# DETERMINERS IN FINNISH-SPEAKING LEARNERS' ENGLISH: A TRANSFER BASED, QUANTITATIVE TEXT ANALYSIS

Juuso Helminen

MA Thesis

English, Degree Programme for Language Learning and Teaching

School of Languages and Translation Studies

Faculty of Humanities

University of Turku

November 2020

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

UNIVERSITY OF TURKU
School of Languages and Translation Studies / Faculty of Humanities
HELMINEN, JUUSO: Determiners in Finnish-speaking Learners' English: A Transfer Based, Quantitative Text Analysis

MA Thesis, 53 p., 9 app.
English, Degree Programme for Language Learning and Teaching
November 2020

---

This thesis studied texts written by Finnish-speaking learners of English. The goal of this study was to examine how Finnish-speaking learners use determiners in their written English. English has traditionally been difficult to learn for native Finnish-speakers, as the languages are quite distant from each other. This study focuses specifically on determiners, as the challenges a Finnish-speaking learner faces can be more comprehensively understood by examining a linguistic system that greatly differs between the two languages.

The texts analyzed in this thesis originated from the YKI-Corpus; a collection of essays written by candidates taking *The National Certificate of Language Proficiency* examination, compiled by the University of Jyväskylä for research purposes. Essays were taken from 30 participants in total, all of whom were native Finnish-speakers taking the examination in English. 15 of the learners took an passed the exam on intermediate level (B level on the *Common European Framework of Reference*), 15 on advanced level (C level on CERF). The texts were analyzed using two Natural Language Processing tools: *Simple Natural Language Processing Tool* (SiNLP) and a Python extension library *spaCY*. All determiners in the texts were first counted and categorized, then analyzed and compared using the statistical analysis tool *SPSS*.

The results of this study indicated that Finnish-speakers prefer articles as their determiners: 57 % of all determiners were some type of an article across both proficiency levels. Definite articles were the most common type of an article on both levels. Advanced level learners used a higher number of determiners per word than intermediate learners, slightly over 18 %, which seemed to be very close to the ratio of native level speakers. Advanced level learners also used zero articles much more often, whereas intermediate level learners preferred possessive determiners instead. Overall, a higher proportion of articles, especially zero articles, as well as an increased number of determiners per word suggested a higher level of proficiency in English. However, it is outside of the scope of this thesis to further analyze the reasons for these differences, as that would require more extensive, perhaps qualitative research.

Keywords: English as a foreign language, determiners, transfer, written language

# Table of Contents

# List of Tables



# List of Figures

# 1 Introduction

The goal of this study is to examine language transfer in the written English of Finnish-speaking learners. If focuses specifically on identifying and analyzing how Finnish as a native language influences determiner use and the concept of definiteness in compositions written in English, as well as how the proficiency in English affects these linguistic choices. As Finnish as a language is very different from English in many ways, including typology, there are many challenges a Finnish-speaking learner must overcome in order to achieve proficiency in English. While this idea has been generally acknowledged and accepted (see e.g. Ringbom 1987, Meriläinen 2010), much of the influence Finnish, or any native language for that matter, has on learning English remains uncertain. The topic of language transfer and the complex psycholinguistic processes related to language learning are in need of further research, and while this study will not offer conclusive answers, it can help alleviate that need and advance the research from its own perspective, both in terms of English language pedagogy in Finland, as well as language learning as a broader, more general subject.

Finland has offered a hospitable environment for transfer research, especially comparative studies, due to its peculiar language situation: Finland has two official languages, Finnish and Swedish, and the speakers of both come from very similar cultural and socioeconomical backgrounds, enabling comparison of English learners with minimal differences other than the native language. Such studies have been conducted by Finnish researchers like Håkan Ringbom (see e.g. 1987, 2007) since the 1970s, the results suggesting that Swedish-speakers have less difficulty in learning English than Finnish-speakers, presumably due to language similarities. In the early 2000s, the transfer research opportunities in Finland also attracted non-Finnish scholars such as Scott Jarvis and Terence Odlin (see e.g. Jarvis and Odlin 2000). However, despite the fruitful environment, Finnish transfer into English has not been extensively studied, relying mostly on Ringbom's work, some of it from over 30 years ago. In the last decade, Lea Meriläinen has done research on transfer with a focus on Finnish-speakers as well

as learner Englishes in general (see e.g. 2010, 2017), bringing forth a fresh perspective on Finnish-to-English transfer.

As determiners (and especially articles) have traditionally been challenging for Finnish-speakers to integrate into their English, it has been a topic of interest in transfer research in Finland. However, Meriläinen does not include articles in her 2010 study of Finns' written English, explaining that studying such a frequent element is better served in a study focusing on articles, and that truly identifying transfer-related issues in article use would require analysis on different types of article use; something that is simply outside the scope of her research (Meriläinen 2010, 114). Furthermore, even though problems in Finnish-speakers' article use are well documented in e.g. Ringbom's work (1987), language teaching and the linguistic landscape has changed in Finland in the last 20–30 years, as can be seen in Jarvis's research (see e.g. 2017). Perhaps as a result of English becoming more prominent in Finland, even the Finnish language has changed: there has been some discussion about whether colloquial Finnish already has what could be identified as an article (see e.g. Laury 1996, Larjavaara 2001). This could indicate a shift in the perspective of Finnish-speakers regarding definiteness and articles. For all the reasons above, this study examines determiners in Finnish-speakers' English texts, as determiners can potentially offer a fresh, broad insight on Finnish-to-English transfer, as it includes but is not limited to articles.

The texts in question (by 30 writers in total) come from the YKI-Corpus (2017) compiled by the University of Jyväskylä. This corpus is intended for research purposes, and it offers a vast collection of essays written by candidates taking the National Certificates of Language Proficiency examination, combined with comprehensive background information and data of the candidates, which allows for selecting Finnish native speakers taking the English examination to fit the purposes of the present study. First, these texts were analyzed by two tools designed for linguistic analysis, *SiNLP* and *spaCY*, in order to parse the essays for determiners of different categories and other important metrics, such as word count. The parsed data were further examined with the statistical analysis tool *SPSS*, which enabled comparison between the two groups of learners.

This thesis will begin with Section 2: an overview on transfer studies and the definition of transfer and terms related to it, followed by a brief comparison of the two

languages relevant to this thesis, Finnish and English, with special focus on definiteness and determiners. Next, Section 3 will introduce the material and methods used in this study: the aforementioned YKI-Corpus with the participants and essays it offers, the tools and programs used in the analytic process, and the reasoning behind all choices made in terms of the material and methods in this study, as well as all principles the analysis will follow in organizing and calculating the data. Section 4 will present the results of the study in tables and figures for both a deep insight into the data, as well as an efficiently interpretable overview of the results. The results will be succeeded by Section 5, discussion, which will further elaborate on the produced results, attempt to achieve the goal set for this study, and find answers to the asked questions. Finally, Section 6 will conclude this study by summarizing it and binding it to a broader context of language research by exploring further areas and opportunities of research.

# 2 Theoretical framework

This section will outline the history of transfer studies in a concise manner from the 1950s until the present day and cite the research of major contributors in the field of transfer, especially with Finnish-speaking language learners. This section will also clarify the purpose for this thesis and demonstrate the gaps in knowledge this study attempts to fill, as well as present the results of previous research on the topic of determiner use and distribution in the English language. Finally, the relevant terms and concepts as well as the research questions for the present study will be defined.

## 2.1 History of transfer studies

Throughout history, the source for language learning and acquisition has been people speaking different languages coming into contact with one another for various reasons (Odlin 1989, 6-14). In such a scenario, three elements are often present: the individual, their native language (NL) and the target language (TL) they have encountered and are now either actively learning or at least passively absorbing. While all three are vital parts

of the process called *second language acquisition*, this thesis as well as transfer studies in general are focused on the language learner's native language.

In the 1950's, language learning was believed to occur through habits formed by imitation and repetition, where the goal was automatization of stimulus–reaction patterns, *behaviors*, as they were called (Ellis 2015, 117). This model of learning was based on research by behavioral psychologists, such as Pavlov and Skinner, whose studies led to these behavioristic methods gaining popularity in language learning and teaching (Järvinen 2014, 78-79). While this repetitive and mechanical model was rather quickly abandoned, as it could not adequately explain, for instance, learners' creativity (Järvinen 2014, 79), it paved the way for other influential language learning models, first of which was Robert Lado's (1957) *Contrastive Analysis Hypothesis*. Like the behaviorists, he focused on habits as well, but from a different perspective: his theory claimed that old habits (formed by learning one's native language) actively prevent new habits (which are needed in order to learn a new language) from developing (Lado, 1957). Two different views on this hypothesis were presented: the strong form of Contrastive Analysis Hypothesis by Lado in 1957 and a weak form by Ronald Wardhaugh in 1970 (Ellis 2015, 117). The strong form proposes that all difficulties in second language learning can be predicted by contrasting the learner's native language and the target language, and identifying the differences, whereas the weak form suggests that instead of predicting the difficulties, contrastive analysis can be used to identify and explain them (ibid.).

Contrastive analysis established that a learner's native language does have an effect on their second language learning and acquisition, but not to the extent Lado had hypothesized: empirical research in the 1960s showed, that not all learning difficulties can be attributed to differences between languages, and even the ones that do were not always correctly predicted by contrastive analysis (Odlin 1989, 17). Moreover, some cross-language differences do not always cause learning difficulties, whereas some cross-language similarities do: for example, Spanish-speaking learners of English often commit errors with the English verb *be*, despite similar grammatical structures in both languages, as documented by multiple studies (Odlin 1989, 17-18). Therefore, contrastive analysis has been deemed not only incapable of predicting learning difficulties, but also inadequately comprehensive at explaining said difficulties.

These shortcomings of contrastive analysis led to the development of *Error Analysis*, which was a model that aims to classify and evaluate learner errors to find explanations to errors not stemming from language transfer (Odlin 1989, 18). Different types of analyses of learner errors were conducted in the late 1960s and 1970s by researchers such as S. P. Corder, Jack C. Richards, Heidi Dulay and Marina Burt (see e.g. Dulay, Burt and Krashen 1982), and they found that many learner errors should be attributed, instead of the native language, to other factors, such as the training the learner had received, overgeneralization of a structure, or the stage of development the person is currently in as a language learner (Gass and Selinker 1994, 6-7; Odlin 1989, 18-19; Ellis and Barkhuizen 2005, 52). Error analysis, while to this day a useful tool for e.g. teachers to evaluate their students, proved ineffective as a comprehensive model to understanding language learning, partly because of how difficult it is to define an error; language can be used in various different contexts and situations by different groups of language users, which makes judging grammaticality or acceptability a very complicated process (Ellis and Barkhuizen 2005, 56). However, out of the error analysis era emerged Dulay and Burt's 1974 theory of *developmental sequence*, which suggested that language learners acquire the linguistic elements of a new language in a similar order children do in their native language (Odlin 1989, 20-21). Dulay and Burt conducted an error analysis, which resulted in an overwhelming majority of errors stemming from the learner's developmental stage (Ellis 2015, 118). From result they inferred that the learner's native language has minimal effect on second language acquisition (ibid.)

The research since the mid-1970s, however, has shown that both Lado's as well as Dulay and Burt's views on a native language's effect on language learning have been too extreme; the notions of contrastive analysis or developmental sequences are not mutually exclusive (Gass and Selinker 1994, 6-7). The contemporary view on the first language highlights its role as one of the components affecting second language acquisition, shaping the learner's mental processes with other factors, both internal and external (Ellis 2015, 118). Instead of only harming the language acquisition process, the learner's native language can also be helpful. Odlin (1989, 36-37) separates positive and negative transfer; positive transfer, i.e. similarities between the two languages can, for example, reduce the time required to develop reading comprehension, make sound identification easier, and lessen the difficulties with syntactic elements (articles, word

order, etc.). Negative transfer represents the older view on native language affecting language learning: differences between the two languages cause learning difficulties, which can manifest themselves in various ways.

## 2.2 Defining transfer

The previous section offered the brief history of transfer studies, from the 1950s full effect transfer through 1970s no effect transfer, all the way until the more balanced view today. This section will further elaborate on the definition of transfer and its different manifestations, focusing especially on issues regarding syntactic elements (such as article use) and written language production.

Learning a second language (L2) implies that the learner already knows another language: their first language or native language (L1). Unlike a toddler learning their L1, an L2 learner has already achieved a native-level competence in one language. This knowledge of a language consequently shapes the learner's thinking and affects how they can learn and use another language (Jarvis 2017, 12; Jarvis and Pavlenko 2008, 11). This process is called *transfer* or *crosslinguistic influence*: transferring features, such as vocabulary, sounds or cultural elements, from one language into another (Pietilä and Lintunen 2014, 14). As the learner has mastered their L1 but the development of the L2 is still in progress, transfer occurs in what is often described as *interlanguage*: a language system the learner constructs at any given moment, independent from their L1 and the target language (Selinker 1972, Ellis 2015). In addition to a learner's L1 affecting their L2 learning, transfer can be used to describe the effects any previously learned language can have on a subsequently learned language: L1 can affect a third language, L3, which can in turn influence L4, and so forth (Jarvis 2017, 12-13). However, the primary focus is often on how L1 affects any additional language learned after the L1 regardless of the chronological order of learning those languages (Jarvis 2017, 13), as it will be in this study as well. A strict focus on chronology is also problematic in a situation where a person has more than one native language (Pietilä and Lintunen 2014, 12). For example, in Finland 5.2 percent of the population (approx. 290,000 people) speak Swedish as a native language (Official Statistics of Finland). Many of these people have also learned Finnish well enough to identify as bilingual (Ringbom 1987, 9); their first foreign

language is technically their L3, which is influenced by both Finnish and Swedish, their native languages.

Transfer manifests itself in different ways. It can affect all linguistic subsystems, but not necessarily at all times or in all contexts (Jarvis 2017, 14). A variety of factors affect the likelihood of the effects of transfer appearing in a learner's language, for example the relation of the two languages, markedness and frequency of the language feature in use, and the learner's stage of development (ibid.). As mentioned previously, the contemporary understanding of transfer includes the notion of beneficial transfer, similarities between languages that help the learner, in addition to detrimental transfer, or *interference*, which denotes the differences between languages that can make learning difficult (Pietilä and Lintunen 2014, 15). A simple example of this is the use of articles in Germanic languages such as Swedish and English but not in Finnish (Ringbom 1987, 19-20); a common feature makes the learning of articles presumably easier for Swedish L1 learners of English than for Finnish L1 learners, who do not benefit from transfer in this particular instance.

Positive L1 influence can be difficult to identify, but its outcome is clear: it makes acquiring a new language either easier, quicker, or both. On the other hand, negative influence often causes distinct divergences from the target language norms and is, therefore, easier to recognize. There are multiple outcomes of negative transfer: *underproduction* and *overproduction* refer to the frequency of use for certain structures or elements in comparison to native speakers, whereas *production errors* fall within the purview of traditional error analysis: a native language form being used in L2 (Odlin 1989, 36-37). *Avoidance* is a phenomenon rather similar to underproduction: if a structure does not occur in a learner's L1, the learner might try to avoid using that structure altogether, resulting in no noticeable errors (Ortega 2009, 39-40). In certain cases of production errors, the learner might be cognizant of their L1 affecting their L2 production and try to incorrectly compensate for that, leading to *hypercorrections* (Odlin 1989, 38). For example, a Finnish-speaking learner of English could use a sentence such as **I went to home*, because they know that the Finnish illative case *-iin* (the basic meaning being *into (the inside of something)*) is often equivalent to the preposition *(in)to*.

As previously mentioned, transfer affects all linguistic subsystems to some extent. The effects of transfer are easier to identify in subsystems that are more common and thus less prone to avoidance; it is quite difficult for the learner to never use a certain difficult phoneme, for example, especially in unplanned speech. Syntactic elements however, as opposed to e.g. phonetic or lexical elements, can be easier to avoid, resulting in few noticeable effects of negative transfer (Meriläinen 2010, 23). In addition to this, the learner's developmental stage and universal learning mechanisms have a rather large influence on the learning of L2 syntax, which makes identifying the effects of transfer more difficult (Odlin 1989, 85-110). As the mental processes of language learning are exceedingly complex and no consensus on, for example, the concept of universal grammar has been reached (White 1994; Meriläinen 2010, 24-26), this thesis will focus on adult learners of English with several years of language learning, to diminish the effect the developmental stage has. Furthermore, although this thesis studies determiners, a clearly syntactic system, avoidance should not pose much of an issue: as far as syntax goes, determiners are a rather common element, present with every noun and noun phrase of English, therefore making their occurrences frequent and analyzing less difficult.

In addition to different linguistic systems, the distance between the two languages, the learner's L1 and L2, is a factor in how transfer affects language learning. This phenomenon was researched extensively in the late 1970s: several researchers, including Ringbom and Sjöholm were able to show, that learners with an L1 distant from their L2 are less likely to transfer forms than learners with less distance between their L1 and L2 (Ellis and Barkhuizen 2005, 65; Ellis 2015, 123). In his study of Dutch university level learners of English translating idioms, Kellerman (1979) argued that larger perceived distance between L1 and L2 makes the learner less likely to rely on their L1 to fill in the gap in knowledge, and that more advanced learners are more likely to recognize the "problem" that is their L1 and have a larger selection of linguistic tools to approach such situation. To put it more simply, transfer occurs less in an advanced learner with distant native language and target language. Comparative studies on Finnish-speaking Finns and Swedish-speaking Finns learning English have produced similar results. Ringbom (1987) describes studies on these two groups of learners conducted by himself, as well as e.g. Rolf Palmberg and Kaj Sjöholm: evidence suggested

that Swedish-speakers benefit from transfer caused by L1-L2 similarity much more than Finnish-speakers do, however, they also manifested more negative transfer in the form of errors.

