UNIVERSITY OF TARTU

DEPARTMENT OF ENGLISH STUDIES

A CORPUS-BASED STUDY OF ADJECTIVE INTENSIFICATION AMONG NATIVE SPEAKERS AND LEARNERS OF ENGLISH

MA thesis

DENYS SAVCHENKO

SUPERVISOR: Assoc. Prof. JANE KLAVAN

TARTU

2022

ABSTRACT

The aim of this MA thesis is to examine the use of intensifiers by advanced learners of English in comparison to English L1 speakers. The focus of this study is the intensification of adjectives in two data sets: the Estonian component of the Louvain International Database of Spoken English Interlanguage (LINDSEI, Gilquin et al. 2010) and the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004). The research questions of this thesis are as follows: What is the distribution of intensifiers in two data sets?; What are the differences between two groups of speakers (English L1 and Estonian L1) in terms of the use of intensifiers? I expect to observe differences in the number of types of intensifiers that are influenced by the types of adjectives intensified.

The thesis consists of three main chapters: introduction, where theoretical background about intensifiers is introduced; empirical analysis, where the methods are introduced, and the results are presented; conclusion and appendices. In the introduction, I discuss proposed classifications of intensifiers based on their function (amplifiers and downtowners) and the types of adjectives they modify. This chapter introduces implications based on the previous research that examined the use of intensifiers in the native speakers' data and in the data of the Learners of English with various L1 backgrounds.

The second chapter explains the procedure of extraction of intensifiers from the data sets in LINDSEI and LOCNEC corpora. The Methodology subsection introduces two methods used in this thesis: hierarchical cluster analysis (HCA) and multiple correspondence analysis (MCA). The following subsection presents the results of the analysis based on the cluster dendrograms and correspondence analysis biplot. The results are followed by a discussion of the main results compared to the previous studies on intensification. The main findings as implications for further research are summarised in the conclusion.

Appendix 1 contains the link to the repository with the Rstudio scrips for conducting HCA and MCA. Appendix 2. contains the tables with a set of the most frequent amplifiers and adjectives they collocate with and the types of the adjectives according to the classification by Paradis (1997).

Table of Contents

INTRODUCTION

In this MA thesis, I will examine the distribution of intensifiers in the Estonian component of the Louvain International Database of Spoken English Interlanguage (LINDSEI-EST). The thesis addresses the following research questions:

1. What is the distribution of the intensifiers in two data sets?

2. What are the differences and similarities between two groups of speakers (English L1 and Estonian L1) in terms of the use of intensifiers?

To my knowledge, there are no previous studies that analyse intensifiers in the data of English learners whose L1 is Estonian or other Finno-Ugric languages. Thus, this study is meant to fill the existing research gap, potentially adding to the general understanding of how learners of English use intensifiers. Intensifiers in the corpus will be analysed according to the frequency and types of adjectives they modify. This will allow me to identify the current trends in the language of EFL speakers. The LINDSEI (Gilquin et al. 2010) corpus contains interviews with L2 speakers of English of various language backgrounds that follow the same structure, which ensures comparability of the data. Alongside the LINDSEI corpus, the Louvain Database contains the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004). The LOCNEC corpus is comprised of 50 interviews with native speakers. The interviews follow the structure of the interviews in LINDSEI. In other words, both Learners of English and English L1 speakers faced the same tasks during the interviewing process. Access to the corpora will allow me to compare data of Estonian EFL learners with the data of English L1. The empirical analysis will be done in order to identify the over- and underuse of intensifiers by Estonian L1 speakers of English.

The primary focus of this MA thesis is the functional category of intensifiers in the spoken language of Estonian EFL speakers. According to Quirk et al. (1985: 589-590),

intensifiers are subjuncts that deal with the category of degree and their primary function is to scale a quality up or down from the assumed norm. Quirk et al. (1985) distinguish two main types of intensifiers: amplifiers and downtoners. Amplifiers scale quality upwards, and downtoners scale quality downwards. Amplifiers include such adverbs as *very, really and so*. Downtoners are adverbs such as *quite, rather* and *pretty*. Intensifiers are characterised by constant renewal and competition in terms of the frequency of the distribution of each member of this class. Moreover, intensifiers differ according to the degree of grammaticalisation. Peters (1994: 271) attributes the changing nature of intensifiers to "speaker's desire to be original". Preliminary analysis of the existing data from LINDSEI-EST provides the following examples of boosters (1), maximisers (2) and downtoners (3):

(1) you have a *really really* modern and great library;
(2) each time we went there she was *super* friendly;
(3) I guess a *pretty* long time;

Various studies conducted in the last two decades have examined different aspects of the use of intensifiers. Ito and Tagliamonte (2003) showed that intensifiers may function as markers of different generational groups within communities of speakers and examined the effect of gender on the distribution of intensifiers. Xiao and Tao (2007), on the other hand, focused on the interaction of extralinguistic factors and pointed out the influence of contextual constraints such as genre, discourse mode etc. on the distribution of intensifiers. Considering the syntactic, semantic, lexical and stylistic restrictions that intensifiers are subjected to (Altenberg 1991), it is assumed that intensifiers present certain difficulties for English learners as they may not always be aware of all the factors that determine the choice of a particular adverb to modify adjectives. Granger (1998) and Lorenz (1999) examined the differences in the use of intensifiers in the written language of both L1 and L2 speakers of English. They concluded that compared to L1 speakers, learners tend to overuse intensifiers that collocate with a wider range of items. For spoken language, a study by Pérez-Paredes

(2010) suggests that amplifiers are not part of learners' active spoken lexical repertoire when compared to native speakers.

In order to answer the research questions, I will use the data from the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI) that is currently being compiled at the English Department of the University of Tartu. The corpus includes the interviews that are structured around three tasks: set topic, free discussion and a picture description task. The duration of interviews varies from interviewee to interviewee; however, the average length of the interview is 15 minutes. Currently, the LINDSEI-EST corpus consists of 25 interviews (348 minutes of speech; 35,117 words of transcribed text). All of the interviewees were native speakers of Estonian in the third or fourth year of the English language and literature programme at the University of Tartu.

The thesis consistss of three main parts: an introduction, a chapter introducing intensifiers, empirical analysis and conclusion. The first chapter is divided into 5 subchapters that introduce classifications of intensifiers, discuss adjectives and gradability, provides an overview of the previous studies and implication for the study described in this thesis. The second chapter is also divided into 5 subchapters that describe the data collection procedure, introduce the methods used in this thesis, present and discuss the result of the study conducted to analyse the use of intensifiers. The final chapter provides a conclusion that summarizes the main findings describe in this thesis. Appendix 1 contains the link to the repository with the scripts for Rstuido for conducting hierarchical cluster analysis and multiple correspondence analysis. Appendix 2 contains the sample of the intensifiers used for hierarchical cluster analysis and multiple correspondence analysis.

1. INTENSIFIERS

This chapter provides an overview of intensifiers as a functional class and defines all the terms and concepts relevant to this thesis. The chapters starts with introduction of classification, discusses adjectives and gradability, gives a short overview of the previous studies that looked at the use of intensifiers, and provides the summery containing the implications for this thesis.

1.1. Classification of intensifiers

The focus of this thesis is intensification and, more specifically, intensification of adjectives. Intensification provides speakers with means of conveying commitment and judgment (Lorenz 1999:24) towards the information expressed, as well as emphasising certain parts of the message. Intensification usually involves the use of adverbs (*very, really, a bit* etc.) that are referred to as intensifiers. Intensifiers are concerned with a property of degree, which allows modulating the degree to which a certain property holds. Modulation of degree is the primary reason why it is possible to talk about commitment, judgment or emphasis that is expressed when the speaker uses intensifiers. However, the term *intensifier* is not clearly defined in the literature about intensification and is somewhat confusing as different authors employ different terms or define intensifiers differently. Thus, the first task is defining intensifiers and identifying parts of speech that this term can label.

Many authors have studied intensifiers from various points of view, and several classifications were proposed. Among them, the most notable is the classification presented in *Comprehensive Grammar of the English language* (Quirk et al. 1985). The authors define intensifiers as subjuncts that function as scaling devices. Moreover, Quirk et al. (1985) draw attention to the fact that intensifiers involve scaling in both up and downward direction. Thus, modification by intensifiers specifies the degree to which a property is applicable on

an “abstractly conceived intensity scale” (ibid). The notion of degree is closely connected with adjectives and adverbs as they undergo comparison. However, Bolinger (1972:15) points out that degree is not only applicable to adverbs and adjectives but also to nouns and verbs. In all the cases, the degree is the primary property that drives the classification of scaling devices. Based on this, Quirk et al. (1985) organise subjuncts in terms of the points on scale and divide them into two main groups: amplifiers and downtoners. In the present model (Figure 1), amplifiers denote the extreme and high degrees and thus scale the property upwards, while downtoners scale in the opposite direction.

<table>
<tr><td rowspan="6">INTENSIFIERS</td><td rowspan="2">AMPLIFIERS</td><td>Maximisers: completely</td></tr>
<tr><td>Boosters: very much</td></tr>
<tr><td rowspan="4">DOWNTONERS</td><td>Approximators: almost</td></tr>
<tr><td>Compromisers: more or less</td></tr>
<tr><td>Diminishers: partly</td></tr>
<tr><td>Minimisers: hardly</td></tr>
</table>

Figure 1. Quirk et al. (1985) classification of intensifiers

Biber et al. (1999), on the other hand, distinguish between only two types, namely, Amplifiers/Intensifiers and Diminishers/Downtoners. The authors of both classifications specify that amplifiers and downtoners modify gradable adjectives. Quirk et al. (1985) point out that one of the features that distinguishes gradable adjectives from non-gradable is the ability to be modified by intensifiers, in particular, the adverb *very*. The models provide a fully grammatical classification based on the function of intensifiers and, as a result, disregard semantic constraints. Based on this classification, it is not entirely clear how intensifiers combine with adjectives they modify. However, Paradis (1997) points out that

the types of gradable adjectives play a role in determining the adjective can be modified by an intensifiers. In her PhD thesis (Paradis 1997: 41), she mentions that the combination *absolutely amazing* is possible, but *absolutely nice* sounds odd. She points out that gradable adjectives can be subdivided into scalar, extreme, and limit (ibid). In this classification, *nice* is a scalar and *amazing* is an extreme adjective. Thus, an extreme adjective can be modified by the maximiser *absolutely*, and a scalar adjective would require the booster *very*.

Paradis (2001) proposes a model (Figure 2) that accommodates the observation of the existing constraints on modification in terms of types of gradable adjectives. In Paradis` analysis, gradability has its own structure that provides predictions regarding the compatibility between intensifiers and adjectives. In the next subchapter, I will discuss gradability in detail and show how the analysis proposed by Paradis (1997, 2001) incorporates the internal structure of gradable (and to a certain extent, non-gradable) adjectives to explain why the semantic compatibility between intensifiers and adjectives is equally important.

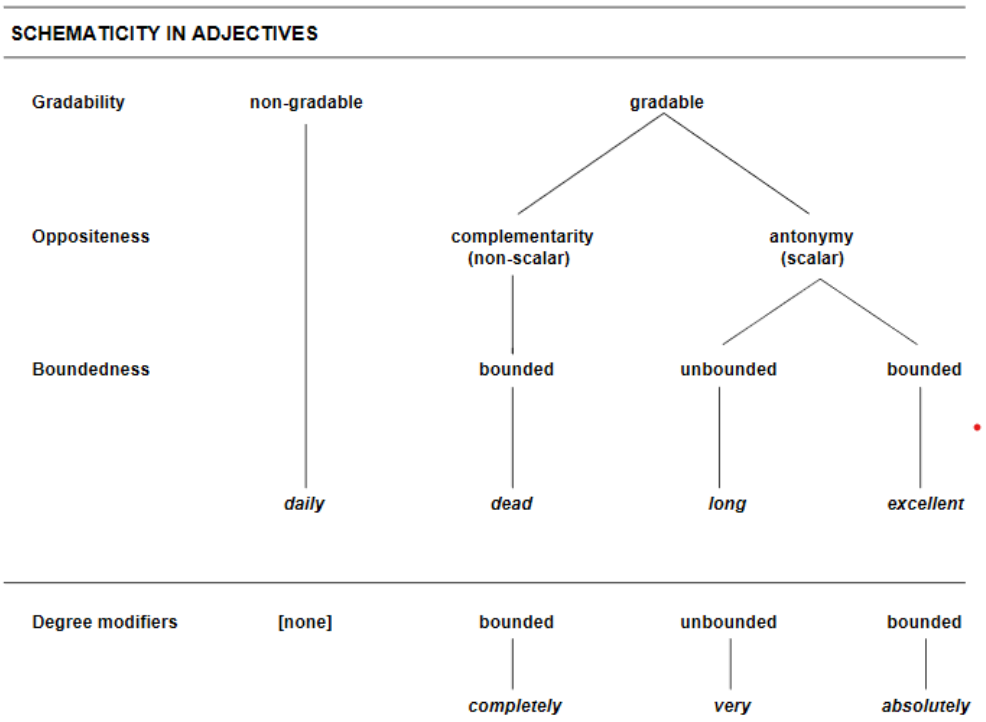


Figure 2. Paradis (2001: 54) Schematisity in adjectives

1.2 Adjectives and Gradability

Gradability is listed as one of the key properties of adjectives (Quirk et al. 1985, Huddleston and Pullum 2002). Quirk et al. (1985) note that the ability to be premodified by *very* in its intensification function is a characteristic of gradable adjectives. Huddleston and Pullum (2002) elaborate on the issue of gradability and non-gradability and point out that the distinction appears to be not as strict as it may seem. Thus, the authors list non-gradable adjectives that include *alphabetical, chief, medical, left, public* etc. (Huddleston and Pullum 2002: 531). However, non-gradable adjectives may acquire a gradable meaning that can be illustrated using the example of differences between pairs *a* and *b* in (1) and (2) (ibid): In the example (1a) and (2a), the non-gradable adjectives *public* and *open* acquire a gradable meaning in (1b) and (2b) respectively:

(1) a. the *public* highway
    b. a *very public* quarrel
(2) a. the door was *open*
    b. You haven't been *very open* with us

Additionally, Huddleston and Pullum (2002) distinguish absolute adjectives that include *ideal, absolute, impossible* etc. The authors point out that a prescriptive bias existed that rejected modification of these adjectives by intensifiers. However, there is a consensus that certain intensifiers can modify absolute adjectives and, thus, the categorisation of adjectives into gradable and non-gradable can not be taken as a rule. The system shows a great degree of flexibility (ibid).

Paradis's model (Figure 2) takes the categorisation proposed by Quirk et al. (1985) and divides modifiers according to two types of degree: totality and scalarity (Paradis 1997: 28). Maximisers and approximators are grouped into totality modifiers, while boosters and downtoners are grouped into scalar modifiers. The author also notes that there are two

readings of *quite,* - it can be a maximiser or a downtoner. The type of the intensifier *quite* is determined by the adjective it modifies. Thus, in combination with scalar adjectives, *quite* functions as a downtoner and specifies a lower degree of an adjective. Limit and extreme adjectives, on the other hand, render the modifier *quite* a maximiser. The modifier *quite* illustrates that the categorisation of intensifiers is not as straightforward. On the one hand, intensifiers like *very*, *really* are stable in terms of the degree they denote. On the other hand, *quite* shows that it is possible for the same modifier to operate on both upper and lower areas of the scale.

