *and stuff like that* referring to a countable noun occurred in five out of twelve instances. This type of incompatibility in the use of *and stuff (like that)* was also observed in the data of NNs and NSs in the LINDSET-SW and LOCNEC corpora (Aijmer 2105, 220). Another notable finding from the data is that there were no instances used with verb phrases. This differs from Aijmer's (2015, 220) study,

in which verb phrases occurred both in the LINDSEI-SW (8 tokens) and LOCNEC (6 tokens) corpora.

The following example illustrates the atypical usage. The speaker uses countable noun in plural in the proceeding context.

(21) if one of them gets stolen or all your . life or savings and bank acc= bank accounts *and stuff like that* gets stolen too

In example 16 the extender tag refers to a clause.

(22) Helsinki in particular in Finland has become really multicultural . bus= busdrivers are . usually black guys *and stuff like that*

The preferred usage of the short variant *and things* resembles that of *and stuff.* This is further shown by the fact that no short variants occurred in the data (Table 12). Four out of the ten tokens occurred with an expected NP, which represents the minority of the tokens *and thing like that.*

Table 12 The use of *and things (like that)* and its environments

	short variant	long variant
NP count	0	4
NP non-count	0	1
VP	0	0
Clause	0	3
Other	0	2
Total	0	10

An illustration of the expected usage of *and things* is exemplified in excerpt (23) in which the head noun refers to an inanimate, countable noun *cakes*. The given topic deals healthy food and the speaker gives cakes as one example of a larger set of unhealthy food.

(23) it's not obligatory to eat just healthy food of course you can eat cakes *and things like that* whenever you want

Next, the example (24) embodies the case where the anaphoric reference is a clause.

(24) they could . pay some expen= expensive swim pool time *and things like that* so that people really start moving

An example of unlikely use of *and things* is illustrated in example (25) in which the head noun refers a non-count noun.

(25) you are what you eat is part true . because if you eat only fat and . sug= sugar *and things like that* you go fat

4.3 And so on

In this section I will present the results of the analysis on *and so on* since it is the most frequently occurring extender tag in the dataset.

As mentioned earlier my analysis on the extender tag *and so on* is based on the research on list-construction by Jefferson (1990). I examined the items that occur in front of the extender tag *and so on* and counted and grouped them according to what number of *and so on* is exhibited on the list. Table 13 shows the breakdown of the items that occur in front of the extender tag *and so on*.

Table 13 The breakdown of the items in front of *and so on*

	1 item	2 items	3 items	4 items	Sum
Basic	0	1	2	0	3
Intermediate	11	24	5	5	45
Advanced	2	3	1	1	7
Total	13	28	8	6	**55**

The analysis revealed that these lists consist of between one and four items plus extender tag *and so on*. Hence, the analysis exposed four cases from which examples are introduced.

The first group contains one item in the preceding context of *and so on*. In the learner corpus 13 instances out of the total 55 belonged to this first group. In example (26) the intermediate level test speaker gives only one example of diseases which can be avoided by exercising. As it was discussed in Section 2.1.4 that one item plus *and so on* signals that a list construction is in progress. The extender tag signals that there are other diseases that can be avoided as well, but for some reason the speaker is not giving more examples whereupon the extender tag closes the list construction process.

(26) motion is important I think you you are in better condition for instance your many diseases you can avoid by moving for instance heart diseases *and so on*

The majority of the usage of *and so on* fall into the cases which corroborate the theory of list-construction by Jefferson (1990) that lists should contain three parts. Two excerpts illustrate this function. Excerpt (27) is from the intermediate level.

(27) me personally I . I could eat more . more healthier but I try to eat more green food . that means veggies and beans *and so on*

(28) I personally like want to hear their opinions about real subject matters serious matter such as (er) . EU (x) poverty issues *and so on*

The third group contained 8 instances of the usage of *and so on* in which three items precede the extender tag *and so on*. This structure was used on all test levels. The following example (29) illustrates the list construction by the speaker on the basic test level.

(29) I like to stay place like . where I can sporting . swimming and climbing *and so on*

Lastly, six instances with four list members in front of *and so on* were found in the dataset. One of them is exemplified here.

(30) of course I find that information from internet and or all magazines and newspapers and television *and so on*

In addition, there were two instances of *and so on and so on* which are not numbered among the amount of *and so on* occurring solely. Both of them are used by the advanced level test participants.

(31) and then the solution for a man . is that . let's solve this problem . what can you do . why aren't you doing this . what could you do better . why don't you talk to him . *and so on and so on*

(32) hm parties should divide their monery for . for example adver= . advertising so . that (er) . every candidate would get the exact same little money and exact little same time on media *and so on and so on*

By using the double construction, the speaker wants to emphasize that the list could be continued with many more examples in addition to those that have been given thus far.

In addition, three tokens of *and so forth* and one token of *and et cetera* were used in the learner data. The usage of the one single *and et cetera* in the data belongs to the first category in which there is one item in front of the extender tag. This case has been treated in Section 4.1.

Two of the three tokens of *and so forth* belong to the group containing two items in the preceding context and thus follow the expected list construction model as discussed in Section 2.1.4. The third one of them is included into the fourth group. The next example illustrates the use of *and so forth* as an alternative choice to *and so on*.

(33) they don't know any foreign people so they are afraid . afraid because they don't know their culture they don't know the language . *and so forth*

4.4 The results of the collocation analysis

I examined the context of disjunctives in the dataset and searched for the phrases in Aijmer's list. The distribution of collocations is presented in Table 14. The result of the search showed that 8 test participants out of 39 participants using or-extender tags employed the following collocations: *like, maybe, some, probably, some kind of.* The most used collocation was *like*. The collocation *some kind of*, which was not listed by Aijmer, occurred two times in the dataset.

Table 14 The distribution of collocations with disjunctives

Collocation		Disjunction
like	3	or something (like that)
maybe	2	or something (like that)
some	2	or something / or something something else
probably	1	or something like that
some kind of	2	or something (like that)
	10	

In addition, five collocations did not occur in the learner data at all: *I don't know (dunno), sort of/kind of, I don't know what it's called, I think (I suppose)….or something, I can't remember*. In my estimation, one reason for their absence in the learner data might

be that the speaking test is not a natural language interaction with other people and since the two-minute time of speech slips away very quickly, test participants do not have too much time to spend pondering possibly missing words. Ultimately, every test participant is aiming at obtaining high grades and a good certificate of their language proficiency.

It is noteworthy that the data analyzed by Aijmer (2015) originated from the interviews in which an interviewee's fastness to talk is not under evaluation. Hence, the learners of her study could use these kinds of hesitation markers more as a fluency device. Since no quantitative results on the usage of these collocations are presented in Aijmer's (2015) article, it is difficult to compare, for example, the proportional shares of the usage of collocates between Finnish and Swedish learners. Following this, a couple of examples of the usage of above mentioned collocates are provided.

The use of *maybe* is illustrated in example (34) in the learner data. The topic is related to healthy food and the speaker is describing what ingredients a healthy meal contains. After first presenting vegetables as a portion of a healthy meal the speaker shows hesitation by employing *maybe* in front of the word chicken and in that manner expresses that also chicken might be healthy, but he is not certain about it.

(34) I think healthy er meal is that kind where is er lots of vege= vegetables and . not lot of fat and **maybe** chicken and rice *or something like that* but not beaf

Example (35) contains some hesitation by a speaker who employs both **like** and **maybe** in the preceding context of or something. There are **maybe** and twice **some** in the sentence, but they are outside the scope of the extender tag; *or something* points to in work.

(35) people maybe people will do some sport things more if they get er some extra advantage **like** in **maybe** in work work *or something*

Both of the above examples represent the intermediate level participants. It is noteworthy that only one test participant at the advanced level used the collocation *some.* It is possible that the test situation and the two-minute time limit forces the advanced level test participants to present their best language skills and thus they do not want to show any hedging in their speech.

4.5 The results of the Orange analysis

Before presenting the results of the Orange analysis, I will explain the major treatment of the files. Prior to the actual Orange analysis, it was de rigueur to edit all six Excel files to correspond to each other before joining all of them to one file which is then used as input on the Orange toolkit. The files retrieved from the YKI Corpus contained information that was not relevant to my research purposes and consequently the following background questions were excluded.

In the old material:
Question 15: Where did you get information about the National Certificates from?
Question 16: How did you choose the test level?
Question 17: For what purpose will you use the certificate?

In the new material:
Question 13: For what purpose do you need the certificate?
Question 16: In your opinion, how well do you use the test language?

When regarding the numbering of the background questions, it can be noticed that there are several differences between the old and new materials, for example, the question numbered 16 are distinctly dissimilar. There were also other disparities that had to be edited in order for the questions and columns in the Excel file to correspond to each other.

The answers for the question about basic education and the question about where the participant has studied English were compared and insertions were made accordingly. For example, if a test participant had answered that s/he had studied English in high school, but the information about high school on the basic education was missing, the information was added to the basic education column.

In addition, eight columns containing extender tag information were inserted to the file. One column contained the information on whether the test participant has used an extender tag (1) or not (0). The six columns containing itemization of the variants of extender tags read as follows: and short, and long, and so on, or short, or long, and other. The total sum of the extender tags used was counted and displayed in the eighth column.

Furthermore, it was necessary to change the numerical information on some answers to correspond to a ranking order so that it was possible for the Orange toolkit to analyze the data numerically. This was done regarding the information on answers to the question 14 (Where and how often do you use the test language?) in the old material. The highest value corresponding to the greatest number 3 was marked on an answer “almost

daily", the second highest value being 2 on the answer "a couple of times a week", the number 1 corresponding the answer "once a month or less frequently" and the number 0 equaling the answer "not at all". The same procedure was done with the information on "basic education".

The background information also contained a question about a participant's socio-economic status. One column with corresponding income in euros was added in order to study the role of the socio-economic class. This was done since it was impossible to rank the numbers that the answers represented in the data, for example, the answer "unemployed" corresponded the number 8 and "entrepreneur" the number 1. Hence, the households' income by socio-economic consumption unit (IPCU) in 2010 was retrieved from the website of Statistics Finland (See Appendix 4). The year 2010 was selected since it is the year that divides the data into two parts. According to Statistics Finland, for example, a student's income per consumption unit in 2010 was 11.700 euros.

4.5.1 Test level and basic education variables

The effect of the test level on the usage of extender tags was examined by comparing the three test levels and the extender tags used by the test participants (See Table 15 for details). On account of the results, the advanced test level does not correlate to the usage of extender tags in this dataset as one might expect. The largest percentage of users who utilized extender tags, 41% (39 out of 64 used extender tags once), is found on the intermediate test level and not on the advanced test level. There are no test participants who used extender tags four- or five-times during their oral performance on the advanced test level, whereas there are 3 test participants on the intermediate level who did this. On the basic test level only 7 test participants out of 48 used extender tags which denotes that a language learner should command rather good language skills in order to be able to use extender tags. The corresponding figure from the Orange toolkit can be seen in Appendix 4.

Table 15 The distribution of test participants according to the test level and the extender frequency in oral performances

Extender frequency/	Test level		
	basic	intermediate	advanced
5	0	1	0
4	0	2	0
3	0	7	1
2	1	15	8
1	6	39	34
0	41	92	113
	48	156	156

The distribution of the participants according to the basic education and the extender usage is shown in Table 16. It is worth mentioning that only 18% of test participants with polytechnic education (7 out of 40) used extender tags in the dataset. Considering the scope of this thesis it is not possible to study all Universities of Applies Sciences in Finland. However, when regarding the internationality at Turku University of Applied Sciences, one reason may be that students of Turku University of Applied Sciences do not widely take advantage of student exchange. According to statistics of Turku University of Applied Sciences, 335-395 students out of over 10,000 students per year between 2016 and 2019 have taken part in a student or internship exchange (Härkönen 2019). The most popular target countries have been other than English speaking countries (Härkönen 2019).