It is important to note, that in terms of transfer, language distance can refer to both the actual, structural similarity between two languages (*typological proximity*), as well as the *perceived distance*, the differences language learners assume exist between their L1 and L2 (Ivaska and Siitonen 2017, 227–228; Ringbom 2007, 28); the objective and subjective distance, respectively. Interestingly, the perceived distance can override the objective distance: if the learner feels like their L1 is very different from their target language, they might act as if that is the case and rely less on their L1, regardless of how distant the two language are in reality (Jarvis and Pavlenko 2008, 177–178). Furthermore, Ringbom (1987) presented that perceived distances are not limited to L1 and L2; Finnish-speaking learners recognized that their L3, Swedish, is much more similar to English than Finnish is, resulting in them attempting to fill the gaps in their English proficiency with Swedish. Therefore, transfer in this instance occurs between Swedish and English, rather than Finnish and English.

The final concept of transfer regarded in this thesis is *markedness*. The term has been defined in different ways, but at its core, it divides linguistic elements into marked and unmarked, "special" and "basic" (Ellis 2015, 124). Generally in all languages, English included, positive and negative meaning are conveyed differently from each other: *I write* versus *I do not write*. In this example, *I write* is the positive, basic, unmarked form, whereas *I do not write* is the negative, special, marked form, which requires a distinct feature to separate it from the unmarked form: a dummy *do* plus the negation *not.* Markedness is higher in features that are, for instance, complex or low in frequency. Kellerman (1979, 53–54) notes the importance of markedness in transfer: marked forms are "potentially less transferable than unmarked ones", i.e. a learner is less likely to use a linguistic feature in their L2, if said feature is uncommon or special in their native language, thus limiting the effects of transfer in these instances. Several studies have showed, that markedness is also a factor in how easy a structure is to learn: unmarked structures are easier and faster to learn (Jarvis and Pavlenko 2008, 186; Ellis 2015, 124–125). Finally, markedness can also affect the choices a learner makes in L2 production. The learner usually relies on the less marked, *prototypical* forms of their L1 to transfer

over to L2, instead of the more figurative choices (Jarvis and Pavlenko 2008, 188). An example would be a learner describing something as *big* instead of *tall, wide* or *long,* even if one of those would have been a more accurate representation.

## 2.3 Determiners and definiteness

This research examines determiners, which are defined in Longman Student Grammar of Spoken and Written English as follows: they are "function words used to specify the kind of reference a noun has" (Biber et al. 2002, 65). A simple example of such reference is the article a noun possesses: an indefinite article *a* or *an* is used when the following noun names something yet unidentified, whereas a definite article *the* indicates that the noun is already known and identified (Krohn 1971, 54). Articles, however, are not the only type of determiners. The following table is modelled after Table 4.1 in Longman Student Grammar of Spoken and Written English (Biber et al. 2002, 65), and it shows different determiner types and examples of how they combine with nouns.

As Table 1 shows, determiners consist of much more than just articles. Articles are a single word class amongst many others, such as nouns, pronouns and numerals, which can all be used as determiners in the correct context and situation. According to Verspoor and Sauter, the present-day English articles derive from the demonstrative pronoun *this* (now *the*) and the cardinal number *one* (now *a(n)*) (Verspoor and Sauter 2000, 99). They also note that "articles are always used dependently as determiners" (Verspoor and Sauter 2000, 100), which further highlights the relationship between articles and determiners: all articles are determiners, but not all determiners are articles.

*Table 1: Determiner types*

| **determiner type** | **countable nouns** | | **uncountable nouns** |
|---|---|---|---|
| | **singular** | **plural** | |
| zero article | – | books | water |
| indefinite article | **a** book | – | – |
| definite article | **the** book | **the** books | **the** water |
| possessive | **my** book | **my** books | **my** water |
| genitive | **Juuso's** book | **Juuso's** books | **Juuso's** water |
| demonstrative | **this** book | **these** books | **this** water |
| | **that** book | **those** books | **that** water |
| quantifier | **every** book | – | – |
| | – | **all (the)** books | **all (the)** water |
| | – | **many** books | **much** water |
| | – | **some** books | **some** water |
| | – | **(a) few** books | **(a) little** water |
| | – | **enough** books | **enough** water |
| | – | **several** books | – |
| | **(n)either** book | **both** books | – |
| | **any** book | **any** books | **any** water |
| | **no** book | **no** books | **no** water |
| Numeral | **one** book | **three** books | – |
| Wh-word | **whose** book | **whose** books | **whose** water |

To understand how and why determiners are used in different languages, the concept of *definiteness* must be examined. As mentioned previously, determiners in the English language specify the kind of reference a noun or a *noun phrase* (NP) has. The nature of that reference dictates the definiteness (or indefiniteness) of the noun phrase in question: generic or specific, countable or uncountable, identified or unidentified (Chesterman 1991). However, in practice definiteness can be difficult to determine, as it is often context-dependent and subjective: what is, for example, already identified, old information changes from conversation to conversation, naturally affecting the

definiteness of the subject matter of the conversation. Furthermore, the systems different languages have for denoting definiteness are often full of exceptions, which makes analyzing definiteness very complicated. For instance, in the English article system articles generally do not occur with proper nouns such as names, however, *the John I knew* is perfectly fine and grammatical given the right context. Almost any noun or noun phrase can be made work with any article in the right context and circumstance (Chesterman 1991, 7). Furthermore, despite the dichotomy of definite and indefinite, the reality is seldom so black and white: Chesterman argues that definiteness is a scalar quantity, and different tools are used to place a noun's reference somewhere on that scale (Chesterman 1991, 182). The function of determiners can be defined in such ways that do not mention definiteness, such as setting specifying or generalizing the meaning of a noun, or providing grammatical status (Master 2013), but for the purposes of the present study, these definitions are close enough to the concept of definiteness to be regarded as similar.

The above examples of definiteness have been from the English system, which denotes definiteness with determiners, most commonly articles. As to how exactly determiners are used in English, especially written English, has been studied by Peter Master: his 2013 study on determiners in research articles of different fields of study written in English found, that articles are the most prominent determiner type in the English language, especially in the genre of research articles (Master 2013). In more detail, Master discovered that the main focus of his study, research articles in science and technology, had an average determiners to words ratio of 0.183 and an articles to total determiners percentage of 90.3 % (ibid.). Out of these articles, the most common type was zero article at 51.2 %, followed by the definite article (37.8 %) and the indefinite article (11.0 %). Master highlights, that while the definite article *the* is the most common word in the English lexicon, the zero article is actually the most common linguistic unit (Master 2013, 7). As for the other determiner categories, demonstrative determiners comprised 4.5 % of the total determiner count, followed by possessive determiners at 2.4 % (ibid.) After these major determiner types, Master categorization diverges from the present study's; Master separates the assertive-nonassertive determiners (*some* and *any*), negative determiners, and universal/dual determiners (*each, every, either, neither*), whereas in the present study, all of these fall within the category of quantifiers.

Together, they cover 2.06 % of the total number of determiners (0.9 + 0.16 + 1.0 %). Finally, wh-words tallied up to 0.14 % in Master's study (ibid.). Master study will be revisited in the material section 3.1 of this study to discuss the benefits and limitations regarding its use as a baseline in this study, as well as in the discussion section 5 to highlight and compare Master's results with the present study's results in more detail, whenever applicable.

Contrary to the English system, where articles are the most common determiner type, there are no articles in Finnish: instead, definiteness is traditionally represented by the case system. The following is an example of countability: *ostin* ***auto\|n*** [*I bought* ***a car (in accusative case)***] as opposed to *ostin* ***ruoka\|a*** [*I bought* ***food (in partitive case)***]. The Finnish linguistic system, similarly to English, can use other determiners, such as pronouns as demonstratives, to mark definiteness: ***eräs*** *mies* [***a*** *man*]. As mentioned before, the Finnish system does not officially use articles to denote definiteness. However, structures similar to articles have been emerging especially in colloquial Finnish for some time: the Finnish grammar book *Iso suomen kielioppi* (2004) states, that pronouns such as *se* [*it*], *tämä* [*this*] and *yks(i)* [*one*] have been used to define a noun in an article-like manner (Hakulinen et al. 2004, § 1418). No consensus has been reached on whether articles exist in the Finnish language, as there is not enough material for a comprehensive diachronic analysis and comparing Finnish to article-languages has proven inadequate for determining one or the other (ibid.). Laury (1996) argues for the Finnish article by claiming that *se [it]* can be used with referents that are new, generic, or identifiable by common knowledge. On the other hand, Larjavaara (2001) sees that *se* cannot be viewed as an article, because its use is not obligatory in the general sense of definiteness, but instead it occurs for special purposes. However, both Laury and Larjavaara agree, that in colloquial Finnish, the use of article-like structures is undeniably commonplace. For the purposes of this study, whether the prescriptive Finnish grammar officially has articles or not is irrelevant; important is that Finnish-speakers are presumably somewhat familiar with the concept of definiteness from an article-language perspective. This study attempts to discover how and to what extent that manifests in their interlanguage.

Section 2 introduced the history of transfer studies, the present study's definition of transfer and the essential terms and concepts in regard to the study. As

established, Finnish and English are intrinsically quite different from each other as languages, but English has an influence on many Finnish-speakers' native language: structures akin to the English determiners have snuck into Finnish as well. This study is interested in whether this manifests as transfer in Finnish-speakers' English, and if so, to what extent. Thereby the research questions for the present study are as follows:

1. How do intermediate and advanced level Finnish-speaking learners of English use determiners?
2. How does their determiner use differ from native level English speakers?

Finding these answers would not only expand the understanding of Finnish-speakers' English use and language transfer in general, but also shed more light on the problems Finnish-speakers encounter as they learn English, and thus help teachers and the learners themselves focus on the most important subjects and challenges. To achieve this, these questions will be answered by utilizing the material and using the methods outlined in the next section.

# 3 Material and methods

Section 3 outlines the data, tools, and methods used in this study. The materials section will introduce the corpora, participants, and the texts that were chosen to be analyzed, as well as explain the reason behind these choices. The methods section will present the tools used in the text and statistical analysis and explain the importance of those tools in the framework of *Natural Language Processing* (NLP), as well as describe the concrete process of the aforementioned analyses.

## 3.1 Material

The data for this study come from the YKI corpus by the University of Jyväskylä (YKI-Corpus 2017). YKI-Corpus is compiled by the University of Jyväskylä from the examinations of The National Certificates of Language Proficiency (*Yleiset kielitutkinnot*). The certificate is a language testing system aimed at adults, and it is independent from any curriculum or syllabus (YKI-Corpus 2017). The examination is offered in many

languages; however, the present study is interested in Finnish native speakers taking the exam in English. The exam comes in three proficiency levels, all of which are further divided into two sublevels: these are the Basic level (1–2), the Intermediate level (3–4), and the Advanced level (5–6) (ibid.). The levels 1–6 correlate with the *Common European Framework of Reference (CEFR)* levels A1–C2. The examination tests four different aspects of language proficiency: reading comprehension, listening comprehension, speaking, and writing. Because the present study is concerned with written language production, all examinations without a written performance are ignored. The written performances include three essays per participant, with various topics such as informal letters, formal letters, application forms, letters to editors, Curricula Vitae, or argumentative pieces. In all topic choices, the most important indicator of success in each task is effectively conveying the message (YKI-Corpus 2017). New material is added to the YKI-Corpus after every testing round, and the current corpus is divided into old (up to 2010) and new (from 2011 onwards) material (ibid.). This study focuses on the new material, as it can provide the most up-to-date view on the topic.

In addition to the qualitative data, i.e. the essays by the examination participants, the corpus also provides ample quantitative data: background information of the participants, such as age, gender, education, and socioeconomic status; the level and grade of the certificate, as well as the participant's self-assessment; the reason for taking the examination, for example demonstrating language proficiency for work or studies; and self-reported frequency of use of the examination language in different situations, such as "I speak the test language with my family" from "Not at all" to "Almost daily". However, filling in the background information is not mandatory and is therefore not available for every participant. This study discards participants without complete background information, in the interest of the best, most comprehensive comparative analysis of different learners.

Both the intermediate level and the advanced level texts come from the YKI-Corpus. Following Zoltan Dörnyei's (2007, 309) principles of practicality and cost-effectiveness, both groups will include 15 learners, each with three essays of approximately 100–200 words each. This number is a compromise between a healthy amount of text to provide a good basis for analysis, and the limited resources for

conducting said analysis, in case of an extended amount of manual analysis is required to achieve reliable results. The sample of 15 learners per proficiency level was selected by picking the first 15 hits from the list generated by the search function of the YKI-Corpus by the following criteria: test in English, test level intermediate/advanced, a passing grade, native language Finnish, age range 18–25. These criteria were chosen to have the two groups be as comparable with each other as possible, excluding their competence level in English. The downside of such search criteria is that the sample of learners became rather homogenous, and thus cannot reliably be considered to represent the whole population. However, one could argue that young adults are one of the premier groups of English learners and users amongst the Finnish population, and that a contemporary review of their language use probably guides us on how Finnish speakers' use of English is developing in the future.

As mentioned, the sample of learners whose texts qualified to be analyzed in the present study is rather homogenous, but the YKI-Corpus' background information reveals a number of noteworthy differences that should be highlighted. All of these are displayed in a table in the appendices section. Out of the intermediate learners, only two are women, 13 are men. For advanced learners the ratio is more balanced, seven women to eight men. Ages of the participants range from 19 to 25 with a mean of 21.1 for the intermediate group, 23.6 for the advanced group. Of the intermediate learners, 13 had mentioned upper secondary school as their basic education, two had reported university level, whereas the ratio is five to ten, respectively, for the advanced group. All but one of the participants reported having studied English either 7–9 years or 10+ years, and as that is very common within the Finnish basic education system, it can be assumed that the one participant with missing information falls somewhere around the nine-year mark. All of the learners had studied English at school, and five participants mentioned they had studied English abroad as well. Finally, the learners differed in their reasons for acquiring the language certificate: 26 of them report taking the test to apply for a job, and six of them need the certificate to prove their language competence in their current job. Two of the learners also mention study purposes as a reason, and three participants took the test to get feedback on their English. As mentioned earlier in this section, there is more background data available in the YKI-Corpus, such as where and how often the

learners use English, but as that is not the focus of this study, these differences are not presented here.

In the present study, only the texts from the YKI-Corpus will be analyzed. However, in order to cover research question two, some sort of a comparison is required. To serve as a point of comparison, Peter Master's (2013) study of determiner use in research articles in the field of science and technology will be used. Master studied texts of researches in e.g. biology, physics, and computer science using many methods similar to the ones the present study will as well, such as a comparable categorization of determiners, and *Natural Language Processing*, which will be more thoroughly defined in the next section of this study. Master's study also provides comparable metrics, such as percentages of different determiners types in relation to the total word count as well as total determiner count. An overview of Master's results was shown in Section 2.3. However, Master adds the number of zero articles into the total word count to accommodate for the fact that zero articles do not show in the total word count otherwise, which will not be the case for the present study. Furthermore, in the texts Master examined, the genre is strictly "research article" regardless of the field of study, whereas the genres in the texts from the YKI-Corpus are varied and distinctly different from an academic text. However, while the YKI essay topics vary in formality, none are pure fiction, which would be the genre opposite of research articles in terms of objectivity, which could cause problems in comparing determiner use. For example, in terms of determiners per word, the total word could bloat in a work of fiction due to increased figurativeness, diluting the ratio of determiners in the text. The gap in style between the YKI-Corpus essays and the research articles should be negligible enough to allow for fair comparison. Another challenge with using Master's study as a point of comparison is that not all of the authors are native English speakers. However, as most of them are (either North American or Indian), and as one could expect peer-reviewed and published articles to be proofread, and thus contain flawless language, Master's study can be deemed to provide results for native-level English.

## 3.2 Methods

This study analyzes the material introduced in section 3.1, texts written in English by Finnish-speaking learners. This analysis is achieved by *Natural Language Processing*, or NLP, which is the use of computers or artificial intelligence to analyze and understand human languages (Crossley et al. 2014; Bird, Klein and Loper 2009). Compared to a fully manual text analysis, NLP can be used to parse a much larger amount of text quicker, which allows for a larger sample size of text. The primary language processing tool used in the present study is the *Simple Natural Language Processing Tool* (SiNLP) which is a tool written in the programming language *Python* (Crossley et al. 2014). SiNLP was chosen for two reasons: firstly, because of its overall simplicity in comparison to other language processing tools, as for the analysis in this study, using an unnecessarily complex tool would be a waste of resources. Secondly, I myself am somewhat familiar with Python, which makes modifying the tool for the needs of this study much easier. However, one shortcoming of SiNLP is its incapability of part-of-speech tagging, which means it cannot distribute words into different categories and thus, cannot recognize parts-of-speech like nouns, for example. This is a problem, as without part-of-speech tagging the case of a zero article as a determiner would be impossible to include without extensive manual analysis, yielding a skewed picture of the learners' determiner use. For this reason, SiNLP was supplemented by *spaCY*, an open-source Python extension library, which allowed for a complete analysis, including instances of a zero article.

Two primary factors make SiNLP simple to use: it operates on simple file types without the need for much computation, and it relies on no external or third-party databases. The tool takes text input in plain text (.txt files) and outputs the parse results as comma separated values file (.csv), which essentially is a plain text file transformed into a table, easily convertible into a Microsoft Excel spreadsheet (.xls file). These file type requirements together with SiNLP functioning as a standalone tool and needing very little computational power make the analysis process quick on a modern device. In addition to requiring no extra tools to be run simultaneously, SiNLP operates on no external databases: simply a user-created dictionary as a text file suffices. This also eliminates the need for an internet connection, which reduces the number factors that could hinder the analysis process.