This categorisation (Paradis 1997, 2001) considers that gradability is a complex phenomenon and is vital for analysing intensifiers. Moreover, Paradis draws attention to the relations between intensifiers and adjectives, namely that adjectives go together with compatible modifiers, and modifiers match with compatible adjectives. The latlatter fact is not clearly reflected in the models proposed by Quirk et al. (1985) and Biber et al. (1999).

The issue of gradability discussed by Paradis (1997) requires further clarification as to how gradability can be conceptualised. Croft and Cruse (2004: 166) classify adjectives in terms of such features as complementaries and antonyms. Complementaries are adjectives that are mutually exclusive and, at the same time, designate the boundaries of certain properties. The most typical complementaries are pairs of adjectives *dead* and *alive, open* and *shut, true* and *false*. They can be viewed as properties that are present or absent. Thus, someone can only be either dead or alive, but not **more alive* or **more dead.* These adjectives are not gradable and cannot be put on a scale, as they do not denote properties that can be graded but only further emphasised by intensifiers. Antonyms, on the other hand, are gradable adjectives that denote different values on a scale. Typical pairs of antonyms are *long* and *short*, *good* and *bad*. These adjectives can be constructed on a scale that would denote the different degree to which a property holds.

Croft and Cruse (2004) divide the scales into monoscalar and biscalar systems. A monoscalar system is described with the example of the pair *short* and *long*. The two properties denote length from zero to an indefinite point. *Short* and *long* on this scale are associated with lower and higher degrees of the property of length. Modified by intensifiers, they scale in opposite directions. A biscalar system, on the other hand, can be constructed as a symmetrical scale that has one zero point. A property denoted by this scale can extend in the opposite direction. An example of this type of a scale is a scale constructed of adjectives *hot* and *cold*. The biscalar system can be characterised as consisting of two monoscalar systems, as on each part of a scale it is possible to talk about degrees of each property separately. The main difference between monoscalar and biscalar systems is that the zero point in biscalar system functions as a dividing line between the properties that are antonyms. Thus, the low degree of the property of "hotness" would at some point cross the dividing line and qualify an object as having the property of "coldness", and vice versa. However, when it comes to *long* and *short*, low degree of the property "long" could be qualified as short, but low degree of the property "short" would mean that the object is approaching the zero point.

Paradis (1997) points out that the conceptualisation of the scales is crucial for understanding how scalar, extreme and limit adjectives combine with intensifiers. In her model, limit adjectives are complementaries, and this is the reason why they can be modified by maximisers. Limit adjectives designate properties that are present or absent, but not easily gradable. Thus, *dead* can be modified by the maximiser *completely*, but not the downtoner *pretty*. Extreme adjectives are antonyms that denote a high degree of a property that can also be only modified by maximisers. For example, Paradis (1997:56) points out that *excellent* is an extreme adjective that resists modification by *very, really,* and instead combines with *absolutely.* Scalar adjectives are antonyms and they can be modified by boosters and

downtoners, as scalar adjectives denote gradable properties. That is why *long* can be modified by the booster *very* or the downtoner *pretty*. This classification explans why certain combinations of intensifiers and adjectives are possible and others sound odd.

It is evident that gradability plays a role in terms of the ability of intensifiers to modify certain adjectives. The classification of adjectivesinto different types in terms of scalar, extreme, and limit adjectives provides a valuable addition to the classification proposed by Quirk et al. (1985) and Biber et al. (1999), as it allows to explain the differences in terms of adjectives that can be modified by intensifiers. However, Paradis (1997) herself draws attention to the fact that limit adjectives can behave as scalar adjectives. Thus, it is possible for limit adjectives like *dead* to be modified by boosters like *very*, even though this would render such utterances as humorous or ironic. The author refers to the concept of contextual modulation that allows to adjust the types of adjectives and give rise to a different reading (Paradis 1997: 59). The later point echoes the observations made by Huddleston and Pullum (2002) regarding the flexibility of distinction between gradable and non-gradable adjectives. The classification presented here allows assuming that the differences in the use of particular intensifiers relate to the types of adjectives that are more readily modified by these intensifiers.

1.3 Previous studies on intensifiers

For the previously mentioned reasons regarding the complexity of the internal structure of intensifiers and the adjectives they modify, intensifiers are a fascinating topic for research. There are a number of studies that examine intensifiers from various points of view. Intensifiers are of particular interest as they "afford a picture of fevered invention and competition that would be hard to come by elsewhere for in their nature they are unsettled" (Bollinger 1972:18). Bollinger points out that intensifiers are the primary means of emphasis

that may lose their strength. This is why speakers are constantly looking for new means of highlighting the importance of information (Bollinger 1972: 18). Lorenz (1999: 24-25) points out that emphatic expressions (intensifiers included) can also serve as markers of particular social groups. For these reasons, many studies have examined such socio-linguistic factors as age, gender and education that influence the distribution of intensifiers (Ito and Tagliamonte 2003, Tagliamonte and Roberts 2005, Tagliamonte 2008, Murphy 2010). Other studies have looked at the distribution of intensifiers across spoken and written registers (Xiao and Tao 2007, Biber et al. 1999) and major varieties of English (Wagner 2017).

Most of the previous studies about intensifiers have been conducted in a corpus linguistic framework and used various corpora. Intensification was examined in the British National Corpus (BNC), GloWbE, London-Lund Corpus (LLC), Corpus of London Teenage Language (COLT), London English Corpus (LEC), the Michigan Corpus of Academic Spoken English (MICASE). Other studies have used the data collected in York (UK), Toronto (Canada) or have compiled corpora based on available transcriptions of TV series. The findings obtained from the above-mentioned corpora provide a comprehensive picture in terms of the distribution of intensifiers and how intensification interacts with a variety of factors. Previous research provides general trends concerning the way intensification presents itself across a collection of text materials available from different corpora.

For example Ito and Tagliamonte (2003) have found that the most common intensifiers used in the data collected in York (UK) are *very*, *really* and *so*. Moreover, the findings of this study show that there are differences in terms of age, as *really* is more frequently used by speakers under the age of 35. In the study conducted on the data from Toronto, Tagliamonte (2008) reports that the frequency of *really* and *so* has significantly increased among the speakers of the younger generation (2008). Similar results were

reported in the study that examined the use of intensifiers in ten seasons of the American television series *Friends.* Tagliamonte and Roberts (2005) show that in *Friends* data *so* is the most frequent intensifier in adjective modification position. The authors of the three above mentioned studies reported that frequencies of intensifiers reflect the ongoing process of change in the use of modifiers. Furthermore, Biber et al. (1999) show that American and British varieties differ in terms of frequencies of intensifiers. The main difference between these two varieties concerns the topmost frequent intensifiers. *Very, absolutely* and *bloody* are more frequent in British English, while *so, really* and *real*, and *totally* in American English. Wagner (2017) confirms this observation reporting that in GloWbE *very* occurs with the rate of 1000 per million words in the British variety, which is more frequent compared to North American varieties, where *very* has a rate of 800 per million words.

The changing norms are especially evident when the age of the speakers is taken into account. Macaulay (2005, 2006), who researched intensifiers in the speech of teenagers of Glasgow found out that *really* and *so* are by far the most frequent intensifiers. Previous research shows that intensifiers are influenced by changes within communities of speakers. This fact has been noted by Lorenz (1999), who suggested that intensifiers may signal group identity. In this case, age is the distinctive factor in terms of the distribution of intensifiers. However, on the other hand, it is not entirely clear whether there are differences between frequencies of intensifiers when the register is taken into account. Thus, Recski (2004) reported that *very* is still the most widely used intensifier in the MICASE corpus that contains data of spoken language used in the academic setting.

Various studies have examined how the use of intensifiers differs in terms of register. Biber et al. (1999) report that there are differences in terms of the distribution of intensifiers between varieties of English and register. Their findings show that intensifiers are by far more frequent in spoken English than in academic prose. Xiao and Tao (2007) compared

spoken and written data of the BNC (British National Corpus) and reported that not only amplifiers are more frequent in spoken English, but they also show which intensifiers are characteristic of each register. The findings are presented in Figure 3.

Category	Amplifier	Spoken	Written	LL score
Significantly more frequent in speech	really	1726.31	327.84	24645.56
	very	2421.36	1083.15	10856.41
	quite	1050.50	332.64	8553.00
	absolutely	182.66	43.34	2110.16
	bloody	88.38	9.62	1863.40
	pretty	107.62	43.70	586.65
	real	21.56	2.29	460.89
	jolly	18.95	7.66	104.08
	terribly	23.11	10.62	97.85
	dead	19.05	8.26	91.29
	damn	8.80	2.58	79.56
	awfully	9.77	3.17	76.69
	totally	77.55	56.04	67.27
Significantly more frequent in writing	highly	26.78	100.01	724.28
	fully	39.65	96.07	404.11
	deeply	7.45	40.16	386.98
	heavily	10.35	43.86	355.18
	greatly	7.35	36.90	339.28
	particularly	153.65	230.24	270.50
	wholly	5.22	24.76	218.61
	considerably	11.02	31.67	171.28
	entirely	40.23	72.46	161.13
	severely	5.41	19.22	132.57
	utterly	3.96	13.86	94.13
	badly	25.53	44.95	93.60
	extremely	46.41	71.14	91.88
	by far	2.42	10.19	82.08
	exceptionally	2.80	9.98	68.97
	thoroughly	11.80	22.04	54.08
	perfectly	33.17	45.61	35.07
	completely	79.48	86.12	4.87
Not statistically significant	enormously	7.06	8.48	2.35
	incredibly	6.87	8.13	1.95
Total		6282.22	2946.74	25294.48

Figure 3. Amplifiers in speech and writing in British English(Xiao and Tao 2007)

The amplifiers in Figure 3 are listed according to their log-likelihood scores, which can be seen in the right most column. The log-likelihood score is widely used in corpus linguistics and measures the effect of variable association (Stefanowitsch 2020: 224). It is apparent that among maximizers that are associated with spoken register in the British

variety of English, the most frequent are *really*, *very*, *quire* and *absolutely* and *totally*. It is possible to conclude that the influence of the register plays a role in the distribution of intensifiers.

The modifiers in Figure 3 are ordered according to statistically significant differences between registers. Thus, while the differences in the use of *really* in spoken and written data are the most significant, *very* remains the most frequent intensifier in BNC in terms of frequency of tokens. Notably, *pretty* as a booster is also relatively frequent in the British National Corpus. Finally, *absolutely* is a maximiser that is associated with the spoken language.

It is essential to take a closer look at the methodological decisions made by previous research and consider the implications relevant to the current study. In sub-chapter 1.1, it was noted that the terminology and categorisation of intensifiers defer, which is why it is hard to compare findings reported in previous studies. Most of the studies mentioned previously use Quirkian classification (Quirk et al. 1985). However, while acknowledging the distinction between different types of intensifiers, some studies (Ito and Tagliamonte 2003, Xiao and Tao 2007) grouped maximizers and boosters into one type, collectively named as amplifiers, to avoid confusion. Xiao and Tao (2007) pointed out that Quirk et al. (1985) include *extremely* in the subgroup of maximisers; however, Kennedy (2003), following Lorenz (1999), classifies *extremely* as a booster. Considering the restrictions that adjectives put on maximisers and boosters, it would be consistent to consider *extremely* a booster. *Extremely* designates a higher degree of property and lacks the same connotation as *absolutely*, which designates the upper extreme, which is why they would easily combine with limit and extreme adjectives. Considering the pairs in (3), it seems that *extremely* behaves as a booster and modifies scalar adjective *good*, but not the extreme adjective *excellent*. A search in the BNC corpus yielded the results that show that *extremely* is more

frequently used as a modifier of *good*. These observations are compatible with Paradis' model (Paradis 1997) and would explain why both Kennedy (2003) and Lorenz (1999) included *extremely* in the subgroup of boosters.

(3) a. *absolutely excellent* (13 tokens in BNC)
b. **extremely excellent* (0 tokens in BNC)
c. **absolutely good* (1 token)
d. *extremely good* (139 tokens)

In their study, Ito and Tagliamonte (2003) have also referred to amplifiers as intensifiers, even though the term includes both amplifiers and downtoners. The authors justify their choice by stating that, according to Mustanoja (1960), amplifiers are more frequent compared to downtoners. Additionally, amplifiers under negation were also excluded by Ito and Tagliamonte (2003) for the reason that negation affects the semantics of the degree of modifiers. Thus, the function of amplifiers is similar to downtoners. Lorenz (1999), on the other hand, has included the negated intensifiers.

Regarding the function of intensifiers, it would be hard to control the environments where modifiers are under the scope of negation as they would be qualified as downtoners, which would mean that the overall frequency of downtoners would be reduced be much higher. Moreover, Lorenz (1999) pointed out the fact that *really* under negation retains a modal meaning. Therefore, the instances of intensifiers under negation were excluded in many studies (Ito and Tagliamonte 2003, Tagliamonte 2008).

Other authors (Xiao and Tao 2007, Kennedy 2003, Wagner 2017) have not excluded intensifiers under negation. However, they do not discuss the frequencies and cases of negated intensifiers and, thus, it will not be possible to compare the results, mainly because most of the studies focus on amplifiers alone. The instances that are analysed as intensifiers under negation include (4) and should exclude (5). In (4) scope of the negation affects *too*

happy and negates the degree to which *happy* holds. Thus, the degree of 'happyness' in (4) is lowered. However, in (5), the complement of the prepositional phrase is not affected by negation, and the adjective phrase *very long* is still analysed as an instance of *very* functioning as a booster. However, examples like (5) were not found in the data used in this thesis.

(4) … she's not **too happy**… <LOCNEC_26_P>[1]

(5) And I didn't have boyfriends **for very long** 10(d)

(example 5 is cited in Ito and Tagliamonte 2003)

Another methodological decision that unites the previous studies mentioned in this sub-chapter is the decision to examine intensifiers as modifiers of adjectives. Previously, it was stated that gradability is one of the properties closely related to adjectives and adverbs. The majority of the studies have focused on intensifiers as modifiers of adjectives, as Bäcklund (1973: 279) reported that 72% of intensifiers are used with adjectival heads. Ito and Tagliamonte(2003: 263) suggest that focusing on the "consistent denominator" will allow getting consistent results. This decision requires the context of the tokens to be checked to decide which part of the speech an intensifier is modifying. This is especially important in the case of those adverbs that are homonymous with adjectives, for example, *hard, real, regular, flat.*

Additionally, Ito and Tagliamonte (2003) restricted instances of intensification only to those that permit intensifiers. However, this decision is not entirely obvious regarding the requirements of excluding such instances. Thus, the authors give the following examples (6).

