UNIVERSITY OF TARTU
FACULTY OF ARTS AND HUMANITIES
INSTITUTE OF ESTONIAN AND GENERAL LINGUISTICS

Ryomei Ueda

Subject pronoun omission in written Estonian

Master's thesis

Supervisors: lecturer Helen Hint and professor Liina Lindström

Tartu 2024

Affirmation of authorship

I confirm that I have written this thesis myself and have correctly cited the contributions of other authors. The work was written based on the thesis requirements of the Institute of Estonian and General Linguistics of the University of Tartu and is in line with good academic practices.

Ryomei Ueda

Short summary

In terms of the system of pronouns, Estonian is regarded as a partial pro-drop language. More concretely, there are three possible pronoun choices. They are namely the short form, the long form and the zero reference. Therefore, this paper will analyze data extracted from a corpus in order to find out when and under which conditions subject pronoun omission occurs in Estonian. Different factors are taken into consideration which could be influential when one opts for an overt pronoun or instead omitting it. In addition, the choice between the short and long pronoun forms will be investigated as well. Estonian National Corpus 2021 was utilized to obtain the data. The results of the study indicate that different factors such as person and verb do play a role when it comes to subject pronoun omission.

Keywords: pronoun, subject pronoun omission, language variation, syntax, Estonian

Table of contents

1. Introduction

Subject pronoun omission is a concept that opposes overt subject pronoun expression. The first term means avoiding the explicit mention of the subject while the second one is the exact opposite, meaning that the subject is expressed with a pronoun. Below is a pair of examples (1 and 2) in Estonian to help better understand the two terms. The first sentence is a clear example of subject pronoun omission as the subject pronoun of the verb *soovima* 'to want' is absent. In the second sentence, on the other hand, overt subject pronoun expression can be seen since the subject pronoun of the verb *kavatsema* 'to plan' is present (*ma*).

(1) *Soovi-n* *pitsat* *süüa.*

want-PRS.1SG pizza.PART.SG eat.INF

'I want to eat pizza'.

(2) *Ma* *kavatse-n* *seal* *käia.*

1SG.NOM plan-PRS.1SG there go.INF

'I am planning on going there'.

Although some other terms such as pro-drop (pronoun-dropping) (e.g. Pešková 2013), null-subject (e.g. Biberauer et al. 2009), zero reference (e.g. Hint 2021) have been used to refer to the same concept as subject pronoun omission, this paper utilizes only the last

term as it is free of terminology used in generativist paradigm and has been employed in many other studies on Estonian (e.g. Lindström & Vihman 2017). However, exceptions will be made when directly paraphrasing previous studies.

From here on, the term *overt* and *zero* will be sometimes utilized for the sake of simplification. *Overt* means *overt subject pronoun expression* whereas *zero* refers to *subject pronoun omission*.

When dealing with different null-subject types, languages can be divided into four groups: expletive, partial, discourse and consistent null-subject languages (Biberauer et al, 2009). One example of expletive null-subject languages is German, in which non-dummy pronouns have to be overtly expressed. Romance languages such as Spanish and Italian, for instance, fall into the consistent null-subject language category. These are inflected languages in which rich verbal agreement can be found. Japanese, on the other hand, is a language in which there is no verbal agreement unlike Romance languages. Still subject pronoun omission occurs anyway, therefore it is regarded as a discourse null-subject language (Biberauer et al, 2009). The target language of this paper is Estonian, which is a partial pro-drop language (Metslang 2009). In addition to Finnish, another Finnic language, Hebrew, Russian, Icelandic and Marathi, for example, belong to the group as well (Biberauer et al, 2009).

Regarding partial pro-drop languages, the omission of pronouns occurs under restricted conditions. Additionally, there are two overt forms of personal subject pronouns in

Estonian which could be employed besides subject pronoun omission, namely the short and long forms. Thus, the main goal of this paper is to investigate when subject pronoun omission occurs in Estonian depending on different factors and also to find out what influences the choice between the use of long pronoun form and that of short one by utilizing Estonian National Corpus 2021. In the past, subject pronoun omission in Estonian has been dealt with by different authors (Duvallon & Chalvin 2004, Lindström et al. 2009). Especially, overt and zero variation regarding third person has been actively addressed (Hint 2015, Hint 2021, Hint, Reile & Kaiser 2023). In addition, a bachelor's thesis was written before where the main topic was the use of pronouns in Messenger texts (Sepp 2010). Nonetheless, there has not been a study where subject pronoun omission across different persons and numbers was thoroughly investigated. There have been studies such as the one conducted by Lindström and Vihman (2017), where subject-like argument omission was analysed (Lindström & Vihman 2017). However, the study did not tackle subject pronoun omission. Therefore, this study will aspire to be the first in exploring the system of subject pronoun omission in Estonian exhaustively. The focus of the study is on written Estonian, which might behave differently from spoken Estonian regarding subject pronoun omission.

The paper will first address previous studies related to subject pronoun omission not only in Estonian but also in general to give a clear idea about what is already known about the topic. Then, it will move onto the part where the methodology employed for the study will be described in detail. The section will be divided into three: 1) corpus, which was utilized for collecting the data, 2) verbs, which were chosen to facilitate the study, and 3) factors, which were taken into consideration to see if any of them are indeed influential

in terms of subject pronoun omission. Afterwards, the results of the study will be shared in detail along with discussions. Finally, conclusions will be made to summarize the thesis.

2. Background

2.1. General description

Before diving into previous studies, below (example 3) are three sentences respectively in Estonian, Spanish and English, all of which mean the same thing. This would help understand different kinds of subject pronoun omission. The pronouns are marked with bold letters, and Ø means zero.

(3)

ET: ***(Mina/ma/Ø)*** *arvan, et* ***(sina/sa/Ø)*** *pead sinna minema, aga* ***(meie/me/Ø)*** *jääme siia.*

ES: ***(Yo/Ø)*** *creo que* ***(tú/Ø)*** *tienes que ir allí, pero* ***(nosotros/nosotras/Ø)*** *nos quedamos aquí.*

EN: ***I*** *think that* ***you*** *have to go there, but* ***we*** *will stay here.*

Although subject pronoun omission can take place in certain contexts in English too (Haegeman & Ihsane 2001), the only plausible English sentence would be the one written above. However, it is possible to construct 12 slightly different sentences with the Spanish one. It is worth mentioning here that Spanish has three options when referring to first person plural entity as *nosotros* denotes an exclusive male group while *nosotras* indicates a group of people of which at least one person is female. As for Estonian, there are three pronoun options (short, long and zero) each, consequently 27 possible sentences in total could be built although sometimes the omission is more natural while at other times it is

less so, which is the reason why this study will try to find out what exactly it is that makes it natural to drop a subject pronoun.

In Estonian, personal pronouns could be either short or long. Below is a table of all the subject pronouns in the language (table 1). The word on the left side is the short form while the one on the right side is the long form.

Table 1. Subject personal pronouns in Estonian

	Singular	Plural
First person	*ma/mina*	*me/meie*
Second person	*sa/sina*	*te/teie*
Third person	*ta/tema*	*nad/nemad*

The subject pronoun system is quite clear based on the table above. As for the difference between the short and long forms, according to Pajusalu and Pajusalu (2004), the choice between them depends mainly on pragmatic reasons as the long one is usually employed when contrast is aimed to be emphasized or when the speaker shows opposition to someone else (Pajusalu & Pajusalu 2004). Moreover, the unmarked and neutral choice to refer to third person singular entities that were already mentioned in the ongoing discourse is the short form *ta* (Pajusalu 2009). Another study by Pajusalu (2005) showed that the long form (*tema*) is used more frequently to refer to animate entities than the short form (*ta*) (Pajusalu 2005). One thing to mention here is that second person plural *te*/*teie* can

also be used to refer to second person singular in a polite way like *vous* in French. This study does not differentiate the two types of *te*/*teie*. Nonetheless, there might be interesting differences between the two, which could be studied in future studies connecting subject pronoun omission to politeness.

Since the main topic of this paper is personal pronouns rather than demonstrative pronouns, the latters will not be dealt with in detail. However, below is a brief description and summary of them in Estonian in table 2. After all, demonstrative pronouns are related to the study in that they are pronouns and can be used to refer to third person singular and plural.