Despite all of its simplicity, choosing SiNLP presented a few challenges. Originally the tool was created for the purposes of discourse analysis and its functionality has been tested primarily in predicting essay quality by finding linguistic features in essays (Crossley et al. 2014, 515-517). As the present study is concerned with a specific linguistic subsystem, determiners, many of the metrics provided by SiNLP by default are irrelevant, as they relate to the text as a whole, assess the quality and cohesion of the text, or are too broad or imprecise. These metrics include, for instance, the number of sentences or paragraphs, the number of unique words, word frequency, type-to-token ratio, or the mean number of words before the main verb. In addition to the unnecessary metrics, the default dictionary in SiNLP is almost entirely unsuitable for the purposes of the present study, as even though it does contain some of the determiners in the English language, it is mainly built to parse references and anaphor use (Crossley et al. 2014, 518-519).

The analysis using SiNLP proceeded as follows: first, the essays were formatted in plain text, combining all three essays per writer into one .txt file. Next, the text files were placed in a folder, which serves as the input folder for SiNLP. An output folder was also generated to house the analysis results. Finally, a text file to serve as the dictionary had to be created. As established before, the default dictionary was insufficient. The new dictionary file was categorized into all the different determiner types outlined in Table 1, with two exceptions: the zero article types, because that is outside of the capabilities of SiNLP, and the genitive determiners, which were manually checked and added to the possessive determiners category, as those were very difficult to separate from other "apostrophe + s" structures in the SiNLP analysis. In addition to the determiner categories, a number of exception categories had to be created in order to exclude counting any determiner type words serving a different function in the text. Examples of exceptions include sentence final positions (*That is **his.** Can you see **that?***), positions linked to an auxiliary or copular verb (***This could** be bad. **What is** it?*), and positions linked to a preposition (*He knows **much about** it. **One of** the guys.*). With the input folder and the dictionary complete, the analysis was run. SiNLP outputs the results of the analysis as a .csv file, which was then converted into a .xls spreadsheet, which was then formatted as follows: the aforementioned irrelevant metrics were removed, leaving only the total number of words per writer. The number of exceptions was then

subtracted from the total number of determiners in the corresponding category (for example, demonstrative exceptions were removed from the total number of demonstrative determiners). That concluded the SiNLP portion of the analysis process, but as noted before, the cases of a zero article used as a determiner were still missing. To include those and complete the analysis, another tool was needed to supplement SiNLP. A Python extension library *spaCY* was chosen for that purpose, and spaCY's setup and analysis will be presented next.

spaCY is an extension library for the programming language Python, and it is designed to be used in Natural Language Processing. Compared to SiNLP, spaCY offers many different options and features for text processing and analysis, such as name recognition, phrase recognition and dependency parsing, visualization and, most importantly for the present study, part-of-speech tagging. Despite its multiple functions, spaCY lacks in ease-of-use, speed, and providing readily reportable results – all of which are reasons why SiNLP was chosen as the primary tool for this study's analysis. However, as discussed previously, spaCY was needed to fill in the gaps left by SiNLP.

At the time of the analysis in early 2020, the latest version of both Python and spaCY were used; Python 3.7.6 and spaCY 2.2.3. spaCY was downloaded and installed through Python's package manager *pip*. All of the code for the text parsing and analysis was written in Python's shell window, which is an interactive mode that waits for the user's input and executes one line of code at a time instead of entire chunks, which allows for trial and error, and quick debugging of the code. As was the case with SiNLP, spaCY required the essays in plain text, as .txt files. The texts were then input and read by spaCY, which separates every word in the text as an independent unit and assigns the words their relevant attributes, for example, which part-of-speech they represent. From this list, all noun phrases were selected and separated into a new list, which was further refined by removing every but the head of each noun phrase. This was done in order to produce a list of words that can be converted from spaCY's attributed entities into strings, which are essentially countable, plain text words. This process has now provided the number of noun phrases in each writer's essays. The number can now be compared with the number of determiners in the essays, calculated by the SiNLP analysis: the remainder is the number of noun phrases with zero article.

To conclude the analysis and to finalize the results spreadsheet, a manual check was still needed. The essays were copied into Microsoft Word to utilize the program's spell checker. First, as mentioned previously, all genitive determiners were counted and added to the possessive determiners column. Next, all spelling errors were checked, and if the nature of the error was such that it would cause miscalculations in the automated analysis, the numbers were manually adjusted in the relevant determiner categories in the results spreadsheet. An example of an error like this would be *taht* car*, where the determiner *that* would have gone unnoticed by SiNLP, as the program searches for exact matches for the word *that.* Finally, all 15 intermediate and 15 advanced writers were added together in the same spreadsheet with counted and adjusted determiner numbers, including the cases of zero article, total number of determiners, the percentage of each type of determiner for each writer, total number of words, and determiners per word ratio. However, in order to get a clearer picture of how the two groups of learners used determiners in their text, statistical analysis is required.

To conduct the statistical analysis of the data in the present study, the software *SPSS* version 25.0 by IBM was used. In addition to SPSS, Microsoft Excel was used to produce the preliminary spreadsheets, which were then be imported into SPSS for the analysis. The statistical analysis consisted of determining the *normality* and *variance* of the results by calculating means, variances and the *standard deviation* of the relevant data, using the Shapiro-Wilk test to determine the normality of each category's distribution. In addition to numeral calculation of normality, a graphical evaluation was used as a supplement due to the limited sample size of 30 participants. This was achieved by examining *quantile-quantile plots* drawn from the data using SPSS. Next, the two proficiency level groups were compared with both *parametric and non-parametric tests*, depending on whether the determiner category in question is normally distributed or not, to find any statistically significant differences between them. The tests chosen for the present study were the independent samples t-test as the parametric test, and its non-parametric counterpart Mann-Whitney U-test as the non-parametric test. These tests were the most suitable for comparing the means of two independent groups, in this case two groups of learners representing two proficiency levels of English. Finally, the *effect size* of the results was analyzed to further supplement the comparison of means, strengthen the validity of this study, and to better highlight

interesting results arisen from the comparison of the two groups. As SPSS does not have a built-in effect size function, these were calculated manually using the following formulas:

$$\text{Cohen's } d = (\text{mean}_1 - \text{mean}_2) / \sqrt{((\text{sd}_1^2 + \text{sd}_2^2) / 2)}$$
$$\text{Correlation coefficient } r = Z / \sqrt{N}$$

The use of these formulas will be explained further in the relevant parts of the results section. To conclude the statistical analysis, the *correlations* between different determiner categories were tested. This was done using Pearson's correlation and Spearman's correlation, a parametric and non-parametric correlation test, respectively. The significant correlations were also graphically presented for easier and viewing, and to supplement the numerical correlation results. All choices in terms of tests were made following the suggestions of Larson-Hall (2010) and Mackey and Gass (2005), and these choices will be further elaborated in the results section.

# 4 Results

This section will present the results of the study. First, the overall frequencies of determiners in the texts of both intermediate and advanced writers will be displayed. Next, the focus will be on the statistical analysis, presenting the results of comparing the two groups as well as highlighting any interesting details, should those emerge from the data. Finally, the results will be summarized, and their reliability and validity will be reviewed. All the results shown in this section will be further reviewed in the discussion section of this study.

## 4.1 Distribution of determiners

Table 2 shows the word and determiner count as well as the determiners per word ratio for all participants on both proficiency levels, intermediate (level B in all subsequent tables) and advanced (level C in all subsequent tables), as they are denoted in both the YKI-Corpus and CEFR. All 30 participants have their three essays coalesced into a single text, and all the numbers henceforth will refer to that text, instead of any of the individual essays of a writer.

*Table 2: Total number of determiners*

| ID | level | words | determiners | determiners per word |
|---|---|---|---|---|
| 1001 | Intermediate | 297 | 38 | 0.13 |
| 1002 | Intermediate | 349 | 51 | 0.15 |
| 1003 | Intermediate | 400 | 48 | 0.12 |
| 1004 | Intermediate | 484 | 67 | 0.14 |
| 1005 | Intermediate | 405 | 65 | 0.16 |
| 1006 | Intermediate | 352 | 69 | 0.20 |
| 1007 | Intermediate | 445 | 77 | 0.17 |
| 1008 | Intermediate | 286 | 23 | 0.08 |
| 1009 | Intermediate | 339 | 53 | 0.16 |
| 1010 | Intermediate | 607 | 106 | 0.17 |
| 1011 | Intermediate | 513 | 83 | 0.16 |
| 1012 | Intermediate | 477 | 89 | 0.19 |
| 1013 | Intermediate | 441 | 56 | 0.13 |
| 1014 | Intermediate | 438 | 72 | 0.16 |
| 1015 | Intermediate | 395 | 61 | 0.15 |
| 2001 | Advanced | 603 | 129 | 0.21 |
| 2002 | Advanced | 532 | 93 | 0.17 |
| 2003 | Advanced | 601 | 124 | 0.21 |
| 2004 | Advanced | 440 | 80 | 0.18 |
| 2005 | Advanced | 485 | 95 | 0.20 |
| 2006 | Advanced | 582 | 111 | 0.19 |
| 2007 | Advanced | 493 | 77 | 0.16 |
| 2008 | Advanced | 633 | 93 | 0.15 |
| 2009 | Advanced | 443 | 66 | 0.15 |
| 2010 | Advanced | 446 | 84 | 0.19 |
| 2011 | Advanced | 330 | 56 | 0.17 |
| 2012 | Advanced | 467 | 88 | 0.19 |
| 2013 | Advanced | 584 | 122 | 0.21 |
| 2014 | Advanced | 323 | 56 | 0.17 |
| 2015 | Advanced | 417 | 74 | 0.18 |

In terms of word count, the texts range from 286 up to 633 words with a mean of 454 words; group B having a slightly lower average at 415 words compared to group C's 492 words. For the determiner count, the range goes from 23 to 129, the mean for all participants being 77 determiners. Again, the mean is lower in group B with 64, as group C reaches 90 determiners per participant on average. Determiners per word is the total determiner count divided by the total word count. The values range from 0.08 up to 0.21, and the means follow a similar trend as with the word and determiner count: 0.17 for the whole group, 0.15 and 0.18 for groups B and C, respectively.

*Table 3: Number of determiners by category I*

| ID | level | indefinite articles | indef% | definite articles | def% | zero article | zero% |
|---|---|---|---|---|---|---|---|
| 1001 | B | 5 | 13 % | 7 | 18 % | 6 | 16 % |
| 1002 | B | 10 | 20 % | 10 | 20 % | 6 | 12 % |
| 1003 | B | 6 | 13 % | 11 | 23 % | 6 | 13 % |
| 1004 | B | 7 | 10 % | 16 | 24 % | 9 | 13 % |
| 1005 | B | 11 | 17 % | 18 | 28 % | 10 | 15 % |
| 1006 | B | 6 | 9 % | 18 | 26 % | 10 | 14 % |
| 1007 | B | 10 | 13 % | 17 | 22 % | 10 | 13 % |
| 1008 | B | 1 | 4 % | 1 | 4 % | 7 | 30 % |
| 1009 | B | 7 | 13 % | 12 | 23 % | 10 | 19 % |
| 1010 | B | 16 | 15 % | 28 | 26 % | 11 | 10 % |
| 1011 | B | 7 | 8 % | 8 | 10 % | 17 | 20 % |
| 1012 | B | 10 | 11 % | 31 | 35 % | 14 | 16 % |
| 1013 | B | 14 | 25 % | 5 | 9 % | 12 | 21 % |
| 1014 | B | 5 | 7 % | 28 | 39 % | 5 | 7 % |
| 1015 | B | 17 | 28 % | 24 | 39 % | 2 | 3 % |
| 2001 | C | 24 | 19 % | 42 | 33 % | 26 | 20 % |
| 2002 | C | 11 | 12 % | 24 | 26 % | 20 | 22 % |
| 2003 | C | 12 | 10 % | 41 | 33 % | 26 | 21 % |
| 2004 | C | 6 | 8 % | 21 | 26 % | 19 | 24 % |
| 2005 | C | 9 | 9 % | 29 | 31 % | 25 | 26 % |
| 2006 | C | 8 | 7 % | 43 | 39 % | 32 | 29 % |
| 2007 | C | 8 | 10 % | 20 | 26 % | 11 | 14 % |
| 2008 | C | 8 | 9 % | 20 | 22 % | 22 | 24 % |
| 2009 | C | 11 | 17 % | 13 | 20 % | 16 | 24 % |
| 2010 | C | 10 | 12 % | 17 | 20 % | 19 | 23 % |
| 2011 | C | 5 | 9 % | 5 | 9 % | 11 | 20 % |
| 2012 | C | 15 | 17 % | 20 | 23 % | 18 | 20 % |
| 2013 | C | 20 | 16 % | 29 | 24 % | 32 | 26 % |
| 2014 | C | 8 | 14 % | 17 | 30 % | 16 | 29 % |
| 2015 | C | 10 | 14 % | 27 | 36 % | 17 | 23 % |

Table 3 shows the number of determiners in the first three determiner categories for each participant, as well as the percentage of the total determiner count for all the types. The first three categories consist of the three different types of articles in the English language: indefinite, definite, and zero article. The mean for indefinite articles across both proficiency levels is 13 % of all determiners, 25 % for definite articles, and 19 % for zero articles. When added together, articles make for a large portion of the total determiners used in the essays: 57 % on average (52 % for group B, 62 % for C). Comparing the two proficiency levels, the means for indefinite articles are 14 % for

group B and 12 % for group C, and for definite articles, 23 % for group B and 26 % for group C. The largest difference was in the zero article category, in which group B's number was 15 % compared to group C's 23 %, on average.

*Table 4: Number of determiners by category II*

| ID | level | possessives | poss% | demonstratives | dem% | quantifiers | quant% |
|---|---|---|---|---|---|---|---|
| 1001 | B | 12 | 32 % | 2 | 5 % | 3 | 8 % |
| 1002 | B | 10 | 20 % | 10 | 20 % | 5 | 10 % |
| 1003 | B | 16 | 33 % | 4 | 8 % | 5 | 10 % |
| 1004 | B | 18 | 27 % | 8 | 12 % | 8 | 12 % |
| 1005 | B | 9 | 14 % | 7 | 11 % | 6 | 9 % |
| 1006 | B | 14 | 20 % | 10 | 14 % | 4 | 6 % |
| 1007 | B | 16 | 21 % | 12 | 16 % | 7 | 9 % |
| 1008 | B | 7 | 30 % | 4 | 17 % | 2 | 9 % |
| 1009 | B | 9 | 17 % | 7 | 13 % | 8 | 15 % |
| 1010 | B | 25 | 24 % | 13 | 12 % | 7 | 7 % |
| 1011 | B | 28 | 34 % | 9 | 11 % | 5 | 6 % |
| 1012 | B | 14 | 16 % | 13 | 15 % | 3 | 3 % |
| 1013 | B | 9 | 16 % | 9 | 16 % | 3 | 5 % |
| 1014 | B | 22 | 31 % | 9 | 13 % | 3 | 4 % |
| 1015 | B | 4 | 7 % | 10 | 16 % | 4 | 7 % |
| 2001 | C | 19 | 15 % | 8 | 6 % | 6 | 5 % |
| 2002 | C | 19 | 20 % | 11 | 12 % | 7 | 8 % |
| 2003 | C | 22 | 18 % | 12 | 10 % | 9 | 7 % |
| 2004 | C | 16 | 20 % | 10 | 13 % | 4 | 5 % |
| 2005 | C | 14 | 15 % | 7 | 7 % | 9 | 9 % |
| 2006 | C | 17 | 15 % | 7 | 6 % | 3 | 3 % |
| 2007 | C | 14 | 18 % | 14 | 18 % | 10 | 13 % |
| 2008 | C | 10 | 11 % | 10 | 11 % | 18 | 19 % |
| 2009 | C | 15 | 23 % | 5 | 8 % | 6 | 9 % |
| 2010 | C | 12 | 14 % | 15 | 18 % | 7 | 8 % |
| 2011 | C | 15 | 27 % | 12 | 21 % | 5 | 9 % |
| 2012 | C | 20 | 23 % | 9 | 10 % | 5 | 6 % |
| 2013 | C | 19 | 16 % | 16 | 13 % | 5 | 4 % |
| 2014 | C | 4 | 7 % | 5 | 9 % | 4 | 7 % |
| 2015 | C | 5 | 7 % | 6 | 8 % | 7 | 9 % |

Table 4 presents the next three determiner categories: possessive and demonstrative determiners, as well as quantifiers. Possessive determiners cover, on average, 20 % of all determiners for all 30 participants, whereas the number is 12 % for demonstrative determiners, and 8 % for quantifiers. For group B, the percentages are 23 % for

possessive and 13 % for demonstrative determiners, 8 % for quantifiers, and for group C, 17 % for possessive and 11 % for demonstrative determiners, 8 % for quantifiers.