Other basic education backgrounds, upper secondary (34%), vocational (32%) or university (35%) rank close to each other. The results of the Orange calculations are illustrated in the scatter plot in Appendix 4.

Table 16 The distribution of test participants according to the basic education and the frequency of extenders in oral performances

Extender frequency	Basic education				
	comprehensive	upper secondary	vocational	polytechnic	university
5	0	0	0	0	1
4	0	0	1	0	1
3	0	4	1	1	2
2	0	7	6	0	11
1	0	16	12	6	45
0	6	53	42	33	112
	6	80	62	40	172

When comparing Tables 15 and 16, it is interesting to note that the one test participant who utilized five extender tags during their speech has notified university as their basic education. This person has, however, taken the test on the intermediate level.

Next, the percental shares of the extender tag users with university education and non-university education are presented below. As shown in Figure 8, 34,9% of test participants with university education used extender tags whereas the corresponding percental share among of the test participants with non-university education is 28,7%.

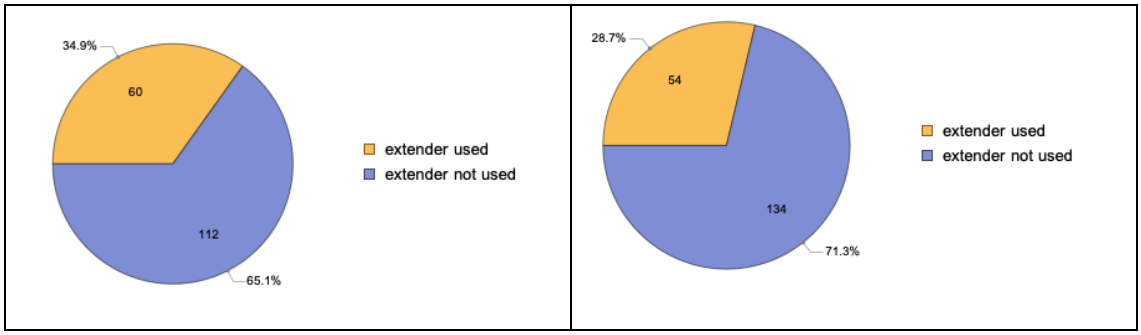


Figure 8 The university and non-university extender users

The results seem to indicate that a person with a university education is more likely to use an extender tag, according to this study. There might be multiple reasons for that. Firstly, the studying material of several fields of study is in English and secondly, many students take part in student exchange programs, where the use of the English language is essential.

4.5.2 Age factor

In addition, correlation between the use of extender tags and age was studied by choosing Pearson's correlation from the Orange toolkit. Pearson's correlation of -0,081 indicates that there was no strong relation between the use of extender tags and the age in the data. However, the negative Pearson correlation refers to a reverse effect of age. That is to say, the older a language learner is, the less likely they are to use extender tags. This applies particularly, with disjunctives. In this case Pearson's correlation of -0,127 is for long disjunctives and Pearson's correlation of -0125 for short disjunctives. The age distribution of extender tag users (Figure 9) points to the direction of inverse correlation since the highest occurrence of extender users can be found among participants between the ages of 24 and 30. The highest number of the extender tag users (18) fall into the age group of 28 to 30, comprising 41,86% of the age group in question.

Age also had a weak inverse correlation on all of the grades of different parts of the test. In other words, the older a test participant is, the lower the grades are. The Pearson correlations are: -0,281 for listening comprehension, -0,273 for reading comprehension, -0,215 for writing and -0,183 for speaking test.

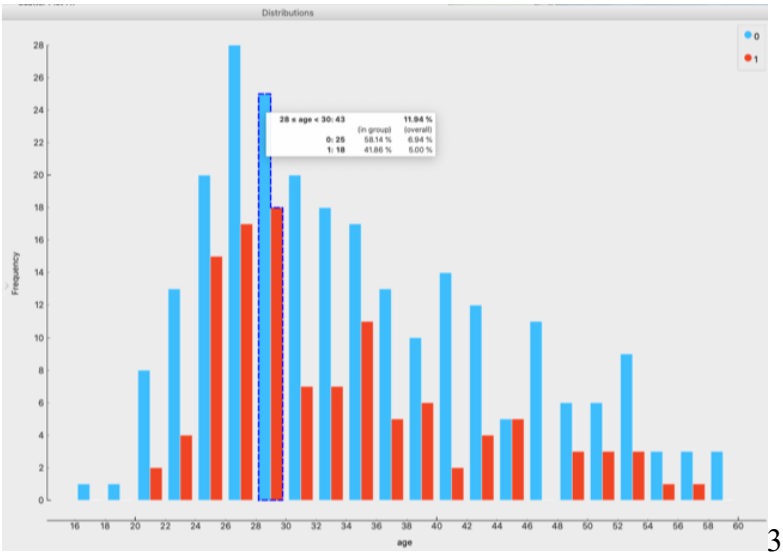


Figure 9 The age distribution of extender tag users (red) and non-users (blue)

Another question related to test participants arose, in addition to the age distribution of extender tag users. Especially, when listening to the recordings for the first time, my attention was drawn to multiple male voices talking about their international experiences in peacekeeper forces. That is why a closer examination on the gender

distribution of extender tag users and national defense as a field of employment was taken. In the old material, the test participants were given the possibility to answer this question with the alternative “national defense”, but in the new material three fields of employment were joined under one answer alternative which contained security, rescue and national defense activities. Consequently, it is impossible to know which one of these three alternatives the test participant meant in answering this question. That is why I excluded the new data in this regard and examined only the old part of the data.

4.5.3 Gender distribution and national defense variable

The gender distribution of users and non-users of extender tags is shown in Figure 10 in which blue (1) represents female participants and red (2) male in the background data. Missing information was presented by the number zero, but there were no such cases in the dataset. The results show that 63,2% consists of male participants who used extender tags. The majority of non-users of extender tags also comprises of male participants. It reflects the overall distribution of gender in the dataset since the majority of test participants, 61,4%, are men.

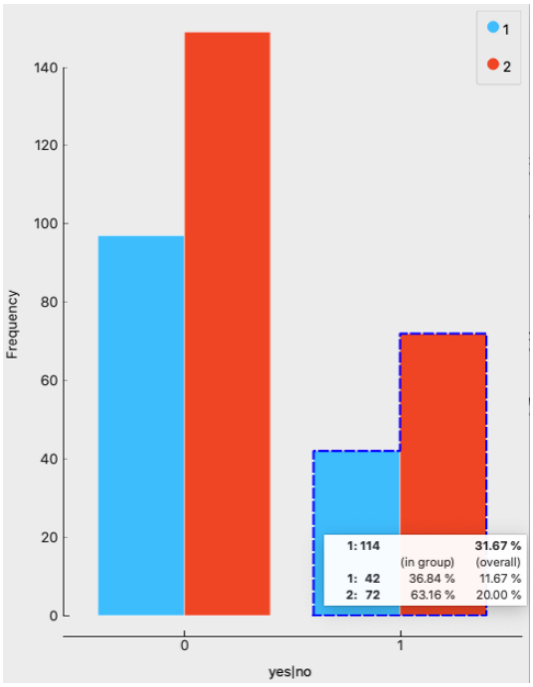


Figure 10 The gender distribution by users (1) and non-users (0) of extender tags

After specifying the gender distribution, the following calculations were made (See Table 17): the number of men in the old data by test levels, the number of men who

had entered national defense as their field of employment, the number of women with national defense activity and finally the number of male extender users with background in national defense. The results show that 34,9% (61) of men in the old data have a history in national defense and 26,2% (16) of them used extender tags in oral performances. It is a slightly larger portion when compared to the men from the whole dataset of the old part (22,9%) who used extender tags.

Table 17 The distribution of male and female extender users by test levels and national defense history

Data ≤ 2010	Men	Defense	Defense men	Defense women	Extender users defense men
Basic	19	4	2	2	0
Intermediate	102	47	47	0	14
Advanced	54	14	12	2	2
Total	175	65	61	4	16

Bearing in mind that the results originate from the data before 2011, however, the portion of men with national defense activities is substantially large. It does not solely explain the large amount of male test participants but indicates that the men working in national defense need a certificate for their English language skills.

4.5.4 Socioeconomic status variable

In order to study the role of socioeconomic status in the usage of extender tags, the frequency of extender tag users and non-users by socioeconomic income per consumption unit and Pearson's correlation were calculated by using the Orange toolkit.

The frequency of extender tag users was fairly high among test participants with high income per consumption unit and on the other hand a cluster of usage of extender tags was found when the IPCU was under 12.000 € (Figure 11). This means that on one hand test participants who earn a good salary and on the other hand students with low income seem to be extender users in this study. Students were on the lowest income level (See the first part of Section 4.5). It may be that employees in higher positions use more English, for example, during working trips abroad and students use English during the study time, for example, being in student exchange. However, it can be stated that academically educated people can also work in low-paid fields.

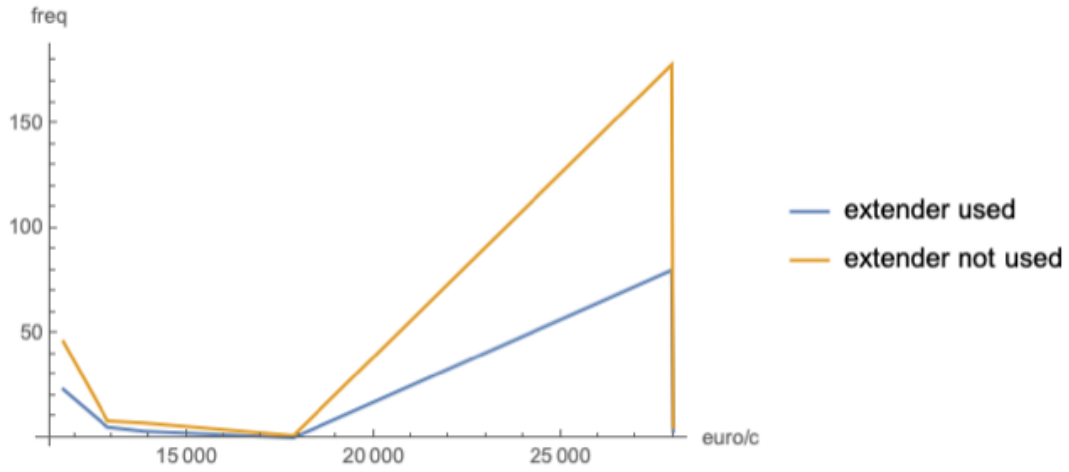


Figure 11 The frequency of extender users and non-users by socioeconomic income per consumption unit

In terms of Pearson's correlation, no correlation between socio-economic status and the usage of extender tags was found in this study. However, a negative Pearson correlation of -0,280 between the socio-economic status and how often the English language was used during studies was discovered. Normally, the majority of students are in their early 20s and often use the English language in their studies. In addition, they do not earn much money during their study time.

Positive Pearson correlations were found between socio-economic status and age and the use of English in work, 0,366 and 0,238 respectively. One explanation may be that employees in these age groups are on the peak of their working career and the working language may be English. For example, it is typical for multinational companies to have a large amount of input in English.

4.5.5 What kind of person uses extender tags?

Since it is possible for the Orange toolkit to predict what kind of person would use extender tags on the basis of this data by using logistic regression, such calculations were conducted and are seen in the workflow in the Orange toolkit in Figure 12. In fact, it presents all of the Orange calculations conducted for this study.