Table 2. Demonstrative pronouns in Estonian

	Proximal	Distal
Singular	*see*	*too*
Plural	*need*	*nood*

Demonstrative pronouns in Estonian are utilized mainly to refer to inanimate entities. (Pajusalu 2009, Hint, Nahkola & Pajusalu 2020)**.** In Estonian, there are two stems when it comes to demonstrative pronouns according to how far the entity/object is from the subject's point of view. The first one is proximal *see*, which would correspond to *this* in English as it is used to refer to something closer. The second one is distal *too*, which would be equivalent to *that* in English. The distal variant *too* is rarely used in Standard Estonian (Reile 2016, Reile 2019, Pajusalu 2009).

Here, Estonian verb endings will be briefly discussed as they are fundamentally related to the topic of the study. The Estonian verb system has inflections, which in general means that the ending of a verb gives clear information about who the subject is without the overt use of subject pronoun. Table 3 and the example (4) below show how to inflect the verb *teadma* 'to know'.

Table 3. Inflections of the verb *teadma* 'to know'

	Singular	Plural
First person	*tea-n*	*tea-me*
Second person	*tea-d*	*tea-te*
Third person	*tea-b*	*tea-vad*

(4) *Tea-me* *seda.*
know-PRS.1PL this.PART
'We know it.'

In example (4), verbal agreement is visible, and the ending *-me* clearly indicates that the subject is first person plural.

2.2. Pronoun omission in languages

From here on, previous studies related to the current study will be looked at. First of all, omitting the use of pronouns would respect the principle that linguistic economy is aimed to be achieved (Hint, Reile & Kaiser 2023). On the other hand, overt subject pronoun

expression is preferred possibly due to the fact that language users ponder the necessity of using a pronoun to help a reader/listener cognitively (Hint 2015).

2.2.1. Pronoun omission in Spanish

According to a study carried out by Otheguy, Zentella and Livert (2007), which investigated different Spanish varieties, the result showed that the percentages of overt subject pronoun were higher in dialects spoken in the Caribbeans such as Puerto Rico and Dominican Republic (Otheguy, Zentella & Livert 2007). This indicates that there are differences between varieties even within the same language.

One study conducted by Pešková (2013), which focused on the relations between pro-drop and persons/numbers in spoken Porteño Spanish revealed that subject pronouns were absent 52% of the times while 48% of the times they were overtly expressed. Another result from the same study showed that the frequency of overt pronoun expression was higher in singular (51%) than in plural (44%). Besides, the result of the study also demonstrates that the percentage of overt pronoun varies depending on the person and number as the percentage of overt was 47% with first person singular, 33% with second person singular (familiar), 70% with second person singular (formal), 55% with third person singular, 36% with first person plural, 47% with second person plural, 48% with third person plural. When it comes to different kinds of verbs in Spanish, the same study also revealed that epistemic verbs (57%) such as *creer* 'to think' and *saber* 'to know' were more likely to be accompanied by a pronoun than perceptive verbs (39%) such as *escuchar* 'to listen to' and *mirar* 'to watch'**.** The study also compared the result based on types of sentences, and it turned out that interrogative sentences were more likely to have

a subject pronoun (wh-interrogatives 53% and absolute interrogatives 52%) than declarative ones (48%) (Pešková 2013).

All in all, Pešková's (2013) study suggested that rich verbal agreement does not have to be a main pro-drop motivator, therefore Pešková (2013) is of the opinion that an indirect relation exists between rich verbal agreement and pro-drop, meaning that the existence of rich agreement encourages pro-drop to occur (Ackema & Neeleman 2007).

2.2.2. Pronoun omission in Finnish

As Estonian is a Finnic language that belongs to Finno-Ugric languages, previous studies related to Finnish, another Finnic language, will be briefly addressed in this part.

According to Metslang (2009), a common feature among typical Standard Average European is that they are non-pro-drop languages. Estonian and Finnish, however, are both pro-drop languages. One thing to note here is that it is not always possible to infer the person based on the verb forms in Estonian. For example, the quotative mood uses the same ending *-vat* for all the persons and numbers. Furthermore, no clue is given with the verb ending when a sentence is negative in Estonian. In Finnish, on the other hand, it is possible to tell who the subject is based on the negative particle used before the verb (Metslang 2009). In addition, it is known that subject pronoun omission is preferred in Standard Finnish while overt subject pronoun expression is more common in colloquial varieties (Helasvuo & Laitinen 2006).

A study by Hint, Nahkola and Pajusalu (2020), which compared Estonian, Finnish and Russian from the point of view of third person singular referential devices, revealed that

the use of personal pronouns (52.8%) was more common than zero reference (37.1%) in Estonian. Finnish, on the other hand, exhibited an interesting contrast to Estonian as zero reference (36.1%) and personal pronouns (35.1%) were used almost equally often. The same study also showed that in general both Estonian and Finnish used zero reference similarly, and it occurs frequently only under certain conditions (Hint, Nahkola & Pajusalu 2020).

One study conducted by Duvallon and Chalvin (2004), which tackled the topic of subject pronoun omission in Finnish and Estonian, the omission occurred more often in second person (27%) than in first person (18%) in Finnish (Duvallon & Chalvin 2004).

According to a study led by Helasvuo and Kyröläinen (2016), where first person pronoun was investigated, overt subject pronoun expression (88.8%) occurred far more often than subject pronoun omission (11.2%). In the same study, other than referential distance, which will be explained later, it was found that syntactic complexity was also relevant to the choice of first person nominative subject as zero was preferred when the context was syntactically simple (Helasvuo & Kyröläinen 2016).

2.2.3. Pronoun omission in Estonian

This section describes the phenomenon of subject pronoun omission in Estonian. The study above mentioned by Duvallon and Chalvin (2004) revealed that the percentage of subject pronoun omission was 18% when the subject was first person and 49% when it was second person singular based on their analysis of spoken data (Duvallon & Chalvin 2004). Additionally, a study led by Lindström et al. (2009), which compared different

dialects, showed that the rates of first person singular subject omission ranged from 10.8% to 54.3% (Lindström et al. 2009). This means that the probability of subject pronoun omission varies a great deal depending on the dialect. It is also known that subject pronoun omission occurs commonly in spoken Estonian, especially if the subject is deictic, in other words, the first and second persons (Vihman 2015).

Concerning third person zero reference, it is known that the occurrence of subject pronoun omission is more restrictive if the referent is third person (Hint 2015). In addition, the tendency that third person is less likely to be omitted is related to the fact that third person instances without an overt pronoun can be read as generic in Estonian (Vihman 2015, Erelt et al. 2017), as in example (5).

(5) *Siin* *saa-b* *joosta.*
Here can-PRS.3SG run.INF
'You (generic) can run here'.

Another study carried out by Hint, Reile and Kaiser (2023) suggested that the organization of sentences such as the use of the conjunction *ja* 'and' or a full stop between two clauses could have an influence on the choice of referential device as the result showed that when *ja* was utilized between two consecutive clauses which have the same subject, subject pronoun omission was preferred (Hint, Reile & Kaiser 2023).

As for different types of clauses, which will not be dealt with in this study, it was revealed that overt pronoun expression is more common than subject pronoun omission when it is

the subject of a subordinate clause. On the other hand, this does not apply to main clauses as the two choices are employed almost the same amount (Hint, Nahkola & Pajusalu 2020).

Although the current study focuses only on pronoun omission that concerns subjects (nominative), object pronoun can be dropped in Estonian, especially when it is possible for the speaker and hearer to understand who the referent is based on its high saliency (Hint 2015, Vihman 2015).

Concerning transitivity, Metslang (2013) found that the subject omission of transitive constructions (39%) was more common than that of intransitive constructions (30%) (Metslang 2013).

Regarding polarity, it is known that subject pronoun omission is unlikely to occur in negative sentences probably because negative verb forms do not indicate explicitly who the subject is (Pajusalu & Pajusalu 2004). However, subject pronoun omission does occur even when the subject is not inferable (Metslang 2009).

As written earlier, factors which could be influential concerning the topic of the paper will be investigated. The approach that takes into account multiple factors including referential distance, which confronts the salience-only perspective (Gundel, Hedberg & Zacharski 1993), has been proclaimed for the proper description of referential devices (Kibrik et al. 2016). According to a study already mentioned above by Hint, Nahkola and Pajusalu (2020), in which referential distance was measured taking utterance as the

counting unit, the result showed that referential distance indeed had an influence on the choice of referential devices (Hint, Nahkola & Pajusalu 2020). Moreover, it was found that referential distance was an important factor in Estonian with first person (Lindström et al. 2009). In another study carried out by Lindström and Vihman (2017), however, it turned out that referential distance was not an influential factor that caused experiencer argument omission. Instead, it affected the choice of case marking (Lindström & Vihman 2017).