*Table 5: Number of determiners by category III*

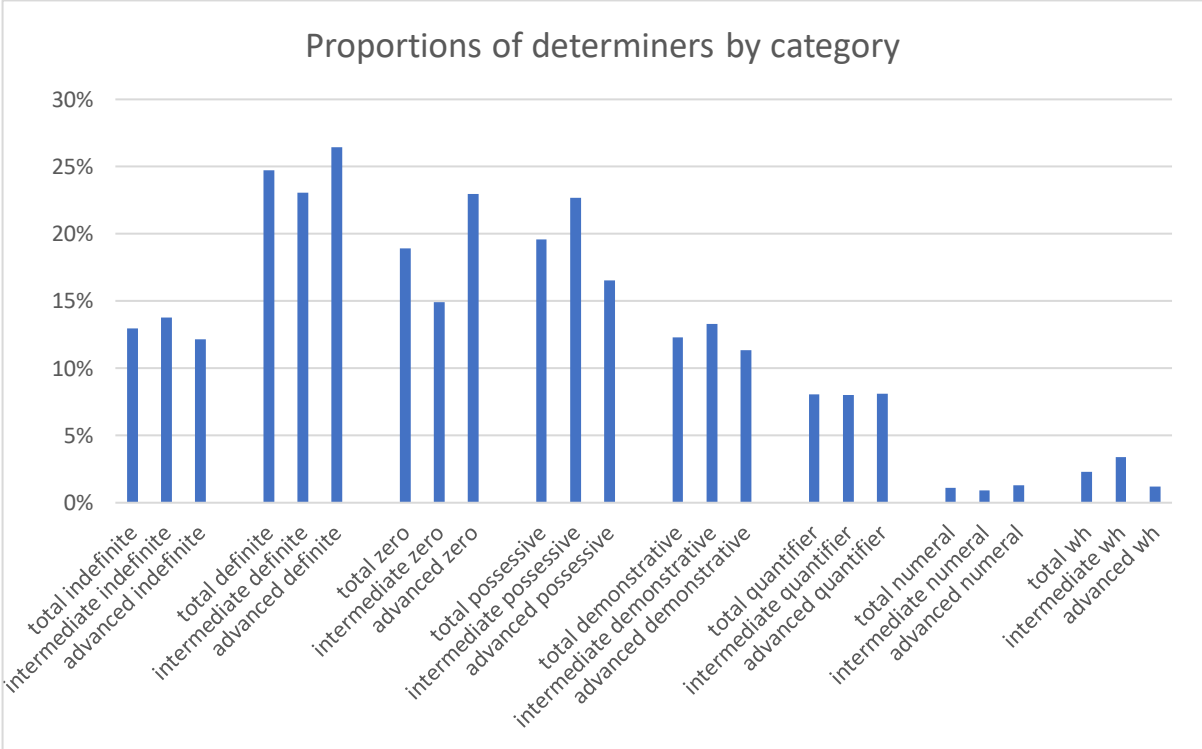
| ID | level | numerals | num% | wh-words | wh% |
|---|---|---|---|---|---|
| 1001 | B | 0 | 0 % | 3 | 8 % |
| 1002 | B | 0 | 0 % | 0 | 0 % |
| 1003 | B | 0 | 0 % | 0 | 0 % |
| 1004 | B | 1 | 1 % | 0 | 0 % |
| 1005 | B | 2 | 3 % | 2 | 3 % |
| 1006 | B | 0 | 0 % | 7 | 10 % |
| 1007 | B | 0 | 0 % | 5 | 6 % |
| 1008 | B | 1 | 4 % | 0 | 0 % |
| 1009 | B | 0 | 0 % | 0 | 0 % |
| 1010 | B | 3 | 3 % | 3 | 3 % |
| 1011 | B | 1 | 1 % | 8 | 10 % |
| 1012 | B | 1 | 1 % | 3 | 3 % |
| 1013 | B | 0 | 0 % | 4 | 7 % |
| 1014 | B | 0 | 0 % | 0 | 0 % |
| 1015 | B | 0 | 0 % | 0 | 0 % |
| 2001 | C | 3 | 2 % | 1 | 1 % |
| 2002 | C | 1 | 1 % | 0 | 0 % |
| 2003 | C | 0 | 0 % | 2 | 2 % |
| 2004 | C | 4 | 5 % | 0 | 0 % |
| 2005 | C | 2 | 2 % | 0 | 0 % |
| 2006 | C | 0 | 0 % | 1 | 1 % |
| 2007 | C | 0 | 0 % | 0 | 0 % |
| 2008 | C | 0 | 0 % | 5 | 5 % |
| 2009 | C | 0 | 0 % | 0 | 0 % |
| 2010 | C | 2 | 2 % | 2 | 2 % |
| 2011 | C | 3 | 5 % | 0 | 0 % |
| 2012 | C | 1 | 1 % | 0 | 0 % |
| 2013 | C | 0 | 0 % | 1 | 1 % |
| 2014 | C | 0 | 0 % | 2 | 4 % |
| 2015 | C | 0 | 0 % | 2 | 3 % |

Finally, Table 5 shows the number and percentages for the last two determiner categories: numerals and wh-words. These two are smallest of all the categories, consisting only 1 % and 2 % of the total number of determiners across all participants for numerals and wh-words, respectively. For proficiency level B, the mean for the

percentage of numerals was 1 %, and for wh-words, 3%. For level C, both categories totalled to 1 % on average.

To conclude, Figure 1 below shows all determiner categories and their percentage of the total determiner count for the whole participant group, as well as both proficiency levels separately. In the next section, the results of the statistical analysis of the determiner data will be presented.

*Figure 1: Proportions of determiners by category*




## 4.2 Statistical analysis

As mentioned previously, the statistical analysis for the present study was done in SPSS. First, the data were tested for normality using two different tests: the Kolmogorov–Smirnov test and the Shapiro–Wilk test. Of these two tests, the Shapiro–Wilk test is considered to be a better fit for sample sizes under 50 (Larson-Hall 2010, 84), and for this reason, it was chosen as the primary test for normality in the present study. Table 6 below shows the results for these tests.

*Table 6: Tests of normality*

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| indefinite articles % | .124 | 30 | .200* | .932 | 30 | .055 |
| definite articles % | .117 | 30 | .200* | .953 | 30 | .206 |
| zero article % | .109 | 30 | .200* | .977 | 30 | .756 |
| possessives % | .105 | 30 | .200* | .957 | 30 | .260 |
| demonstratives % | .082 | 30 | .200* | .977 | 30 | .753 |
| quantifiers % | .145 | 30 | .109 | .920 | 30 | .027 |
| numerals % | .322 | 30 | .000 | .733 | 30 | .000 |
| wh-words % | .239 | 30 | .000 | .761 | 30 | .000 |
| determiners per word | .091 | 30 | .200* | .961 | 30 | .320 |

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The results in Table 6 suggest that for the first five determiner categories (from indefinite articles to demonstratives) as well as for the determiners per word category, the data are normally distributed (*Sig.* > 0.05) across all 30 participants. For the two smallest categories, numerals and wh-words, the opposite appears to be true, as can be expected due to the low number of determiners in these categories in total. As for quantifiers, the Kolmogorov–Smirnov test suggests the data to be normally distributed, however, the Shapiro–Wilk test does not. Thus, that category appears not normally distributed. In addition to numerical data, graphics should be examined as well to achieve a more robust sense of normality in data (Larson-Hall 2010, 84). For this reason, quantile-quantile plots (Q-Q plots) were drawn of all the categories to supplement the formal normality calculations with a visual check. These Q-Q plots are shown in Figure 2. These graphs strengthen the conclusions drawn from numerical data: the Q-Q plot for quantifiers seems to be skewing right, as the plots at the top end of the graph deviate from the expected normal, whereas for numerals and wh-words, the plots are altogether too scattered to be considered following the line of a normal distribution. The top end of the graph shows two outliers in the indefinite articles category as well, which is in line with the numerical data: the category was just barely within the boundaries of normality (*Sig.* = 0.055). For the rest of the categories, the graphical representation seems to be in agreement with the formal data, and they can be considered normally distributed for the purposes of statistical analysis going forward in this study.

*Figure 2: Q-Q plots of determiner category distribution*

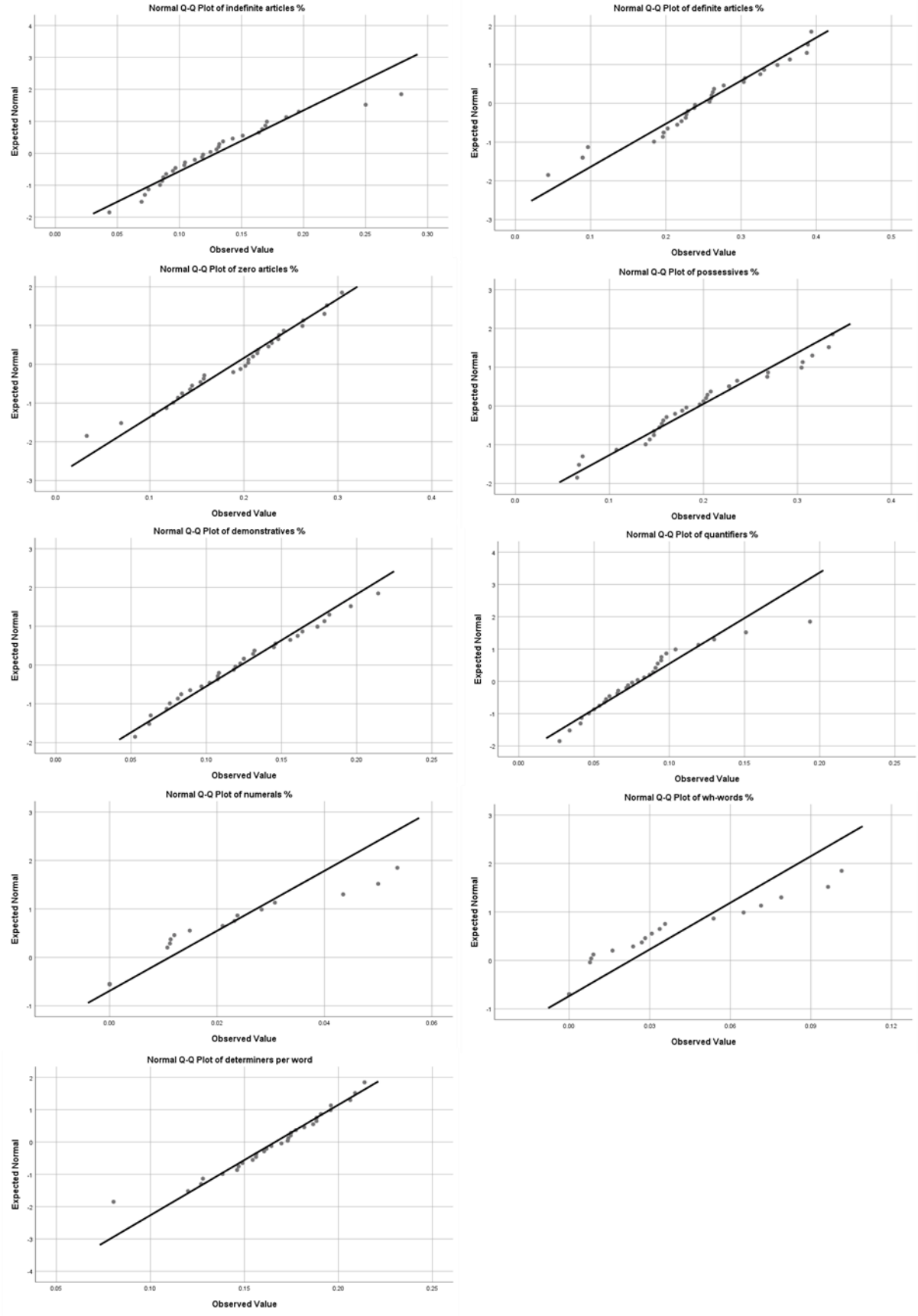


*Table 7: T-test for articles*

**Group Statistics**

| | level | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| indefinite articles % | B | 15 | .137634627287 | .064400090469 | .016628031858 |
| | C | 15 | .121347203212 | .037417500412 | .009661157063 |
| definite articles % | B | 15 | .230479675404 | .102170100565 | .026380206531 |
| | C | 15 | .264419062907 | .075689525399 | .019542951423 |
| zero articles % | B | 15 | .149263621464 | .063971374627 | .016517337904 |
| | C | 15 | .229464791870 | .037533392153 | .009691080182 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | |
|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) |
| indefinite articles % | Equal variances assumed | 1.863 | .183 | .847 | 28 | .404 |
| | Equal variances not assumed | | | .847 | 22.485 | .406 |
| definite articles % | Equal variances assumed | .644 | .429 | -1.034 | 28 | .310 |
| | Equal variances not assumed | | | -1.034 | 25.810 | .311 |
| zero articles % | Equal variances assumed | 1.699 | .203 | -4.188 | 28 | .000 |
| | Equal variances not assumed | | | -4.188 | 22.618 | .000 |

| | | | | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|
| | | Mean Difference | Std. Error Difference | Lower | Upper |
| indefinite articles % | Equal variances assumed | .0162874240743 | .0192309489964 | -.02310538 | .055680237 |
| | Equal variances not assumed | .0162874240743 | .0192309489964 | -.02354527 | .056120127 |
| definite articles % | Equal variances assumed | -.033939387502 | .0328305078698 | -.10118963 | .033310859 |
| | Equal variances not assumed | -.0339393875029 | .0328305078698 | -.10144768 | .033568910 |
| zero articles % | Equal variances assumed | -.080201170406 | .0191504435079 | -.11942907 | -.04097326 |
| | Equal variances not assumed | -.0802011704060 | .0191504435079 | -.11985397 | -.04054836 |

With the normality of the data established, the proficiency level groups B and C can be compared. For this purpose, an independent-samples t-test was chosen for suitable categories, i.e. the three different article types, as well as possessives and demonstratives. An independent-samples t-test was conducted to see if Finnish-speaking learners of English at two different levels of proficiency use determiners differently in their written English. Table 7 shows the results of the t-test for the three article categories of determiners. In this two-tailed test, an alpha value of 5 % was used, as it is the common and preferred value in linguistic research (Larson-Hall 2010, Mackey

and Gass 2005). According to Levene's test for equality of variances, for all three article types, the equality of variances can be assumed (*Sig.* > 0.05). For indefinite articles, group B's mean is 0.14, standard deviation (sd) is 0.06, and the number of samples (N) is 15. Group C's mean = 0.12, sd = 0.04, N = 15. The 95 % confidence interval (CI) for the difference in means is from -0.02 to 0.06, t-value (*t*) is 0.847, p-value (*p*) is 0.404, and the degree of freedom (df) is 28. This means that there is no statistically significant difference in means of the two groups using indefinite articles (*p* > 0.05), and that the confidence interval is very narrow regardless, stretching only slightly around zero, which would indicate that the result is relatively precise and that the difference is minor. Both proficiency levels, B and C, seem to use indefinite articles to a similar capacity.

The next category to observe is the definite articles. Group B's mean = 0.23, sd = 0.10, N = 15, and group C's mean = 0.26, sd = 0.08, N = 15. The 95 % CI is from -0.10 to 0.03, *t* = -1.034, *p* = 0.310, df = 28. As with indefinite articles, there is not statistical significance in the difference in means (*p* > 0.05), however, the confidence interval is much wider, indicating higher potential disparity, but also a higher probability of a sampling error. Regardless, it cannot be considered a trustworthy result for the purposes of this study, due to lower significance. The final determiner category of the article variety is zero articles. Here, group B's mean = 0.14, sd = 0.06, N = 15, whereas group C's mean = 0.23, sd = 0.04, N = 15. The 95 % confidence interval for the difference in means ranges from -0.12 to -0.04, *t* = -4.188, *p* < 0.01, df = 28. For the zero article category, the low p-value suggests that there is a statistically significant difference in how the two proficiency groups differ in their use of the zero article type determiner. Furthermore, the 95 % CI only covering negative numbers would indicate that there is an increase in zero article use in group C's mean, compared to group B. Combined with a noticeable difference in means (-0.08) and a relatively low difference in standard deviation (0.02), the statistical analysis suggests that intermediate level Finnish-speaking learners of English use fewer zero articles in their writing than their advanced level counterparts, on average.

*Table 8: T-test for possessive and demonstrative determiners*

**Group Statistics**

| | level | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| possessives % | B | 15 | .226634322358 | .081791543337 | .021118485680 |
| | C | 15 | .165257831582 | .056336976989 | .014546144910 |
| demonstratives % | B | 15 | .132846014381 | .036585694599 | .009446385726 |
| | C | 15 | .113370892445 | .046143279117 | .011914143437 |

**Independent Samples Test**

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | |
|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) |
| possessives % | Equal variances assumed | 3.318 | .079 | 2.393 | 28 | .024 |
| | Equal variances not assumed | | | 2.393 | 24.843 | .025 |
| demonstratives % | Equal variances assumed | .708 | .407 | 1.281 | 28 | .211 |
| | Equal variances not assumed | | | 1.281 | 26.616 | .211 |

| | | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
|---|---|---|---|---|---|
| | | | | Lower | Upper |
| possessives % | Equal variances assumed | .06137649 | .0256433377 | .008848494 | .11390448 |
| | Equal variances not assumed | .06137649 | .0256433377 | .008546156 | .11420682 |
| demonstratives % | Equal variances assumed | .01947512 | .0152046380 | -.01167016 | .05062041 |
| | Equal variances not assumed | .01947512 | .0152046380 | -.01174328 | .05069357 |

In addition to different article types, possessive and demonstrative determiners were suitable for an independent samples t-test. The results for these two categories are presented in Table 8. As with the articles, Levene's test for equality of variances shows that variances between groups B and C for these two determiner types can be assumed equal ($p > 0.05$). For possessive determiners, group B's mean = 0.23, sd = 0.08, N = 15, and for group C, mean = 0.16, sd = 0.06, N = 15. The 95 % confidence interval goes from 0.01 to 0.11, $t = 2.393$, $p = 0.024$, df = 28. Thus, possessive determiners seem to be another category that displays a significant difference between the two intermediate and advanced levels ($p < 0.05$). The CI staying above zero would suggest that the participants in group B use more possessive determiners than the members of group C, however the difference could be as little as less than 0.01, or 1 %. To conclude the categories analysed using a t-test, demonstrative determiners presented the following numbers: group B's mean = 0.13, sd = 0.04, N = 15, and group C's mean = 0.11, sd = 0.05,

N = 15. Comparing the means of the two groups, the 95 % CI ranges from -0.01 to 0.05, $t = 1.281$, $p = 0.211$, df = 28. The demonstratives category shows no significant difference between the means, and the confidence interval follows suit: it shows only minor divergences from zero. Demonstrative determiners appear to be used in a similar fashion across proficiency levels.