(6) You didn't have any really scary teachers.

[1] The tag designates that the example is take from the interview number 26 in Louvain Corpus of Native English conversation. The letters S, F and P stand for the task used for the interwies: set topic, free discussion, picture description.

Their justification for excluding this example is that in (6) a speaker does not inquire about the degree of "scariness", but rather about the kind of a teacher. It seems that the authors refer to the fact that *scary* is used as a classifying, non-gradable adjective and, thus, does not permit modification by intensifiers. However, as pointed out in sub-chapter 1.1, classifying adjectives can acquire gradable meaning. Moreover, *scary* is usually a scalar adjective, and that is why it readily accepts modification by intensifiers even when used in a classifying sense.

Ito and Tagliamonte (2003) pointed out that the intensifiers *so* and *too* can appear in constructions that do not strictly qualify their function as intensifiers. Thus, the constructions in which the intensifiers *so* and *too* introduce the result clause were excluded by the authors. Quirk et al. (1999) mention that in this construction, a degree reading is still present. However, when it comes to *too* in (7), the adjective seems to get a degree meaning, as it rather says it is hot to the degree that is not possible to do something. In my opinion, the intensifier *too* + result clause provides the degree reading, and that is why I am going to include those examples.

(7) a. … too hot to do anything so ... <LOCNEC_25_F>

Lastly, Tagliamonte (2005) reports that in her data there were instances of double intensification, repetition of intensifiers in particular. These instances involved examples like *so so happy*, *very very sorry*. Rescki (2004) reported that combination *very very* is the most frequent combination in MICASE corpus. Kennedy and McNally (2005) analysed the construction of multiple degree intensification different from the ones reported by Ito and Tagliamonte (2005) and Rescki (2004). These constructions involve modification by different intensifiers and are analysed in a nested manner as in (8).

(8) … the job market was so very bad … <LOCNEC_E01_F>

The analysis of Kennedy and McNally (2005) suggests that construction like these have to be treated as the case where the adjective phrase modified by the intensifier *very* designates a degree of the adjective *bad* that is further modified by the modifier *so*. Construction like (8) were rather frequent in the data used in this study. However, these instances of double intensification were not analysed separately from other examples of intensification because there is no sufficient theoretical background on this topic.

The previous research that investigated intensifiers in native speakers' data has yielded valuable insights into the ways intensification presents itself in natural language data. The authors have used various corpora and investigated different syntactic patterns of intensifies. These findings provide important points for consideration when designing a study that deals with intensifiers in spoken language data.

1.4 Intensifiers and Learner language

As shown in the previous subsection, the use of intensifiers among native speakers is influenced by various grammatical, semantic and socio-linguistic factors. It is clear that intensifiers are a difficult topic for learners of English. Collocational preferences of intensifiers are an especially difficult task as learners may not be aware of the factors that play a role in combining intensifiers with adjectives.

Intensification has received a substantial amount of attention in the studies on learner language. Various studies have examined the use of intensifiers in written language of advanced learners of English. (Lorenz 1999, Granger 1998, Rescki 2004, Pérez-Paredes 2010, Schweinberger 2020). In the studies conducted by Granger (1998) and Lorenz (1999) on the differences in the use of intensifiers in the written language of both L1 and L2 speakers of English, several generalisations were made. The authors concluded that

compared to L1 speakers, learners tend to overuse *very* and *really* in argumentative texts. Granger (1998) investigated the differences between native and learner data in ICLE (International Corpus of Learner English). The corpus contains material that consists of samples of written language produced by native speakers and learners of English who are native French speakers.

Granger (1998) reported significant underuse of amplifiers in the learner corpus. In this study, the focus was on adverbs that end in *-ly* in comparison to bare adverbs like *very*. The author pointed out that the underuse of amplifiers in terms of tokens and types is due to the overuse of maximisers *completely* and *totally* and significant overuse of *very*. It was concluded in this study that *very* is not only the most frequent intensifier but also a strategy that the learners used extensively due to the fact that *very* is one of the most versatile intensifiers. *Very* collocates with a significant number of adjectives and, according to Granger (1998), is the main reason learners underused -ly adverbs compared to native speakers. However, the author only examined intensifiers that end in *-ly*.

Lorenz (1999) investigated the use of intensifiers in terms of frequencies and their syntactic structures. His approach differs from the one undertaken by Granger (1998). Lorenz (1999) conducted an extensive study that looked at all the intensifiers present in the sample of written data produced by English learners, who are native speakers of German, and compared it to the use of intensifiers in native speakers' essays. His findings suggest, contrary to what was reported by Granger (1998), that learners of English use adjective intensification more frequently compared to English L1 speakers. The study conducted by Lorenz (1999) takes into account a great number of variables: syntactic position, idiomaticity and information structure. The main differences between learners and native speakers were reported in terms of adverb + adjective collocations and "information overcharge", as learners have used intensifiers as modifiers of adjectives that are part of a subject noun

phrase. Learners deviate from English L1 speakers as they tend to modify adjectives in the thematic and rhematic role, leading to an overall "overcharge" of information in their texts. Lorenz (1999: 209) argues that this creates an impression of "wordiness" and "overstatement".

Schweinberger (2020) examined the differences between the learners of English with various L1 backgrounds and English L1 speakers in terms of the use of the intensifiers *very*. This study is based on LINDSEI and LOCNEC data sets. The author used a complex statistical model that allowed them to control such factors as individual speakers' differences, priming in the use of intensifiers, and syntactic position of adjectives. According to the author, a combination of many factors did not help to identify substantial statistical differences between the Learners of English and English L1 speakers. However, the method allowed to identify that the Learners of English are likely to deviate from English L1 speakers when intensifying high frequency adjectives in predicative syntactic position.

For spoken language, a study by Pérez-Paredes (2010) suggests that amplifiers are not part of the active spoken lexical repertoire of learners when compared to native speakers. In this study, Pérez-Paredes (2010) analyzes the use of adverbs in the Spanish component of the Louvain International Database of Spoken English Interlanguage (LINDSEI) in comparison to the native speaker component of this database. He focused on the last part of the interviews, namely the picture description task, as both learners and native speakers were faced with an identical set of pictures. However, it is worth mentioning that the three tasks of LINDESI and LOCNEC (set topic, free conversation, and picture description) are identical and only differ in terms of the topic participants choose in the first and the second parts. The author reports that Spanish speakers in the data almost exclusively neglected downtoners, and among 59 interviewees, 29 have not used intensifiers at all. However, Pérez-Paredes (2010) points out that native speakers alike did not use a lot of intensifiers,

which correlates with the way informants have approached the task of picture description. Similar findings were reported in Perez-Parades and Camino Bueno-Alastuey (2019) for the spoken data of Chinese, German and Spanish learners of English. In this data set, the authors pointed out that there are differences in terms of how participants approached the task. Thus, German and Spanish learners understood the task as requiring a high degree of involvement. The native speakers, on the other hand, showed a low degree of involvement and tried to report on the events described in pictures as concisely as possible and as the result used fewer adverbs. This point is important as it suggests that the task itself influences the distribution of intensifiers.

1.5 Implications for the study

Previous research provides valuable insights that help approach the topic of intensification, paying attention to various factors that influence the distribution of intensifiers. As shown in the study conducted by Paradis (1997), there is an interaction between intensifiers and adjectives in terms of the semantics of adjectives. Using the classification of adjectives can explain the over- or underuse of particular intensifiers. The previous research on the learner language (Perez-Parades and Camino Bueno-Alastuey 2019) points out that the tasks influence the distribution of intensifiers. Thus, the intensifiers from LINDSEI-EST and LOCNEC will be analysed according to the tasks. Schweinberger (2020) suggests that statistical models that incorporate multiple factors provide significantly more fine-grained analysis. Thus, the analysis will try to incorporate multifactorial analyses in order to approach the over- or underuse in the way that would provide statistically informed results.

2. EMPIRICAL ANALYSIS

This study aims to analyse the use of degree adverbs by native and non-native speakers of English. The interviews presented in the LINDSEI (can be accessed through the LINDSEI website) and LOCNEC corpora provide comparable data as both native and non-native speakers are asked to talk on the same topics presented with the same visual prompts. The corpus study allows to analyse natural language in use and avoid collecting data by conducting an experiment, which may provide ecologically less valid data. Additionally, the chosen corpus contains audio that would be particularly helpful in the study as it provides clues for potentially disambiguous language use

The corpus I am using for my study is the LINDSEI corpus which contains the spoken data of EFL learners of different language backgrounds. Degree adverbs are especially frequent in the spoken language. Thus, this factor serves as the primary reason for choosing the corpus. Additionally, the LINDSEI corpus contains an Estonian component that allows conducting the research as there are no previous studies that examined the use of degree adverbs by EFL learners with Estonian as L1.

The sample used in my study is representative of advanced Estonia EFL speakers. The focus is on the comparison of the use of degree adverbs by native and non-native speakers, and the corpora provide representative data to analyse. The Estonian subcorpus contains transcribed interviews of 25 speakers. The native speaker's subcorpus (LOCNEC) contains 50 interviews.

The data is stored in the .csv format as it is a quite versatile and convenient format for the analysis and statistical representation. Adverbs, adjectives and the left and right context of tokens are included in different columns.

The main data gathering tool is Python (Van Rossum and Drake Jr. 1995). Modules that allow language processing and statistics are NLTK (Bird, Loper and Klein 2009), Numpy (Harris, Millman, van der Walt 2020) and RStudio (RStudio Team 2020, version 1.4.1717) Additionally, I am using *Notepad++* to manually clean the .txt files and prepare them for being used in Python. The hierarchical cluster analysis and multiple correspondence analysis require the use of the following packages in Rstudio: *cluster* (Maechler et al. 2022), *pvclust* (Suzuki et al. 2019), *FactoMiner* (Le et al. 2008), *ca* (Nenadec et al. 2007) and *rms* (Harrell 2022). The cluster and pvclust are used to conduct hierarchical cluster analysis and validate the clustering results. FactoMiner package allows to conduct multiple correspondence analyses and extract valuable for the interpretation of the results information. Rms package provides functions used to conduct logistic linear regression that is used to validate the results of multiple correspondence analyses.

In order to extract Adverb-Adjective collocations with the context, I manually cleaned the files containing the transcriptions. Pauses, extralinguistic information, and interviewee turns were removed to ensure proper data analysis for the frequency distribution. The data was POS-tagged in order to allow the extraction of the ADV-ADJ pairs. The extracted collocations were annotated for the type of degree adverbs. Additionally, adjectives in the data were annotated according to the classification proposed by Croft and Cruse (2004) and used in Paradis (1997). Namely, adjectives were divided into three groups: scalar, extreme and limit.

The scalar adjectives among many include such adjectives as *good, fast, long, difficult, nasty, interesting, nice, embarrassed, tricky, sunny, pleasant, hardworking, good, boring, interesting, nasty, glad, simple, friendly, big, unpleasant, small.*

The group of extreme adjectives contains the following adjectives: *excellent, huge, minute, terrific, disastrous, brilliant, extraordinary, shattered, astounding, crowded, super, splendid, barmy, terrifying.*

The limit adjectives include the following adjectives: *true, sober, sufficient, dead, identical, possible, sure, clear, different, right, certain, true, impossible, sufficient, wrong, normal, possible.*

Given the fact that LINDSEI-EST contains 25 interviews and LOCNEC contains 50 interviews, I decide to create a random sample of 25 interviews from LOCNEC. This was done in order to have a comparable data set for hierarchical cluster analysis and multiple correspondence analysis. The random sample contains interviews of the following speaker codes from LOCNEC: eng022,eng035, eng012, eng051, eng045, eng050, eng031, eng007, eng004, eng010, eng011, eng019, eng038, eng005, eng055, eng049, eng003, eng042, eng013, eng027, eng021, eng018, eng016, eng037, eng009.

The previous research on intensifiers largely focused on drawing conclusions that are based on raw frequencies and lexical collocations. However, this approach has its limitations as it disregards the fact that certain adjectives can be overall frequent in the corpus and that would lead to higher frequencies of intensifiers that modify these overall frequent adjectives. Thus, in order to provide a more thorough analysis, I am using two exploratory methods: hierarchical cluster analysis (HCA) and multiple correspondence analysis (MCA). Detailed overview and guidelines of application for both of the methods are described in Levshina (2015) and Desagulier (2017).

To identify differences between intensifiers, I divided annotated collocations in terms of types of intensifiers: boosters and maximizers. According to Paradis (1997, 2001), there is a close connection between the types of intensifiers and adjectives. Namely,

maximisers modify limit and extreme adjectives, and boosters modify scalar adjectives. The methods used in the analysis make it possible to identify the differences in the use of intensifiers based on adjectives that are intensified, type of adjectives (scalar, extreme and limit) and speakers (English L1 and Estonian L1).

2.1 LINDSEI and LOCNEC Corpora

The corpora used in this study contains the Estonian component of the Louvain International Database of Spoken English Interlanguage (LINDSEI-EST) and the English L1 component (LOCNEC). LINDSEI-EST contains 25 fully transcribed interviews that amount to 348 minutes of speech and each interview is from 10 to 17 minutes long. The procedure of recording interviews was as follows. The interviewer asked each participant to choose one of the set topics: an experience that had taught the speaker an important lesson, a country that had impressed the contributor or a film or play that had attracted the speaker's attention. The first part of the interview was followed by a free discussion. In the end, the participants were asked to construct a story based on the set of pictures.

The LOCNEC component contains 50 interviews conducted with university students in the UK. All the participants were native speakers of English. Twenty-five interview transcriptions of LINDSEI-EST data contain 35,117 words, and fifty interviews of LOCNEC – 121,483 words.

The interviews in the LINDSEI-EST were recorded at the University of Tartu. The participants were advanced learners of English with the Estonian language as L1.

2.2 Data extraction and annotation

In this section, I will discuss the process of extracting the data from LINDSEI and LOCNEC corpora. The transcriptions of interviews present in both corpora are not annotated

for part of speech, so it is impossible to extract intensifiers and adjectives using conventional corpus analysis software (for example, *AntConc*). The task requires the use of tools available for text analysis in programming languages, such as *Python*. *Python* programming language is chosen because it is widely used for analysing natural language data and provides a variety of tools that allow corpus analysis of unannotated data. However, the validity of the outcome of the analysis depends on a well-planned data preparation process. Additionally, the extracted data needs to be manually annotated to include relevant information used for interpreting the results. For these reasons, the process of data extraction and annotation consists of the following steps:

1. Data preparation
2. POS-tagging
3. Extraction of target items
4. Manual annotation
5. Data storage

As previously stated, the transcriptions included in both corpora are not tagged for the part of speech. Moreover, according to LINDSEI and LOCNEC corpus guidelines, the data includes tags that designate speaker turns, overlapping speech, empty and filled pauses, unclear passages, phonetic features, prosodic information, nonverbal vocal sounds, and contextual comments. The transcriptions contain the following tags:

Speaker turns: <A> and </A> designate interviewer's turns.

<B> and </B> designate interviewee's turns.

Filled pauses: (eh) , (er), (em), (erm), (mm), (uhu) and (mhm).