Lastly, linguistic factors are not the only ones which affect referential choices as non-linguistic factors play a role as well (Heine 2019, Vogels, Krahmer & Maes 2018).

2.3. Hypotheses

Based on the previous studies mentioned above, common knowledge and the author's intuition, eight hypotheses have been formulated to see if they will be confirmed in this study. The hypotheses are listed below in a random order.

1. Subject pronoun omission would be less likely to occur with third person.
2. The long pronoun form would be utilized more often after the verb.
3. Referential distance would be an influential factor, meaning subject pronoun omission would be likely to occur when the referential distance is smaller.
4. Verbs which have to do with emotions and opinions would prefer overt pronoun expression.
5. Subject pronoun omission would occur more often with transitive verbs than intransitive ones.

6. Verb endings which do not indicate who the subject is would be likely to be accompanied by an overt pronoun.
7. Tense would be unlikely to have a huge influence on the choice between subject pronoun omission and overt subject pronoun expression.
8. The conditional mood is over all more likely to use an overt pronoun.

3. Methodology

This section will give details of the methodology of the study. Concretely, 1) how the data was collected; 2) the nine verbs, which were chosen for the study; 3) the factors, which were taken into account will be explained here.

3.1. Data

In order to collect the data, Estonian national corpus 2021, which is accessible on Sketch Engine (https://www.sketchengine.eu/), was utilized. The corpus contains 2,945,431,278 tokens, and it is one of the most versatile Estonian language corpora online (Koppel & Kallas 2022). Data extraction was conducted by getting instances of nine verbs which had been chosen beforehand. Except for the two verbs *muutma* and *muutuma*, 200 instances were collected for each verb. With the two verbs mentioned above, only 100 instances were taken out as the number of instances suitable for the study was rather small. Thus, 1,600 instances in total were extracted from the corpus. Through the process, a useful feature on the corpus called CQL in concordance was utilized to facilitate the research as it helps exclude the instances which would not fit into the study. For example, as the topic of this paper is subject pronoun omission, impersonal instances such as *öeldake* should not be included, and it is possible to exclude them by making the most of the feature. In order to optimize data extraction, the tag [lemma="verb"& features!="da|des|ge|gem|gu|ma|maks|mas|mast|mata|ta|takse|ti|tud"] was utilized to leave out instances related to, for example, infinitive, imperative and impersonal.

To give an example, the following tag was used when searching for the results with the verb *teadma*:

[lemma="teadma"&features!="da|des|ge|gem|gu|ma|maks|mas|mast|mata|ta|takse|ti|tud"].

Despite the complexity of the tag, there were still some irrelevant or unsuitable instances left which were manually removed by the author.

After having the corpus show all the suitable instances by using the tag above, genre was restricted, and blog was selected as the target genre as it was likely that it contained sentences including different persons and numbers. In terms of register, both formal and informal texts were found in this genre like politics (formal) and casual blogging (informal). After this, random instances were chosen by utilizing a feature within the corpus which allows its users to get results arbitrarily.

In short, 1) all the suitable instances were shown on the corpus after making the most of the CQL feature; 2) the genre was restricted only to blog; 3) 200 (or 100) instances were randomly chosen. Later, the data collected on the corpus were coded on an Excel file based on different factors that could be considered influential as far as subject pronoun omission is concerned. After this, pivot tables based on the coded data were created and analyzed.

3.2. Verbs

As briefly mentioned above, nine verbs were chosen to conduct the study. They were namely *pidama*, *arvama*, *teadma*, *muutma*, *muutuma*, *jooksma*, *nägema*, *tahtma* and

ütlema. Each verb represents respectively obligation, thought, knowledge, transitive, intransitive, motion, perception, emotion, communication verbs. All of the verbs are put together in table 4 below with an example.

Table 4. Verbs and their characteristics

Verb	Characteristics	Example
Pidama 'to have to'	Obligation	*Nüüd **pean** seda tegema.* 'Now, I have to do it'.
Arvama 'to think'	Thought	***Arvad**, et see on tore.* 'You think it is great'.
Teadma 'to know'	Knowledge	*Me **ei teadnud** tõde.* 'We didn't know the truth'.
Muutma 'to change'	Transitive counterpart	*Sa **muutsid** suhtumist.* 'You changed the attitude'.
Muutuma 'to change'	Intransitive counterpart	***Muutume** iga päev.* 'We change every day'.
Jooksma 'to run'	Motion	*Teie **jooksete** kogu aeg.* 'You run all the time'.
Nägema 'to see'	Perception	*Nad **nägid** meid eile.* 'They saw us yesterday'.
Tahtma 'to want'	Emotion	*Ta **tahab** puhata.* 'He (or She) wants to rest'.
Ütlema 'to say'	Communication	*Ma **ei ütle** mitte midagi.* 'I won't say anything'.

This study is largely based on these verb choices as one of the goals is to try to find out if certain verbs prefer subject pronoun omission or not due to the characteristics peculiar to each verb. All the verbs selected for the research are among the top 50 most frequent verbs on the corpus except for *muutuma* (the 51st) and *jooksma* (the 137th). The main reason why frequent verbs were chosen is that this way there would be enough instances to carry out the study smoothly. Some of the verbs above were the only plausible choice that represents the characteristics such as the verb *teadma* while others were selected among several possible choices by the author. In addition, *muutma* and *muutuma* as a pair were chosen to see if transitivity has any effect on pro-drop. *Jooksma* ended up representing the motion verbs because it turned out to be the best candidate. The verb *tulema*, for example, was not the most convenient choice since the third person singular form of the verb has another meaning, which is obligation like *mul tuleb tööd teha* 'I must work'. Moreover, some instances which have nothing to do with the main characteristics of the verbs were removed like the example (6) below, which contains the verb *pidama*, but it is not related to obligation in this case.

(6) *Ma* *pea-n* *sind*
1SG.NOM consider-PRS.1SG 2SG.PART
targa-ks.
smart.TRANSL.SG
'I consider you smart'.

Furthermore, when the verb *nägema* is accompanied by the word *välja*, it means *to look*, which does not qualify for the study either (example 7).

(7) *Sa* *näe-d* *hea* *välja.*

2SG.NOM see-PRS.2SG good.NOM.SG out

'You look good'.

Lastly, the judgement of generic reading was done by the author, and instances with third person singular which could be read as generic were not included. An example (8) is given below.

(8) *Pea-b* *kooli-s* *käima.*

have to-PRS.3SG school-INE.SG go.SUP

'You (generic) have to go to school'.

However, second person singular generic reading was not specifically excluded as it was rather difficult to judge such instances. In addition, ambiguous instances were excluded as well to increase the accuracy of the analysis.

3.3. Factors

First of all, two dependent variables (pronoun type and pronoun form) were coded. **Pronoun type** was coded either zero or overt. Overt means the use of either short or long form of pronouns (example 9). When the coding for pronoun type was zero, the referent was inferred based on verb endings or contextual evidence (example 10).

(9) *Nad tahavad Rootsis käia.* (overt)

'They want to go to Sweden'.

(10) *Arvame, et pole vaja seda õppida.* (zero)

'We think that it is not necessary to study it'.

Pronoun form was coded short (example 11), long (example 12) and NA, when zero was used.

(11) *Sa oled natuke muutunud.* (short)

'You have changed a little bit'.

(12) *Mina ütleksin, et nad on süüdi.* (long)

'I would say that they are guilty'.

The factors listed below are the ones which this study takes into consideration that are meant to explain the system of the dependent variables above mentioned. Those that are considered to have an influence are position, person, number, mood, form of conditional, tense, negation, referential distance, case marking of the previous mention, animacy and verb lemma.

Concerning the position of the pronoun, typical word order in Estonian is SVO (Lindström 2001), but the order is relatively free (Lindström 2017), therefore a pronoun could be before the verb (SV) or after the verb (VS). **Position** was then coded by left, right and NA. If the pronoun was before the verb, the instance was coded as left (SV, see example 13), and if it was after the verb, it was coded as right (VS, see example 14).

(13) *Ma pidin vett jooma.* (SV)

'I had to drink water'.

(14) *Eile jooksite teie pargis koos.* (VS)

'Yesterday you ran together in the park'.