*Table 9: Mann-Whitney test*

**Ranks**

| | level | N | Mean Rank | Sum of Ranks |
|---|---|---|---|---|
| quantifiers % | B | 15 | 15.90 | 238.50 |
| | C | 15 | 15.10 | 226.50 |
| | Total | 30 | | |
| numerals % | B | 15 | 14.93 | 224.00 |
| | C | 15 | 16.07 | 241.00 |
| | Total | 30 | | |
| wh-words % | B | 15 | 17.23 | 258.50 |
| | C | 15 | 13.77 | 206.50 |
| | Total | 30 | | |

**Test Statistics**[a]

| | quantifiers % | numerals % | wh-words % |
|---|---|---|---|
| Mann-Whitney U | 106.500 | 104.000 | 86.500 |
| Wilcoxon W | 226.500 | 224.000 | 206.500 |
| Z | -.249 | -.390 | -1.138 |
| Asymp. Sig. (2-tailed) | .803 | .697 | .255 |
| Exact Sig. [2*(1-tailed Sig.)] | .806[b] | .744[b] | .285[b] |

a. Grouping Variable: level

b. Not corrected for ties.

As stated previously, not all determiner categories in the present study were a good fit for an independent samples t-test: quantifiers, numerals and wh-words had to be analysed in a different manner. The reason is two-fold: the number of occurrences in all three categories, especially the last two, was low compared to the others, and they are not normally distributed. To statistically analyse these categories, the Mann-Whitney U-test was chosen, because it is the closest non-parametric equivalent to a t-test as it compares two independent groups in a similar fashion, however, it does not use the mean of the group, but instead divides the data into mean ranks (Mackey and Gass 2005,

279; Larson-Hall 2010, 377). The results of the Mann-Whitney U -test are shown in Table 9. The determiner category of quantifiers displayed a mean rank (MR) of 15.90 for group B, and 15.10 for group C, with the number of participants (N) being 15 for both groups. The comparison of groups B and C showed a U-value (U) of 106.5, a Z-value (Z) of -0.249 and an exact (one-tailed, multiplied by two) significance, or p-value (*p*), of 0.806. It appears that there exists no statistically significant difference in how the two groups use quantifiers in their written language, as neither the p-value ($p > 0.05$) or the Z-value ($|Z| < 1.96$) suggest there is.

For numerals, the Mann-Whitney test output the following results: groups B's MR = 14.93, N = 15, and group C's MR = 16.07, N = 15. The comparison of the proficiency levels showed that U = 104.0, Z = -0.390, $p = 0.744$. Again, all indicators propose that there is no significant difference between the groups ($p > 0.05$), and that both groups use numerals as determiners in a very similar fashion. Finally, in the wh-words category, the numbers for group B were MR = 17.23, N = 15, and for group C, MR = 13.77, N = 15. The test's U = 86.5, Z = -1.138, $p = 0.285$. Following the trend of quantifiers and numerals, there is no statistically significant difference in this category either, despite the p-value being lower compared to the other two categories ($p = 0.285$).

Examining differences between two independent groups strictly from a viewpoint of statistical significance ultimately cannot provide the full picture of the matter: even if a difference exists, it could be too minor to be interesting enough for further study. Conversely, even a statistically insignificant difference could have potential for more research, if the difference seems large. In order to supplement the statistical significance analysis with information about the magnitude of the analyzed difference, the effect size had to be calculated. As mentioned in the methods section, this was done manually, as SPSS has no built-in function for these calculations, using the following formulas:

Cohen's $d = (\text{mean}_1 - \text{mean}_2) / \sqrt{((\text{sd}_1^2 + \text{sd}_2^2) / 2)}$
Correlation coefficient $r = Z / \sqrt{N}$

Cohen's *d* is used to measure the difference in means by standard deviation (sd in the formula), meaning that the result is the number of standard deviations the means differ from one another (Larson-Hall 2010, 115). As a guideline, Cohen proposed the following magnitudes for interpreting effect size: 0.2 = small, 0.5 = medium, 0.8 = large (Larson-

Hall 2010, 113). This study will use the same interpretation; however, it should be noted that these are not strict lines but instead rough directions on a scale of effect size. In addition, instances of Cohen's *d* greatly exceeding the top end of the scale (> 0.8) will be highlighted. The result can also be negative, which does not affect the scale, only the direction of the effect; if the first mean is larger, the result will be positive, and vice versa. Correlation coefficient (or relationship index) *r* measures the variance percentage of the statistic in question (Larson-Hall 2010, 377-378). This measure is used to determine the effect size of a Mann-Whitney U-test. The Z in the formula is the Z-value given by SPSS as a result for the Mann-Whitney test, and N is the total number of observations. The correlation coefficient will range from -1.0 to 1.0, with a smaller effect size close to zero, growing as it ventures further into either direction. The results of the effect size calculations are displayed in Tables 10 and 11 below.

*Table 10: Effect sizes for parametric test categories*

| | Group B | | | Group C | | | |
|---|---|---|---|---|---|---|---|
| | Mean | sd | N | Mean | sd | N | Cohen's d |
| Definite | 0.14 | 0.06 | 15 | 0.12 | 0.04 | 15 | 0.31 |
| Indefinite | 0.23 | 0.10 | 15 | 0.26 | 0.08 | 15 | -0.38 |
| Zero | 0.15 | 0.06 | 15 | 0.23 | 0.04 | 15 | -1.53 |
| Possessives | 0.23 | 0.08 | 15 | 0.17 | 0.06 | 15 | 0.87 |
| Demonstratives | 0.13 | 0.04 | 15 | 0.11 | 0.05 | 15 | 0.47 |

*Table 11: Effect sizes for non-parametric test categories*

| | Z | N | r |
|---|---|---|---|
| Quantifiers | -0.249 | 30 | -0.05 |
| Numerals | -0.390 | 30 | -0.07 |
| Wh-words | -1.138 | 30 | -0.21 |

Table 10 shows the mean, standard deviation and number of observations (rounded to two decimal points) for both proficiency levels in every parametric test category, as well as the effect size, Cohen's *d*. As mentioned before, *d* can be positive or negative, depending on whether the first or second group's mean is higher. The difference in means for definite and indefinite articles was 0.31 and 0.38 standard deviations, respectively. Participants at proficiency level C seemed to use fewer definite articles but more indefinite articles in their writing than participants at level B. When these results

are placed on Cohen's suggested scale, they fall between small and medium effect. The zero article category did not follow the trend of the other two article groups: the associated *d* is -1.53. This result far exceeded Cohen's suggested scale, where a large effect begins at $d > 0.8$. Therefore, there seems to be a very large and noticeable increase in the use of zero article in group C's texts compared to group B; indeed, a plain visual analysis of Table 3 would suggest that the percentage of zero articles rises from group B's 10-19 % bracket into the 20s in group C. The possessive determiners category also displayed sizable effect: 0.87 decrease from group B to C. Following Cohen's guidelines, this is a large effect. Finally, demonstratives resulted in Cohen's $d = 0.47$; just slightly below medium effect size. In conclusion for parametric test effect sizes, none of the categories showed a negligible effect size; they ranged from small to very large, which suggests the results are potentially interesting, should there also exist a statistically significant difference in addition to a sizable one. Next, the effect sizes of non-parametric test will be discussed.

The correlation coefficients (*r*) for the Mann-Whitney U-tests are presented in Table 11. The Z-value is output in SPSS as a negative number in all three categories, regardless of which group possessed a higher mean rank (MR); as can be seen in the *r* formula, a negative Z also produces a negative *r*. Because of this, the direction of the effect size is found in the mean ranks in Table 9: an increase in the determiner type in question is towards the proficiency group with the higher MR. As mentioned before, *r* ranges from zero to one, or from no linear relation between the variables, to a perfect linear relation: higher the number, higher the effect size. Again, Cohen's guideline for the magnitude of *r* can be used here: 0.1 = small, 0.3 = medium, 0.5 = large (Larson-Hall 2010, 112). As this scale suggests, *r* is not directly comparable with Cohen's *d*, as they scale differently: *d* can exceed the value of 1, *r* cannot. However, they can be used to roughly determine whether the effect sizes are orders of magnitude apart or not. The determiner category of quantifiers had an *r* of -0.05, an increase of quantifiers in group B's texts. According to Cohen's scale, this is a very minor effect ($r < 0.1$). Numerals look very similar to quantifiers: a coefficient of -0.07, again a tiny increase, this time in favor of group C. The wh-words category, however, differs from the previous two: at an $r = -0.21$, there was a higher magnitude of increase in this type of determiner in group B's written English. As $0.1 < r < 0.3$, the effect is no longer minor, albeit still small.

In conclusion, this section presented the effect sizes calculations and results for both parametric and non-parametric tests comparing the participants of two proficiency levels, B and C. Four out of the eight determiner categories resulted in a small effect size, two were less than small, and two were large or very large effects. The discussion section of this study will explain how the differences in means between the numbers of the two groups and their effect sizes relate and how they can be interpreted in order to answer the question of how Finnish-speaking learners of English use determiners in their written language.

After comparing the means between members of proficiency levels B and C in all of the determiner categories, more information about how they use determiners can still be distilled from examining if using one type of determiner relates to the use of another. In order to study these relationships, correlations between the different categories must be calculated. As there are both parametric and non-parametric categories to compare, both Pearson's (parametric) and Spearman's (non-parametric) correlation were used. In addition to the determiner categories, the determiners per word ratio was also included in the correlation analysis to discover if an increase in determiner frequency is related to a growing number of a certain type of determiner. At the end of this section, the most interesting and significant of the correlations are presented graphically for a better visual representation of the examined phenomena.

Table 12 shows the results of the parametric correlation analysis. In addition to Pearson's correlation value ($r$) and significance ($p$), SPSS outputs the number of observations (N) of each category; as N = 30 in every case, these rows are omitted from the report in order to save space. The three categories with the most statistically significant occurrences of correlation are definite articles, possessives, and determiners per word: all of which appear to correlate with three other categories. Definite articles show a very significant negative correlation with possessives ($r$ = -0.510, $p$ = 0.004), a significant negative correlation with demonstratives ($r$ = -0.395, $p$ = 0.031), and a very significant positive correlation with determiners per word ($r$ = 0.558, $p$ = 0.001). This suggests that as the relative number of determiners in a learner's text increases, so does the relative number of definite articles, however, the increase in definite articles coincides with a decrease in demonstratives and possessives.

*Table 12: Parametric correlations*

| | | indefinite articles % | definite articles % | zero articles % | possessives % | demonstratives % | quantifiers % | numerals % | wh-words % | determiners per word |
|---|---|---|---|---|---|---|---|---|---|---|
| indefinite articles % | Pearson Correlation | 1 | .084 | -.338 | -.409* | .067 | -.121 | -.294 | -.019 | -.008 |
| | Sig. (2-tailed) | | .658 | .067 | .025 | .724 | .525 | .115 | .921 | .965 |
| definite articles % | Pearson Correlation | .084 | 1 | -.254 | -.510** | -.395* | -.254 | -.336 | -.257 | .558** |
| | Sig. (2-tailed) | .658 | | .175 | .004 | .031 | .176 | .069 | .170 | .001 |
| zero articles % | Pearson Correlation | -.338 | -.254 | 1 | -.194 | -.278 | -.025 | .209 | -.039 | .116 |
| | Sig. (2-tailed) | .067 | .175 | | .304 | .138 | .897 | .268 | .837 | .542 |
| possessives % | Pearson Correlation | -.409* | -.510** | -.194 | 1 | .030 | -.070 | .212 | .070 | -.438* |
| | Sig. (2-tailed) | .025 | .004 | .304 | | .873 | .714 | .261 | .714 | .016 |
| demonstratives % | Pearson Correlation | .067 | -.395* | -.278 | .030 | 1 | .109 | .269 | -.076 | -.208 |
| | Sig. (2-tailed) | .724 | .031 | .138 | .873 | | .566 | .151 | .689 | .270 |
| quantifiers % | Pearson Correlation | -.121 | -.254 | -.025 | -.070 | .109 | 1 | -.107 | -.071 | -.409* |
| | Sig. (2-tailed) | .525 | .176 | .897 | .714 | .566 | | .572 | .709 | .025 |
| numerals % | Pearson Correlation | -.294 | -.336 | .209 | .212 | .269 | -.107 | 1 | -.249 | -.027 |
| | Sig. (2-tailed) | .115 | .069 | .268 | .261 | .151 | .572 | | .185 | .888 |
| wh-words % | Pearson Correlation | -.019 | -.257 | -.039 | .070 | -.076 | -.071 | -.249 | 1 | -.018 |
| | Sig. (2-tailed) | .921 | .170 | .837 | .714 | .689 | .709 | .185 | | .927 |
| determiners per word | Pearson Correlation | -.008 | .558** | .116 | -.438* | -.208 | -.409* | -.027 | -.018 | 1 |
| | Sig. (2-tailed) | .965 | .001 | .542 | .016 | .270 | .025 | .888 | .927 | |

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

The possessive determiners category displays the aforementioned correlation with definite articles, but also a significant negative correlation with indefinite articles ($r$ = -0.409, $p$ = 0.025), and a significant negative correlation with determiners per word ($r$ = -0.438, $p$ = 0.016). The determiners per word category seems to correlate with definite articles and possessives, as mentioned, and quantifiers ($r$ = -0.409, $p$ = 0.025) as well. In terms of other noteworthy results, the zero articles category displays a moderate negative correlation with indefinite articles, even though the result is not statistically significant according to the 0.05 alpha level ($r$ = -0.338, $p$ = 0.067).

Table 13 presents the results of the non-parametric correlation analysis. Again, N = 30 for all categories, allowing for the removal of that information from the report. The correlation coefficient is reported as Spearman's rho ($\rho$) and the significance as $p$. The same categories can be highlighted in the non-parametric analysis as were in the parametric one as well: definite articles, possessives, and determiners per word. The definite articles category suggested a very significant negative correlation with the possessive determiners category ($\rho$ = -0.474, $p$ = 0.008) and an equally significant positive correlation with determiners per word ($\rho$ = 0.524, $p$ = 0.003). The possessive determiners category showed, in addition to definite articles one, a significant negative correlation with determiners per word ($\rho$ = -0.362, $p$ = 0.049). Finally, determiners per word seemed to correlate with the aforementioned definite articles and possessive determiners, but also with quantifiers to a very significant degree ($\rho$ = -0.524, $p$ = 0.003). The non-parametric correlation analysis also showed a number of instances of potential correlation that were, however, outside the boundaries of statistical significance, such as indefinite articles and possessives ($\rho$ = -0.336, $p$ = 0.069), zero articles and demonstratives ($\rho$ = -0.329, $p$ = 0.076), and definite articles and quantifiers ($\rho$ = -0.334, $p$ = 0.071).

To conclude the analysis of correlations in Finnish-speaking English learner's determiner use, Figure 3 presents every significant correlation found in the data as a scatterplot graph. The visual data highlights the observations already made from the numerical data: the strongest correlation seems to exist between definite articles and possessives, and the only significant positive correlation is between definite articles and determiners per word.