Unclear passages: <X> designates unclear sounds up to one word.

<?> is used when a transcriber is not sure about the word.

Tasks: <S>, <F>, <P> are used to mark the beginnings and the ends of tasks.

The focus of this study is the use of intensifiers by learners of English compared to native speakers. Thus, I have chosen to analyse interviewee turns only. For this reason, the interviewer's turns marked by <A> </A> had to be excluded using regular expressions in *Notepad++*. Regular expression is a feature included in many text editing programs that allows to construct search patterns, that make it possible to delete different strings, for example all the strings of text that are enclosed by the speakers turns' tags <A></A>. The resulting text includes all the interviewees' turns. Additionally, the data was divided according to the three tasks: set topic, free discussion, and picture description. Dividing the data by the tasks will allow to analyse the distribution of intensifiers according to each of the three conversational situations. As a result, the interviews had to be separated into different files according to the tasks. In the end, I have compiled a dataset with 75 files of the LINDSEI-EST transcriptions and 150 files of the LOCNEC transcriptions.

The tags present useful information for the data annotation. The data in the corpora is an example of spoken language and transcribed in a way that disregards punctuation, as it is quite difficult to designate utterance boundaries in spoken data. Thus, preserving all the tags that mark the beginning and the end of utterances allows not to include the instances of target items that belong to separate turns as words like *too* and *so* are quite frequent at the end of utterances.

In order to count the overall number of words present in the corpora, the tags have to be excluded. For these reasons, I have decided to create two sets of data: with and without transcription tags. The first one is used solely for the data annotation. Additionally, the POS-tagger turned out to be able to handle the tags, and, for the most part, it did not result in inaccurate POS-tag assignment.

The POS-tagging process was conducted using NLTK in Python. NLTK is the freely available library of text analysis tools that includes a POS-tagger. POS-tagging is an essential step for text analysis and data extraction. The tagger assigns part of speech tags for each word in a text and allows the extraction of adverb-adjective combinations. The NLTK tagger requires the sequence of code to be added to the script in *Python* that is used for data extraction. The text files with interviews are used as input for the POS tagger, and the result is a string of text that contains part of speech tags next to each word of the text. The POS-tagged text is later used as the input for the piece of code that handles the extraction of target items (ADV-ADJ combination). The output of POS-tagging can be seen here:

```
<p nt=A nr=E55>
<B> ... he’s er he’s painting this young lady . and then .. he’s er she’s he’s finished and . she comes over and looks at the picture and .. doesn’t like it she’s frowning . and she’s frowning on the picture as well which is why she doesn’t like and her hair is all drab . and boring old dress

[('<', 'JJ'), ('p', 'NN'), ('nt=A', 'NN'), ('nr=E55', 'JJ'), ('>', 'NNP'), ('<', 'NNP'), ('B', 'NNP'), ('>', 'NN'), ('...', ':'), ('he', 'PRP'), ("'s", 'VBZ'), ('er', 'CC'), ('he', 'PRP'), ("'s", 'VBZ'), ('painting', 'VBG'), ('this', 'DT'), ('young', 'JJ'), ('lady', 'NN'), ('.', '.'), ('and', 'CC'), ('then', 'RB'), ('..', 'VB'), ('he', 'PRP'), ("'s", 'VBZ'), ('er', 'JJ'), ('she', 'PRP'), ("'s", 'VBZ'), ('he', 'PRP'), ("'s", 'VBZ'), ('finished', 'VBN'), ('and', 'CC'), ('.', '.'), ('she', 'PRP'), ('comes', 'VBZ'), ('over', 'RB'), ('and', 'CC'), ('looks', 'VBZ'), ('at', 'IN'), ('the', 'DT'), ('picture', 'NN'), ('and', 'CC'), ('..', 'NN'), ('does', 'VBZ'), ("n't", 'RB'), ('like', 'VB'), ('it', 'PRP'), ('she', 'PRP'), ("'s", 'VBZ'), ('frowning', 'VBG'), ('.', '.'), ('and', 'CC'), ('she', 'PRP'), ("'s", 'VBZ'), ('frowning', 'VBG'), ('on', 'IN'), ('the', 'DT'), ('picture', 'NN'), ('as', 'IN'), ('well', 'RB'), ('which', 'WDT'), ('is', 'VBZ'), ('why', 'WRB'), ('she', 'PRP'), ('does', 'VBZ'), ("n't", 'RB'), ('like', 'VB'), ('and', 'CC'), ('her', 'PRP$'), ('hair', 'NN'), ('is', 'VBZ'), ('all', 'DT'), ('drab', 'NN'), ('.', '.'), ('and', 'CC'), ('boring', 'VBG'), ('old', 'JJ'), ('dress', 'NN')
```

The data extraction step is done in *Python* using the code that allows specification of target items and takes as an input the POS-tagged text. The code was specifically written to allow the extraction of adverbs that modify adjectives and five items to the left and right in order to include context. The result is saved in an excel document in the form of a table, where each token occupies a separate cell. The procedure was done separately for each of

the files in order to ensure that the output is correct, as automated processing of large sets of data may lead to errors, especially as the output is saved in the spreadsheet format .csv.

Finally, the extracted data in the csv. format had to be manually annotated for the types of intensifiers, predicative and attributive positions of adjectives. This process included disregarding instances where adverbs do not function as intensifiers or are in the scope of negation. Thus, the following examples and similar were excluded:

Negation:

… yeah it is it's not too bad at all … <LOCNEC_45_F>

Nontarget items:

… we had this tiny little house … <LOCNEC_50_S>

… a little tiny melting pot … <LOCNEC_50_S>

… I had a little flat (tagged as adjective) <LOCNEC_36_S>

For the reasons that the extracted data contained many false tokens , as shown above, mainly due to the mistakes made by POS-tagging, I had to go through the material and ensure that the examples included in the analysis were genuine instances of intensification.

The extracted instances of intensification were saved in the .csv format. This format is particularly suitable for the analysis done in Rstudio.

2.3 Methodology

In this subsection, I introduce the two methods used in this study: hierarchical cluster analysis (HCA) and multiple correspondence analysis (MCA). Here, the main procedures required for applying both of the methods are introduced.

2.3.1 Hierarchical cluster analysis

The first method used in this thesis is hierarchical cluster analysis. This type of analysis belongs to a family of exploratory methods (Levshina 2015: 320) and allows observation of certain data structure patterns. The results of such types of analysis help to outline particular trends exhibited in the data; however, they should be validated using hypothesis-testing methods, such as regression analysis, among others (ibid.). The data used in this thesis is limited to 25 interviews in LINDSEI-EST corpus and 50 interviews in LOCNEC. Thus, exploratory methods are more suitable, as the small number of observations regarding intensifiers would yield insufficient statistical results in regression analysis or distinctive collexeme analysis (Gries and. Stefanowitsch 2004).

The main purpose of cluster analysis is to identify groups or clusters of data points that have similar profiles (Levshina 2015:309). In other words, the lexical items that behave similarly would be clustered together. The result of cluster analysis is a tree or dendrogram where items are situated at the ends of nodes. Each item in the data represents its own distinct cluster which is compared to all the other items. The basis for identifying similarities between data points is a distance matrix that allows representing the data in terms of numerical distances between the data points. The distances of similar profiles are smaller and for the identical items equals 0. The distance matrix and the values presented in it serve as bases for visualising the dendrogram and the branches in the cluster analysis trees.

The study on intensifiers (Desagulier 2014) described in Desagulier (2017: 278-283) uses the hierarcical cluster analysis to identify groups by which intensifiers can be classified into boosters, maximixers, diminishers and moderators. The results of the study reveal that intensifiers in the Corpus of Contemporary American English (Davies 2008-2012) can be grouped according to the classification proposed in Paradis(1997). These findings provide additional grounds for applying this method to the data used in this thesis.

The application of cluster analysis consists of various steps and includes methodological decisions on the part of the researcher. The dendrogram that visualises the results of the analysis requires a thorough interpretation of all the factors that contribute to the clustering of data points. Thus, these steps are worth being discussed here. These steps are as follows: data preparation, application, and validation of the clustering results.

The initial step requires the data to be transformed into a table of two columns representing the data points. In the case of the analysis described in this thesis, four tables were created. These tables include intensifiers and adjectives from both LONCEC and LINDSEI-EST corpora. The intensifiers with types of adjectives were classified according to the classification proposed by Paradis (1997). These four tables allow analysing clusters that are based on adjectives modified by intensifiers in both corpora and compare the results to the clustering based on types of adjectives. This decision allows for inspecting whether there is an observable correlation between adjectives and the types they represent. In turn, it also allows identifying which adjectives and types of adjectives contribute to the clustering of intensifiers.

The data preparation procedure requires creating a distance matrix that serves as a primary basis for grouping the data points together. Desagulier (2017) and Levshina (2015) point out that there exist several types of distance matrices based on different distance

measures: Euclidean, Manhattan and Canberra. According to the authors, Canberra is the most suitable type of distance matrix for linguistic data containing so-called "zero occurrences" (Desagulier 2017: 278). These are the cases when certain types of items (intensifiers, adjectives or types of adjectives, in this study) do not co-occur in the data. Levshina (2015: 307-308). points out that Canberra is better at handling the cases when frequencies of particular items in the data are higher compared to the rest of the data set

The output of the transformation of the table containing intensifiers and types of adjectives into the matrix can be displayed in Rstudio. Inspection of the matrix reveals that such intensifiers as *absolutely*, completely, *terribly* and *totally* do not co-occur with scalar types in the LOCNEC data set. Thus, following the recommendations of Levshina (2015) and Desagulier (2017), I chose to use the Canberra distance measure.

```
ns.matrix

##
##                extreme limit scalar
##   absolutely         5     2      0
##   completely         3     5      0
##   particularly       0     0      7
##   really            10     6     91
##   so                11     7     17
##   terribly           1     1      0
##   too                1     3      8
##   totally            0     2      0
##   very              21    35     98
```

The next step after creating the matrix is calculating distances between the data points and stroying it in the format of the distance matrix. In Rstudio this is done by using function *dist* from the package *cluster*. The function allows for the specification of the distance measure for calculating distances and requires a matrix to be supplied. The result can be examined in the console of RStuido.

```
distance <- dist(ns.matrix,method="canberra", diag=T, upper=T)
distance
##             absolutely completely particularly    really        so  terribly
## absolutely   0.0000000  1.0178571    3.0000000 1.8333333 1.9305556 1.5000000
```

```
## completely    1.0178571  0.0000000    3.0000000 1.6293706 1.7380952 1.7500000
## particularly  3.0000000  3.0000000    0.0000000 2.8571429 2.4166667 3.0000000
## really        1.8333333  1.6293706    2.8571429 0.0000000 0.8097273 2.5324675
## so            1.9305556  1.7380952    2.4166667 0.8097273 0.0000000 2.5833333
## terribly      1.5000000  1.7500000    3.0000000 2.5324675 2.5833333 0.0000000
```

The output provides a glimpse of the distance values that are used to produce the results of hierarchical cluster analysis. The results are shown for the first six intensifiers out of nine for the reason of space. The lower the distance values, the more similar items are. The values are presented in the distance matrix format, where the lowest value, which is always equal to zero, is between an intensifier in comparison with itself. The matrix reveals that intensifier *absolutely* has the lowest distance value in comparison to *completely*, which suggests that they are likely to be grouped in one cluster.

The next step requires submitting the matrix to the function *hclust* in Rstudio that analyses the distances and divides items into clusters. The procedure requires specification of the method of clustering, which according to Levshina (2015) and Desagulier (2017:279) is Ward's method. This method is one of the most popularones used in linguistics as it allows the creation of clusters of small sizes, which is suitable when dealing with linguistic data where the variance is usually quite high. The result of the clustering procedure can be outputted by using the function *cutree* in Rsudio. The number under the intensifiers represents different clusters according to which items were grouped. Thus, as observed previously by examining the distance matrix, intensifiers *absolutely* and *completely* are grouped together.

```
clusters <- hclust(distance, method="ward.D2")
cutree(clusters,4)
```

```
##   absolutely completely particularly really so terribly too totally very
##        1          1          2          3    3    4      4     1     3
```

In addition to the steps mentioned above in cluster analysis, Levshina (2015) suggests a couple of ways the clusters produced by the analysis can be further examined. The clustering can be examined for the optimal number of groups that would yield the results

worth being reported for a particular data set. Computation of the silhouette widths allows for an informed decision regarding the optimal number of distinct clusters (Levshina 2015: 312)

```
asw <- sapply(2:7, function(x) summary(silhouette(cutree(clusters,k = x)
, distance))$avg.width)
asw
```

```
## [1] 0.2423905 0.2896215 0.2490490 0.2445117 0.1932182 0.1635612
```

The result of the computation of silhouette widths provides a set of widths according to the number of clusters specified, and the largest widths correspond to the optimal number of clusters. In this case, testing the widths for 2 to 7 clusters reveals that 3 clusters are optimal for this data set, as 0.28 is in the second position that corresponds to 3 clusters.

The other technique mentioned by Levshina (2015: 313) is the computation of absolute differences between the scores of a particular cluster. This procedure provides means for identifying distinct features that contribute to particular clusters being formed. For this, the average proportions for every feature of the cluster are computed using the function *colMeans*. Computation of the average proportions of the features of the cluster containing intensifiers *absolutely*, *completely*, *terribly* and *totally* shows that the distinctive characteristic of this cluster is a very low proportion of scalar adjectives and a higher proportion of extreme and limit adjectives.

```
diff
```

```
## extreme   limit  scalar
##   -6.35   -7.70  -44.20
```

The final technique suggested by Levshina (2015) is the validation of the cluster solution (Levshina 2015: 315). The aim of validation is to check how well the provided data support the clusters. The function *pvclust* takes the data as input and calculates Approximately Unbiased p-value (Levshina 2015: 316). The closer the p-value of each cluster to 100, the greater the support for the cluster in the data.

The procedures mentioned above ensure the correct application of the hierarchical clustering analysis and robust interpretation of the results. The method allows for computing and visualising patterns observed in the data. As pointed out by Levshina (2015: 306), the clustering method provides a reliable alternative to examining frequency tables and drawing conclusions that are not informed by supplementary statistical methods.

2.3.2 Multiple Correspondence Analysis

This section introduces Multiple Correspondence Analysis (MCA) and how it was applied to the data sets used in this thesis.

MCA is another exploratory is used to analyse data that contains more than two categorical variables. Categorical variable is a type of variable that is characterised by having categories, which differs it from numerical variables that have numerical values. Thus, in the data set, compiled by extracting all instances of intensifiers in LOCNEC and LINDSEI-EST, contains such variables as *intensifier*, *adjective type* and *speaker*. The variable intensifiers has each unique intensifier as categories. Adjective types has three categories: scalar, extreme and limit. Finally, the variable *speaker* has two categories: English L1 speaker and Estonian L1 speaker. According to Desagulier (2017: 269), MCA was initially designed for the data obtained from surveys. Thus, this method is designed to handle data that contains categorical variables.