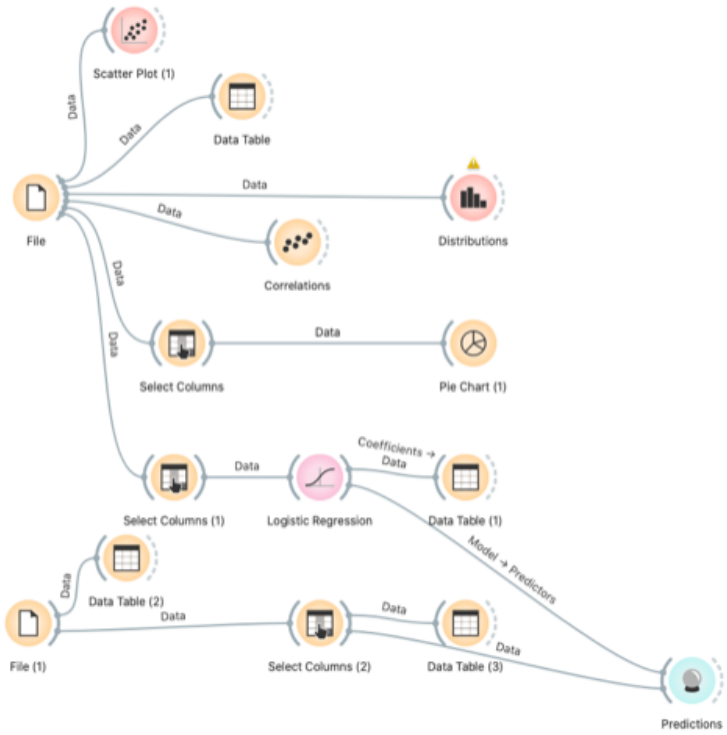


Figure 12 The workflow in the Orange toolkit used in this study

Firstly, a test dataset (File 1 in the workflow) with three imaginary test participants was invented. The first one was given features that supposedly are related to a person who uses extender tags. These features include, for example, a university education, young age, an advanced test level and good test grades. The second imaginary test person was the opposite of the extender tag user with features such as high age and a low-level education. The third person was given features that are located in between those two. Secondly, by choosing logistic regression from the Orange toolkit the prediction calculations were made. The results are seen in Figure 13.

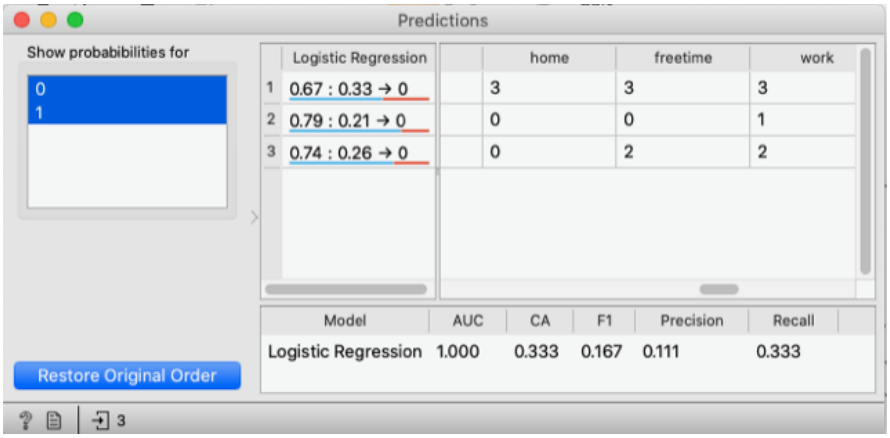


Figure 13 The prediction results from logistic regression

The following indicators are used to describe the prediction results: Classification accuracy (CA), F1, and Precision. "CA is the proportion of correctly classified examples." (Orange.biolab, n.d.). The indicator F1 (F-1) "is a weighted harmonic mean of precision and recall." Recall "is the proportion of true positives among all positive instances in the data" and precision "is the proportion of true positives among instances classified as positives." (ibid.). According to the calculations, the indicators CA, F1 and Precision (See Figure 13) are weak which can be interpreted in such a way that a person using extender tags could not be predicted. The best test person received a probability of 33% and the weakest test person was provided with the weakest probability 21% of the three. The preciseness of 0,111 shows that predictions were weak. Thus, on the base of this dataset it is not possible to define exact features of extender tag users. Regarding the causes for this, the following three main issues can explain the results:

1) a small size of dataset
2) missing background information
3) questions that would have given more detailed information on test participants.

5 Discussion

This section will summarize the findings presented in the previous section. In Section 5.1, the general discussion will follow the order of the research questions and the analysis. Some limitations of the study will be reflected in Section 5.2. In Section 5.3, I will offer some suggestions for future research.

5.1 General discussion

This study is the first one to examine the use of extender tags in oral performances of test participants of The National Certificates of Language Proficiency in Finland and to compare the results to the findings of other similar types of studies conducted outside Finland. The research questions for the present study were as follows:

1. What is the distribution of extender tags in oral performances of the YKI-Corpus?
2. Which structures of extender tags are used by Finnish EFL learners?
3. What is the possible influence of social variables such as age, education level and test level in the use of extender tags by Finnish EFL learners?

When comparing the findings of the present study to the findings of Buysse (2014) and Aijmer's (2015) studies, both similarities and disparities were found. First of all, the high occurrence of adjunctives in the YKI Corpus was not in line with the results of Buysse (2014) and Aijmer (2015). It is explained by the most frequent extender tag *and so on* that lies behind the high usage of adjunctives by the Finnish EFL learners. *And so on* was also widely used in the Dutch learner data and it came in the third place among the three most popular extenders in Buysse's (2014) study whereas the third position in the Finnish learner data was occupied by *or something*. It is also worth noting that in both the Dutch (Buysse 2014) and Swedish (Aijmer 2015) learner data the extender tag *or something (like that)* was the most popular extender tag. As discussed in the analysis, the use of *and so on* can be considered as overuse since it has been typically reported in written (Biber et al. 1999, 117) and formal speech (Overstreet 1999, 7) and occurred only twice in the native speaker corpus in Buysse's study (2014).

Another odd finding in the Finnish learner data was the use of *et cetera* and its variants. These extender tags did not occur in the learner corpora in Buysse's (2014) and Aijmer's (2015) research. The high occurrence of *and so on* and *et cetera* can be explained by crosslinguistic influence which was addressed in relation to learner language in Section 2.1.6.

Similarly, as learner language develops gradually, so does a good command of native-like phrases develop progressively. This also applies to the usage of extender tags. When using extender tags, a language learner creates new variants that deviate from the basic structure of extender tags. This was also visible when comparing the variants of extender tags used by the Finnish language learners to those in the learner data of Buysse (2014) and Aijmer (2015). To give an illustration of these variants (Table 7), an example of them is given here: *and healthy stuffs like that*.

When I examined the use of *and stuff/things (like that)* more closely, the Finnish language learners' preference for the long variants of *and stuff/things* was revealed. The same phenomenon was reported by Buysse (2014), but not by Aijmer (2015). It is contrary to what has been reported for native speakers (Aijmer 2015) who prefer to use short variants not only regarding *and stuff/things* but all extenders.

Another surprising finding was that the use of *and stuff (like that)* was not highly common in the Finnish learner corpus as might have been expected on the basis of arguments by Buysse (2014). According to him, the vast supply of American programs may be the reason for the high occurrence of *and stuff (like that)*. It can be stated that the supply of spoken American English is nearly equal in both countries, Finland and Belgium, nonetheless the Finnish EFL learners used *and stuff (like that)* substantially less than the Dutch and Swedish EFL learners.

In addition to the structure of extender tags, the referents of the extender tag *and stuff/things (like that)* and the collocates that occur with disjunctives were also studied. The results of referent analysis revealed that there are such incompatibilities as, for example, the head noun *stuff* referring back to a countable noun. This is also reported by Aijmer (2015). The collocation analysis showed that the minority of test participants used less than half of the collocations listed by Aijmer. The collocations that have not been used express major hesitation such as *I can't remember*. The test situation itself may be the reason for a minor use of these collocations. Test participants are given only two minutes to speak about the topics and they should present their best language skills within that time in order to obtain good grades in the test.

Since *and so on* is the most frequent extender tag in the dataset, I examined its use from the list construction point of view. The results of the analysis support Jefferson's (1990) proposition according to which list are mostly constructed as three-part units. Next, the results of the Orange analysis are discussed.

The third research question was aimed at studying the possible effect of social variables on the usage of extender tags. It was made possible by the background information that had been collected from the test participants. In this connection, one aim of this thesis was to test how applicable the Orange toolkit is in this type of research. The results retrieved from the Orange toolkit are possible to be easily presented in distinct figures, for example, as a split pie or a scattered plot. Thus, specific features can draw the researcher's attention which then leads to studying the data more thoroughly. When it comes to correlations, the correlation widget calculates all correlations between all variables in the data. It can be concluded that the Orange toolkit is highly recommendable for the end user who commands no programming skills.

The results in terms of educational background indicated that test participants with university education were likely to use extender tags. The test level did not correlate on the use of extender tags as expected since the majority of extender tag users was found on the intermediate level.

When studying the age factor, no strong correlation between the use of extender tags and age was found. However, the age distribution of extender tag users indicated that the use of extender tags was most common among younger test participants between the ages of 24 and 30.

Pearson's correlation showed no correlation between socio-economic status and the use of extender tags in this dataset. Despite this fact, it is interesting to note that on one hand the extender tag users are found among students and on the other hand among test participants with high income per consumption unit.

Finally, the Orange toolkit was used to predict a possible extender tag user by using logistic regression. This proved difficult and such a person could not be predicted. As for the causes, there are several reasons which can result in failure to predict and are discussed in the following subsection. The present study showed that there are many variables that are affecting the usage of the extender tags. Thus, it is difficult to name a single dominant factor.

When it comes to pedagogical implications of the study, it can be stated that the use of extender tags should be taken into consideration in the teaching of language.

Teachers need to increase their students' awareness for this linguistic phenomenon. Students should learn to recognize different variants of extender tags and gradually also learn how to use them or, as the case may be, to limit their use in more official speaking situations. As the findings of Metsä-Ketelä (2016) showed, the use of extender tags helped spoken interaction in international academic settings. It seems to me that it would facilitate communication in discourse in general, not just in academic settings. Particularly, if Finnish EFL learners learned to use extender tags to fill in silent pauses in speech, for example, when they are thinking what to say next. It would make them feel more fluent in speaking English.

5.2 Limitations of the present study

In this subsection I will discuss some of the limitations that might have affected the results of the present study. Firstly, the transcriptions were not provided by the YKI Corpus which is the reason why the size of corpus in terms of word count was not calculated. It would be of high importance if generalizations and statistical significance are to be determined. Secondly, the questions and the numbering of answers for the background information of the participants on the new material were somewhat different compared to the old material (See Appendix 3). This caused a considerable amount of challenges and additional work when modifying the files for the Orange analysis. In addition, some of the questions were too broad in nature. In particular, when asking about basic education, or where a test participant had studied English. For example, answering the latter question with "abroad" is too unspecific. More detailed answers in terms of time spent studying, the length of work or of the stay abroad would have given valuable information that would have allowed the relation between time spent abroad and the use of extender tags to be explored.

In addition, there was insufficient data on the background, up to a lack of 11% for some parts, which affected, for example, the preciseness of the predictions performed by logistic regression. Furthermore, the dataset from the basic level would have been larger with additional background data properly collected.