Person was coded by 1st, 2nd and 3rd. As mentioned earlier, third person singular generic reading was not applicable to the paper. Additionally, proper nouns and demonstrative nouns like *see* and *need* were out of the scope as well. See example (15) as an instance of first person plural pronoun,

(15) *Me ei jookse homme.* (1st)

'We won't run tomorrow'.

Number was coded by singular and plural (example 16).

(16) *Ta ütles et, see ei ole oluline.* (singular)

'He (or She) said that it was not important'.

In Estonian, there are five moods in total, namely indicative, imperative, conditional, quotative and jussive. In this study, **mood** was coded by indicative (example 17), conditional (example 18) and quotative (example 19). Since pronouns tend to be omitted when the imperative mood is utilized, this mood was not included in the study. Jussive

was left out too since there were not enough instances for analysis.

(17) *Meie nägime sind tantsimas.* (indicative)

'We saw you dancing'.

(18) *Ma tahaksin ujuda.* (conditional)

'I would like to swim'.

(19) *Sa teadvat seda.* (quotative)

'You are reported to know it'.

Speaking of moods, **form of conditional** was coded specifically to see if the shorter form of the mood would make it less likely for subject pronoun omission to occur. The coding for this factor was shot, long, NA and NR. The longer form like *peaksid* was coded as long (see also example 20) whereas the shorter form like *peaks* (see also example 21) was coded as short. NA refers to the instances whose mood was not conditional. NR was used to indicate the instances where the conditional mood was used but at the same time where further analysis seemed meaningless.

(20) *Ma peaksin nendega rääkima.* (long)

'I should talk to them'.

(21) *Nad näeks seda täna õhtul.* (short)

'They would see it tonight'.

In Estonian, tenses can be divided into four. Accordingly, the category of **tense** was coded by present, perfect (example 22), past and pluperfect.

(22) *Sa oled seda enne mulle öelnud.* (perfect)

'You have said it to me before'.

Negation was coded by either yes or NA. Yes means negative (example 23) while NA equals affirmative.

(23) *Tema ei muutu kunagi igavaks.* (negative)

'He (or She) will never become boring'.

As mentioned in the backgroud section, **referential distance** means the distance between the subject and the previous mention of the referent. The distance was measured and later categorized as 1, 2, 3, 4+ and unclear. A number bigger than 4 was categorized together as 4+ to simplify the analysis. Here, unclear means such situations where the referent was not trackable when the data was coded. In this study, clause was the unit which was utilized to measure the distance. Below, examples (24 and 25) demonstrate how referential distance was counted. The subject in question is in the bold letters.

(24) *Mul oli lõbus, aga **ma** olen praegu väsinud.*

'I had fun, but I am tired now'.

Referential distance: 1 (*mul* as the previous mention)

The referential distance above is counted as one as the clause right before the subject *ma* uses *mul*, which refers to the same referent.

(25) *Me käisime seal ja vaatasime filmi koos. Pärast seda **me** sõime restranis.*

'We went there and watched a movie together. After this, we ate at a restaurant'.

Referential distance: 2 (*me* as the previous mention)

In example (25), the clause before the first plural subject *me* does not have an overt pronoun. The clause before it, however, has the same referent *me*, therefore the referential distance is two.

The next factor which was coded is **case marking of the previous mention**. Below is an example (26) where the subject in question is in the bold letters.

(26) *Mul oli seda vaja, seega **ma** tegin seda.*

'I needed it, therefore I made it'.

Case: adessive (*mul*)

In this case, the previous mention is in the form of *mul* (adessive), thus it is coded as adessive.

In Estonian, there are 14 cases. However, not all of them were utilized as the last form of mention. Out of the total 14, 11 cases were found. Namely, nominative, genitive, partitive, illative, inessive, elative, allative, adessive, ablative, abessive, comitative excluding

translative, terminative and essive. In the coding, NA (the ones coded as unclear in the referential distance section) was included as well.

Animacy was coded by yes when the subject was animate (example 27) and by NA when it was an inanimate entity.

(27) *Te teate, et me oleme poes.* (animate)

'You know that we are in the shop'.

Verbs were coded by the nine verbs already mentioned above (example 28).

(28) *Nad tahtsid neid raamatuid.* (*tahtma*)

'They wanted those books'.

Table 5, which summarizes all the factors that were coded in the study, is presented in the next page.

Table 5. Factors

Factor	Coding
Pronoun type	Zero, overt
Pronoun form	Short, long, NA
Position	Left, right, NA
Person	1st, 2nd, 3rd
Number	Singular, plural
Mood	Indicative, conditional, quotative
Form of conditional	Short, long, NA, NR
Tense	Present, perfect, past, pluperfect
Negation	Yes, NA
Referential distance	1, 2, 3, 4+, unclear
Case marking of the previous mention	Nominative, genitive, partitive, illative, inessive, elative, allative, adessive, ablative, abessive, comitative, NA
Animacy	Yes, NA
Verb lemma	*Pidama*, *arvama*, *teadma*, *muutma*, *muutuma*, *jooksma*, *nägema*, *tahtma*, *ütlema*

4. Results and discussion

The effect of each factor was estimated by using Excel pivot tables. In this section, the results are presented as contingency tables.

4.1. General

Before diving into the detailed results, general results will be given here. As mentioned earlier, there were in total 1,600 instances extracted from the corpus for this study. Out of them, it turned out that 790 were instances in which an overt pronoun, either short or long, was found while 810 instances appeared without a pronoun. This results in an almost even percentage of the two variants (overt 49.4% and zero 50.6%). It is in fact interesting that the percentage of zero turned out to be even slightly higher than that of overt.

When it comes to pronoun form, out of the 790 instances which had an overt pronoun, 702 utilized the short form while 88 employed the long one. This shows that the short form was significantly dominant over the long one (short 88.9% and long 11.1%). This is not a surprising result as the common use is the short form in Estonian. Either way, the current study succeeded in showing the distribution of the two options. Over all, the results above mean that out of all the instances extracted for the study, the distribution of the three possible choices was zero 50.6%, short 43.9% and long 5.5%.

Concerning the position, there were 678 instances of left from the verb (SV) and 112 instances of right from the verb (VS) out of the 790 overt instances, resulting in the subject

usually appearing before the verb (left 85.8% and right 14.2%). This result reflects the typical Estonian word order as Estonian is an SVO language which is quite flexible with word order (Lindström 2001, Lindström 2017). Hypothesis no.2 was not confirmed by the data (the long pronoun form would be utilized more often after the verb). When the long form was utilized, the percentage of left was 84.1% (74 instances) while that of right was 15.9% (14 instances). When the short form was employed, on the other hand, the percentage of left was 86.0% (604 instances) while that of right was 14.0% (98 instances). See table 6 below,

Table 6. Pronoun form and position

	Total count	SV count (%)	VS count (%)
Short	702	604 (86.0%)	98 (14.0%)
Long	88	74 (84.1%)	14 (15.9%)

Therefore, it seems that there is no high correlation between pronoun form and position. As for the VS instances, the reversion sometimes occurred due to the V2 principle. One such sentence is presented below in example (29).

(29) *Samuti teadsid nad, et rannas on 40+ inimest, kes otsivad maja, kus pidu panna.*
'They also knew that there were 40+ people on the beach who look for a house to throw a party'.

In the sections to follow, different factors will be analyzed in detail to better understand the patterns of subject pronoun omission in Estonian.

4.2. Person

In general, the instances with first person were dominant in the dataset, taking up more than two-thirds of the total 1,600 instances (71.6%). Nonetheless, there were enough instances with second and third persons to analyze tendency based on person. The result can be seen in table 7 below. According to the data, it turns out that third person seems to incline towards overt subject pronoun expression while second person shows a slight inclination towards subject pronoun omission. This finding is in accordance with previous studies (e.g. Hint, Reile & Kaiser 2023). The hypothesis no.1 (subject pronoun omission would be less likely to occur with third person) was confirmed by the result.

Table 7. Person and pronoun type

	Total count	Overt count (%)	Zero count (%)
First	1145	559 (48.8%)	586 (51.2%)
Second	202	89 (44.1%)	113 (55.9%)
Third	253	142 (56.1%)	111 (43.9%)

4.3. Number

In the dataset, a dominant number of instances was in singular (82.3%) while a small yet sufficient number of instances was in plural. Based on the data, plural appears to incline slightly towards the omission of subject pronoun compared to singular. The result is

summarized below in table 8. It resembles the result of one Spanish study previously mentioned (Pešková 2013). As for Estonian, it is possible that the percentage of overt is lower due to the fact that the overt use of first person plural *me* and second person plural *te* is preferably avoided in written language.