MCA is suitable for the exploration of the relations between various aspects of the data. Moreover, Levshina (2017: 375) points out that the advantage of the method is the possibility of the representation of each individual item and information that describes them in the same 2D or 3D space. Thus, it is possible to examine to what extent LONCEC and LINDSEI-EST data differ in terms of a set of intensifiers and adjective types modified by them.

MCA uses data frames with variables coded as factors as input for the analysis. Data frames are a common data structure used in RStudio that are similar to spreadsheets. However, data frames permit changing the type of variables presented in columns. Thus, the transformation of columns of a data frame into factors makes it possible to use columns as categories. Each such categories contain levels. For example, a table containing columns intensifiers, adjective types, and speaker has each unique instance of these columns as levels of categories intensifier, adjective type and speaker. Data frames containing factor variables can be submitted to various analyses, including MCA.

The results of MCA are visualised as a plot that represents two dimensions calculated on the bases of variables in the data. The goal of MCA is to represent the data using fewer dimensions, maintaining a high explanatory level. Unlike hierarchical cluster analysis described in the previous session, MCA calculates the relations between multiple variables describing each item in the dataset submitted to this type of analysis. Thus, it is possible to observe which intensifiers behave similarly and which intensifiers and adjective types are characteristic of two groups of speakers in the dataset.

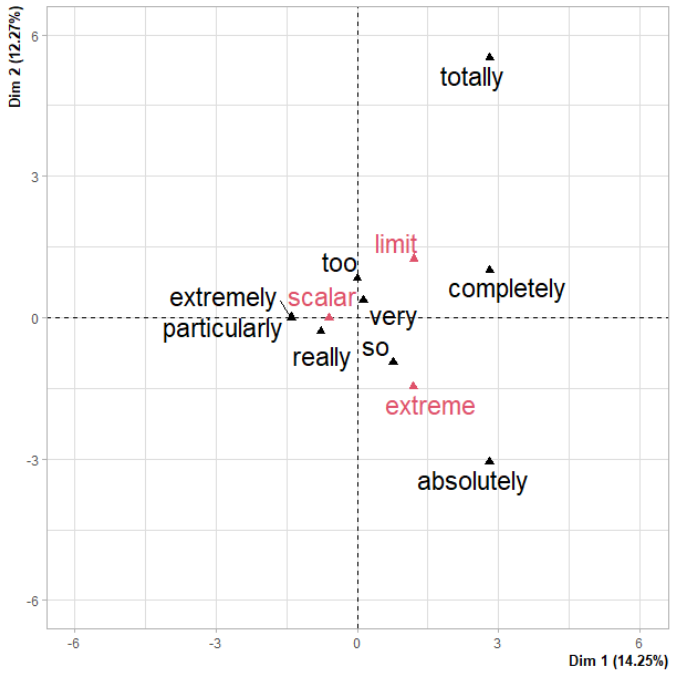


Figure 2. MCA biplot: intensifiers, type of adjectives and speakers in LOCNEC

Figure 2 demonstrates the visualisation based on the results of MCA. This biplot presents the results of MCA applied to the LOCNEC dataset. The biplot is constructed on the basis of two dimensions. The aim of the analysis is to explain the data in the most concise way using these two dimensions. Levshina (2015: 382) and Desaguiler (2017: 272) explain that one of the important values that should be reported in terms of the results of MCA is the proportion of explained variance. As can be seen in Figure 2, horizontal and vertical axes have the titles of dimension and percentages of explained variance in the brackets. Desaguiler (ibid) points out that the percentages are usually quite low in linguistic data. Thus, the percentages of explained variance of dimensions 1 and 2 in Figure 2 are expected to be low.

However, Levshina (2015:382) mentions that due to the fact that usually the data for MCA contains a large number of variables and the values of explained variance should be adjusted. Levshina (2017: 382) cites the solution proposed by Greenacre (2007), which is implemented in the package *ca* and allows for a more realistic calculation of explained variance . The output of the adjusted values can be seen below. According to the adjusted values of explained variance, the first dimension accounts for 77.8%. This number shows that the explenatory power of the first dimension is much higher compared to what was calculated initially in Figure 2.

```
summary(mjca(df.mca))

## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.180360  77.8  77.8  ******************
##  2      0.051366  22.2 100.0  ******
##  3      00000000   0.0 100.0
##  4      00000000   0.0 100.0
##  5      00000000   0.0 100.0
##  6      00000000   0.0 100.0
```

Another measure that Levshina (2015: 380-381) and Desaguiler (2017: 275) recommend reporting is the 95% confidence ellipsis for the categories. This provides the

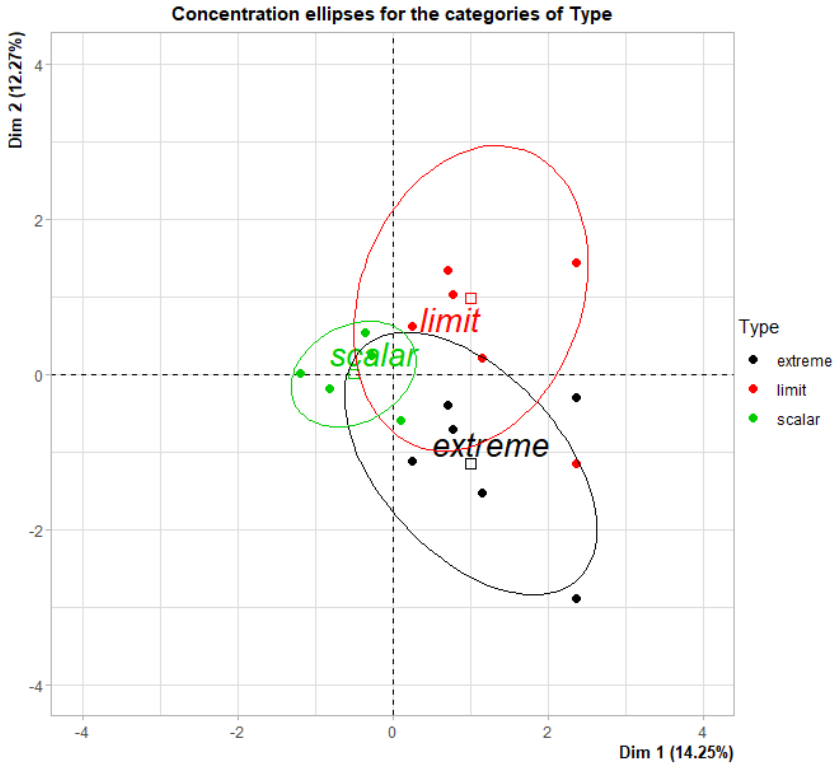


Figure 3. Confidence ellipses of the types of adjectives in LOCNEC

means of examining how distinct categories are. Here, I apply *plotellipses* specifying the category of types of adjectives. The intensifiers are represented as dots on the biplot.

The confidence ellipses of types of adjectives in LOCNEC in Figure 3 show the overlap between the categories. This fact makes it possible to conclude that the types of adjectives do not exclusively correlate with particular types of intensifiers.

Levshina (2015: 383) provides another useful technique for interpreting the MCA results. The author suggests that the dimensions calculated in MCA can be used to examine their predictive power in a logistic regression model. Logistic regression allows for calculating relationships between response and explanatory variables. Thus, it is possible to calculate to which extent the dimensions predict two categories of speakers in LOCNEC and LINDSEI-EST. The category of speakers consists of English L1 and Estonian L1 categorical variables. These categorical variables are chosen as response variables, and the dimensions of MCA – as explanatory variables. The logistic regression calculates how well intensifiers

and types of adjectives can explain the categories of speakers. The assumption is that if English L1 and Estonian L1 speakers differ in terms of the use of intensifiers in relation to types of adjectives, then the results of logistic regression will support this by statistical measures. Levshina (2017: 259) explains that the most important measure in the statistic C. According to the scale proposed by Hosmer &Lemeshow (2000: 162, as cited in Levshina 2017:259), if C equal 0.5, then the model does not support concluding substantial differences between the response categories. Values of C that lie between 0.7 and 0.8 are acceptable, and all the values above 0.8 confirm the substantial effect of explanatory variables (ibid).

In this subchapter, multiple correspondence analysis was introduced. This type of analysis is suitable for exploring the data in order to identify whether there are substantial observable differences between English L1 and Estonian L1 speakers in LOCNEC and LINDSEI-EST datasets.

2.4 Results

In this section, I present the analysis results of the data analysis in the form of frequency tables and figures and provide a detailed explanation of the findings. The subsection is divided into sections in which I describe the distribution of intensifiers in LINDSEI-EST and LOCNEC data sets, provide the results of hierarchical clustering analysis and multiple correspondence analysis.

2.4.1 Distribution of intensifiers in LINDSEI-EST and LOCNEC

The data extraction yielded a total of 1317 instances of adjective intensification in LINDSEI-EST and LOCNEC combined. This number of intensifiers is divided into 318 for LINDSEI-EST and 999 for LOCNEC. Due to the differences in size between the two data sets, the frequency numbers reported in this thesis were normalised. In order to draw

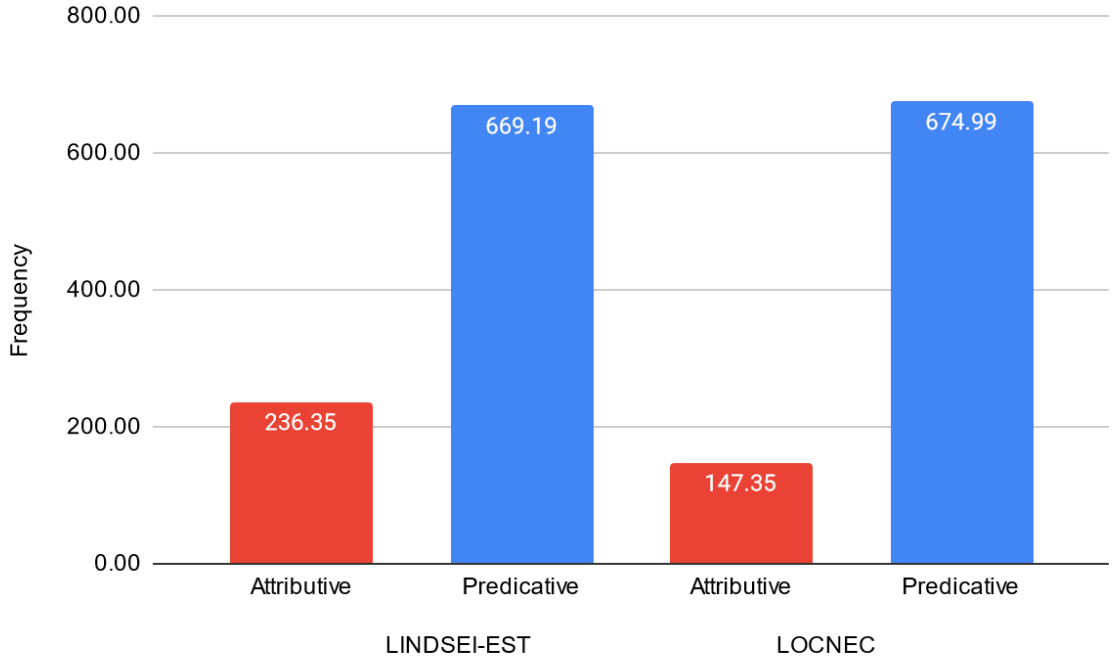


Figure 4. Normalised frequency of intensifiers according to positions of adjectives

conclusions based on relatively comparable frequencies of intensifiers, 10,000 was chosen as the basis of normalisation. In terms of normalised frequencies, there are 905.54 tokens in LINDSEI-EST and 822.34 in LOCNEC. The normalised frequencies point to the fact that the datasets do not differ significantly in terms of overall frequencies.

LINDSEI-EST contains 28 unique intensifiers, and a total of 160 different adjectives are modified. LOCNEC data set includes 35 different intensifiers, and a total of 317 different adjectives are intensified. The most frequent adjectives in the scope of intensification in both LINDSEI-EST and LOCNEC are *good, different, nice, interesting, impressive, happy, long, interested, friendly, cool, beautiful, small, lovely, funny, difficult, big, bad, hot*. The only difference between the most frequent adjectives in the two data sets is that *bad* and *hot* are absent in LINDSEI-EST.

Intensified adjectives in both data sets are found in attributive and predicative positions. The frequencies by the position of adjectives can be seen in Figure 1. Overall, the

INTENSIFIERS	Raw Frequency	Normalised Frequency
very	94	267.68
really	81	230.66
so	42	119.60
quite_downtoner	30	85.43
particularly	9	25.63
completely	8	22.78
too	8	22.78
kind_of	6	17.09
rather	6	17.09
pretty_downtoner	5	14.24
quite_maximizer	4	11.39
such	4	11.39
a_bit	3	8.54
enough	2	5.70
extremely	2	5.70
sort_of	2	5.70
a_little_bit	1	2.85
crazy	1	2.85
entirely	1	2.85
fairly	1	2.85
highly	1	2.85
just	1	2.85
kinda	1	2.85
quite_well	1	2.85
super	1	2.85
terribly	1	2.85
totally	1	2.85
vastly	1	2.85
Grand Total	**318**	**905.54**

INTENSIFIERS	Raw Frequency	Normalised Frequency
very	299	246.12
really	219	180.27
quite_downtoner	106	87.26
so	93	76.55
quite_maximizer	86	70.79
a_bit	36	29.63
too	28	23.05
completely	13	10.70
fairly	12	9.88
absolutely	12	9.88
particularly	11	9.05
sort_of	11	9.05
pretty_downtoner	10	8.23
rather	9	7.41
totally	9	7.41
enough	8	6.59
pretty_booster	7	5.76
a_little_bit	5	4.12
kind_of	4	3.29
a_little	3	2.47
incredibly	2	1.65
a_lot	2	1.65
terribly	2	1.65
slightly	1	0.82
relatively	1	0.82
immensely	1	0.82
highly	1	0.82
extremely	1	0.82
exceptionally	1	0.82
dead	1	0.82
bloody	1	0.82
awful	1	0.82
all	1	0.82
whole	1	0.82
wonderfully	1	0.82
Grand Total	**999**	**822.34**

Table 1. Frequency of intensifiers in LINDSEI-EST (on the left) and LOCNEC (on the right)

two data sets are similar in terms of total frequencies. The only difference that can be observed is the higher frequency of adjectives in attributive position in LINDSEI-EST. Thus, among the frequencies of the top adjectives in attributive position (*good, different, long*),

one of the interviewees used these adjectives in attributive position twice as many compared to other speakers. Overall, the results show a consistent trend towards the preferences for modification of predicative adjectives.

Forty-two different intensifiers and their frequencies are presented in alphabetical order in Table 1. LINDSEI-EST appears to have a higher normalised frequency of intensifiers with lower raw frequencies. This is the result of the fact that interviews in the data sets are not of equal size. Normalised frequencies cannot capture the influence of individual interviews on the overall frequency of intensifiers. Despite the differences in the size of the two data sets, we can still study the main differences between the two sets of interviews.