It should be also noted that the two-minute time limit puts pressure on test participants and the situation does therefore not correspond to natural discourse. This type of oral performance does not serve as the ideal research setting, particularly, when the

function aspects of extender tags are examined. The following section will be dedicated to some suggestions for further study.

5.3 Call for future research

Extender tags used particularly by Finnish EFL learners is an object of study that still needs to be explored. Many interesting ideas for future research surfaced during the analysis. Since extender tags are characterized by multifunctionality, their functional role in conversation would be an excellent object of study. That type of study would benefit from the method of self-recording in order to collect naturally occurring spoken data. The method has been applied in sociolinguistics in recent years, for example, in collecting spoken data for the British National Corpus (Andersen 2010, 549).

In addition to spoken data, a questionnaire or an interview of participants should be included in the study. This would make it possible to collect background information in more detail and to investigate the possible effects of social variables on the usage of extender tags. Moreover, it would be important to gather detailed background information on the participants' studies at university and the time spent abroad, as discussed above. For example, from which faculty does a participant have a degree from and how many student credits a participant has completed. When examining the correlations between distinct variables, the study should include comparisons between the results of Pearson's and Spearman's correlation tests.

Furthermore, the role of the teaching materials, textbooks, listening comprehension and workbooks in the English language should be investigated. It would be particularly interesting to examine how often extender tags appear in the teaching material and how much teachers pay attention to extender tags while teaching.

6 Conclusion

The main goal of the study was to investigate how native speakers of Finnish use extender tags in oral performances. I selected the YKI Corpus as the source of the data firstly, because it includes the English-speaking tests of The National Certificates of Language Proficiency in Finland and secondly, since no studies on extender tags in terms of the YKI Corpus have previously been made.

My aim was to investigate the distribution of extender tags used in the Finnish EFL learner corpus and compare it to the recent studies conducted in the field. In particular, studies from Buysse (2014) and Aijmer (2015) were selected as material for this comparison since they had a learner language perspective. The distribution of extender tags revealed the overuse of the extender tag *and so on* which was not totally unique since it was found in the Dutch learner data by Buysse (2014) as well. The high occurrence of *and so on* in the dataset increased the share of adjunctives over the proportion of disjunctives. According to the frequency of occurrence, the top three extender tags are *and so on*, *or something like that* and *or something*.

My examination of the structures of the extender tags used in the Finnish learner data confirmed that there are deviations that are characteristic to learner language. This is illustrated with an example from the data, the extender tag *and healthy stuffs like that*. When studying this tag and the context of the extender tags *and stuff/things (like that),* more closely, other incompatibilities were also found. For example, the head noun *stuff* was used to refer back to referents that are countable nouns whereupon there develops a contradiction in properties between the head noun *stuff* (uncountable) and its referent (countable). These findings are similar to the results of previous studies on extender tags (Cheshire 2007, Tagliamonte & Denis 2010). The results of the analysis of *and so on* as part of list construction corroborated Jefferson's (1990) proposition, according to which speakers prefer to construct lists as three-parts.

My final aim was to investigate the use of extender tags by the Finnish EFL learners from a sociolinguistic point of view and to find out what effect different social variables have on the use of extender tags. In addition, the applicability of the Orange toolkit was tested. Based on the analysis no single factor can be named. However, the study indicates that a higher level of education and youth promote the usage of extender tags. As for the Orange toolkit, it is applicable and highly recommendable as a tool for similar studies since it does not require any programming skills whatsoever.

To sum up the foregoing, it may be concluded that the results of the present study can be used as a starting point for more extensive research on extender tags used by the Finnish EFL learners. As for language learners, a good command of extender tags would provide them with another tool for their language production.

7 References

Primary sources

YKI-Corpus. 2019. Accessed 1 August 2019. http://yki-korpus.jyu.fi.

Secondary sources

Aijmer, Karin. 2015. “General extenders in learner language.” In *Corpora, Grammar and Discourse. In honour of Susan Hunston*, edited by Groom, Nicholas, Maggie Charles and Suganthi John, 211-233. Accessed 21 August 2019. Research Gate.

Andersen, Gisle. 2010. “How to use corpus linguistics in sociolinguistics.” In *The Routledge Handbook of Corpus Linguistics*, edited by Michael McCarthy and Anne O'Keeffe, 547-562. Accessed August 23, 2020. ProQuest Ebook Central.

Arnon, Inbal, and Morten H. Christiansen. 2017. “The Role of Multiword Building Blocks in Explaining L1–L2 Differences.” *Topics in Cognitive Science* 9: 621–636. Accessed 1 July 2020. Wiley Online Library.

Avramidis, Eleftherios. 2012. “Quality Estimation for Machine Translation output using linguistic analysis and decoding features.” *Proceedings of the 7th Workshop on Statistical Machine Translation*, 84–90. Accessed 19 May 2020. https://www.aclweb.org/anthology/W12-3108.pdf.

Ball, Catherine N., and Mira Ariel. 1978. “Or something, Etc.” *Penn Review of Linguistics* 3: 35-45.

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

Bertrand, Roxane, and Beatrice Priego-Valverde. 2017. “Listing Practice in French Conversation: From Collaborative Achievement to Interactional Convergence.” *Discours. A journal of linguistics, psycholinguistics and computational linguistics*, 20: 4-33. Accessed 20 July 2020. Openedition Journals.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3: 993-1022. Accessed 26 May 2020. http://www.jmlr.org/papers/v3/blei03a.

Brown, Penelope, and Stephen C., Levinson, 1987. *Politeness. Some Universals in* Language *Usage*. Cambridge: Cambridge University Press.

Brunni, Sisko, and Jarmo H. Jantunen. 2015. “Learner Language.” In *The Routledge Handbook of Educational Linguistics*, edited by Martha Bigelow and Johanna Ennser-Kananen. Accessed 11 May 2020. EBSCOhost.

Buysse, Lieven. 2014. “We went to the restroom or something. General extenders and stuff in the speech of Dutch lerners of English.” In *The Yearbook of Corpus Linguistics and Pragmatics 2014: New Emprical and Theoretical Paradigms*, edited by J. Romero-Trillo, 213-237. Accessed 21 August 2019. Springer Link.

Carroll, Ruth. 2008. Historical English Phraseology and the Extender Tag. *SELIM. Journal of the Spanish Society for Medieval English Language and Literature* 15: 7-37. Accessed 19 October 2019. https://www.unioviedo.es/reunido/index.php/SELIM/article/view/13404.

Channell, Joanna. 1994. *Vague Language*. Oxford: Oxford University Press.

Cheshire, Jenny. 2007. “Discourse variation, grammaticalisation and stuff like that.” *Journal of Sociolinguistics* 11, no. 2: 155-93. Accessed 8 March 2019. Wiley Online Library.

Cook, Vivian. 2007. “Multi-Competence: Black Hole or Wormhole for Second Language Acquisition Research?” in *Understanding Second Language Process*, edited by Han, ZhaoHong, 16-26. Accessed 16 September 2020. De Gruyter.

Cortés Rodriguez, Luis. 2006. LOS ELEMENTOS DE FINAL DE SERIE ENUMERATIVA DEL TIPO Y *TODO ESO, O COSAS ASÍ, Y TAL*, ETC. PERSPECTIVA INTERACTIVA. [Enumerative listing sequences with a closing marker of the type *and all that*, *things like that*, and *and so on*, among others. Interactive perspective.] *Boletín de Lingüística* 18:102-129. Accessed 29 October 2019. https://www.redalyc.org/articulo.oa?id=34702604.

Crystal, David, and Derek Davy. 1975. *Advanced Conversational English*. London: Longman.

Cucchi, Costanza. 2007. “An Investigation of General Extenders in a Corpus of EU Parliamentary Debates.” *Proceeding of The*. Accessed 13 October 2019. https://www.academia.edu/.

Dines, Elizabeth R. 1980. ”Variaton in discourse–“and stuff like that”.” *Language in Society* 9: 13-33. Accessed 31 October 2019. JSTOR.

Dubois, Sylvie. 1993. “Extension particles, etc.” *Language Variation and Change* 4: 179-203. Accessed 28 October 2019. Cambridge University Press Journals Digital.

Ellis, Nick C., Ute Römer, and Matthew Brook O’Donell. 2016. “Constructions and Usage-based Approaches to Language Acquisition.” *Language Learning. A Journal of Research in Language Studies* 66, Suppl. 1: 23-44. Accessed 1 June 2020. Wiley Online Library.

Farrús, Mireia, Marta Ruiz, Maja Popovic, and Carlos Henríquez. 2012. “Study and correlation analysis of linguistic, perceptual and automatic machine translation evaluations.” *Journal of the Association for Information Science and Technology* 63, no. 1: 174-184. Accessed 22 September 2020. Wiley Online Library.

Fernandez, Julieta, and Aziz Yuldashev. 2011. “Variation in the use of general extenders and stuff in instant messaging interactions.” *Journal of Pragmatics* 43: 2610-2626. Accessed 21 August 2019. Elsevier.

Fraser, Bruce. 1996. “Pragmatic markers.” *Pragmatics* 6, no. 2: 167-190. Accessed 18 May 2020. John Benjamins Journals.

Halliday, M. A. K., et al. 2003. *On Language and Linguistics: Volume 3*. Accessed 14 May 2020. ProQuest Ebook Central.

Härkönen, Anu. 2019. *Turun AMK:n kansainvälisen toiminnan tilannekatsaus*. PowerPoint. Turku University of Applied Sciences.

Immonen, Pinja. 2020. “The Ashamed Selves: Autobiographical Study of Finnish University Students’ Shame of Speaking English.” MA Thesis, University of Turku. Accessed 27 May 2020. http://urn.fi/URN:NBN:fi-fe202003047298.

Jefferson, Gail. [1973] 1990. “List-construction as a task and resource.” In *Interaction Competence*, edited by George Psathas, 63-92. Lanham, MD: University Press of America and The International Institute for Ethnomethodology and Conversation Analysis. Accessed 29 September 2020. http://liso-archives.liso.ucsb.edu/Jefferson/List%20construction.pdf.

Lakoff, Robin. 1973. ”Language and Woman's Place.” *Language in Society 2*, no.1: 45-80. Accessed 28 August 2020. JSTOR.

Leech, Geoffrey N. 1980. *Explorations in Semantics and Pragmatics*. Amsterdam: John Benjamins Publishing Company. Accessed 23 May 2020. ProQuest Ebook Central.

Leech, Geoffrey. 1983. *Principles of Pragmatics*. London and New York. Routledge.

Lehtonen, Jaakko. 1979. "Speech Rate and Pauses in the English of Finns, Swedish-speaking Finns, and Swedes". In *The Kuopio-Vaasa-Stockholm Tests. Perception and Production of English: Papers on Interlanguage*, edited by Rolf Palmberg, Håkan Ringbom, and Jaakko Lehtonen, 35-52. Turku: Åbo Akademi. Accessed 18 September 2020. https://eric.ed.gov/?id=ED177880.

Lerner, Gene H. 1994. "Responsive List Construction: A Conversational Resource for Accomplishing Multifaceted Social Action." *Journal of Language and Social Psychology* 13, no. 1: 20-33. Accessed 27 October 2019. https://utuvolter.fi/permalink/358FIN_UTUR/1rsgc7g/cdi_crossref_primary_10_1177_0261927X94131002.

Macaulay, Ronald. 1985. The Narrative Skills of a Scottish Coal Miner. In *Varieties of English Around the World: Focus on Scotland*, edited by Manfred Görlach, 101-124. Accessed 28 October 2019. ProQuest Ebook Central.