Table 8. Number and pronoun type

	Total count	Overt count (%)	Zero count (%)
Singular	1317	663 (50.3%)	654 (49.7%)
Plural	283	127 (44.9%)	156 (55.1%)

4.4. Person and number

When person and number are combined, the result looks like the table 9 in the next page. There are three eye-catching findings to pay attention to here. First, third person plural is the one that utilizes an overt pronoun the most even though plural in general seems to prefer omitting a pronoun as noted above. This is not in accordance with the study conducted by Hint (2015) as the result indicated the opposite tendency showing that third person plural zero was more often than third person plural overt (Hint 2015). However, it has to be mentioned that the number of third person plural instances in total was quite small (62 instances), especially compared to the instances of first person singular. The second peculiarity in the table is that second person plural utilizes subject pronoun omission very often compared to the singular equivalent and generally speaking as well. Although the fact that the total number of second person plural instances was rather small (73 instances) has to be kept in mind in this case too, it is the most outstanding result among other things here. As written above (See section 4.3.), the avoidance of overt *te* in

written language could be considered as a reason. Although this study does not differentiate the two types of second person plural in Estonian, if they were distinguished, it would be possible that politeness has something to do with this result like it was the case with Spanish (Pešková 2013). The third point to mention is that the percentage of overt in first plural is slightly lower than the singular counterpart. This is probably owing to the fact that first person singular tends to be used more often to express feelings or opinions, for which the overt use makes sense.

Table 9. Person/number and pronoun type

	Total count	Overt count (%)	Zero count (%)
First singular	997	493 (49.4%)	504 (50.6%)
Second singular	129	64 (49.6%)	65 (50.4%)
Third singular	191	106 (55.5%)	85 (44.5%)
First plural	148	66 (44.6%)	82 (55.4%)
Second plural	73	25 (34.2%)	48 (65.8%)
Third plural	62	36 (58.1%)	26 (41.9%)

Table 10 in the next page focuses on the relation between overt personal pronoun and its position in regard to the verb. It is clear that the third person preferred having a pronoun after the verb compared to the other two persons (example 30).

(30) *Kuid arvatavasti on tal mingi salaplaan või tahab **ta** lihtsalt endale ja teistele mujlet avaldada.*

‘But supposedly he (or she) has a secret plan or simply wants to impress himself (or herself) and others’.

Table 10. Person and position

	Total count	SV count (%)	VS count (%)
First	559	491 (87.8%)	68 (12.2%)
Second	89	76 (85.4%)	13 (14.6%)
Third	142	111 (78.2%)	31 (21.8%)

When the forms were compared, third person behaved differently this time again. Below in table 11, it appears that third person slightly prefers the use of the short form more than the other two persons.

Table 11. Person and pronoun form

	Total count	Short count (%)	Long count (%)
First	559	493 (88.2%)	66 (11.8%)
Second	89	79 (88.8%)	10 (11.2%)
Third	142	130 (91.5%)	12 (8.5%)

Comparing the difference between singular and plural, it turned out that the use of the long form is more common in the plural section. The result is presented in table 12. See also example (31),

(31) *Mida teie muideks sellest teooriast arvate?*

‘By the way, what do you think about this theory’?

Table 12: Number and pronoun form

	Total count	Short count (%)	Long count (%)
Singular	663	596 (89.9%)	67 (10.1%)
Plural	127	106 (83.5%)	21 (16.5%)

4.5. Mood

Most of the instances in the data were either in the indicative or conditional mood, and there were only five occurrences with the quotative mood, all of which were accompanied by an overt pronoun. It has to be stated, however, that the total number of instances was too small to claim anything concrete here. Either way, it is logical that the presence of an overt pronoun is preferred to make it clear who the subject is as all the persons/numbers with the quotative mood utilize the same ending *-vat* (see example 32). The hypothesis no.6 (verb endings which do not indicate who the subject is would be likely to be accompanied by an overt pronoun) is in accordance with the result.

(32) *Sest ta olevat ikka mööda tervet Rutu tänavat jooksnud ja nurga tagant tulnud, sest kaamera on nurga pääl.*

‘Because he (or she) is reported to have run along the entire Rutu street and come from behind the corner since the camera was on the corner’.

Moving onto the two main moods of the study, which are indicative and conditional

(example 33), there were much more indicative instances than conditional ones as imagined. The differences between these two moods were rather clear. Table 13 clearly exhibits this contrast. As mentioned earlier, Estonian conditional mood has two variants, either long or short ending, which will be analyzed later. As the short ending does not give any clear idea about who the subject is, it is expected that the use of an overt pronoun would be preferred in this case. This is very likely the main reason why the percentage of overt turned out to be relatively high in conditional. The following section was established in order to make sure that the speculation, which sounds reasonable, is indeed correct. Anyway, this result confirms the hypothesis no.8 (the conditional mood is over all more likely to use an overt pronoun.).

(33) *Ma tahaks seda lugeda, ka see oleks väga vajalik üles riputada.*

'I would like to read it, also it would be very necessary to upload it'.

Table 13. Mood and pronoun type

	Total count	Overt count (%)	Zero count (%)
Indicative	1471	48.2% (709)	51.8% (762)
Conditional	124	61.3% (76)	38.7% (48)
Quotative	5	100% (5)	0% (0)

4.6. Form of conditional

By focusing only on the conditional mood in the dataset, the short and long endings were compared. See examples (34 and 35) in the next page,

(34) *Ma tahaks seda lugeda.*

'I would like to read it'.

(35) *Millises koolis tahaksin õppida?*

'In what kind of school would I like to study'?

Out of the total 124 conditional instances, there were 106 instances which were appropriate to be analyzed here to compare the two different endings. For example, the sentence (36) below is a negative sentence which is irrelevant and not suitable.

(36) *Aga elada ma tõesti ei tahaks ei maal ega linna lähedal oma majas.*

'But I really would not like to live neither in the countryside nor in a house near the city'.

After the manual removal of such instances, there were 71 instances with the long ending and 35 instances with the short ending. Table 14 in the next page shows the clear difference between the two variants. Again, the number of instances was not big at all, but still the difference has to be noticed. In a nutshell, the use of the short ending means a higher probability of overt subject pronoun expression as it is not clear who the subject is based on the ending like the quotative instances, confirming again hypothesis no. 6. However, 25.7% is a considerable proportion, and it is interesting that quite often the subject of a conditional sentence with the short ending seems to be omitted. Looking back at the previous section and comparing the result of the indicative mood and the long ending, the percentages are almost the same, meaning that the mood itself was apparently

not a relevant factor regarding subject pronoun omission. It is largely because of the existence of the short ending that the conditional mood over all is more likely to employ an overt pronoun.

Table 14. Form of conditional and pronoun type

	Total count	Overt count (%)	Zero count (%)
Short ending	35	26 (74.3%)	9 (25.7%)
Long ending	71	33 (46.5%)	38 (53.5%)

Additionally, in comparison to the result obtained by Pajusalu and Pajusalu (2004), where pronouns always accompanied the long endings, this study showed that there were combinations of the long ending and zero. As a matter of fact, the percentage of such case was quite high (53.5%) (example 37).

(37) *Ütleksin, et meeldid mulle sõbrana.*

'I would say that I like you as a friend'.

4.7. Tense

Most instances in the collected data were either in the present or past tense while the occurrences of perfect and pluperfect were rather rare in the dataset. In table 15, the four tenses are compared. It was speculated that tense would not be influential, but surprisingly the data suggested that it was. Thus, hypothesis no.7 (tense would be unlikely to have a huge influence on the choice between subject pronoun omission and overt subject

pronoun expression.) turned out to be incorrect. When comparing the two dominant tenses, present and past, it is clear that the absence of a pronoun was more common in the past tense (61.2%) than in the present tense (44.4%). This is probably due to the fact that the past tense tends to be narrative, and ellipsis is likely to happen in narrative contexts (Kivik 2010, Hint 2015). For example, the sentence (38) below is a narrative which does not utilize a pronoun.

(38) *Seega kirjutasin ta õpetajale ja ütlesin, et Joosep reedel kooli ei tule.*

'Therefore I wrote to his teacher and said that Joosep would not come to school on Friday'.

The percentage of perfect was even (50% and 50% each), and pluperfect preferred having a pronoun even though there were only 18 (10 overt) instances.