*Table 13: Non-parametric correlations*

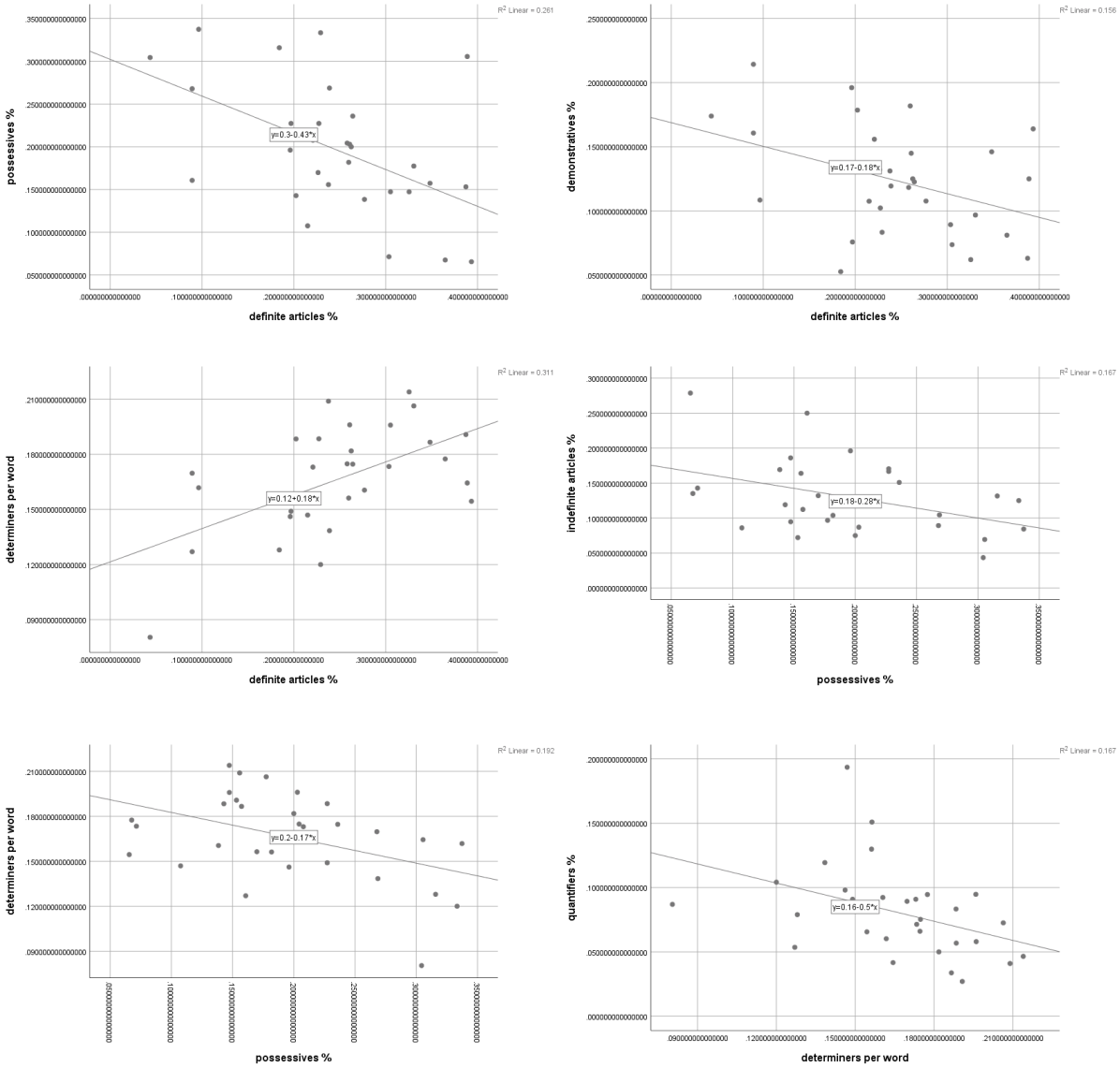
| | | indefinite articles % | definite articles % | zero articles % | possessives % | demonstratives % | quantifiers % | numerals % | wh-words % | determiners per word |
|---|---|---|---|---|---|---|---|---|---|---|
| indefinite articles % | Correlation Coefficient | 1.000 | -.012 | -.248 | -.336 | -.035 | .025 | -.198 | .028 | -.074 |
| | Sig. (2-tailed) | | .948 | .185 | .069 | .855 | .894 | .294 | .882 | .697 |
| definite articles % | Correlation Coefficient | -.012 | 1.000 | -.130 | -.474** | -.318 | -.334 | -.130 | -.065 | .524** |
| | Sig. (2-tailed) | .948 | | .492 | .008 | .087 | .071 | .494 | .734 | .003 |
| zero articles % | Correlation Coefficient | -.248 | -.130 | 1.000 | -.293 | -.329 | -.134 | .124 | .052 | .269 |
| | Sig. (2-tailed) | .185 | .492 | | .116 | .076 | .479 | .513 | .785 | .150 |
| possessives % | Correlation Coefficient | -.336 | -.474** | -.293 | 1.000 | .062 | .019 | .153 | -.167 | -.362* |
| | Sig. (2-tailed) | .069 | .008 | .116 | | .745 | .921 | .418 | .379 | .049 |
| demonstratives % | Correlation Coefficient | -.035 | -.318 | -.329 | .062 | 1.000 | .065 | .130 | -.147 | -.219 |
| | Sig. (2-tailed) | .855 | .087 | .076 | .745 | | .732 | .493 | .438 | .244 |
| quantifiers % | Correlation Coefficient | .025 | -.334 | -.134 | .019 | .065 | 1.000 | -.081 | -.237 | -.524** |
| | Sig. (2-tailed) | .894 | .071 | .479 | .921 | .732 | | .671 | .207 | .003 |
| numerals % | Correlation Coefficient | -.198 | -.130 | .124 | .153 | .130 | -.081 | 1.000 | -.187 | .153 |
| | Sig. (2-tailed) | .294 | .494 | .513 | .418 | .493 | .671 | | .323 | .419 |
| wh-words % | Correlation Coefficient | .028 | -.065 | .052 | -.167 | -.147 | -.237 | -.187 | 1.000 | .132 |
| | Sig. (2-tailed) | .882 | .734 | .785 | .379 | .438 | .207 | .323 | | .486 |
| determiners per word | Correlation Coefficient | -.074 | .524** | .269 | -.362* | -.219 | -.524** | .153 | .132 | 1.000 |
| | Sig. (2-tailed) | .697 | .003 | .150 | .049 | .244 | .003 | .419 | .486 | |

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

As a summary of the whole Section 4, it began by displaying the distribution of the different types of determiners for all 30 English learners in the present study in section 4.1, then proceeded to the statistical analysis in section 4.2 by presenting the results of tests of normality, mean and mean rank comparisons, and effect sizes between the two groups of participants, before finally concluding in the correlation analysis between the different determiner categories. How these results are interwoven, how they compare to earlier research on the topic, and ultimately, how and whether or not they answer the research questions of the present study will be discussed next in Section 5.

*Figure 3: Graphs of significant correlations*



# 5 Discussion

In this section, the results presented in Section 4 will be further reviewed to discover possible explanations for the results, compare them with previous research in the field, and to ultimately find answers to the research questions of the present study: how do Finnish-speaking learners of English at two different proficiency levels use determiners in their English texts, and how does their determiner use differ from native-level English. Section 5.1 will discuss the results in comparison to the results in Peter Master's study, as this will set the context for determiner use in native-level English texts. Next in Section 5.2, the results of the statistical analysis will be discussed in order to discover the differences between intermediate and advanced learners and find possible explanations for said differences. The discussion will proceed in the same order as the results were presented in Section 4. After this, the whole study will be summarized and concluded in the next chapter, Section 6.

## 5.1 Comparison to Master's study

Table 2 shows the results of the text analysis per participant in very basic metrics: total words, total determiners, determiners per word ratio. As the tasks the essays in the YKI-Corpus are based on naturally produce short texts, the text lengths begin at as low as 286 words with averages around 450 words, it is clear that the sample size could be larger to minimize the effect of outliers skewing the results; for example, using wh-word determiners several times in a text, as was the case for participants *1006* and *1011*, as can be seen in Table 5. As mentioned in Section 3.1, Master added the number of zero articles into the total word count in his study, which was not done in the present study. I deem the resulting discrepancies too minor to invalidate the comparison of results between Master's study and mine, but it is something worth noting. In Master's text, the total word count and determiner count of the texts analysed were naturally massively higher, given the nature of the texts in question, but the determiners per word ratio can be compared. In the present study, the ratios were 0.15 for group B, 0.18 for group C, and 0.17 for the two combined. In Master's data, the number was 18.3 %, or

0.183 (Master 2013, 6); around the same as group C in my data. It would seem that the number of determiners per word increases as the competence level of the writer increases, however, this is purely from eyeballing the numbers, with no deeper analysis behind these conclusions. It is, of course, also clear that the determiners to words ratio cannot keep increasing perpetually, as one noun can only take one determiner. Additionally, as I will discuss in the next section, the number of zero articles advanced level learners use is higher, which causes a slight increase in determiners per words, because zero articles add to the determiner count but not to the word count. All of this considered, I would suggest the number for a proficient English speaker is somewhere slightly over 18 % of words being determiners.

Tables 3, 4, and 5 in the results section present the distribution of different determiner categories per participant. Table 3 shows the three different article types: indefinite, definite, and zero article. As is explained in Section 4.1, different article types are the most common determiners used by participants of both levels, B and C. As mentioned in Section 2.3, Master found similar results: in his analysis of research articles in science and technology, 90.3 % of all determiners were articles, and the percentage was correspondingly high in previous research Master cites (Master 2013, 14). In my data, the number was only 52 % for intermediate writers, 62 % for advanced writers, and 57 % for all participants. This is clearly lower than Master's results, but a part of the reason for that could be the different genres of writing: Master analysed research articles; an objective, scientific style of text, whereas the text from the YKI-Corpus are more personal in style (letters, messages, opinion pieces, etc.), resulting in different, more personal determiners, such as possessive determiners. The ideal would be, of course, to compare texts of similar length and genre, but for reason outlined in Section 3.1, this was not possible in the present study. Nevertheless, it seems that the increased use of articles instead of other determiners is another sign of higher competence of English, as the percentage rises from group B's 52 % to group C's 62 % all the way up to native level, proofread articles' 90.3 %. I will return to this interpretation in the discussion regarding the correlation analysis later in Section 5.

Looking at Table 4 next, it shows the numbers for the possessive, demonstrative, and quantifier categories. In the present study, possessive determiners appear in 20 % of all determiner use cases, demonstratives in 12 %, and quantifiers in 8 %. Master found

that only 2.4 % of all determiners in his data were possessive determiners (Master 2013, 9). This is vastly smaller than the 20 % in my data, however, as speculated before, this could be in part due to the differences between the style of writing in a research article versus a more personal type of text, e.g. a letter. Master would further reinforce that interpretation; he discovered that possessive structures were much more common in fields more concerned with people, such as non-science/technology or clinical psychology, as opposed to geology or chemistry, which are less preoccupied with personal experience and people (Master 2013, 31). Even though the discrepancies in Master results are smaller than from 2.4 to 20 %, it seems like the genre and subject of writing affects determiner use. As for demonstratives, Master's study produced a result of 4.5 % (Master 2013, 8), whereas in my data the number is noticeably higher at 12 %. This could be the result of Finnish L1 transfer: as there are no articles in the Finnish grammar officially, the closest natural alternative is to use words such as *tuo [that]*, i.e. demonstratives, as discussed in Section 2.3. Some of the demonstratives appear in situations, where a native English-speaker would probably choose to use a definite article instead of a demonstrative determiner, but the transfer effect pushes a native Finnish-speaker towards a demonstrative, such as in this example:

> *Maybe you should buy new windows for **those** rooms [on the highway-facing side of the hotel]*.

In the present study, quantifiers covered 8 % of all determiners for all 30 participants combined, as well as for both groups B and C separately. As was shown in Table 1, I chose to combine many different subcategories of determiners under quantifiers, however, Master's study does not do that, which makes comparing the two studies challenging regarding this metric. Therefore, I have combined Master's categories to have a roughly equivalent group of determiners, which includes Master's assertive-nonassertive, negative, universal, and dual determiners. The sum of these categories is 4.8 % (0.9 + 2.9 + 1.0 %), which means that there seems to be a difference in the results, however, this could be due to the differences in categorizing the determiners. Another potential reason to a higher percentage of quantifiers amongst English learners is, again, transfer: in some cases in the analyzed learners texts, a native level English speaker might use an indefinite or zero article instead of a quantifier, for example:

> *Put them in **some** ice, will you.*

On the other hand, a native level speaker could potentially use more quantifiers than a learner, as there is no direct equivalent to the negative determiner *no* in the Finnish language, but it is rather denoted as a negative verb structure, or with the help of a preposition, such as *without*. Consider the following examples (my own, not from the analyzed texts):

> *How can we survive, if we have **no** water?*
> *How can we survive, if we **don't** have water?*
> *How can we survive **without** water?*

A Finnish-speaker could favor the latter two options, as both structures exist in Finnish grammar. Nevertheless, I can only speculate on the effect of transfer on this matter, as it was not in the scope of this study to test this in such detail, however, there is room for further study in this field.

Master's study excludes numerals altogether (2013, 3), which means that category is also omitted from this section's comparison. Wh-words were analyzed, however, and Master found them to account for 0.14 % of total determiners (2013, 13). In the present study, the average of wh-words was 1 %, as can be seen in Table 5. While not as low a number as Master's, it should be noted that there seem to be outliers in this category, as some participants, especially intermediate ones, had a wh-word percentage of 7–10 %, while many participants had 0 %. Therefore, I would not consider this a sample large and comprehensive enough to be reliable, however, even in the present study's data, it seems that both numerals and wh-words see very infrequent use in the English language across proficiency levels. In conclusion, the comparison with Master's 2013 study presented in this section has suggested that slightly over 18 % of total words are determiners in a competent English-speakers' written English, articles are, by a considerable margin, the most common determiner type in native level English, and that the genre of the text probably affects the determiner choices in said text. In the following section, I will discuss the results of the statistical analysis of the data presented in Section 4.2.

## 5.2 Discussion of comparison and correlation

In this section, I will highlight and discuss the most interesting and important results of the statistical tests introduced in Section 4.2. Normality will be considered only briefly, as it is mostly a prerequisite test in order to conduct the comparison of means analysis, however, both types of mean comparisons for the determiner categories will be fully explored and discussed. Alongside the comparisons of means, effect sizes will be discussed to achieve a comprehensive understanding of how the two proficiency groups differ in their determiner use. Finally, this section will be concluded with a discussion of the correlation analysis results and their relationship with the rest of the analysis.

Both Table 6 and Figure 2 present whether the data is normally distributed; numerically in the table, graphically in the figure. The content of Table 6 was explained in Section 4.2: the first five determiner categories, including all articles, possessives, and determiners, as well as determiners per word were normally distributed, whereas the rest of the categories were not. As highlighted there, quantifiers were normally distributed in the Kolmogorov-Smirnov test, but not in the Shapiro-Wilk test, leading to it being considered not following a normal distribution. Looking at Figure 2, the Q-Q plots of quantifiers skew right, confirming this decision. As alluded to earlier, both numerals' and wh-words' plots look very scattered, and as such, their results, even in a non-parametric test, should be approached with caution.

The comparison of means began with t-tests for the suitable categories, first of which was the indefinite articles, followed by the other two article types. As discussed in Section 5.1, articles were the most common type of determiner, and their frequency seemed to increase in more proficient English speakers' texts. The results of the t-tests showed that there is no statistical significance in how the participants of the two levels, intermediate and advanced, use indefinite or definite articles, however, there is a significant difference in the use of zero articles. The effect sizes in Table 10 show that in addition to being insignificant, the differences between groups B and C in definite and indefinite article use are small. Zero articles, on the other hand, display a very large effect size at $d$ = -1.53. This, together with the statistical significance, suggests that there is a large increase in zero article use in advanced level Finnish-speaking English learners' texts compared to intermediate level writers. As it appears there are more articles in

general in more proficient writers' texts, and as the two groups do not significantly differ in their indefinite or definite article use, the data in the present study suggests that increased use of the zero article determiners is one of the hallmarks of an English learner approaching a native-level competency. Unfortunately, it is outside of the scope of the present study to discover why this is the case, as that would require qualitative analysis of the choices made by the learners, or a much more specialized test designed to gauge article use in different contexts. If I were to speculate, both intermediate and advanced level learners are at a stage of their language development, where they recognize the English language articles system and how it differs from Finnish. However, as learners progress, they realize that not every noun has to be accompanied by either definite or indefinite article, and instead, using a zero article is the correct option and not the effect of their Finnish L1, as was discussed in Section 2.2. While not outright omitting an obligatory structure, they resort to what they already know, resulting in overuse of other structures, such as definite articles (Jarvis and Pavlenko 2008, 192). Some degree of avoidance can also affect the choice of a determiner, as has traditionally been the strategy for Finnish-speakers (Ringbom 1987, 108): less proficient learners can opt to use another determiner type entirely, if they are unsure of which article type to choose. This possibility will be examined further next, as I discuss the comparisons of means for other determiner categories.

Table 8 showed the t-test results for possessive and demonstrative determiners. Possessive determiners did significantly differ between groups B and C, whereas demonstratives did not. In terms of effect sizes in Table 10, possessives had a large effect with group B using more possessives, demonstratives only displayed a small effect, also in favor of group B's participants. It seems that both proficiency levels use determiners to a largely similar capacity, however, intermediate learners use more possessive determiners in their English writing than advanced learners. Considering what I wrote above, the increased use of possessive pronouns could be a way to avoid making a choice between different article types, as choosing a possessive, a determiner type that functions similarly in both Finnish and English, relieves the learner of having to make the difficult choice. A more advanced student should be, first of all, more confident in their ability to choose in general, but also aware of the fact that leaving the article slot empty, i.e. choosing the zero article option, is not an error. This could be connected to the

concept of perceived distance, which was introduced in Section 2.2: as Finnish-speakers perceive English to be very different from Finnish, they often consciously rely less on structures that appear in their L1, such as a noun without an article (Ringbom 2007, 7-8; Meriläinen 2010, 28). This effect, too, seems to diminish with increased proficiency. To conclude the discussion of the t-test results, as Finnish-speaking learners progress in their English development, they seem to use fewer possessive determiners, and opt for more zero article determiners instead. As was outlined in Section 5.1, interpolating this direction of development seems correct, as native-level speakers of English do use more articles, less possessive determiners.

In addition to the parametric t-tests, the data also yielded results for the non-parametric Mann-Whitney tests. These included the results for quantifiers, numerals, and wh-words, and these are presented in Table 9. As the table shows, none of the three non-parametric categories displayed a statistically significant difference between the two proficiency levels. The effect sizes for these categories are found in Table 11: all three categories show either minor or small effect size. This combined with statistically non-significant differences, it would appear that the proficiency level of the participants in the present study does not affect how these non-parametric determiner categories are used in writing. As I alluded to earlier, regardless of the significance or effect size of these results, I would not be confident in the reliability of the results for numerals and wh-words in particular, given their low sample size, scattered distribution, and distinct outliers. To gain more insight into the use of these determiner types, a more refined and specialized study would be necessary. Even then, it seems unlikely the result would noticeably from the present study in the grand scheme of determiner use in the English language: these determiner types would still probably be a tiny minority of the total determiners used in any given text.