Martinez, Ignacio M. 2011. ""I might, I might go I mean it depends on money things and stuff". A preliminary analysis of general extenders in British teenagers' discourse." *Journal of Pragmatics* 43, no. 9: 2452-470. Accessed 8 March 2019. Elsevier.

Mauranen, Anna. 2009. "Chunking in ELF: Expressions for managing interaction." *Intercultural Pragmatics* 6-2: 217–233. Accessed 1 July 2020. EBSCOhost.

Mauranen, Anna. 2011. "Learners and users – Who do we want corpus data from?" In *Taste for Corpora: In honour of Sylviane Granger*, edited by Fanny Meunier, et al., 155-171. Accessed 1 July 2020. ProQuest.

Metsä-Ketelä, Maria. 2016. "Pragmatic Vagueness: Exploring General Extenders in English as a Lingua Franca." *Intercultural Pragmatics* 13, no. 3: 325–51. Accessed 21 August 2019. EBSCOhost search.

Meyerhoff, Miriam. 1992. "'A Sort of Something'-Hedging Strategies on Nouns." In *Working Papers on Language, Gender and Sexism*, edited by Joanne Winter, 59-73. Accessed 28 October 2019. Research Gate.

Morgan-Lopez, Antonio A., Kim Annice E., Robert F. Chew, and Paul Ruddle. 2017. "Predicting age groups of Twitter users based on language and metadata features." *PLOS ONE* 12, no. 8: 1-12. Accessed 24 August 2020. https://doi.org/10.1371/journal.pone.0183537.

Nguyen, Doug, Noah A. Smith, and Carolyn P. Rosé. 2011. "Author Age Prediction from Text using Linear Regression." *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 115-123. Accessed 24 August 2020. https://www.aclweb.org/anthology/W11-1515.pdf.

O'Keeffe, Anne. 2006. *Investigating Media Discourse*. London: Routledge.

Orange. n.d. "Topic modelling of recent tweets." *Orange biolab Screenshots*. Accessed 31 May 2020. https://orange.biolab.si.

———. n.d. "Test and Score." *Orange biolab Widget catalog*. Accessed 30 September 2020. https://orange.biolab.si/widget-catalog/evaluate/testandscore/.

Overstreet, Maryann. 1999. *Whales, Candlelight, and Stuff Like That: General Extenders in English Discourse*. New York: Oxford University Press.

Overstreet, Maryann. 2005. "And stuff und so: Investigating pragmatic expressions in English and German." *Journal of Pragmatics* 37: 1845–1864. Accessed 10 May 2020. Elsevier.

Overstreet, Maryann. 2014. “The role of pragmatic function in the grammaticalization of English general extenders.” Pragmatics 24, no. 1: 105-29. Accessed 3 March 2019. John Benjamins Journals.

Oxford English Dictionary. 2020. s.v. “thing” n.1. Accessed 21 May 2020.

Parvaresh, Vahid, Manoochehr Tavangar, Abbas Eslami Rasekh, and Dariush Izadi. 2011. “About his friend, how good she is, and this and that: General extenders in native Persian and non-native English discourse.” *Journal of Pragmatics* 44, no. 3: 261-279. Accessed 30 July 2019. Elsevier.

Pichler, Heike, and Stephen Levey. 2011. “In search of grammaticalization in synchronic dialect data: general extenders in northeast England.” *English Language and Linguistics* 15, no. 3: 441-71. Accessed 21 August 2019. ProQuest Central.

Pollak, Senja, Nejc Trdin, Anze Vavpetic, and Tomaz Erjavec. 2012. “NLP Web Services for Slovene and English: Morphosyntactic Tagging, Lemmatisation and Definition Extraction.” *Informatica* 36, no. 4: 441-449. Accessed 19 May 2020. ProQuest.

Sayer, Peter. 2008. “Interlanguage.” In *Encyclopedia of bilingual education*, edited by Josué M. González, 405-407. Accessed 16 September 2020. Sage Reference.

Searle, John. [1965] 2004. “What is a Speech Act?” In *Philosophy in America* edited by Max Black, 221-239. Accessed 30 August 2020. ProQuest.

StatisticsFinland. 2010. “Households’ income by Information, Person’s socioeconomic group and Year.” *StatisticsFinland*. Accessed 16 July 2020. http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/StatFin__tul__tjt__kotitalouksien/statfin_tjt_pxt_12ci.px/table/tableViewLayout1/.

Statisticssolutions. n.d. “What is logistic regression?” *Statisticssolutions*. Accessed 16 August 2020. https://www.statisticssolutions.com/what-is-logistic-regression/.

Tagliamonte, Sali A., and Derek Denis. 2010. “The Stuff of Change: General Extenders in Toronto, Canada.” *Journal of English Linguistics* 38, no. 4: 335-68. Accessed 21 August 2019. Sage Premier 2020.

Terraschke, Agnes. 2010. “Or so, oder so, and stuff like that— General extenders in New Zealand English, German and in learner language.” *Intercultural Pragmatics* 7-3: 449–469. Accessed 21 August 2019. EBSCOhost.

Vetschinnikova, Svetlana. 2019. *Phraseology and the Advance Language Learner*. Cambridge University Press. Adobe Digital Edition.

Wagner, Suzanne Evans, Ashley Hesson, Kali Bybel, and Heidi Little. 2015. “Quantifying the Referential Function of General Extenders in North American English.” *Language in Society* 44, no. 5, 11: 705-731. ProQuest Central.

Ward, Gregory, and Betty Birner. 1993. “The semantics and pragmatics of and everything.” *Journal of Pragmatics* 19, no. 3: 205-214. Accessed 28 October 2019. ScienceDirect.

Watanabe, Tomoko. 2014. “A corpus-based study of the use of English general extenders spoken by Japanese users of English across speaking proficiency levels and task types.” Doctoral Thesis, Edinburgh University. Accessed 29 October 2019. Semantic Scholar.

Winter, Joanne, and Catrin Norrby. 1999. “Set Marking Tags – ‘and stuff’.” *Prodeedings of the 1999 Conference of the Australian Linguistic Society,* 1-8. Accessed 29 October 2019. Researchgate.

Wray, Alison. 2012. “What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play.” *Annual Review of Applied Linguistics* 32: 231–254. Accessed 28 June 2020. Cambridge University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Zorčič, Sabina. 2019. “Linguistic habitus, different capitals and the identity of bilingual youth: the case of Austrian Carinthia.” *Journal of multilingual and multicultural development* 40, no. 9: 787-800. Accessed 7 June 2020. Taylor&Francis Online.

Appendix 1: Transcription conventions

. . . Empty pause: . (short), . . (medium), . . . (long)

= Truncation

(x) unintelligible word

Appendix 2: Speaking test, Intermediate level, Fall 2008

A language laboratory part

YLEISET KIELITUTKINNOT

ENGLANTI
KESKITASO
SYKSY 2008

STUDIO-OSA

Nimi:

Syntymäaika:

Tutkinnon järjestäjä:

Studio-osan tehtävät:

Puheen ymmärtäminen
- Tehtävä 1. Lyhyitä puhetilanteita
- Tehtävä 2. Haastatteluja
- Tehtävä 3. Radiotilanteita
- Tehtävä 4. Uutisia
- Tehtävä 5. Puheenvuoroja

Puhuminen
- Tehtävä 6. Kertominen
- Tehtävä 7. Keskusteluja
- Tehtävä 8. Tilanteita
- Tehtävä 9. Puheenvuoro

Optisen lomakkeen täyttäminen: 10 minuuttia

Puhuminen

Tehtävät 6–9

Tehtävä 6. Kertominen

TAPAHTUMA TAI KOKEMUS, JOTA EN VOI UNOHTAA.
Kerro englanniksi esimerkiksi:

- Mitä tapahtui?
- Missä tapahtui?
- Milloin tapahtui?
- Miksi tapahtui?
- Mitä tapahtumasta seurasi?
- Miksi tapahtuma on sinulle tärkeä?

Sinulla on 1 minuutti aikaa miettiä, mitä sanot. Voit tehdä valmisteluaikana muistiinpanoja, jos haluat. Aloita puhuminen vasta sitten, kun kuulet nauhalta: ”*Start speaking now, please.”* Sinulla on 1½ minuuttia aikaa puhua. Yritä puhua koko ajan!

Valmisteluaika: 1 minuutti
Puhumisaika: 1½ minuuttia

Muistiinpanoja:

Tähän päättyy tehtävä 6.

Tehtävä 7. Keskusteluja

Alla on kuvattu kaksi keskustelutilannetta. Ennen kutakin tilannetta sinulla on aikaa tutustua siihen. Näet sulkeissa vihjeen siitä, mitä sinun pitää sanoa **englanniksi**. Kuulet keskustelun toisen osapuolen puheenvuorot nauhalta, ja ne on merkitty tehtävävihkoon tähdillä (esim. Puhuja: *****). Vastaa puheenvuoroihin mahdollisimman luontevasti **englanniksi**. Näet puhumisajan sulkeissa.

Tilanne 1. ***Aiot vierailla Yhdysvalloissa, mutta tarvitset lisätietoja matkustusasiakirjoista. Soitat USA:n suurlähetystöön.***

Telephone Operator: *****

SINÄ: [Kerro tarvitsevasi tietoa matkustamisesta Yhdysvaltoihin.] *(15 sek)*

Telephone Operator: *****

Consulate: *****

SINÄ: [Kysy tarvittavista asiakirjoista.] *(15 sek)*

Consulate: *****

SINÄ: [Vastaa.] *(15 sek)*

Consulate: *****

SINÄ: [Vastaa.] *(15 sek)*

Consulate: *****

SINÄ: [Vastaa.] *(15 sek)*

Consulate: *****

SINÄ: [Vastaa.] *(15 sek)*

Consulate: *****

SINÄ: [Vastaa kielteisesti ja hyvästele lopuksi.] *(15 sek)*

⋆ ⋆ ⋆

Tilanne 2. ***Aiot järjestää syntymäpäiväjuhlat. Soitat pitopalveluun ja tilaat ruokatarjoilun juhliisi.***

Catering: *****

SINÄ: [Kerro juhlista ja niiden aikataulusta. Keksi ajankohta] *(15 sek)*

Catering: *****

SINÄ: [Vastaa. Keksi itse] *(10 sek)*

Catering: *****

SINÄ: [Vastaa. Keksi itse.] *(15 sek)*

Catering: *****

SINÄ: [Kysy ruoka-allergioiden tai rajoitusten hoitumisesta.] *(20 sek)*

Catering: *****

SINÄ: [Vastaa. Keksi itse.] *(15 sek)*

Catering: *****

SINÄ: [Tiedustele vielä hintaa.] *(10 sek)*

Catering: *****

SINÄ: [Vastaa. Lopeta puhelu kohteliaasti] *(15 sek)*

Catering: *****

⋆ ⋆ ⋆

Tähän päättyy tehtävä 7.

Tehtävä 8. Tilanteita

Alla on kuvattu kuusi tilannetta, joissa voisit joutua puhumaan englantia. Tilanteet käydään läpi yksi kerrallaan.

Jokaisessa tilanteessa:

- Kuulet nauhalta tilanteen numeron, esimerkiksi *"Look at situation one."*
- Tutustu kyseiseen tilanteeseen.
- Kuulet nauhalta kehotuksen *"Start speaking now, please".*
- Tämän jälkeen nauhalla on tauko.
- Tauon aikana sanot englanniksi sen, mitä sinua pyydetään sanomaan kuvatussa tilanteessa.

Näet puhumisajan sulkeissa.