Table 15. Tense and pronoun type

	Total count	Overt count (%)	Zero count (%)
Present	928	516 (55.6%)	412 (44.4%)
Perfect	92	46 (50.0%)	46 (50.0%)
Past	562	218 (38.8%)	344 (61.2%)
Pluperfect	18	10 (55.6%)	8 (44.4%)

According to table 16, where only the two main tenses are listed to compare position, the past tense seems to prefer the right position. The sentence (39) is a typical example in the past tense where V2 is respected by the use of an adverbial at the beginning of a sentence.

(39) *Eile nägin ma Eestimaad.*

'Yesterday I saw Estonia'.

Table 16. Tense and position

	Total count	Left count (%)	Right count (%)
Present	516	464 (89.9%)	52 (10.1%)
Past	218	171 (78.4%)	47 (21.6%)

4.8. Negation

Negative sentences are known to prefer overt subject pronoun expression (Pajusalu & Pajusalu 2004) although subject pronoun omission can occur in negative sentences as well (Metslang 2009). The result of the study goes along with this fact, and table 17 in the next page clearly shows the tendency. Although the overt percentage of 87.2% under negation might be a high percentage, this result also indicates that significantly often enough subject pronoun omission takes place under negation. Either way, this also proves the hypothesis no.6 correct. Below is an example (40) of a negative sentence without a pronoun.

(40) *Ja kui ma kohale tulin, ei öelnud samuti, et ma võin osaleda.*

'And when I came to the place, I also didn't say that I can participate'.

Table 17. Negation and pronoun type

	Total count	Overt count (%)	Zero count (%)
Affirmative	1381	599 (43.4%)	782 (56.6%)
Negative	219	191 (87.2%)	28 (12.8%)

When person is taken into consideration, the result seems to be striking as it can be seen in table 18 below. The probability of subject pronoun omission was considerably higher in the third person (example 41) compared to the other two persons although the number of total instances was rather small with the second and third persons.

(41) *Ei muuda ju.*

'It will not change that, you know'.

Table 18. Person/negation and pronoun type

	Total count	Overt count (%)	Zero count (%)
First negation	179	161 (89.9%)	18 (10.1%)
Second negation	15	14 (93.3%)	1 (6.7%)
Third negation	25	16 (64%)	9 (36%)

4.9. Referential distance

The hypothesis about referential distance was that it is an influential factor that encourages subject pronoun omission to occur. Based on the hypothesis, the smaller the number of referential distance is, the more likely it is for subject pronoun omission to occur. As mentioned above, the counting of referential distance in this thesis is based on

clause. The number 3, for example, means that the referent of the subject pronoun was also mentioned in three clauses before. Below, an example (42) illustrates an instance where referential distance is 1, as the pronoun *ma* before the verb *mõtlen* 'to think' is the previous mention which, in turn, is one clause away from the *ma* of the verb *jooksma*.

(42) *Ja ma mõtlesin, et äkki **ma** jooksen natuke aeglasemalt.*

'And I thought that maybe I run a little bit more slowly'.

The result of the study is presented in table 19 in the next page. Keeping the hypothesis in mind and looking at the result of referential distance categorized as 1 and 2, it seems as if the speculation was right. However, the column below (coded by 3) proves that wrong as the percentage of subject pronoun omission increases by 13.5%. Interestingly, it decreases again when observing the result of referential distance categorized as 3 and 4 by even more than 20% (21.2%). This might be due to the fact that distances 1 and 2 are close enough, therefore it does not require a lot of efforts to remember who the referent is. The abnormality of 3-clause distance is inexplicable, but it makes sense that distance of 4 or more is the one that prefers overt the most in the table. Hence, the result partly supports the hypothesis no.3 (referential distance would be an influential factor, meaning subject pronoun omission would be likely to occur when the referential distance is smaller.) while it partly denies it at the same time. There is unfortunately no clear explanation why the percentage did not keep decreasing in the zero section, which would have proven the hypothesis right.

Table 19. Referential distance and pronoun type

	Total count	Overt count (%)	Zero count (%)
1	462	232 (50.2%)	230 (49.8%)
2	184	99 (53.8%)	85 (46.2%)
3	67	27 (40.3%)	40 (59.7%)
4+	26	16 (61.5%)	9 (38.5%)

4.10. Case marking of the previous mention

There were 11 cases found in the dataset among all the 14 cases excluding translative, terminative and essive. Below, table 20 shows the result.

Table 20. Case and pronoun type

	Total count	Overt count (%)	Zero count (%)
Nominative	467	209 (44.8%)	258 (55.2%)
Genitive	93	57 (61.3%)	36 (38.7%)
Partitive	31	15 (48.4%)	16 (51.6%)
Illative	1	1 (100%)	0 (0%)
Inessive	2	1 (50.0%)	1 (50.0%)
Elative	11	7 (63.6%)	4 (36.4%)
Allative	52	30 (57.7%)	22 (42.3%)
Adessive	72	45 (62.5%)	27 (37.5%)
Ablative	2	2 (100%)	0 (0%)
Abessive	1	1 (100%)	0 (0%)
Comitative	7	6 (85.7%)	1 (14.3%)

It turned out that previous mentions in a case other than nominative led to a higher probability of overt pronoun expression except for partitive and inessive. Only in the nominative, partitive and inessive columns, it is noticeable that the percentage of the overt option was higher than 50.0%. As for the inessive, however, there were only two instances in total. A few possible reasons which could explain the result above are that, first, when the previous pronoun mention is done in the same case as the subject, it is likely that it is not cognitively difficult to remember who the subject is. Second, the reason why the partitive case preferred subject pronoun omission might be related to arguments as nominative and partitive, both of which are core arguments, are the only ones that prefer zero unlike oblique arguments.

4.11. Animacy

There were in total 1,566 instances in the data which referred to an animate entity. Therefore, the analysis mainly focuses on the minority, that is, inanimate entities. Out of the 34 instances that had an inanimate entity as the referent, only one had an overt pronoun, meaning that an inanimate entity is not likely to have a pronoun. This is an outstanding result as the percentages were almost even when an animate entity was the referent as can be seen in table 21. The only inanimate instance which used a pronoun had European Union (*Euroopa Liit*) as the referent. Both the table and the example are in the next page.

(43) *Euroopa Liit sai Lissaboni lepinguga suured volitused, millega ta peaks arendama edasi oma välispoliitika eesmärke ja kaitsma oma huve kogu maailam üldusema eesmärgiga edendada rahu.*

'With Lisbon treaty European Union got big authorities, with which it has to keep developing the goals of its foreign politics and protect its interest in the whole world with the more general goal to promote peace'.

Table 21. Animacy and pronoun type

	Total count	Overt count (%)	Zero count (%)
Animate	1566	789 (50.4%)	777 (49.6%)
Animate (third person)	219	141 (64.4%)	78 (35.6%)
Inanimate	34	1 (2.9%)	33 (97.1%)

Consequently, the frequency of overt was even higher (64.4%) with animate third person (64.4%) than third person in general (56.1%) (See section 4.2.).

4.12. Verb lemma

This section presents the result across the nine verbs chosen for the study. The verbs in table 22 are in the order of higher probability of overt pronoun expression. *Teadma* and *arvama* are the ones that marked a percentage higher than 60% in the overt section, meaning these verbs seem to prefer having an overt pronoun. The percentage of *muutma* and *jooksma*, on the other hand, was lower than 40% in the same section. This result could be compared to the Spanish study, where epistemic verbs were more likely to go for overt

pronoun expression (Pešková 2013). Verbs such as *arvama* and *ütlema* are used to convey opinions and knowledge, which might justify the result that these verbs require an overt pronoun to make it clear who the subject is (See example 44) as the hypothesis no.4 stated (verbs which have to do with emotions and opinions would prefer overt pronoun expression.). The result with the verb *pidama* seems logical as well since it is related to what a person has to do, in which case it seems natural to mention overtly the subject of the verb. The reason why the percentage of overt with the verb *tahtma* was not that high is an interesting result as it can be imagined that the subject would be likely to be overtly mentioned owing to the fact that it is a verb which conveys a strong feeling that the subject has. Based on the result with this verb, it seems that the hypothesis no.4 does not appear to be completely correct.

(44) *Ma arvan, et kõik osavõtjad jäid päevaga rahule.*

'I think that all the participants were satisfied with the day'.