The final section in Results presented the correlation analysis in Tables 12 and 13, displaying parametric and non-parametric correlations, respectively. In the parametric correlation analysis, two determiner categories as well as the determiners per word category emerged as the ones with the most significant correlations; the determiners in questions were definite articles and possessives. As mentioned in Section 4.2, the increase in determiners per words seems to correlate with an increase in the relative number of definite articles. Concurrently, the frequency the two most

prominent non-article determiners, possessives and demonstratives, decreases. As there are undoubtedly multiple interrelating factors at play, I cannot conclusively claim that causation exists between these numbers. However, it does seem like a higher number of articles in a learner's text also means a lower number of possessives and demonstratives, maybe even all other determiners types, however, that is unclear from the data. This assessment is further reinforced by possessive determiners negatively correlating with indefinite articles: again, an increase in an article-type determiner correlates with a decrease in another type. Furthermore, examining the correlations for the determiners per word category shows that as the number of determiners per word grows, the number of definite articles grows as well, whereas the number of possessives and quantifiers decreases. All of this appears to signal overapplication of simpler, more L1-like structures instead of the more difficult ones, as is common for language learners (Jarvis and Pavlenko 2008, 192). What was discussed in Section 5.1 as well as earlier in Section 5.2 seems to hold true: advanced level Finnish-speaking learners of English use more determiners per word than their intermediate level peers, and a higher percentage of those determiners are articles. The statistically non-significant results of the correlation analysis also seem to support this view, as they showed negative correlations between indefinite articles and possessives, zero articles and demonstratives, and definite articles and quantifiers.

This discussion section has examined the data presented in Section 4 and analyzed it further. Section 5.1 established that native-level English texts tend to have slightly over 18 % determiners out of the total number of words, and that participants in group C in the present study, the advanced learners, were closer to that number than group B. Native-level writer also chose an overwhelming portion of articles as their determiners. The Finnish-speaking learners did not reach equally high percentages, however, advanced level learners did use more articles than intermediate level writers. All other determiner categories saw much less frequent use, although the contrast between articles and other types was not as drastic in Finnish-speaking learner's texts. Section 5.2 suggested that there is a large difference in zero article use between groups B and C; advanced level learners use zero articles much more often. Intermediate learners, in turn, seem to use more possessive determiners. In conclusion, the results of this study suggest that more advanced learners of English choose more native-level style

determiners, namely articles, than their intermediate level peers, and that these are more often zero articles instead of other article types.

# 6 Conclusion

This study had the goal of examining the determiner use in Finnish-speaking English learners' written texts through the lens of language transfer. This goal was set out to be achieved by analyzing and comparing texts by two groups of learners at intermediate and advanced level of proficiency in English, as well as comparing them to native-level English speakers. First, the central terms and concepts were defined: what transfer is and how it is studied, especially in the context of Finnish-speakers learning English, as well as what determiners are and how they function in both of those languages were the first matters discussed in this thesis. It was determined that there is often a negative transfer effect, or interference, hindering Finnish-speakers learning English at varying proficiency levels, as the languages are very different from one another. A very common obstacle for learners is the English article system, as nothing equivalent exists in the Finnish language, at least in formal Finnish. Articles are one of the many types of determiners, noun-modifying words, in English. Studying transfer and language learning by examining a system with features both common between languages as well as specific to each language can improve our understanding of how Finnish speakers learn English, or how second languages are learnt in general. Thus, the questions this thesis sought to answer were:

1. How do intermediate and advanced level Finnish-speaking learners of English use determiners?
2. How does their determiner use differ from native-level English speakers?

To find answer to these research questions, a collection of essays from the YKI-Corpus by the University of Jyväskylä were used. Essays by 15 intermediate level and 15 advanced level Finnish-speaking English learners were compiled and analyzed using multiple Natural Language Processing tools to parse the determiners from the texts, and statistical analysis tools to compare the determiner use of the two groups of learners,

as well as compare the results to research on determiner use in native-level English speakers' texts.

From this material and through these methods, results were produced. Amongst both intermediate and advanced learners, definite articles were the most commonly used type of determiner. All three article types combined (definite, indefinite, and zero article) cover the majority of all determiners used by participants at both levels. Possessives, demonstratives, and quantifiers were all also seeing regular use, all totalling over 5 % of all determiners. Numerals and wh-words were left in the minority with infrequent use by both intermediate and advanced learners. As previous research (e.g. Master 2013) in determiner use has shown, native-level speakers of English also prefer different articles as their determiners of choice, however, the margin between articles and other determiners is potentially even larger. While definite articles were used the most by both groups, the distribution of all articles types was not as similar between the groups: advanced learners used zero articles much more frequently than intermediate learners. On the other hand, intermediate learners preferred possessive determiners more often than advanced learners did. Viewing the determiner use of Finnish-speaking English learners as a whole, it seems that more proficient learners use determiners more frequently in general, a larger portion of the determiners are articles, especially zero article types, and fewer are other types of determiners.

A number of limitations must be kept in mind regarding this study, the most glaring of which is the sample size: a larger amount of text, especially across different genres, would yield a more comprehensive picture of the matter at hand. This is most feasible done by increasing the number of participants, as acquiring sufficiently lengthy texts from learners is difficult and time-consuming. Furthermore, while the texts are analyzed from the point-of-view of transfer, and it is undoubtedly one of the primary effects driving learners' determiner or language use, its precise effects are left to mere speculation in a quantitative study such as this one. A qualitative or a mixed-methods study would be required to further examine the learners' choices and thought process in deciding how and why they modify nouns in their texts. Finally, in order to make a comparison between learners and native-level speakers of English more consistent, the cohort, genre of texts, and the amount and categorization of data should be more similar than was possible in the present study.

This thesis has provided a fresh view on how Finnish-speaking language learners produce written English. Focusing on a single linguistic element, determiners, has helped in painting the full picture of English learning and the effects Finnish as a native language has on it, one brush stroke at a time. While the subject as a whole is too complex and multi-faceted to tackle in this thesis, the questions this study has answered can help guide language teaching for Finnish-speaking students, especially regarding the learners' written English production. Perhaps in the future, an even more comprehensive understanding of English learning can be attained, piece by piece.

# References


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

Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Boston: O'Reilly Media.

Chesterman, Andrew. 1991. *On Definiteness: A Study with Special Reference to English and Finnish*. New York: Cambridge University Press.

Crossley, Scott A., Laura K. Allen, Kristopher Kyle and Danielle S. McNamara. 2014. "Analyzing Discourse Processing Using a Simple Natural Language Processing Tool (SiNLP)". *Discourse Processes* 51(5–6): 511–534. DOI: 10.1080/0163853X.2014.910723

Dulay, Heidi, Marina Burt and Stephen Krashen. 1982. *Language Two*. New York: Oxford University Press.

Dörnyei, Zoltán. 2007. *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Ellis, Rod. 2015. *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Ellis, Rod, and Gary Barkhuizen. 2005. *Analysing Learner Language.* Oxford: Oxford University Press.

Gass, Susan and Larry Selinker. 1994. *Language Transfer in Language Learning*. Rev. ed., repr. with corr. Amsterdam: Benjamins.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho. 2004. *Iso suomen kielioppi* [The great book of Finnish grammar]. Helsinki: Suomalaisen Kirjallisuuden Seura. Accessed 11 November 2019. http://scripta.kotus.fi/visk

Ivaska, Ilmari and Kirsti Siitonen. 2017. "Learner language morphology as a window to crosslinguistic influences: A key structure analysis". *Nordic Journal of Linguistics* 40(2): 225–253.

Jarvis, Scott. 2017. "Transfer: An Overview with an Expanded Scope". In *Crosslinguistic Influence and Distinctive Patterns of Language Learning: Findings and Insights from a Learner Corpus*, edited by Anne Golden, Scott Jarvis and Kari Tenfjord, 12–28. Bristol: Multilingual Matters.

Jarvis, Scott and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition.* New York: Routledge.

Jarvis, Scott and Terence Odlin. 2000. "Morphological Type, Spatial Reference, and Language Transfer" In *Studies in Second Language Acquisition* 22(4), 535–556.

Järvinen, Heini-Marja. 2014. "Katsaus kielenoppimisen teorioihin" In *Kuinka Kieltä Opitaan: Opas Vieraan Kielen Opettajalle Ja Opiskelijalle* [How a language is learned: A guide for a foreign language teacher and student], edited by Pekka Lintunen and Päivi Pietilä, 68–88. Helsinki: Gaudeamus.

Kellerman, Eric. 1979. "Transfer and non-transfer: where are we now?" In *Studies in Second Language Acquisition* 2: 37–57.

Krohn, Robert. 1971. *English Sentence Structure*. Ann Arbor: The University of Michigan Press.

Lado, Robert. 1957. *Linguistics Across Cultures*. Ann Arbor: The University of Michigan Press.

Larjavaara, Matti. 2001. "Määräinen artikkeli – suomessa?" [A definite article – in the Finnish language?] *Kielikello* 4. Accessed 3 October 2019. https://www.kielikello.fi/-/maarainen-artikkeli-suomessa-

Larson-Hall, Jenifer. 2010. *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York; London: Routledge.

Laury, Ritva. 1996. "Sen kategoriasta – onko suomessa jo artikkeli?" [About the category of it – does an article already exist in the Finnish language?] *Virittäjä* 2: 162–181.

Lintunen, Pekka and Päivi Pietilä. 2014. *Kuinka Kieltä Opitaan: Opas Vieraan Kielen Opettajalle Ja Opiskelijalle* [How a language is learned: A guide for a foreign language teacher and student]. Helsinki: Gaudeamus.

Mackey, Alison and Susan M. Gass. 2005. *Second Language Research: Methodology and Design.* Mahwah, New Jersey: Lawrence Erlbaum.

Master, Peter. 2013. "A contrastive study of determiner usage in EST research articles" In *International Journal of Language Studies* 7(1): 33–58

Meriläinen, Lea. 2010. *Language transfer in the written English of Finnish students*. Joensuu: University of Eastern Finland.

Meriläinen, Lea. 2017. "The progressive form in learner Englishes: Examining variation across corpora" In *World Englishes* 36(4): 760–783.

Odlin, Terence. 1989. *Language Transfer: Cross-linguistic influence in language learning*. Oxford: Oxford University Press.

Official Statistics of Finland. 2018. *Population structure*. Helsinki: Statistics Finland. Accessed 14 October 2019. https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html

Ortega, Lourdes. 2009. *Understanding Second Language Acquisition*. London: Hodder Education.

Ringbom, Håkan. 1987. *The Role of the First Language in Foreign Language Learning*. Clevedon: Multilingual Matters.

Ringbom, Håkan. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon [England]; Buffalo: Multilingual Matters.

Selinker, Larry. 1972. "Interlanguage." *IRAL* X(3): 219-231.

Verspoor, Marjolyn and Kim Sauter. 2000. *English sentence analysis: an introductory course.* Amsterdam: J. Benjamins cop.

YKI-Corpus. 2017. Soveltavan kielentutkimuksen keskus, University of Jyväskylä. Accessed 29 April 2019. http://yki-korpus.jyu.fi/

White, Lydia. 1994. "Is it Just a New Name for Old Problems?" In *Language Transfer in Language Learning*, edited by Susan Gass and Larry Selinker, 217–232. Amsterdam: Benjamins.

# Appendices

APPENDIX 1: Learners' background information and differences in Section 3.1

Part 1, basic information

*ID 10XX = intermediate level, 20XX = advanced level*

| ID | GENDER | AGE | EDUCATION | YEARS OF STUDY | STUDIED ABROAD |
|---|---|---|---|---|---|
| 1001 | male | 19 | upper secondary school | 10+ | 0 |
| 1002 | male | 25 | university | missing | 0 |
| 1003 | male | 19 | upper secondary school | 10+ | 0 |
| 1004 | male | 23 | upper secondary school | 10+ | 0 |
| 1005 | male | 19 | upper secondary school | 7–9 | 0 |
| 1006 | male | 22 | upper secondary school | 7–9 | 0 |
| 1007 | male | 20 | upper secondary school | 7–9 | 0 |
| 1008 | male | 20 | upper secondary school | 10+ | 0 |
| 1009 | male | 20 | upper secondary school | 10+ | 0 |
| 1010 | female | 25 | university | 10+ | 0 |
| 1011 | male | 19 | upper secondary school | 10+ | 0 |
| 1012 | male | 20 | upper secondary school | 10+ | 1 |
| 1013 | female | 22 | upper secondary school | 10+ | 0 |
| 1014 | male | 21 | upper secondary school | 10+ | 0 |
| 1015 | male | 23 | upper secondary school | 7–9 | 0 |
| 2001 | male | 25 | university | 10+ | 0 |
| 2002 | female | 24 | university | 10+ | 0 |
| 2003 | female | 25 | university | 10+ | 0 |
| 2004 | male | 25 | university | 10+ | 0 |
| 2005 | female | 25 | university | 10+ | 1 |
| 2006 | male | 25 | university | 7–9 | 1 |
| 2007 | female | 23 | upper secondary school | 10+ | 0 |
| 2008 | male | 21 | upper secondary school | 10+ | 0 |
| 2009 | female | 22 | university | 10+ | 0 |
| 2010 | female | 23 | upper secondary school | 10+ | 0 |
| 2011 | female | 25 | university | 10+ | 0 |
| 2012 | male | 25 | university | 10+ | 0 |
| 2013 | male | 24 | university | 10+ | 1 |
| 2014 | male | 21 | upper secondary school | 7–9 | 1 |
| 2015 | male | 21 | upper secondary school | 7–9 | 0 |

Part 2, reasons for acquiring the language certificate

| ID | JOB APPLICATION | CURRENT JOB | STUDY PURPOSES | FEEDBACK |
|---|---|---|---|---|
| 1001 | 1 | 0 | 0 | 0 |
| 1002 | 1 | 0 | 0 | 0 |
| 1003 | 1 | 0 | 0 | 1 |
| 1004 | 1 | 1 | 0 | 0 |
| 1005 | 1 | 0 | 0 | 0 |
| 1006 | 1 | 0 | 0 | 0 |
| 1007 | 1 | 0 | 0 | 0 |
| 1008 | 1 | 0 | 0 | 0 |
| 1009 | 1 | 0 | 0 | 0 |
| 1010 | 1 | 0 | 0 | 1 |
| 1011 | 1 | 0 | 0 | 0 |
| 1012 | 1 | 1 | 0 | 0 |
| 1013 | 1 | 1 | 1 | 0 |
| 1014 | 1 | 0 | 0 | 0 |
| 1015 | 1 | 0 | 0 | 0 |
| 2001 | 0 | 1 | 0 | 0 |
| 2002 | 1 | 0 | 0 | 0 |
| 2003 | 1 | 0 | 0 | 0 |
| 2004 | 1 | 0 | 0 | 1 |
| 2005 | 1 | 0 | 0 | 0 |
| 2006 | 1 | 0 | 0 | 0 |
| 2007 | 0 | 1 | 0 | 0 |
| 2008 | 1 | 0 | 0 | 0 |
| 2009 | 1 | 0 | 0 | 0 |
| 2010 | 1 | 0 | 1 | 0 |
| 2011 | 1 | 0 | 0 | 0 |
| 2012 | 0 | 1 | 0 | 0 |
| 2013 | 0 | 0 | 0 | 0 |
| 2014 | 1 | 0 | 0 | 0 |
| 2015 | 1 | 0 | 0 | 0 |

APPENDIX 2: Finnish summary

## Johdanto

Tämän tutkimuksen tarkoituksena on tutkia kielen siirtovaikutusta suomenkielisten englanninoppijoiden kirjoitetuissa teksteissä. Tutkimus keskittyy erityisesti analysoimaan, miten oppijan äidinkieli, suomi, sekä oppijan englannin kielen taito vaikuttavat determinanttien, eli substantiivilauseketta määrittävien sanojen käyttöön ja määräisyyden käsitteeseen kirjoitetussa kielessä. Suomi ja englanti ovat kielinä hyvin erilaisia, joten niiden vaikutus toisiinsa on aiheena kiinnostanut jo pitkään (ks. Ringbom 1987), ja siksi Suomessa onkin kielten välistä siirtovaikutusta tutkittu runsaasti. Erityisesti suomen ja ruotsin puhujien englanninoppimista on vertailtu, ja näin on havaittu, että typologisesti toisiaan lähempänä olevat ruotsi ja englanti helpottavat ruotsinkielisiä oppijoita (ks. esim. Ringbom 1987, Jarvis ja Odlin 2000). Erityisesti ruotsille ja englannille tyypillinen artikkelijärjestelmä puuttuu suomen kielestä kokonaan, ja artikkelit ovatkin perinteisesti olleet suomenkieliselle oppijalle haasteellisia. Tässä tutkimuksessa huomiota kiinnitetään vielä artikkeleja laajemmin determinantteihin, joihin artikkelit toki sisältyvät.

Tässä Pro Gradu -tutkielmassa materiaalina käytettiin Jyväskylän yliopiston koostaman YKI-korpuksen (2017) esseitä. Esseitä oli yhteensä 30 eri kirjoittajalta, jotka sijoittuivat kahdelle kielitaitotasolle: keskitaso B, ja edistynyt taso C. YKI-korpuksesta löytyy esseiden lisäksi osallistujien taustatiedot, joiden avulla osallistujat rajattiin tähän tutkimukseen sopiviksi, esimerkiksi äidinkielen mukaan. Esseet analysoitiin kahdella automaattiseen kielelliseen analysointiin suunnitellulla työkalulla, SiNLP:llä ja spaCY:llä. Lisäksi kielellisen analyysin jälkeen tulosten tilastollinen analyysi suoritettiin SPSS:llä. Tutkimuksen teoreettinen viitekehys, käytetyt materiaalit ja tutkimusmetodit, tulokset, tulosten pohdinta ja loppupäätelmät esitellään tässä tutkielmassa kukin omana kappaleenaan.