Mitä sanoisit **englanniksi** seuraavissa tilanteissa:

Englanninkielinen tuttava haluaa tietää suomalaisista asioista:

Tilanne 1. Kerrot lyhyesti jotain suomalaisista kirjastoista. *(20 sek)*

Tilanne 2. Kerrot myös, mitä suomalaiset yleensä harrastavat syksyllä ja miksi.
(20 sek)

Tilanne 3. Kerrot vielä jostakin asiasta, joka mielestäsi on hyvin suomalaista ja kerro miksi. *(20 sek)*

Työtä hakemassa:

Tilanne 4. Kerro, miksi juuri sinut pitäisi valita työhön. Keksi itse työ. *(30 sek)*

Tilanne 5. Kerro jotain aikaisemmasta työkokemuksestasi. *(30 sek)*

Tilanne 6. Kerro yhdestä asiasta, jossa voisit kehittyä. *(20 sek)*

Tähän päättyy tehtävä 8.

Tehtävä 9. Puheenvuoro

Alla on kaksi aihevaihtoehtoa. Valitse **yksi** aiheista ja **esitä siitä mielipiteesi perusteluineen englanniksi**. Yritä puhua mahdollisimman luontevasti.

Sinulla on kaksi minuuttia aikaa miettiä, mitä sanot. Voit tehdä valmisteluaikana muistiinpanoja, jos haluat. Aloita puhuminen vasta sitten, kun kuulet nauhalta: "*Start speaking now, please"*. Sinulla on kaksi minuuttia aikaa puhua. Yritä puhua koko ajan!

1. Car pooling needs to be promoted!

- Kannatatko kimppa-autoilua? Miksi?
- Millaisissa tilanteissa voisit käyttää kimppa-autoilua?
- Millaisista säännöistä pitäisi kimppa-autoilussa sopia?
- Ketkä voisivat hyötyä kimppa-autoilusta?
- Mitä hyötyä / haittaa kimppa-autoilusta voisi olla kuluttajille?
- Mitä hyötyä / haittaa siitä voisi olla yhteiskunnalle?

2. Would strong marriages be better for society?

- Kuuluuko avioliitto olennaisesti nyky-yhteiskuntaan? Jos kyllä, miten?
- Mitä avioerot aiheuttavat yhteiskunnalle? Yksityisille ihmisille?
- Millaiset avioliitot ovat yhteiskunnalle hyväksi? Millaiset haitaksi?
- Kuka voisi hyötyä pysyvistä avioliitoista?
- Sääteleekö yhteiskunta avioliittoja liikaa / liian vähän?
- Millaista säätelyä tarvitaan?
- Millainen liitto sopii nyky-yhteiskuntaan / tulevaisuuteen?

Valmisteluaika: 2 minuuttia
Puhumisaika: 2 minuuttia

Muistiinpanoja:

Tähän päättyy studio-osa. Kiitos osallistumisesta!

Nyt sinulla on 10 minuuttia aikaa rastittaa (☒) puheen ymmärtämisen tehtävien vastaukset **(M63–M82)** optiseen lomakkeeseen.

Appendix 3: Variable lists 1-2

Search older material (-2010) - List of variables

Variable list of the National Certificates of Language Proficiency Corpus

Variable	Meaning
ID	*Participant's ID number*
language	*test language* 1 = English, 2 = Spanish, 3 = Italian, 4 = French, 5 = Swedish, 6 = Lappish (Sami), 7 = German, 8 = Finnish, 9 = Russian
level	*test level* 1 = basic level 2 = intermediate level 3 = advanced level
gpl	*general proficiency level*
	basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* / intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* / advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*
li	*listening*
	basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* / intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* / advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*
sp	*speaking*
	basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* / intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* / advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*
re	*reading*
	basic: 0 = <1, 1 = 1, 2 = 2 / intermediate: 2 = <3, 3 = 3, 4 = 4 / advanced: 4 = <5, 5 = 5, 6 = 6

	9 = cba*		9 = cba*		9 = cba*

wr *writing*

basic	0 = <1	intermediate	2 = <3	advanced	4 = <5
	1 = 1		3 = 3		5 = 5
	2 = 2		4 = 4		6 = 6
	9 = cba*		9 = cba*		9 = cba*

sv *structures and vocabulary*

basic	0 = <1	intermediate	2 = <3	advanced	4 = <5
	1 = 1		3 = 3		5 = 5
	2 = 2		4 = 4		6 = 6
	9 = cba*		9 = cba*		9 = cba*

*cba = can't be assessed

age *age in years*

gender *gender*

1 = female
2 = male
0 = missing information

ml *mother language*

1	=	Finnish	11	=	Kurdish
2	=	Swedish	12	=	French
3	=	Russian	13	=	Lappish (Sami)
4	=	Albanian	14	=	German
5	=	Arabian	15	=	Somali
6	=	Dari	16	=	Turkish
7	=	English	17	=	Estonian
8	=	Spanish	18	=	Other
9	=	Italian	0	=	missing
10	=	Chinese			information

be *basic education*

1 = comprehensive school / basic education
2 = upper secondary school
3 = vocational school
4 = polytechnic

5 = university
6 = other
0 = missing information

socec *socio-economic status*

1 = entrepreneur
2 = upper-level salaried employee
3 = lower-level salaried employee
4 = manual worker
5 = student
6 = student in labor market training
7 = retired
8 = unemployed
9 = other
0 = missing information

foe *field of employment*

1 = agriculture and forestry
2 = manufacturing
3 = construction
4 = retail trade, accommodation and food service activities
5 = transport
6 = financial, insurance and business activities
7 = national defense
8 = public and other services
9 = other
10 = missing information

Question 13: Where and for how long have you studied the test language?

0 = missing information, 1= less than one year, 2 = 1-2 years, 3 = 3-5 years, 4 = 6-9 years, 5 = 10 years or more, 6 = not at all

kys13_1	comprehensive school / basic education
kys13_2	upper secondary school
kys13_3	vocational school
kys13_4	polytechnic
kys13_5	university

kys13_6	folk high school / adult education center
kys13_7	self-study
kys13_8	courses organized by employer
kys13_9	labor market training
kys13_10	abroad
kys13_11	other

Question 14: Where and how often do you use the test language?

0 = missing information, 1 = not at all, 2 = almost daily, 3 = a couple of times a week, 4 = once a month or less frequently

kys14_1	home
kys14_2	work
kys14_3	leisure
kys14_4	study

Question 15: Where did you get information about the National Certificates from?

0 = no, 1 = yes

kys15_1	my employer
kys15_2	teacher
kys15_3	friend
kys15_4	internet
kys15_5	my school
kys15_6	institution organizing the test
kys15_7	employment office
kys15_8	public notice board
kys15_9	test brochure
kys15_10	ad in a paper
kys15_11	other

Question 16: How did you choose the test level?

0 = no, 1 = yes

kys16_1	my employer required it
kys16_2	recommended by my teacher
kys16_3	based on my own assessment
kys16_4	with the help of a course organizer

kys16_5	it is part of a language course
kys16_6	based on a test brochure
kys16_7	based on test level descriptions
kys16_8	cannot say
kys16_9	other

Question 17: For what purpose will you use the certificate?

0 = no, 1 = yes

kys17_1	to apply for a job
kys17_2	for study purposes
kys17_3	to get feedback on my language skills
kys17_4	to prove my language skills to my employer
kys17_5	to be given duties that require proficiency in the test language
kys17_6	to get a language bonus in my salary
kys17_7	to apply for Finnish citizenship
kys17_8	other

Search new material (2011-) - List of variables

Variable list of the National Certificates of Language Proficiency Corpus

Variable	**Meaning**
ID	*Participant's ID number*
language	*test language* 1 = English, 2 = Spanish, 3 = Italian, 4 = French, 5 = Swedish, 6 = Lappish (Sami), 7 = German, 8 = Finnish, 9 = Russian
level	*test level* 1 = basic level 2 = intermediate level 3 = advanced level
li	*listening* basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*
sp	*speaking* basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*
re	*reading* basic: 0 = <1, 1 = 1, 2 = 2, 9 = cba* intermediate: 2 = <3, 3 = 3, 4 = 4, 9 = cba* advanced: 4 = <5, 5 = 5, 6 = 6, 9 = cba*

wr — *writing*

basic	intermediate	advanced
0 = <1	2 = <3	4 = <5
1 = 1	3 = 3	5 = 5
2 = 2	4 = 4	6 = 6
9 = cba*	9 = cba*	9 = cba*

*cba = can't be assessed

age — *age in years*

gender — *gender*

1 = female
2 = male
0 = missing information

ml — *mother language*

1	=	Finnish	11	=	Kurdish
2	=	Swedish	12	=	French
3	=	Russian	13	=	Lappish (Sami)
4	=	Albanian	14	=	German
5	=	Arabian	15	=	Somali
6	=	Dari	16	=	Turkish
7	=	English	17	=	Estonian
8	=	Spanish	18	=	Other
9	=	Italian	19	=	Thai
10	=	Chinese	0	=	missing information

be — *basic education*

1 = comprehensive school / basic education
2 = upper secondary school
3 = vocational school
4 = polytechnic
5 = university
6 = other
0 = missing information

socec — *socio-economic status*

1 = employed
2 = entrepreneur

3	=	unemployed
4	=	student
5	=	student in labor market training
6	=	retired
7	=	stay at home mom / dad
8	=	other
0	=	missing information

foe *field of employment*

1	=	agriculture and forestry
2	=	manufacturing
3	=	construction and real estate activities
4	=	transportation and storage
5	=	retail trade, accommodation and food service activities
6	=	administrative and support service activities
7	=	financial and insurance activities
8	=	information and communication
9	=	human health and social work activities
10	=	education and research
11	=	security, rescue and national defense activities
12	=	arts, entertainment and recreation
13	=	other
0	=	missing information

Question 13: For what purpose do you need the certificate?

0 = no, 1 = yes

kys13_1	to apply for Finnish citizenship
kys13_2	to apply for a job
kys13_3	to prove my language skills to my current employer
kys13_4	for study purposes
kys13_5	to get feedback on my language skills
kys13_6	other

Question 14: How long have you studied the test language?

1	=	less than one year
2	=	1-3 years
3	=	4-6 years
4	=	7-9 years
5	=	10 years or more
6	=	not at all
0	=	missing information

Question 15: Where have you studied the test language? You may choose several alternatives.

	0 = no, 1 = yes
kys15_1	comprehensive school / basic education
kys15_2	upper secondary school
kys15_3	vocational school
kys15_4	polytechnic
kys15_5	university
kys15_6	folk high school / adult education center
kys15_7	labor market training
kys15_8	courses organized by employer
kys15_9	self-study
kys15_10	abroad
kys15_11	other

Question 16: In your opinion, how well do you use the test language? Choose one alternative from each line.

	1 = poorly, 2 = quite poorly, 3 = quite well, 4 = well, 0 = missing information
kys16_1	I speak the test language
kys16_2	I write in the test language
kys16_3	I understand the spoken test language
kys16_4	I understand written texts in the test language

Question 17: How often do you use the test language? Choose one alternative from each line.