By comparing the results of the verbs *muutuma* and *muutma*, it becomes evident that the intransitive counterpart was more likely to have a pronoun than the transitive counterpart. This goes along with the study by Metslang (2013), and the hypothesis no.5 (subject pronoun omission would occur more often with transitive verbs than intransitive ones.). Thus, transitivity does seem to be influential as far as subject pronoun omission is concerned. *Nägema* was chosen to be the representative of perception verbs, and it turned out that it does not seem inclined towards the use of an overt subject pronoun, which goes along with the study by Pešková (2013). The verb which prefered not having a pronoun the most turned out to be *jooksma*. This is partly because of the fact that the verb tends to

appear in a series of sequences, where the subject is not repeatedly mentioned. As already known, zero can be found the most frequently when a referent is mentioned in successive utterances and there are no other competing referents (Vihman 2015, Hint 2015). Below is an example (45) with the verb *jooksma*.

(45) *Seepeale kargasin ma autost välja ja jooksin talle järele.*

'After that, I jumped out from the car and run after him'.

Table 22. Verb lemma and pronoun type

	Overt count (%)	Zero count (%)
Teadma (knowledge)	133 (66.5%)	67 (33.5%)
Arvama (thought)	121 (60.5%)	79 (39.5%)
Pidama (obligation)	114 (57.0%)	86 (43%)
Ütlema (communication)	109 (54.5%)	91 (45.5%)
Tahtma (emotion)	95 (47.5%)	105 (52.5%)
Muutuma (intransitive)	43 (43%)	57 (57%)
Nägema (perception)	84 (42%)	116 (58%)
Muutma (transitive)	33 (33%)	67 (67%)
Jooksma (motion)	58 (29%)	142 (71%)

4.13. Personal preference

Apart from the analysis which has been written down, one thing which the author noticed that was not visible in the dataset as numbers is that it seems that individual preferrences seem to play a role when it comes to this topic as some sentences from the same context

consistently avoided utilizing subject pronouns while others did use them all the time. This was also mentioned in Pešková's (2013) study, in which it was revelaed that personal preference ranged from 35% to 79% (Pešková 2013). This study was not able to do such an analysis as the data were rather anonymous. However, it would be natural to imagine that the tendency is the same across languages and that individual preferences do exist in Estonian too as Hint (2021) suggests.

5. Conclusion

This paper strived to investigate the phenomenon called subject pronoun omission along with the choice between the short and long pronoun forms in written Estonian. For the study, 1,600 intances with nine verbs were taken out from Estonian National Corpus 2021. The verbs were *pidama*, *arvama*, *teadma*, *muutma*, *muutuma*, *jooksma*, *nägema*, *tahtma* and *ütlema*. The factors, which were expected to be influential in the study, were position, person, number, mood, form of conditional, tense, negation, referential distance, case marking of the previous mention, animacy and verb lemma. Based on previous studies, common knowledge and the author's intuition, eight hypotheses were formulated. All of them are listed below with the result given in brackets.

1. Subject pronoun omission would be less likely to occur with third person. (correct)
2. The long pronoun form would be utilized more often after the verb. (incorrect)
3. Referential distance would be an influential factor, meaning subject pronoun omission would be likely to occur when the referential distance is smaller. (partly correct and partly incorrect)
4. Verbs which have to do with emotions and opinions would prefer overt pronoun expression. (mostly correct)
5. Subject pronoun omission would occur more often with transitive verbs than intransitive ones. (correct)
6. Verb endings which do not indicate who the subject is would be likely to be accompanied by an overt pronoun. (correct)

7. Tense would be unlikely to have a huge influence on the choice between subject pronoun omission and overt subject pronoun expression. (incorrect)
8. The conditional mood is over all more likely to use an overt pronoun. (correct)

Apart from the hypotheses, the results of the study in general showed that different factors did play a certain role. Besides, there are two results which are especially worth mentioning here. The first one is not related to any factors per se, but it was a good finding that the frequency of overt subject pronoun expression turned out to be almost the same as that of subject pronoun omission (overt 49.4% and zero 50.6%). The second one is that the ratio of subject pronoun omission with second person plural was rather high (65.8%). There would be multiple reasons which could explain the second result well, one of which might be related to politeness. This topic could be investigated deeply in the future by differentiating the two types of *te/teie*.

Speaking of possible future studies, although the current study was conducted in a way that the data were as authentic as possible so that the results would reflect the reality, the study had certain limitations. First of all, the amount of instances was 1600, which was enough for carrrying out the study. However, more instances would have made the study more accurate as there were some parts in the paper where the analysis relied on a limited amount of instances. Concerning the verb choices, only one verb was chosen to represent a certain characteristic for this study. However, it would be a better idea to select several verbs as there is a possibility that certain single verbs behave abnormally, and relying only on one verb could result in misleading results and analysis. One suggestion would be, for example, to compare *valmistama* ‘to prepare (transitive)’ and *valmistuma* ‘to

prepare (intransitive)' to see if transitivity indeed has an influence. Moreover, perception, which the verb *nägema* represented, could be further divided as well. For example, the verb *tundma* 'to feel' could be a good candidate as it is a perception verb, but it is related to feelings unlike *nägema*. Furthermore, subject, in other words, the nominative case was the focus of the study, but other cases could be looked into as well like Lindström and Vihman (2017) did before. Although it is expected that investigating object pronoun omission, for example, would be much more complicated. Another suggestion for future studies is that there might be generational/chronological differences when it comes to subject pronoun omission. Nowadays, the language which is influential around the world including Estonia is English, in which subject pronoun omission is rare. Since especially the young are closely in touch with the language, it is possible that English is affecting the use of Estonian among them, which leads to a possible hypothesis that the younger someone is, the more likely it is for the person to prefer overt subject pronoun expression owing to the English influence. In addition, since the study by Lindström et al. (2009) revealed that there were dialectal differences with first person singular, it would make sense to do a similar study, but this time including all the persons and numbers. As for methods, using a book which is translated into multiple languages such as Finnish and analyzing them in terms of how pronouns are differently utilized depending on the language might be a good idea. Moreover, although this study did not include imperative, investigating when subject pronoun omission occurs and whether the short or long form is employed in this mood must be interesting as well. Lastly, it would be a great idea to conduct a study similar to this one but in spoken Estonian, which would surely give different results from the ones found above.

Before finishing up the paper, it has to be mentioned that the author of the paper is not a native Estonian speaker, therefore there is a possibility that the data collection and analysis were not absolutely precise. Nevertheless, the author hopes that this paper will be helpful in deciphering the system of subject pronoun omission in Estonian.

References

Ackema, P. and Neeleman, A., 2007. Restricted pro drop in early modern Dutch. *The Journal of Comparative Germanic Linguistics*, *10*, pp.81-107.

Biberauer, T., Holmberg, A., Roberts, I. and Sheehan, M., 2009. *Parametric variation: Null subjects in minimalist theory*. Cambridge University Press.

Duvallon, O. and Chalvin, A., 2004. La réalisation zéro du pronom sujet de première et de deuxième personne du singulier en finnois et en estonien parlés. *Linguistica Uralica*, *40*(4), pp.270-286.

Erelt, M. and Metslang, H. eds., 2017. *Eesti keele süntaks*. Tartu Ülikooli Kirjastus.

Gundel, J.K., Hedberg, N. and Zacharski, R., 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pp.274-307.

Haegeman, L. and Ihsane, T., 2001. Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language acquisition*, *9*(4), pp.329-346.

Heine, B., 2019. Some observations on the dualistic nature of discourse processing. *Folia Linguistica*, *53*(2), pp.411-442.

Helasvuo, M.L. and Kyröläinen, A.J., 2016. Choosing between zero and pronominal subject: modeling subject expression in the 1st person singular in Finnish conversation. *Corpus linguistics and linguistic theory*, *12*(2), pp.263-299.

Hint, H., 2015. Third person pronoun forms in Estonian in the light of centering theory. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, *6*(2), pp.105-135.

Hint, H., 2021. From full phrase to zero: a multifactorial, form-specific and crosslinguistic analysis of Estonian referential system (Dissertationes Linguisticae Universitatis Tartuensis 42). Tartu: University of Tartu Press.

Hint, H., Nahkola, T. and Pajusalu, R., 2020. Pronouns as referential devices in Estonian, Finnish, and Russian. *Journal of Pragmatics*, *155*, pp.43-63.