The distribution of the most frequent intensifiers varies from speaker to speaker. Thus, in LINDSEI-EST boosters, *very* and *really* are present in almost all of the interviews, and their frequency ranges between 12 and 1 instance. Half of the total 42 instances of *so* are found in four out of 25 interviews. A similar situation is observed with the downtoner *quite* as two-thirds of its total occurrences is divided between five interviews.

The most frequent intensifiers in the LOCNEC data *very* and *really* are found in all of the interviews, except that *very* is completely absent in three interviews. *So* is distributed unequally, and its frequency is greater than 5 in only five out of fifty interviews. *Quite* as a maximiser and booster is not present in a total of four interviews. Four of the speakers used

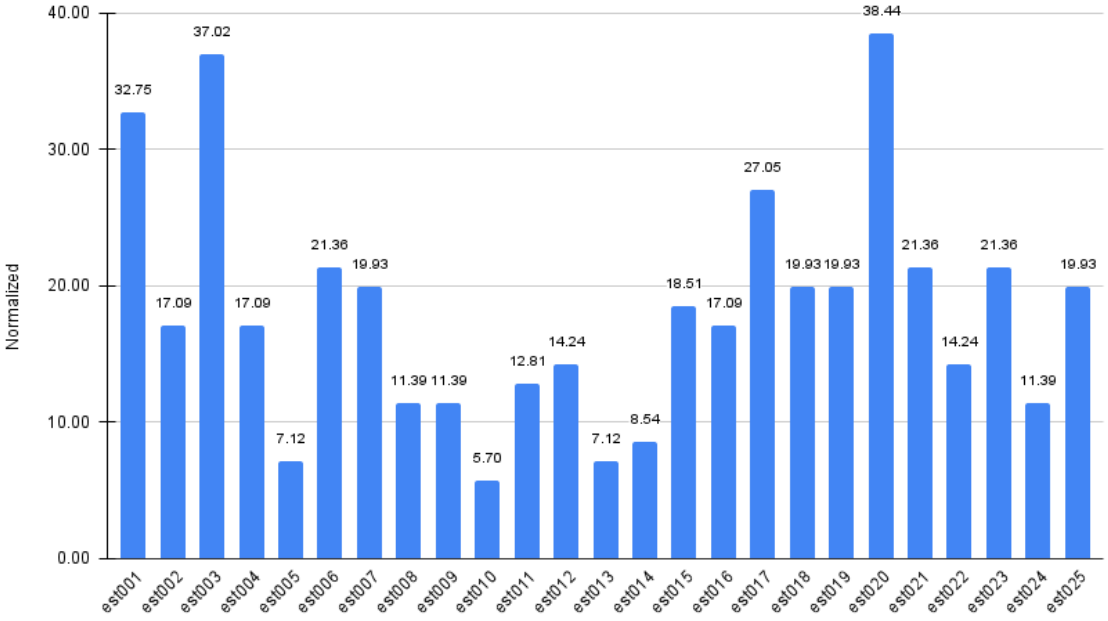


Figure 5. Distribution of intensifiers among speakers in LOCNEC

quite predominantly as a maximiser, and in these interviews frequency of this maximiser is far greater than its downtoner equivalent. The overall frequencies of intensifiers are not equally distributed in the data.

The total frequencies of intensifiers by speakers in LINDSEI-EST range between a maximum of 27 and 4 per speaker. There are only 3 speakers in whose interviews a total number of intensifiers exceed 20 instances. The frequencies of intensifiers by speakers can

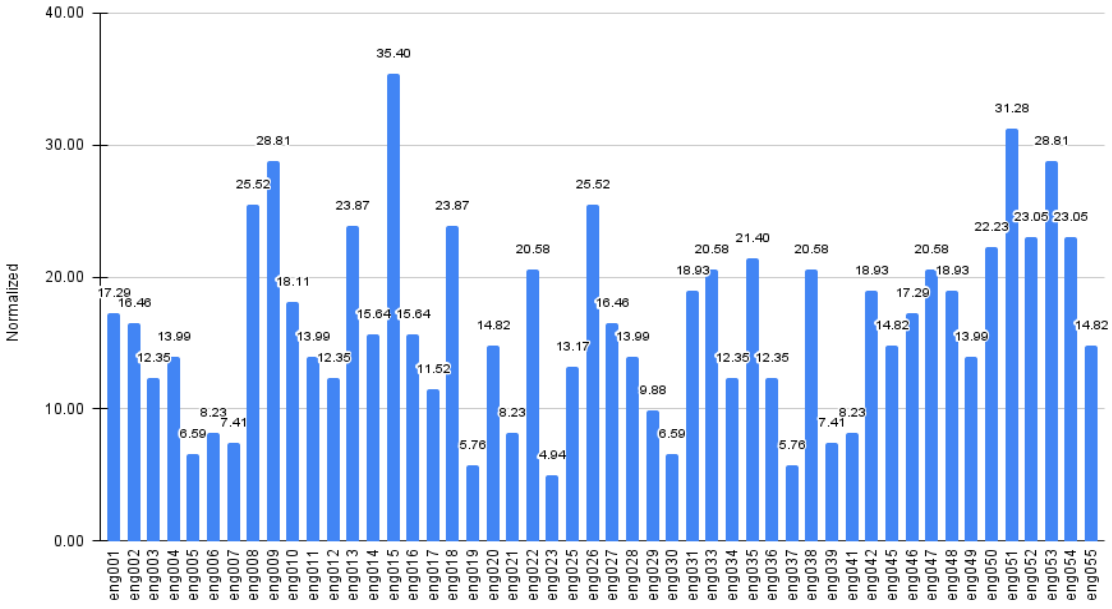


Figure 6. Distribution of intensifiers among speakers in LOCNEC

be observed in Figure 5. LOCNEC data shows a similar pattern with frequencies ranging between a maximum of 43 to a minimum of 6 instances per speaker. Figure 5 and Figure 6 show high variability in terms of the frequencies of use of intensifiers by speakers.

When it comes to the types of intensifiers LINDSEI-EST boosters are by far the most frequent amplifiers of 245 (697.67) tokens. In LINDSEI-EST, downtoners are the second intensifiers by the frequency of 57 (162.31) tokens. Maximisers are the smallest group of 14 (39.87) tokens in total. A similar picture is observed in LOCNEC. Boosters are the most frequent intensifiers as their frequency is 670 (551.52) tokens. Downtoners are the second by the frequency with 206 (169.57) tokens. Maximisers have the frequency of 123 (101.25) tokens in LOCNEC. The types of intensifiers vary according to the three tasks: set topic, free discussion, and picture description. The distribution of intensifiers according to the tasks is shown in Figure 7. The frequency of intensifiers in the figure varies from speaker to speaker.

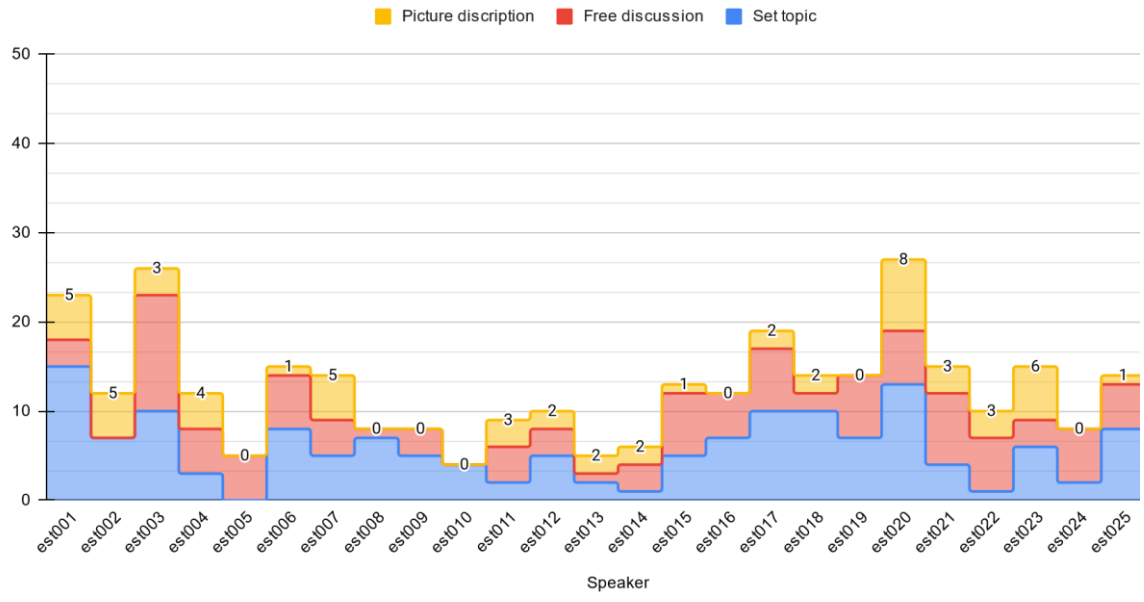


Figure 7. Frequency of types of intensifiers in LINDSEI-EST by tasks and speakers

(frequency of intensifiers in picture description task are on top of the bars)

Table 2. Normalised frequency of types of intensifiers in LINDSEI-EST by tasks

INTENSIFIERS	Set topic	Free discussion	Picture description
booster	347.41	233.51	116.75
downtoner	65.50	68.34	28.48
maximizer	5.70	14.24	19.93
Total	**418.60**	**316.09**	**165.16**

However in most of the interviews, the frequency decreases with each task from set topic to picture description. There are 147 (418.6) intensifiers found in set topic task. This is contrasted with 111 (316.09) instances in free discussion and 58 (165.16) in picture description tasks. However, the picture description task is also the shortest task. Frequencies of types of intensifiers by tasks in Table 2 show that while the frequency of boosters decreases sharply with each task, the number of downtoners remains similar in set topic and free discussion. On the other hand, the frequency of maximisers increases moderately from the first to the last task of the interviews.

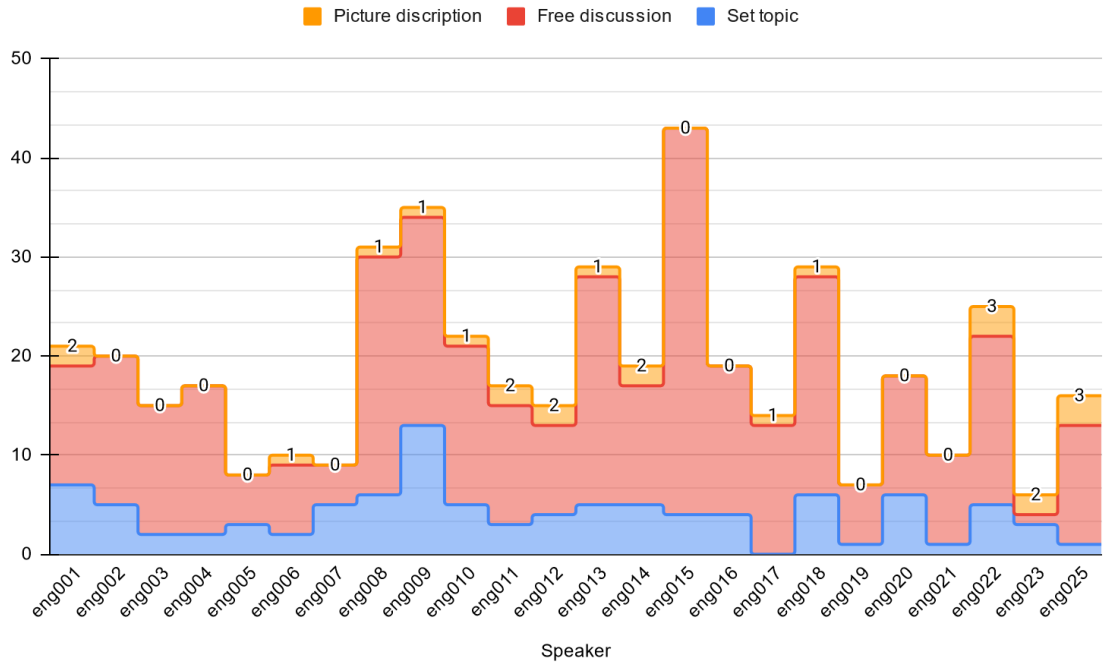


Figure 8. Frequency of types of intensifiers in LOCNEC by tasks and speakers (1-25)

(frequency of intensifiers in picture description task are on top of the bars)

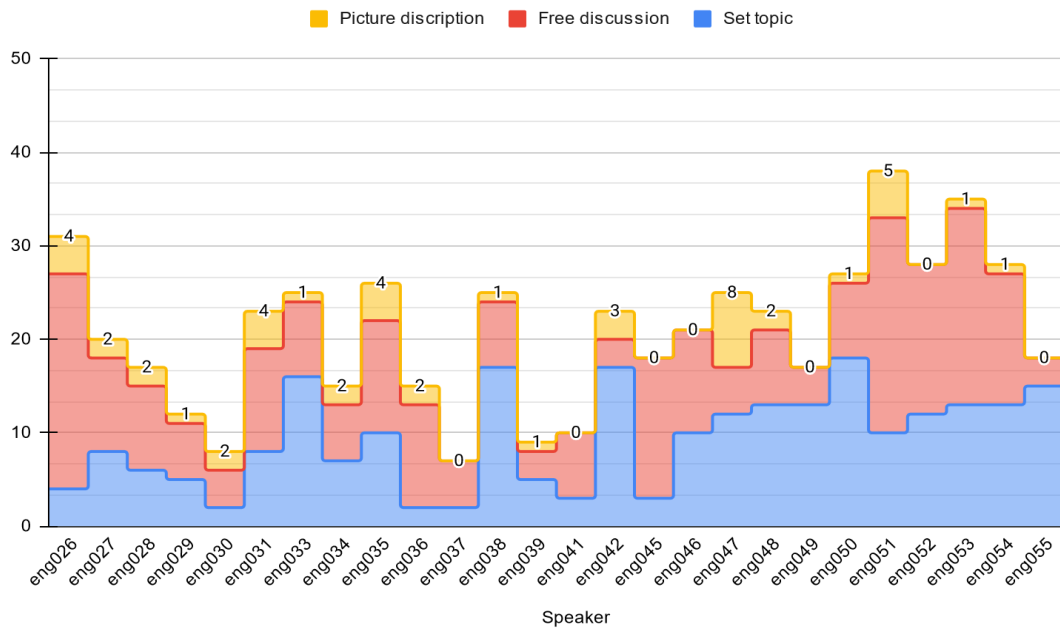


Figure 9. Frequency of types of intensifiers in LOCNEC by tasks and speakers (26-55)

(frequency of intensifiers in picture description task are on top of the bars)

The frequencies of intensifiers by tasks and speakers in LOCNEC can be seen in Figure 8 and 9. In this data set, the frequency of intensifiers in the free discussion task is noticeable higher than in set topic and picture description tasks. The distribution of intensifiers by speakers shows similarities. In most of the interviews, frequencies are increasing from set topic to free discussion. Noticeably, intensifiers are low in frequency in picture description task. There are only four interviews in which the frequency of intensifiers is greater than 3 instances. The frequencies of types of intensifiers presented in Table 3 show the correlation between the increase of intensifiers in general and the types. LOCNEC shows a fairly equal distribution of types intensifiers compared to LINDSEI, especially in terms of

Table 3. Normalised frequency of types of intensifiers in LOCNEC by tasks

INTENSIFIERS	Set topic	Free discussion	Picture description
booster	199.20	322.68	29.63
downtoner	55.15	97.96	16.46
maximizer	27.16	62.56	11.52
Grand Total	**281.52**	**483.20**	**57.62**

maximisers. However, in LOCNEC data set, frequencies of intensifiers in picture description task are twice as low compared to LINDSEI-EST data set.