1 = not at all, 2 = once a month or less frequently, 3 = a couple of times a week, 4 = almost daily, 0 = missing information

kys17_1	I speak the test language with my family
kys17_2	I speak the test language with my friends and acquaintances
kys17_3	I read for example magazines and web pages in the test language
kys17_4	I write messages in the test language, such as emails and text messages
kys17_5	I follow the media in the test language (tv, radio, internet)
kys17_6	I use the test language to take care of everyday matters (shops, doctor's appointments, offices)
kys17_7	I use the test language at work
kys17_8	I use the test language in my studies

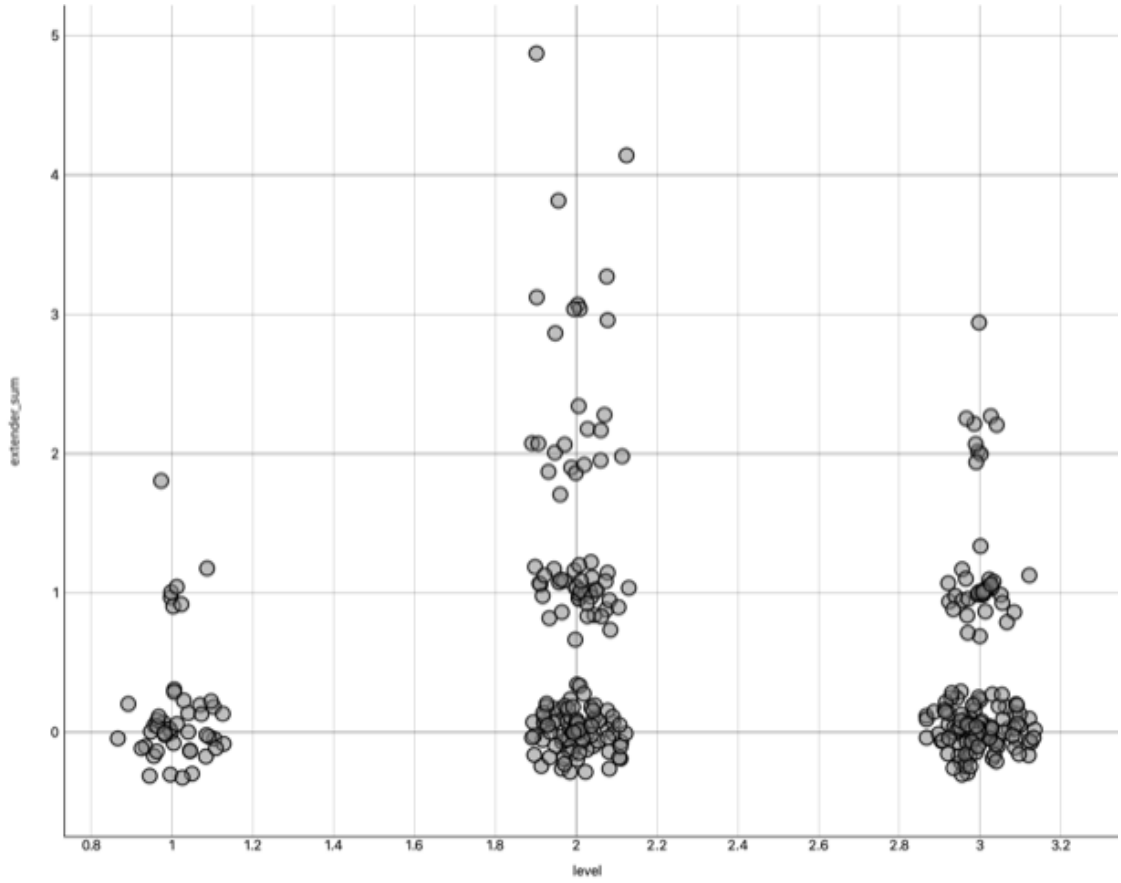
Appendix 4: Disposable income per consumption unit 2010

Households' income by Information, Person's socioeconomic group and Year

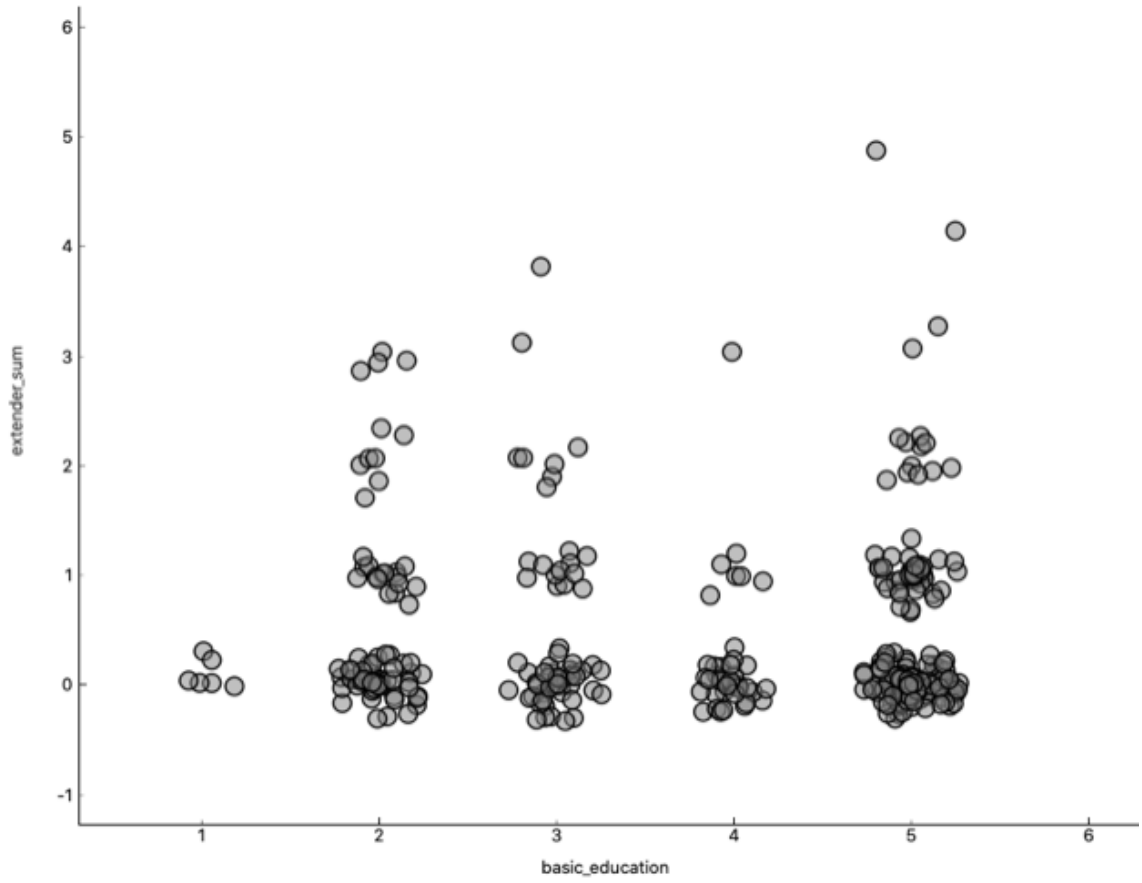
	2010
Disposable cash income per consumption unit, median	
Total	23,046
1 Self-employed persons	28,002
1.1 Agricultural entrepreneurs	25,145
1.2 Self-employed persons excluding farmers	28,911
2 Employees	27,964
2.1 Upper-level employees with administrative, managerial, professional and related occupations	34,844
2.2 Lower-level employees with administrative and clerical occupations	25,751
2.3 Manual workers	24,858
3 Students	11,700
4 Pensioners	17,851
4.1 Retired farmers	14,034
4.2 Retired self-employed persons	17,376
4.3 Retired upper-level employees	28,012
4.4 Retired lower-level employees	18,884
4.5 Retired manual workers	16,607
4.6 Other pensioners	14,400
5 Long-term unemployed	12,885
6 Others	13,878

Appendix 5: Orange outputs 1-3

1. The distribution of the test participants according to the test level and the extender frequency (scatter plot)



2. The distribution of the test participants according to the basic education and the frequency of extenders (scatter plot)



3. Pearson's Correlations

Correlations between age and other variables

1	+0.366	age	euro/c
2	-0.281	age	grade_list_compr
3	-0.273	age	grade_reading_compr
4	-0.243	age	where_ranking
5	-0.237	age	freetime
6	-0.232	age	study_time
7	-0.226	age	level
8	-0.215	age	grade_writing
9	-0.183	age	grade_speaking
10	-0.140	age	sumtotal
11	-0.127	age	or_long
12	-0.125	age	or_short
13	+0.118	age	work

Correlations between socio-economic status and other variables

1	+0.366	age	euro/c
2	-0.280	euro/c	studies
3	+0.238	euro/c	work
4	-0.166	euro/c	freetime
5	-0.145	euro/c	grade_writing
6	-0.140	euro/c	grade_list_compr
7	-0.124	euro/c	level
8	-0.119	euro/c	grade_reading_compr
9	-0.116	euro/c	grade_speaking
10	-0.094	euro/c	sumtotal
11	-0.091	and_short	euro/c
12	-0.086	euro/c	home
13	-0.071	euro/c	where_ranking

Correlations between the frequency of extender tags and other variables

1	+0.575	and_so_on	extender_sum
2	+0.504	and_long	extender_sum
3	+0.477	extender_sum	or_long
4	+0.437	extender_sum	or_short
5	+0.292	extender_sum	other
6	+0.259	and_short	extender_sum
7	+0.093	extender_sum	study_time
8	-0.077	age	extender_sum

Correlations between the test levels and other variables

1	+0.946	grade_list_compr	level
2	+0.942	grade_reading_compr	level
3	+0.932	grade_speaking	level
4	+0.911	grade_writing	level
5	+0.541	basic_education	level
6	+0.369	freetime	level
7	+0.341	level	sumtotal
8	+0.242	level	study_time
9	+0.228	level	where_ranking
10	-0.226	age	level

Correlations between the use of extender tags and other variables

1	+0.839	extender_sum	yes|no
2	+0.519	and_so_on	yes|no
3	+0.436	and_long	yes|no
4	+0.366	or_short	yes|no
5	+0.345	or_long	yes|no
6	+0.221	other	yes|no
7	+0.209	and_short	yes|no
8	+0.113	study_time	yes|no
9	-0.081	age	yes|no
10	+0.039	basic_education	yes|no

Appendix 6: Finnish summary

Suomenkielinen tiivistelmä

Johdanto

Koska vakiintunutta käännöstä englannin *extender tag* -ilmaisulle ei ole suomeksi olemassa, olen suomentanut ilmaisun sanalla *laajennusliite,* jota käytän tässä tiivistelmässä.

Tämä pro gradu -tutkielma kartoittaa, miten suomea äidinkielenään puhuvat englannin oppijat käyttävät englannin laajennusliitteitä (*extender tag)* puheessaan. Kyseessä oleva englannin kielen ilmiö on itse asiassa yhtä vanha kuin itse englanninkieli (Carroll 2008, 7). Vanhimpia mainintoja laajennusliitteen käytöstä löytyy Shakespearen teksteistä, joissa esiintyy *and things* -laajennusliite vuoden 1596 tietämillä (Tagliamonte & Denis 2010, 340). Oxfordin englannin sanakirjassa ensimmäinen maininta *and things* -laajennusliitteen käytöstä on vuodelta 1601. Varhaisimmat yritykset analysoida tätä kielellistä ilmiötä juontavat juurensa 1970-luvulle (Cristal & Davy 1975 ja Ball & Ariel 1978). Sittemmin laajennusliitteen käytön tutkimus on yleistynyt ja 2000-luvulla mukaan on tullut vieraankielen oppijan näkökulma. Tutkimus on myös laajentunut käsittämään englannin lisäksi muita kieliä, kuten esim. ruotsi (Winter & Norrby 2000), saksa (Overstreet 2005), espanja (Cortés Rodríguez 2006), japani (Watanabe 2014) ja persiankieli (Parvaresh et al. 2012).