## Teoreettinen viitekehys

Kielten oppiminen on kiehtonut tutkijoita jo pitkään, ja jo aikaa sitten on oivallettu, että kielen oppimiseen ja sen haastavuuteen vaikuttaa paitsi kohdekieli, myös oppijan äidinkieli. Äidinkielen vaikutusta on tutkittu 1950-luvulta saakka, ja käsitykset aiheesta ovat muuttuneet ja muovautuneet ajan myötä (Ellis 2015). Kun kielenoppiminen nähtiin tapojen muodostamisena ja vahvistamisena, koettiin äidinkielen oppimisesta seuranneet tavat ja tottumukset uuden kielen oppimisen haittana ja hidastajana. Nähtiin, että äidinkieltä ja kohdekieltä vertailemalla pystyttäisiin ennustamaan, tai vähintäänkin tunnistamaan ja selittämään oppimisen ongelmakohtia (Ellis 2015). Ennustaminen osoittautuikin hankalaksi, joten 1970-luvulla keskityttiin analysoimaan oppijoiden virheitä ja selittämään niitä äidinkielen vaikutuksen avulla (Odlin 1989). Virheiden määrittely oli kuitenkin vaikeaa, eikä oppijan äidinkieli vaikuttanut niitä tarpeeksi aiheuttaneen, joten päädyttiin näkemykseen, jonka mukaan äidinkielen vaikutus kielenoppimiseen olisi minimaalinen (Gass ja Selinker 1994). Niinpä 1980-luvulta eteenpäin äidinkielen vaikutus kielenoppimiseen on nähty yhtenä sen osatekijänä, joka voi sekä helpottaa että vaikeuttaa uuden kielen oppimista, kielten eroavaisuuksista riippuen (Ellis 2015).

Kielten välinen siirtovaikutus eli *transfer* voidaan määritellä usealla tavalla. Oleellista on huomata, että siirtovaikutus toimii kaikkien oppijan tuntemien kielten välillä, ja voi toimia moneen suuntaan: uudet kieletkin voivat vaikuttaa oppijan äidinkieleen tai toisiinsa (Jarvis 2017). Usein kuitenkin, kuin myös tässä tutkimuksessa, siirtovaikutus käsitetään äidinkielen vaikutuksena uuteen, opittavaan kieleen. Siirtovaikutus on kokonaisvaltainen: se voi vaikuttaa moniin kielen osa-alueisiin, muttei välttämättä samanaikaisesti tai joka tilanteessa (Jarvis 2017). Kuten jo mainittiinkin, siirtovaikutus voi olla positiivista tai negatiivista; oppimista helpottavaa tai haittaavaa (Pietilä ja Lintunen 2014). Positiivista siirtovaikutusta voi olla vaikea huomata, sillä se ilmenee helpompana tai nopeampana oppimisena. Negatiivisesta siirtovaikutuksesta taas jää usein jokin jälki; virhe tai kielen kummallisuus, joka on helpommin tunnistettavissa. Tällaisia merkkejä ovat esimerkiksi jonkin rakenteen yli- tai alikäyttö, äidinkielisen rakenteen käytöstä aiheutuneet virheet tai jonkin rakenteen käytön välttäminen (Odlin 1989, Ortega 2009). Tämän takia onkin helpompi tutkia kielen yleistä,

vaikeasti vältettävää elementtiä, minkä vuoksi tämän tutkimuksen kohteeksi on valittu determinantit.

Kuten aiemmin mainittiin, myös äidin- ja kohdekielen keskinäinen etäisyys vaikuttaa: on tutkittu, että hyvin erilaisten kielten välillä oppijat pyrkivät välttämään rakenteiden siirtämistä kielestä toiseen (Ellis ja Barkhuizen 2005). Siirtovaikutuksesta ei tällöin ole apua, muttei kovin todennäköisesti haittaakaan. Huomionarvoista on myös, ettei kielten todellisuudessa tarvitse erota toisistaan, vaan siirtovaikutuksen välttämiseksi riittää, että oppija kokee kielten olevan erilaisia (Ringbom 2007). Oppija voi myös suosia muita osaamiaan kieliä rakenteiden siirtämisessä, jos hän kokee sen edulliseksi (Ringbom 1987). Myös siirrettävä rakenne vaikuttaa: siirtovaikutusta tapahtuu helpommin yleisissä perusrakenteissa kuin harvinaisissa erikoisrakenteissa (Kellerman 1979). Siksi vaikkapa verbitaivutus siirtyy helpommin äidinkielestä uuteen kieleen kuin esimerkiksi kielikuvat.

Tämän tutkielman tutkittava rakenne, determinantit, on substantiivilauseketta määrittävä osa, esimerkiksi englannin kielen epämääräinen artikkeli *a* tai *an* (esim. ***a*** *car* [*auto*]). Englannin kielessä jokainen substantiivilauseke vaatii determinantin, minkä vuoksi ne ovat kielen yleisimpiä sanaryhmiä (Biber et al. 2002). Jo mainittujen epämääräisten artikkelien lisäksi substantiivia määrittämään voidaan käyttää määräistä artikkelia, nolla-artikkelia, omistussanaa, demonstratiivipronominia, paljoussanaa, numeraalia tai kysymyssanaa. Determinantin valintaan liittyy vahvasti määräisyyden käsite. Määräisyys tarkoittaa substantiivin tai substantiivilausekkeen viittaussuhdetta, eli esimerkiksi tuttuutta, laskettavuutta tai lukumäärää (Chesterman 1991). Määräisyys on hyvin kontekstisidonnaista ja vaikeasti lokeroitavissa. Englannin kielen määräisyyttä on tutkinut Peter Master (2013), jonka tutkimuksen tuloksia käytetään tässäkin tutkielmassa vertailun pohjana. Masterin mukaan natiivitasoiset englannin puhujat käyttävät teksteissään determinantteina ylivoimaisesti eniten erilaisia artikkeleita: yli 90 %. Demonstratiivit kattoivat 4,5 % kaikista determinanteista, omistussanat 2,4 %, muut sanaryhmän vielä vähemmän (ibid.). Suomen kielessä, toisin kuin englannissa, määräisyys on pienemmässä roolissa, eikä artikkelijärjestelmää virallisesti ole. Varsinkin puhuttuun suomeen on kuitenkin syntynyt, ehkä englannin siirtovaikutuksesta, artikkelin kaltaisia rakenteita, kuten *se* (***se*** *kirja on pöydällä* [***the*** *book is on the table*])

(Hakulinen et al. 2004). Tämä kielten välinen vaikutus toimi yhtenä innoittajana tässä tutkielmassa, ja Pro Gradu -tutkielmani pyrkiikin vastaamaan seuraaviin kysymyksiin:

1. Miten keskitason ja edistyneen tason suomenkieliset englanninoppijat käyttävät determinantteja?
2. Miten determinanttien käyttö eroaa natiivitasoisista englannin puhujista?

## Tutkimusmateriaali ja -metodit

Tutkielman materiaali saatiin Jyväskylän yliopiston (2017) tutkimuskäyttöön keräämästä YKI-korpuksesta. Korpus sisältää Yleisen kielitutkinnon osallistujien suorituksia, muun muassa tässä tutkimuksessa käytettyjä kirjoitettuja esseitä. Suoritusten lisäksi korpus tarjoaa kattavaa taustatietoa osallistujista, mikä helpottaa osallistujien valintaa ja rajaamista. Korpuksen esseet ovat pääsääntöisesti lyhyitä, noin 200 sanan mittaisia vastauksia erilaisiin tehtävänantoihin, kuten muodollisen valituskirjeen kirjoittaminen. Tähän tutkielmaan valittiin 30 osallistujaa seuraavin kriteerein: 15 osallistujaa eurooppalaisen viitekehyksen taitotasolta B (keskitaso) ja 15 osallistujaa taitotasolta C (edistynyt taso). Osallistujien äidinkieli on suomi, kielitutkinnon testikieli englanti, testi suoritettu hyväksytysti, ikähaarukka 18–25. Näiden kriteerien takia osallistujajoukko oli muut taustatekijätkin huomioon ottaen melko homogeeninen, mikä on syytä huomioida tulosten yleistämisessä. Lisäksi vertailumateriaalina tässä tutkielmassa toimii jo mainittu Masterin tutkimus (2013), jossa hän tutkii natiivitasoisten englannin puhujien teksteissä käytettyjä determinantteja. Tutkimuksen vertailukelpoisuutta vähentää tutkittavien tekstien erilaiset tekstilajit ja kategorisointi, mutta se toimii riittävän hyvin tämän tutkielman tarkoituksiin.

Tutkimusmetodi on kvantitatiivinen, automaattinen tekstin analysointi. Analyysin työkaluna käytettiin SiNLP-työkalua (Crossley et al. 2014), joka laskee asetettujen kriteerien mukaiset tekstin osat tekstimassasta. SiNLP ei kuitenkaan itsenäisesti pysty lauseenjäsenien tunnistamiseen, joten nolla-artikkelien laskemiseksi analyysiä täydennettiin spaCY-työkalulla, josta kyseinen toiminto löytyy. Tekstitiedostoiksi muunnetut esseet syötettiin SiNLP:hen, joka tuotti Microsoft Excel -tiedostona eri kategorioiden determinanttien lukumäärien lisäksi sanamäärät, sanojen ja determinanttien suhdeluvun ja muita tunnuslukuja, joita ei tässä tutkielmassa

käytetty. Helppokäyttöisen SiNLP:n lisäksi tutkimusta täydentämään tarvittiin spaCY-työkalu, jonka käyttö vaati enemmän manuaalista tekstin jakamista ja erittelyä Python-ohjelmointikielen avulla. spaCY:llä laskettiin nolla-artikkelien lukumäärä, jolloin tekstianalyysin tulostaulukko saatiin täydennettyä tarvittavilla tiedoilla, kuten determinanttien kokonaismäärällä ja determinanttityyppien prosenttiosuuksilla. Automaattisen tekstianalyysin laadun varmistamiseksi esseet käytiin vielä manuaalisesti läpi tarkistamalla oikeinkirjoitus virheiden välttämiseksi. Manuaalisen tarkistuksen tulokset lisättiin ja muokattiin taulukkoon.

Tekstianalyysin tuottaman valmiin kvantitatiivisen datan tarkempaan tutkimiseen ja vertailuun tarvitaan tilastollista analyysiä, joka tässä tutkielmassa toteutettiin SPSS-työkalun avulla. Erilaisten determinanttien normaalijakauma testattiin Kolmogorov–Smirnovin ja Shapiro–Wilkin testeillä. Normaalijakauman perusteella keskitason ja edistyneen tason oppijoiden tuloksia vertailtiin riippumattomien otosten t-testillä tai Mann-Whitneyn U-testillä. Näillä testeillä selvitettiin ryhmien välisten erojen tilastollista merkittävyyttä, mutta erojen vaikutusten kokoa ei pysty SPSS:n valmiilla toiminnoilla laskemaan. Sitä varten käytettiin Cohenin d-arvon ja Pearsonin r-arvon laskukaavoja, ja laskut suoritettiin manuaalisesti. Tilastollisen analyysin lopuksi testattiin erilaisten determinanttityyppien välisiä korrelaatioita Pearsonin ja Spearmanin korrelaatiotesteillä.

## Tulokset ja pohdinta

Taulukko 2 (Table 2) näyttää jokaisen 30 osallistujan esseiden sanamäärän, determinanttien määrän ja determinanttien osuuden sanojen kokonaismäärästä. Sana- ja determinanttimäärät olivat pääsääntöisesti hieman suuremmat edistyneen tason oppijoilla, samoin determinanttien osuus sanoista. Taulukot 3, 4 ja 5 (Tables 3, 4, and 5) esittävät jokaisen determinanttikategorian determinanttien lukumäärät ja prosenttiosuudet osallistujittain. Taulukoista näkee, miten kolmen artikkelikategorian osuus determinanteista on molemmilla taitotasoilla suuri, keskimäärin 57 %. Määräiset artikkelit olivat suurin ryhmä molemmilla taitotasoilla, 23 % kaikista determinanteista ryhmällä B, 26 % ryhmällä C. Omistussanojen, demonstratiivien ja paljoussanojen osuudet kahdella taitotasolla sijoittuivat 8 ja 23 % välille, numeraalien ja kysymyssanojen osuudet jäivät pariin prosenttiin. Masterin (2013) tuloksia

tarkastellessa huomaa, että natiivitasoisten puhujien teksteissä determinanttien osuus sanamäärästä on hieman yli 18 %; näin oli myös edistyneen tason oppijoilla. Keskitason oppilailla kyseinen luku jäi 15 prosenttiin. Masterin tutkimuksessa artikkelien osuus determinanteista oli vielä tätä tutkielmaa suurempi: 90,3 %. Edistyneen tason oppijat olivat tätä lukua lähempänä kuin keskitason oppijat, mutta jäivät silti kauas natiivitasosta. Eroa selittää tosin kielitaidon lisäksi myös erot tekstilajeissa; oppijoiden tekstit olivat Yleisen kielitutkinnon tehtäviin vastaamista, natiivitason tekstit tutkimusartikkeleita. Siitä huolimatta artikkelien suurta osuutta voi pitää merkkinä korkeammasta kielitaidosta. Natiivitason puhujien teksteistä artikkelit vievät niin suuren osuudet determinanttipaikoista, ettei muille jää juuri sijaa; muiden tyyppien osuudet olivat kautta linjan pienemmät kuin omassa tutkimuksessani. Suuruusjärjestykset ovat kuitenkin omistussanoja lukuun ottamatta samanlaiset. Molempien taitotasojen oppijat käyttivät omistussanoja yhteensä keskimäärin 20 % ajasta, kun taas Masterin tutkimuksessa kyseinen luku on 2,4 %. Jälleen tekstilaji varmasti vaikuttaa valittuihin determinantteihin, sillä YKI-korpuksen teksteissä on pääsääntöisesti tutkimusartikkeleita subjektiivisempaa asiaa. Siitä huolimatta natiivitason puhujia alhaisemman taitotason oppijoiden kohdalla vähäisempi artikkelien määrä vaikuttaa korvaantuneen osin omistussanoilla.

Taulukko 6 (Table 6) näyttää determinanttikategorioiden normaalijakaumatestien tulokset. Viisi ensimmäistä kategoriaa (kolme artikkelityyppiä, omistussanat ja demonstratiivit) olivat normaalisti jakautuneita, muut eivät. Normaalijakauman perusteella viittä ensimmäistä kategoriaa vertailtiin ryhmien B ja C välillä parametrisin testein. Taulukot 7 ja 8 (Tables 7 and 8) näyttävät itsenäisten otosten t-testin tulokset näille determinanttityypeille. Artikkelityyppien osalta tilastollisesti merkittävää eroa löytyi nolla-artikkelien käytössä, ja taulukkoa 10 (Table 10) tarkastellessa huomataan, että edistyneen tason oppijat käyttivät huomattavasti enemmän ($d = -1,53$) nolla-artikkeleita teksteissään kuin vertailuryhmä. Tämän voi spekuloida johtuvan edistyneiden oppijoiden laajemmasta repertuaarista käsitellä omaan äidinkieleen kuulumatonta rakennetta, mutta tämän tutkielman puitteissa syihin ei mennä tämän tarkemmin. Taulukoista 8 ja 10 näkee, että siinä missä nolla-artikkelit olivat selvästi yleisempiä edistyneiden teksteissä, ovat omistussanat useammin keskitason oppijoiden valinta. Ehkä keinovalikoiman ollessa vielä rajallinen, keskitason

oppija turvautuu helpommin äidinkielestään tuttuun, vieraassa kielessä jo opittuun rakenteeseen. Taulukon 9 (Table 9) ei-parametrisen Mann-Whitney -testin tulokset näyttävät, ettei muissa determinanttikategorioissa ollut tilastollisesti merkittäviä eroja taitotasoryhmien välillä. Viimeisenä taulukoiden 12 ja 13 (Tables 12 and 13) korrelaatioanalyysin tulokset näyttävät, että kasvanut artikkelien, erityisesti epämääräisten ja määräisten artikkelien määrä vaikuttaa korreloivan omistussanojen määrän laskun kanssa. Toisaalta artikkelien määrän kasvaessa myös determinanttien osuus sanamäärästä näyttää nousevan.

## Johtopäätökset

Tämän Pro Gradu -tutkielman päämääränä oli selvittää, miten suomenkieliset englanninoppijat käyttävät determinantteja teksteissään. Tähän tavoitteeseen pyrittiin analysoimalla kahden eri taitotason oppijoiden kirjoitettuja esseitä ja vertaamalla näitä ryhmiä keskenään. Suomi äidinkielenä ei perinteisesti ole helpottanut englannin oppimista, ja äidinkielen vaikutusta kielenoppimiseen tämäkin tutkielma kartoitti esittämällä ja vastaamalla seuraaviin tutkimuskysymyksiin:

1. Miten keskitason ja edistyneen tason suomenkieliset englanninoppijat käyttävät determinantteja?
2. Miten determinanttien käyttö eroaa natiivitasoisista englannin puhujista?

Molempien taitotasojen oppijat käyttivät artikkeleita enemmän kuin muita determinantteja. Ryhmien välisistä eroista korostui edistyneiden oppijoiden enemmän käyttämät nolla-artikkelit ja keskitason oppijoiden suosimat omistussanat. Yleisesti artikkelien suuri osuus determinanteista ja determinanttien vähän yli 18 % osuus sanojen kokonaismäärästä vaikuttaa kielivän korkeammasta kielitaidosta, ja edistynyt ryhmä oli molemmilla mittareilla natiivitasoa lähempänä. On kuitenkin huomattava, että tämän tutkielman suhteellisen pieni ja homogeeninen otanta ei mahdollista tulosten laajempaa yleistämistä. Lisäksi ryhmien välisten erojen syihin ja aiheuttajiin ei tässä tutkimuksessa ollut resursseja perehtyä tarkemmin, vaan niiden tutkiminen vaatisi laajempaa, ehkä kvalitatiivista tutkimusta aiheesta. Tutkielma on kuitenkin osaltaan edistänyt ymmärrystä suomen kielen siirtovaikutuksesta englannin oppimiseen ja tarjonnut mahdollisuuden hyödyntää tätä tietoa oppimisen ja opettamisen kehitystyössä.