2.4.2 Clusters of amplifiers in LINDSEI-EST and LOCNEC

In this section, I present the results of hierarchical cluster analysis applied to the sets of the most frequent amplifiers found in LOCNEC and LINDSEI-EST datasets. The most frequent amplifiers were chosen on the bases of occurrences with adjectives. Thus, those intensifiers that cooccurred with less than two adjectives were disregarded. This approach allowed for more representative results. The subset is presented in Appendix 2. The analysis applied to combinations of intensifiers and adjectives and intensifiers plus types of adjectives. I discuss the main differences between the dendograms applied to the datasets. Examples of adjectives and types of adjectives are provided to illustrate what items contribute to the clustering of intensifiers.

The first question regarding the differences in terms of the use of intensifiers is whether intensifiers in the dataset form distinct groups. For this purpose, the HCA has been employed. Adjectives and their types were chosen as the basis for clustering. In order to check how clustering based on the classification proposed by Paradis (1997) is comparable to the clustering based on adjectives modified by intensifiers, two groups of hierarchical cluster analyses were performed. The results of the analysis are presented in this subchapter.

According to Levshina (2017) one of the measures for the dendogram produced based on the results of hierarchical cluster analysis is the values of multiscale bootstrap resampling. This validation method allows for identifying how stable to the clusters in the dendogram are. However, the problem with the bootstrap sampling is that is requires the presence of a minimum of four different categories. In case of the data used in this thesis, types of adjectives consist of three categories: scalar, extreme and limit adjectives. This fact

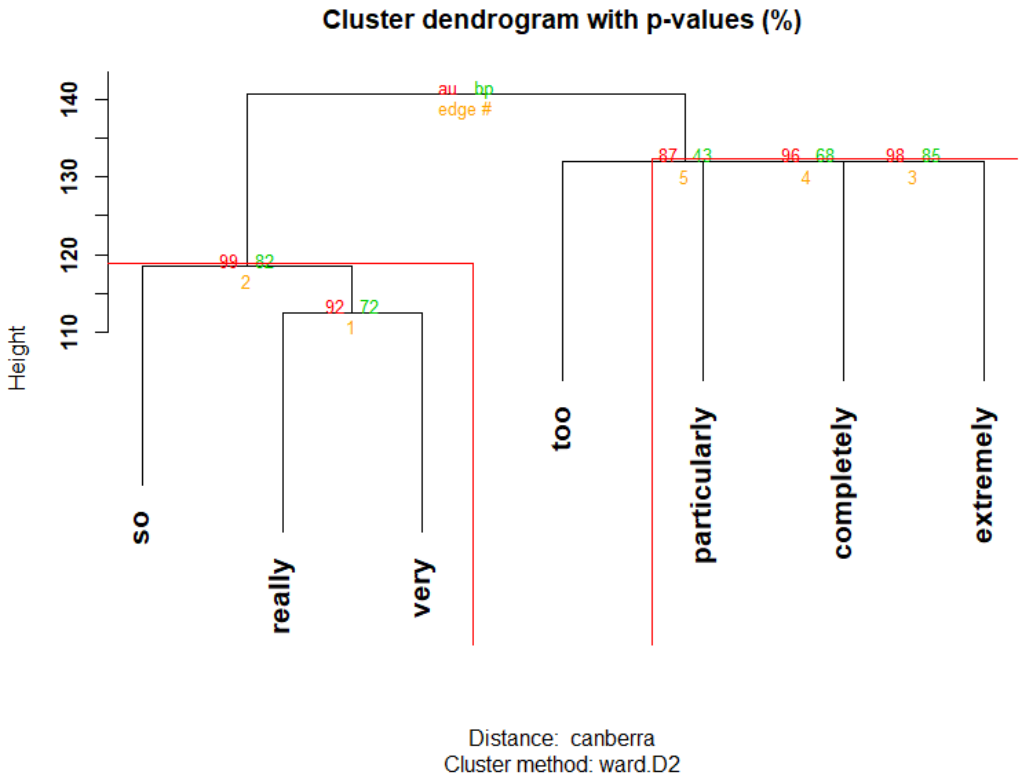


Figure 10. Hierarchical clustering with multiscale bootstrap values for LINDSEI-EST

does not permit validation of the clusters in the dendogram based on the types of adjectives. However, the assumption is that if the dendogram for adjectives is similar to the one for types, then it is possible to conclude that the data in principal should support the clusters in both of the trees.

The visualisation for clustering for the LINDSEI-EST data based on adjectives is presented in Figure 1. The clustering identified two main branches. The first contains intensifiers *very*, *really* and *so*. The second cluster consists of intensifiers *too*, *particularly*, *completely* and *extremely*. The clustering is not particularly intuitive taking into consideration the assumption that *completely* is a maximizer and thus belongs to a separate group. Hoewever, after examining the adjectives that contribute to the clustering of intensifiers into two branches, a trend can be identified. The intensifiers *really*, *so* and *very* share the fact that they are modifying a larger number of adjectives such as *good*, *nice*, *interesting*, *happy* and *cool*. Consequently, these adjectives are among the most frequent

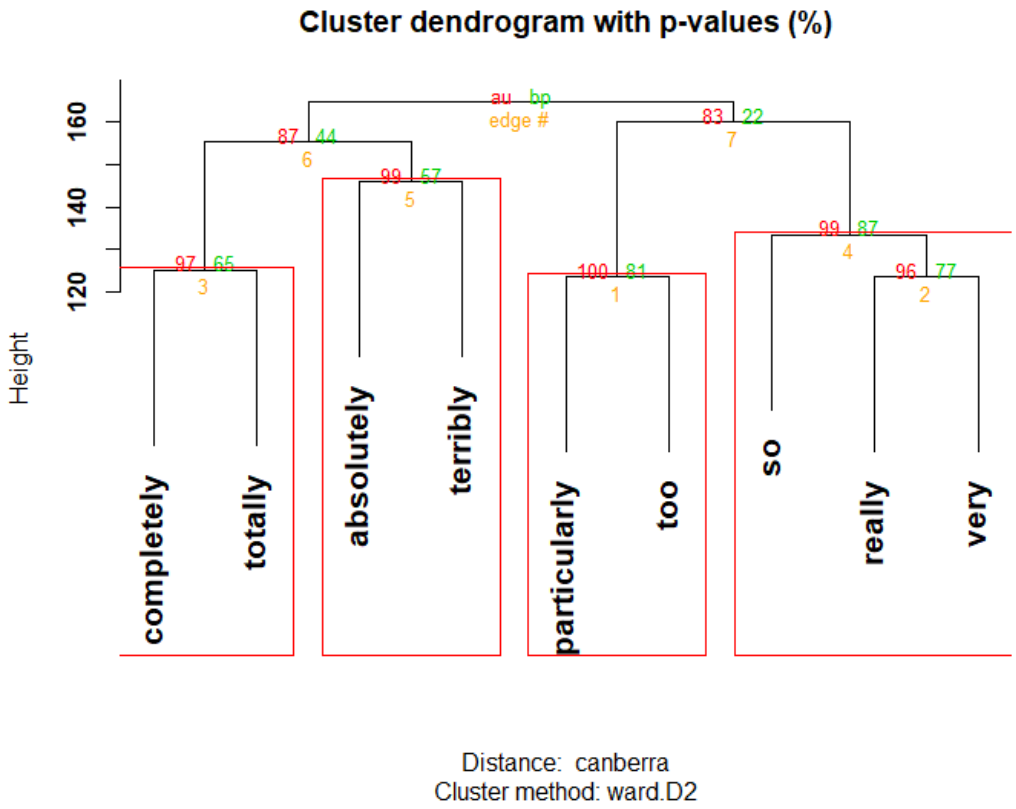


Figure 11. Hierarchical clustering with multiscale bootstrap values for LOCNEC

adjectives in the data. This most likely contributes to the clustering of intensifiers. Overall, intensifiers *so*, *really* and *very* modify the largest number of adjectives in the data, which resulted in the fact that they are grouped together. The results of the bootstrap sampling reaveal that the data supports two branches that are *very*, *really* and *so* as a distinct cluster and *particularly*, *extremely* and *completely*.

The reults of hierarchical clutering analysis for the LOCNEC data are visualised in Figure 11. Similarly to LOCNEC-EST dendogram, the analysis intedtified two main branches. However, in case of LOCNEC data these branches differ in terms of types of intensifiers. Maximizer *completely*, *totally*, *absolutely* form one cluster. This cluster is contrasted to the cluster that contains boosters *really*, *very*, *so*, *too* and *particularly*. The division is not as homogenious as the booster *terribly* is grouped together with maxinizers. This can be explained by the fact that booster *terribly* only modifies two adjectives *modern* and t*ragic*.These adjectives are only found collocating with the intensifier *terribly*. The

clustering of intensifiers in the LOCNEC data is also influenced to a large extent by the number of adjectives found in each branch. Thus, the cluster containing boosters has a larger proportion of frequent adjectives *good, nice, interesting, different, beautiful, big, happy* and *hot*. The branch of maximizers is divided in smaller clusters of *completely* and *totally,* and *absolutely* and *terribly*.

The cluster of *completely* and *totally* is charachterised by such adjectives as *foreign, new, false, gorgeous, horrible* and *wild*. Among these adjectives *false* only collocates with *totally* . The adjective *different* is modifed by both of the maximers and given that adjective is found 15 times in the dataset, it sets this cluster appart from the *absolutely* and *terribly* as this adjective does not collocate with these intensifiers. The rest of the adjectives charachteristic for this cluster are modified by *completely*.

The cluster of *absolutely* and *terribly* is similar to the previous cluster of maximizers as it shares a small proportion of high frequent adjectives collocating with the boosters. Characteristic adjectives for this group are *amazing, brilliant, fantastic, flabbergasted, modern, sure*. Among these adjectives, only *modern* is found with *terribly*. This adjective is found only once in the dataset. The same applies to the adjective *tragic* collocating with *terribly*. Thus, it is difficult to conclude that *absolutely* and *terribly* indeed form a cluster in terms of their collocational preferences. The one distinct feature of the cluster containing maximizers and the booster *terribly* is the higher proportion of negative value concerning the characteristic adjectives of the cluster of boosters. These negative values are based on the total number of adjectives in the dataset. The comparison of distinct features of the clusters point to the fact that mazimizers and boosters exhibit different collocational patterns in the LOCNEC dataset compared to the LINDSEI-EST.

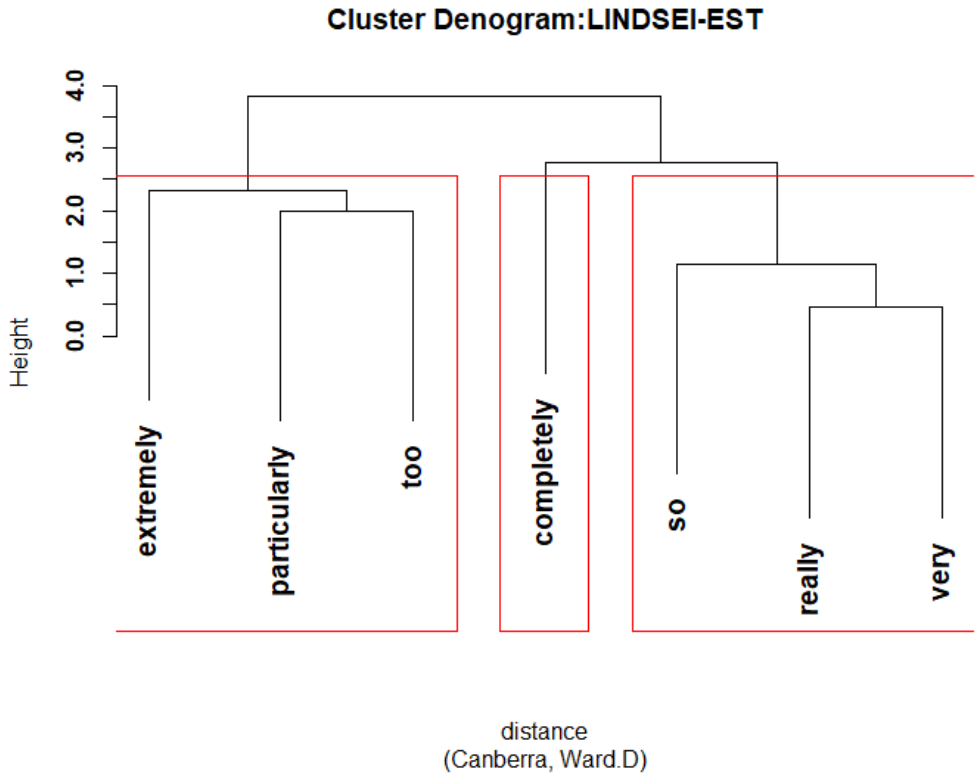
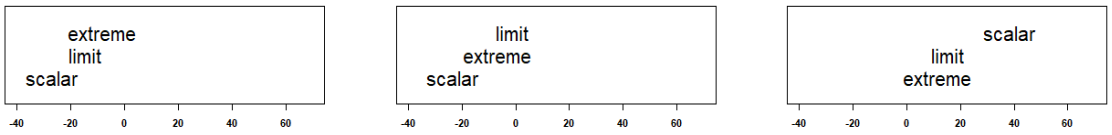


Figure 12. Hierarchical clustering dendrogram of LINDSEI-EST based on adjective types

The obvious question arises whether these collocational patterns correlate with the reults of hierarchical cluster analysis when classification of adjectives is based on their types. The results of the analysis for LINDSEI-EST dataset are presented in Figure 12. The dendogram differs from the one presented in Figure 10 as intensifiers are grouped in three clusters. The main difference is the maximizer *completely* forms a separate cluster. The reason that I decided to highlight three clusters is that the computation of average silhouette widths for this dendrogram has the largest number (0.33) for the three clusters. According to Levshina (2017: 312) this measure serves as the basis for motivating the number of clusters for a particular dendrogram. The widths for two and four clusters is relatively smaller compared to the three cluster solution and equals 0.28 and 0.29 respectively. Additionally, the three clusters seem more intuitive. Two clusters solution woud group *completely* in one group with *very*, *really* and *so*, which would disregard the fact that *completely* is a maximizer and it is more likely to form its own group.