Tutkielman tarkoitus on myös selvittää, mitkä sosiaaliset tekijät mahdollisesti vaikuttavat laajennusliitteiden käyttöön. Suomessa aiheeseen liittyvää vastaavanlaista tutkimusta ei ole tehty. Mari Metsä-Ketelän (2016) tutkimus poikkeaa tästä tutkielmasta siinä, että hän tutki *and so on*, *et cetera* ja *or something (like that)* -laajennusliitteiden käyttöä kansainvälisen kommunikoinnin yhteydessä akateemisessa ympäristössä. Tämä tutkielma hyödyntää YKI-korpusta, jonka Soveltavan kielentutkimuksen keskus on koostanut Yleisten kielitutkintojen testiaineistosta.

Tutkielma pyrkii vastaamaan kolmeen tutkimuskysymykseen, jotka ovat:

1. Mikä on laajennusliitteiden esiintymistiheys YKI-korpuksen englanninkielen suullisessa testiosuudessa?
2. Millaisia laajennusliitteitä rakenteeltaan suomalaiset englannin kielen oppijat käyttävät?
3. Mikä on erilaisten sosiaalisten muuttujien kuten esim. iän, koulutustaustan ja testitason vaikutus laajennusliitteiden käyttöön?

Tässä tiivistelmässä käydään läpi teoreettista taustaa sekä itse laajennusliitteitä havainnollistaen, että aiempiin tutkimuksiin viitaten, jonka jälkeen tutkimuksen materiaali ja menetelmät selostetaan. Lopuksi keskeisimmät tutkimustulokset esitellään johtopäätöksiä samanaikaisesti pohtien.

Teoria ja aikaisemmat tutkimukset

Laajennusliitteitä luonnehtii monimuotoisuus. Toisaalta niiden rakenne ja pituus vaihtelevat suuresti ja toisaalta ne voivat esiintyä kiinteinä kokonaisuuksina kuten esim. *and so on* tai *et cetera*. Tämän johdosta seuraava kuvaus pyrkii esittämään yleisimmin esiintyvät muodot. Perusmuodoltaan laajennusliitteet rakentuvat *and/or* -konjunktioista, määreistä ja pääsubstantiivista (*head noun*) (Overstreet 1999, Carroll 2008, Wagner et al. 2015). Pääsubstantiiveina voivat *things* ja *stuff* -substantiivien lisäksi toimia seuraavat substantiivit: *one*, *people*, *shit* ja *crap* (Tagliamonte & Denis 2010, 337). Laajennusliitteet jaetaan kahteen ryhmään sen mukaan, mikä konjunktio aloittaa liitteen (*adjunctives/disjunctives*). Yksi tärkeimmistä laajennusliitteisiin liittyvistä ominaisuuksista on anaforinen viittaus (*anaphoric reference*).

Kielitieteellinen tutkimus on lähestynyt laajennusliitteiden käyttöä monesta eri näkökulmasta, mikä heijastuu terminologian moninaisuutena. Kielen oppijan näkökulman myötä esiin on noussut sanarypäs-käsite (*multi word units/ MWU*). Kielenoppija omaksuu kieltä sanaryhmittäin eikä pelkästään sana kerrallaan (Vetschinnikova 2019, 1). On kuitenkin huomattava, että laajennusliitteet ovat vain yksi sanarypäs-käsitteen alle kuuluva kielen sanaston osa-alue.

Kuten edellä on mainittu, aiheeseen liittyvää tutkimustietoa löytyy melko runsaasti. Tähän pro-gradu-tutkielmaan otettiin mukaan kahden tutkijan tutkimustulokset, joihin tämän tutkielman tuloksia verrattiin. Vertailun avulla selvitettiin, miten suomalaiset englanninkielen oppijat sijoittuvat kansainvälisesti. Buysse (2014) tutki hollantia äidinkielenään puhuvien englanninkielen oppijoiden korpusta ja Aijmer (2015) puolestaan käytti tutkimuksessaan ruotsalaisten englannin oppijoiden korpusta.

Tutkimuksen aineisto ja menetelmät

YKI-korpus koostuu aikuisille tarkoitetusta kielitaitotestistä, joka voidaan suorittaa kolmella eri testitasolla (*basic, intermediate, advanced*). YKI-korpuksen aineisto on jaettu vanhempaan ja uudempaan osaan vuoden 2010 ollessa jakolinjana. Tutkimuksen aineistoksi valittiin 360 äänitettä Yleisen kielitutkinnon englannin testin suullisen osion neljännen tehtävän suorituksista. Kyseisen suullisen testiosuuden pituus on 2 minuuttia. Kaiken kaikkiaan aineiston kesto on 12 tuntia. YKI-korpus ei sisällä äänitteiden transkriptioita, joten kuuntelun jälkeen laajennusliitteitä sisältävät lauseet sekä niitä edeltävä konteksti transkriboitiin. Löydetyt laajennusliitteet jaettiin kolmeen ryhmään testitasoittain sekä vanhan että uuden aineiston osalta. Ensimmäinen ryhmä alkaa *and*-konjuntiolla, toinen *or*-konjuntiolla. Kolmannessa ryhmässä ovat ne laajennusliitteet, jotka eivät sisällä konjunktiota. Laajennusliitteiden ryhmittely, transkribointi ja esiintymistiheyksien laskeminen on suoritettu Orange-ohjelmaa - käyttäen.

Tarkempaan analyysiin valittiin *and stuff/thing (like that)* ja *and so on*, joista ensimmäisten kohdalla tutkittiin anaforista viittausta sekä jälkimmäisen kohdalla laajennusliitteen käyttöä listan muodostuksen osatekijänä. Lisäksi *or*-laajennusliitteiden konteksti kartoitettiin Aijmerin (2015) luetteloimien kollokaatioiden avulla.

YKI-korpus sisältää myös testiin osallistuneiden antamia taustatieoja, joita hyödynnettiin sosiaalisten muuttujien tutkimuksessa. Tämän aineiston analysointiin käytettiin datan louhintaan kehitettyä Orange-ohjelmistoa. Lisäksi Mathematica-ohjelmistoa on käytetty lähinnä tutkimustuloksia esittävien kaavioiden kuvaamiseen.

Keskeisimmät tutkimustulokset

Analyysin tulosten mukaan suomalaiset englanninkielen oppijat käyttivät eniten *and*-konjunktiolla alkavia laajennusliitteitä. *And*-konjunktioiden alkavien osuus selittyy suurella *and so on* -laajennusliitteen käytöllä. Kolme suosituinta laajennusliitettä olivat: *and so on, or something like that* ja *or something*. Aivan poikkeuksellista *and so on* -käyttö ei ole kielenoppijoiden keskuudessa sillä Buyssen (2014) hollantia äidinkielenään puhuvien oppijoiden korpuksessa *and so on* -käyttö löytyy kolmannelta sijalta. Tutkimusten mukaan natiivit englanninkielen puhujat käyttävät *and so on* -laajennusliitettä yleensä joko kirjoitetussa tai virallisessa puheessa (Overstreet 1999, 7). Suomalaisten kielenoppijoiden suosima *and so on* -laajennusliitteen käyttö on osaltaan selitettävissä äidinkielen siirtovaikutuksella. Suomen kielen vastine *ja niin edelleen* on käännökseltään englantilaista versiota vastaava sekä rakenteeltaan että käytöltään.

Koska *and so on* -laajennusliitettä esiintyi eniten korpuksessa, sitä tutkittiin listan muodostuksen osatekijänä. Analyysin tulokset vahvistavat Jeffersonin (1990) teorian listan muodostamisesta. Teorian mukaan puhujat pyrkivät muodostamaan listan, joka koostuu kolmesta osasta.

Kun *and stuff/things (like that)* -laajennusliitteen käyttöä analysoitiin tarkemmin, osoittautui, että suomalaiset englannin kielen oppijat suosivat laajennusliitteen pidempää muotoa puheessaan. Buysse (2014) raportoi samasta ilmiöstä, mutta Aijmer (2015) ei nähnyt vastaavaa tutkimuksessaan. Yllättävä löydös oli se, että suomalaiset eivät käyttäneet puheessaan suuremmassa määrin *and stuff (like that)* -laajennusliitettä, mikä taas oli Buyssenin (2014) tutkimuksessa ominaista hollantia äidinkielenään puhuvien englannin kielen oppijoille. Buyssen mukaan runsas käyttö juontaa juurensa suuresta amerikkalaisten sarjojen tarjonnasta, jolloin *and stuff (like that)* -laajennusliitettä kuullaan paljon. Siitä huolimatta, että Suomessa on myös laaja amerikkalaisten sarjojen ja elokuvien tarjonta, eivät suomalaiset näyttäneen käyttävän *and stuff (like that)* niin paljon kuin hollantia tai ruotsia äidinkielenään puhuvat englanninkielen oppijat.

And stuff/thing (like that) -laajennusliitteen kontekstin tutkiminen paljasti ristiriitaisuuksia *stuff/things* -pääsubstantiiviin ja sen anaforisen viittauksen kohteena olevan substantiivin välillä. Esimerkiksi pääsubstantiivia *stuff* oli käytetty viittamaan

laskettavissa olevaan substantiiviin. Edellä mainitun tyyppistä laajennusliitteen käyttöä on todettu aikaisemmissa tutkimuksissa (Tagliamonte & Denis 2010).

Sosiaalisten muuttujien analyysi osoitti, että mitään yksittäistä sosiaalista tekijä, joka vaikuttaa laajennusliitteiden käyttöön, ei voitu nimetä. Tutkimus kuitenkin antoi viitteitä siitä, että korkeampi koulutuksen taso sekä nuoruus edistivät laajennusliitteiden käyttöä. Aineiston analyysissä käytetty Orange-ohjelmiston käyttö osoittautui tämän tutkimuksen tekemiseen hyvin käyttökelpoiseksi lähinnä siksi, että sen käyttö ei vaadi ohjelmointitaitoja. Laskutoimituksen tuloksia on mahdollista tarkastella erilaisen graafisten kaavioiden avulla, jotka visualisoivat tuloksia. Tällöin aineistosta saattaa nousta esiin asioita, joiden tarkempi tarkastelu avaa uusia näkökulmia aineiston tulkitsemiseen.

Tutkimusten perusteella voidaan todeta, että laajennusliitteiden käyttö tulisi ottaa huomioon opetuksessa. Opettajien tulee lisätä oppilaidensa tietoisuutta tämän kielellisen ilmiön suhteen. Oppilaiden tulisi oppia tunnistamaan erilaiset laajennusliitteiden erilaiset muunnelmat ja oppia käyttämään niitä oikeassa rekisterissä.

Lopuksi voidaan todeta, että tämä tutkielma voi toimia lähtökohtana laajemmalle laajennusliitteiden tutkimukselle suomalaisten englanninkielen oppijoiden kielenkäytössä. Erityisesti syvällinen kvalitatiivinen laajennusliitteiden funktionaalisten ominaisuuksien tutkimus luonnollisissa keskustelutilanteissa olisi erinomainen tutkimuskohde. Tutkimukseen olisi syytä sisällyttää kyselylomake tai haastattelu, jotta yksityiskohtaisemmat taustatiedot saataisiin kartoitettua. Toinen tutkimus olisi hyvä kohdistaa oppimismateriaalien kartoittamiseen peruskouluissa ja lukioissa. Mielenkiintoista olisi selvittää, kuinka paljon laajennusliitteitä esiintyy opetusmateriaaleissa ja kuinka paljon opettajat itse kiinnittivät ilmiöön huomiota opettaessaan. Kielen oppijoille hyvä laajennusliitteiden hallinta tarjoaa vielä yhden työkalun sujuvaan kielen tuottamiseen.