Hint, H., Reile, M. and Kaiser, E., 2023. Third-person overt pronoun and zero reference in Estonian. Insights from two experiments. *Eesti ja soome-ugri keeleteaduse ajakiri. = Journal of Estonian and Finno-Ugric Linguistics*, *14*(2), pp.75-109.

Kibrik, A.A., Khudyakova, M.V., Dobrov, G.B., Linnik, A. and Zalmanov, D.A., 2016. Referential choice: Predictability and its limits. *Frontiers in psychology*, *7*, 1429.

Kivik, P-K. 2010. Personal pronoun variation in language contact: Estonian in the United States. In Muriel Norde, Bob de Jonge and Cornelius Hasselblatt (eds.), Language Contact: New Perspectives, pp. 63–86. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Koppel, K. and Kallas, J., 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, *18*, pp.207-228.

Lindström, L., 2001. Verb-initial clauses in narrative. *Estonian: Typological Studies*, *5*, pp.138-168.

Lindström, L., 2017. Lause infostruktuur ja sõnajärg. *Eesti keele süntaks*, pp.547-565.

Lindström, L., Kalmus, M., Klaus, A., Bakhoff, L. and Pajusalu, K., 2009. Ainsuse 1. isikule viitamine eesti murretes. *Emakeele Seltsi aastaraamat*, *54*, pp.159-185.

Lindström, L. and Vihman, V.A., 2017. Who needs it? Variation in experiencer marking in Estonian 'need'-constructions1. *Journal of Linguistics*, *53*(4), pp.789-822.

Metslang, H., 2009. Estonian grammar between Finnic and SAE: some comparisons. *Language Typology and Universals*, *62*(1-2), pp.49-71.

Metslang, H., 2013. Coding and behaviour of Estonian subjects. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, *4*(2), pp.217-293.

Otheguy, R., Zentella, A.C. and Livert, D., 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language*, pp.770-802.

Pajusalu, R., 2005. Anaphoric pronouns in spoken Estonian. *Minimal reference. The use of pronouns in Finnish and Estonian discourse*, (Studia Fennica Linguistica 12.), pp.107-134. Helsinki: Suomalaisen Kirjallisuuden Seura.

Pajusalu, R., 2009. Pronouns and reference in Estonian. *Language Typology and Universals*, *62*(1-2), pp.122-139.

Pajusalu, R. and Pajusalu, K., 2004. The conditional in everyday Estonian: Its form and functions. *Linguistica Uralica*, *4*, pp.257-269.

Pešková, A., 2013. Experimenting with pro-drop in Spanish. *SKY Journal of Linguistics*, *26*, pp.117-149.

Reile, M., 2016. Distance, visual salience, and contrast expressed through different demonstrative systems: An experimental study in Estonian. *SKY Journal of Linguistics*, *29*, pp. 63–94.

Reile, M., 2019. *Estonian demonstratives in exophoric use: An experimental approach*. (Dissertationes Linguisticae Universitatis Tartuensis 34). Tartu: University of Tartu Press.

Sepp, P., 2010. Pronoomeni kasutus MSN-vestlustes. Bakalaureusetöö. Tartu Ülikool.

Vihman, V.A., 2015. Pick it up: a look at referential devices in Estonian child-directed speech. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, *6*(2), pp.63-83.

Vogels, J., Krahmer, E. and Maes, A., 2018. Accessibility and reference production: the interplay between linguistic and non-linguistic factors. In Jeanette K. Gundel and Barbara Abbott (eds.), The Oxford handbook of reference. Oxford, UK: Oxford University Press.

Summary: Subject pronoun omission in written Estonian

Estonian is considered to be a partial pro-drop language, which means that it is possible to leave out subject pronouns. In Estonian, there are three possible choices, namely the short form, the long form and the zero reference. In this paper, the data which were collected from a corpus will be analyzed to find out when and under what conditions subject pronoun omission occurs in Estonian. Apart from that, the use of the short and long forms will be investigated as well.

The corpus which was used to collect data is Estonian National Corpus 2021. There were in total 1,600 instances with nine verbs, and they were analyzed on Excel. These nine verbs are *pidama* 'to have to', *arvama* 'to think', *teadma* 'to know', *muutma* 'to change (transitive)', *muutuma* 'to change (intransitive)', *jooksma* 'to run', *nägema* 'to see', *tahtma* 'to want' and *ütlema* 'to say.' Subject pronoun omission could be affected by different factors. In this paper, the influence of the following factors was measured: position, person, number, mood, form of conditional, tense, negation, referential distance, case marking of the previous mention, animacy and verb lemma.

Based on previous studies, common knowledge and the author's intuition, hypotheses were formulated. Some of them turned out to be correct while others turned out to be incorrect. In general, the results showed that different factors play an important role in the choice between overt subject pronoun expression and subject pronoun omission. Especially, three results from the study could be regarded as outstanding. Firstly, subject pronoun omission (50.6%) occurred nearly as often as overt subject pronoun expression

(49.4%). Secondly, as for second person plural, the ratio of subject pronoun omission (65.8%) was quite high compared to the others. (first person singular 50.6%, second person singular 50.4%, third person singular 44.5%, first person plural 55.4%, third person plural 41.9%). Thirdly, it was preferred to avoid the use of subject pronoun in the past tense. The last one, for example, was not in accordance with one of the hypotheses.

The goal of the study was to analyze subject pronoun omission and its conditions thoroughly since this topic has not been dealt with minutely before. Although this study has some limitations, hopefully it will contribute to the science and future investigations.

Kokkuvõte: Subjektpronoomeni väljajätt eesti kirjakeeles

Eesti keelt peetakse osaliseks *pro-drop*-keeleks, s.t eesti keeles on võimalik subjetina esinevat asesõna lausest välja jätta. Eesti keeles on kolm võimalikku valikut, nimelt asesõna lühivorm, pikk vorm ja väljajätt. Selles töös analüüsitakse korpusest saadud andmeid, et teada saada, millal ja millistel tingimustel subjektpronoomeni väljajätt eesti keeles toimub. Lisaks sellele uuritakse asesõna pika ja lühikese vormi kasutust.

Korpus, mida andmete kätte saamiseks kasutati, on Estonian National Corpus 2021. Lauseid oli kokku 1,600 üheksa valitud verbiga ja neid analüüsiti Excel'is. Need üheksa verbi on *pidama*, *arvama*, *teadma*, *muutma*, *muutuma*, *jooksma*, *nägema*, *tahtma* and *ütlema*. Asesõna väljajättu võivad mõjutada erinevad tegurid. Selles töös vaadeldi järgnevate tegurite mõju: asend, grammatiline isik, arv, kõneviis, tingiva kõneviisi lõppude esinemine, ajavorm, eitus, referentsiaalne kaugus, eelneva samaviitelise nimisõnafraasi kääne, elusus ja verb.

Varasemate uuringute, üldiste teadmiste ja autori intuitsiooni põhjal mõeldi välja hüpoteesid, millele töös vastust otsiti. Mõned neist osutusid õigeks ja teised ebaõigeks. Üldiselt tulemused näitasid, et erinevad faktorid mängivad pronoomeni esinemise või väljajätu valikul suurt rolli. Eriti kolme tulemust sellest uuringust võiks silmapaistvaks pidada. Esiteks seda, et subjektpronoomeni väljajätt (50.6%) toimus peaaegu sama tihti kui pronoomeni esinemine (49.4%). Teiseks seda, et mitmuse teise isiku puhul oli pronoomeni väljajätu osakaal (65.8%) võrreldes teiste isikutega üsna kõrge (ainsuse esimene isik 50.6%, ainsuse teine isik 50.4%, ainsuse kolmas isik 44.5%, mitmuse

esimene isik 55.4%, mitmuse kolmas isik 41.9%). Kolmandaks seda, et minevikus eelistati asesõna kasutuse vältimist. See viimane oli üks tulemus, mis ei olnud hüpoteesiga kooskõlas.

Selle uuringu eesmärgiks oli analüüsida põhjalikult asõesõnalise subjekti väljajättu ja selle tingimusi, kuna seda teemat ei ole varem nii üksikasjalikult käsitletud. Kuigi sellel tööl on olulisi piiranguid, siis loodetavasti see panustab teadusesse ja tulevastesse uurimustesse.

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Ryomei Ueda,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis “Subject pronoun omission in written Estonian”, supervised by Helen Hint and Liina Lindström.

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons’ intellectual property rights or rights arising from the personal data protection legislation.

Ryomei Ueda

26/05/2024