Extremely -Particularly-Too Completely Really-Very-So

Figure 13. Snake plot of the differences of 3 clusters in LINDSEI-EST

The differences of the three clusters identified in the LINDSEI-EST are visualised using the so-called snake plots (Levshina 2017:314) that allow for examining the differences of each cluster compared to all the other clusters. Thus, the cluster of *completely*, that can be seen in the midle in Figure 4, points out that this cluster has relatively higher proportion of limit adjectives. In fact, after examining the adjectives modified by the maximizer, it apparent that all the adjectives are clasified as limit. These adjectives are *different*, *shifted*, *normal* and *isolated*. Among these, only the adjective *different* is found more than once in the data. Moreover, the total number of occurances of *different* in the dataset is 12, five of which are found with the maximizer *completely*. The other instances of this adjective are found collocating with boosters *so*, *very* and *really*.This explains why on one hand *completely* forms a cluster of its own, and on the other hand why it is found in the same branch with boosters *so*, *really* and *very* on the other hand.

The types that contribute to clustering intensifiers *particularly*, *too* and *extremely* together can be seen on the left of Figure 13. It is charachterised by a higher proportion of extreme adjectives. The extreme adjective *impressive*, that is found 8 times in the data, is the only extreme adjective in this cluster. *Impressive* is modified by the booster *particularly* and coocures with the booster 3 times. Limit adjective *critical* coocures with the booster *too* and is only found once in the dataset. The rest of the adjectives collocating with boosters in this cluster are scalar adjectives. The most frequent among the adjectives found in this cluster is *good*, which is one of the most frequent adjectives in the whole LINDSEI-EST dataset.

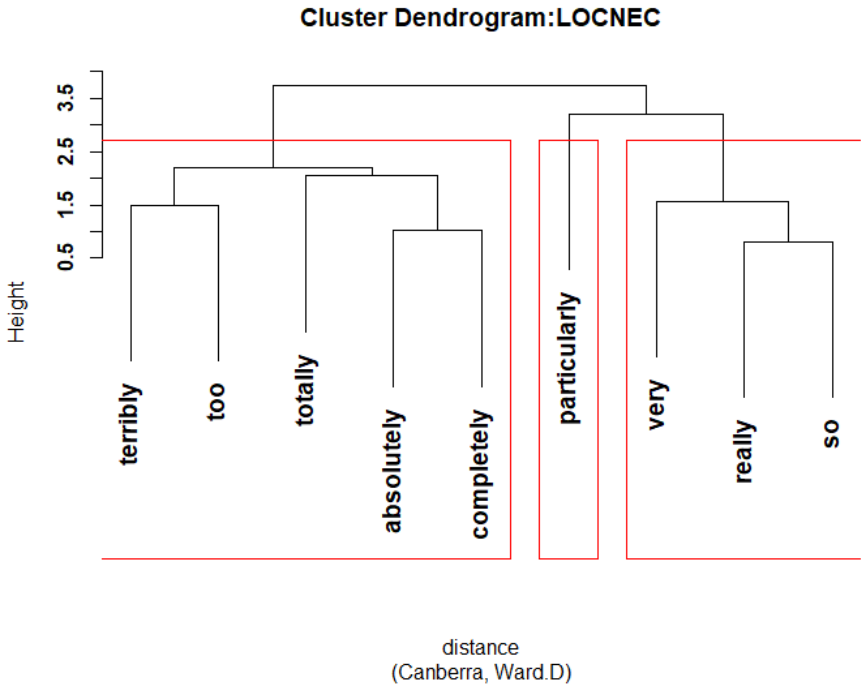


Figure 14. Hierarchical clustering dendrogram of LOCNEC based on adjective types

The results of hierarchical cluster analysis applied to LOCNEC dataset are presented in Figure 14. The dendogram is similar to the dendogram in Figure 12, where the clustering was produced for the English L1 speakers' data is based on adjectives. One of the few differences is that in Figure 14, booster *particularly* now forms a separate cluster within the branch of boosters *so*, *really* and *very*. The other difference between the dendograms is the clustering of booster *too* in the branch containing maximizers *absolutely*, *completely* and *totally* and the booster *terribly*. However, the clustering groups *too* and *terribly* together, which is rather intuitive as these two intensifiers are boosters. Additionally, dividing the denogram into three groups is supported by calculating the average silhouette widths. The output assigns the largest average widths, 0.29, to the three cluster solution. This number is contrasted with two and four cluster options, which are 0.245 and 0.248, respectively. The four clusters option would reveal such separate clusters as *terribly* and *too*; *totally*, *absolutely* and *completely*; *particlularly*; *so*, *very* and *really*. Considering the size of the data, this option will not reveal any valuable information.

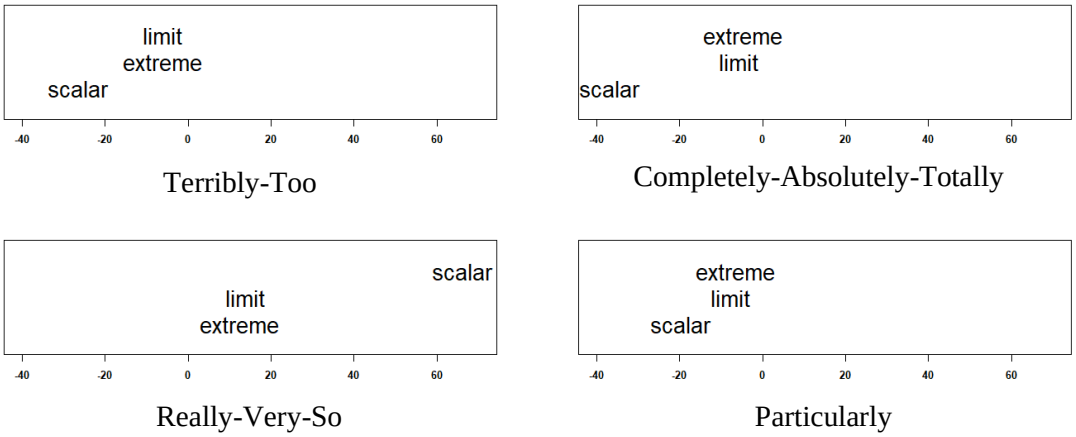


Figure 15. Snake plot of the differences of 3 clusters in LOCNEC

The fact that the booster *particularly* forms a separate cluster can be explained by examining the types of adjectives it collocates in the data. The booster particularly is found exclusively modifying adjectives of the scalar type. This contributes to the clustering of this booster into the same branch as boosters *so*, *really* and *very*. However, since it colocates with scalar adjectives only, it was grouped into a node of its own. The booster *so*, *really* and *very* are characterised by modification of adjectives of all three types. However, the scalar type adjectives constitute the biggest number of adjectives colocating with these boosters. This can be observed in Figure 6 on the bottom left and right.

The snake plot reveals that booster *particularly*, has a smaller negative proportion of scalar adjectives compared to maximizers *completely*, *absolutely* and *totally* on the one hand, and boosters *too* and *terribly* on the other. The most frequent adjective collocating with *particularly* is *good*. This adjective is one of the most frequent adjectives in the dataset and found to be predominantly modified by *very* and *really*. The other adjective modified by *particularly* is *nice*, which is frequent in the data. The fact that *good* and *nice* are modified by *very* and *really* is characteristic for this branch and is the reason why *particularly* is grouped with these frequent boosters.

The boosters *terribly* is found modifying extreme adjective *tragic* and limit adjective *modern* . *Too* is modifying the limit adjectives *realistic*, *worthwhile* and *bored*. Among these adjectives, only *realistic* is found in the data, and it is modified by the booster *so*. It is possible to conclude that these limit adjectives are not distinct characteristics of this cluster. However, taking into account the fact that limit and extreme are overall not as frequent in the data compared to scalar adjectives, it is not surprising that these boosters were clustered within the branch containing maximizers.

The cluster containing maximizers *totally, completely* and *absolutely* is distinct compared to all the other clusters by the overall high proportion of limit and extreme adjectives. The intensifiers of this cluster are not found modifying scalar adjectives. The snake plot in Figure 6 shows that extreme type is ranked higher in this cluster compared to all the other clusters. Despite the cluster containing an almost equal amount of extreme and limit adjectives, the fact that extreme adjective *amazing* occurs 5 times in the dataset and 3 of them with the intensifier *absolutely* renders the extreme type as a distinctive feature of the cluster.

The results of hierarchical cluster analysis for the two datasets reveal certain tendencies in the data. The analysis identified two major clusters containing boosters *so, really* and *very* in LINDSEI-EST and LOCNEC. The distinct characteristic of these boosters is that they modify the most frequent scalar adjectives. However, they are also found colocating with limit and scalar adjectives. This suggests that they have a larger collocational profile. These intensifiers are not only restricted to particular types of adjectives. The LOCNEC data shows a difference between the use of intensifiers by English L1 and Estonian L1 speakers. English L1 use more maximizers compared to Estonian L1 and these maximizers form a separate cluster due to the fact that they exclusively cooccur with limit and extreme adjectives. That is also true for the the maximizer *compeltely* in LINDSEI-EST,

which found modifying limit adjective *different* and other infrequent limit adjectives such as *shifted, normal and isolated*. This fact indicates that in both LINDSEI-EST and LOCNEC datasets, boosters and maximizers have different collocational profiles in terms of adjectives they modify and the types of these adjectives.

The HCA allowed to identify the main tendencies in both of the datasets. It is time to turn to the result of multiple correspondence analysis to examine which of the features of LINDSEI-EST and LOCNEC correlate with each of the two groups of speakers.

2.4.3 The result of Multiple correspondence analysis

In this section, the results of MCA are presented and discussed. For the purpose of identifying which intensifies and features correlate with the two groups of speakers in the data used in this thesis, MCA was conducted. The dataset submitted to MCA was a subset containing the most frequent boosters and maximizers found in LONCEC and LINDSEI-EST. In order to identify similarities and differences between the use of intensifiers by English L1 and Estonian L1 speakers, intensifiers were tagged depending on in which dataset they were found. The tags are “eng” and “est” attached to each intensifier item. After this procedure, the datasets were merged to create a subset containing all instances of intensification from LOCNEC and LINDSEI-EST. The subset contains a total of 578 observations: 333 from LOCNEC and 245 from LINDSEI-EST.

The results of MCA for the LINDSEI-EST-LOCNEC subset are visualised in the biplot presented in Figure 7. The results show the visualisation based on the first two dimensions.

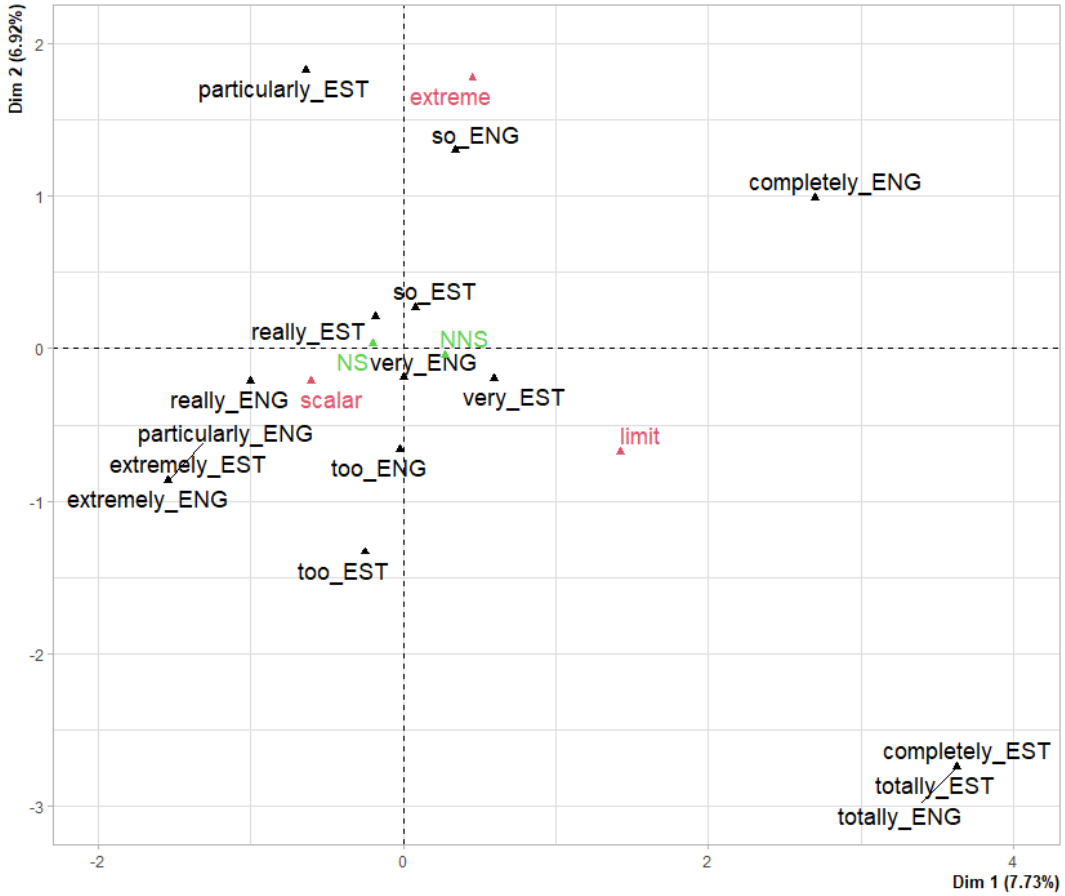


Figure 16. MCA biplot of simultenious representation of intensifers, types of adjectives and speakers in joint LOCNEC-LINDSEI-EST data subset

The biplot shows that majority of the boosters are attracted to the scalar adjectives. These boosters include *really*, *extremely*, *too*, for both groups of speakers. The booster *particularly* in the English L1 data correlates with scalar adjectives, and it is contrasted with the same booster in the Estonian L1 data, where a higher proportion of extreme adjectives are modified by *particularly*. The booster *so* cooccurs with a larger proportion of extreme adjectives and this is more prominent in the LOCNEC data. Intensifier *very* shows that the proportion of adjectives of limit type are modified by this booster in LINDSEI-EST data compared to the LOCNEC dataset. The maximizer *totally* behavious similarly in both datasets and is correlating with adjectives of limit type. The maximizer *completely* contrasts in terms of the proportions of extreme adjectives. As it was shown in the previous subchapter, *completely* exclusively cooccurs with the limit adjectives in the LINDSEI-EST data. This observation is reflected in Figure 7 as *completely* from the LINDSEI-EST data is plotted in the bottom right part of the biplot next to the limit type. *Completely* from the

English L1 data is plotted closer to extreme type. However, there is no high correlation between this maximizer and the extreme type as it also modifies limit adjectives in the LOCNEC data. The fact that extreme type is plotted further up in the plot reflects the lower proportion of extreme adjectives compared to the other types in this subset.

The rationale for choosing two dimensions is the percentages of the total variance explained by each of the dimensions. The output of adjusted percentages of the total variance shows that the first two dimensions explain 72% of the data. The subsequent dimensions do not add up much to the biplot of Figure 7.

```
summary(mjca(df.mca))

##
## Principal inertias (eigenvalues):
##
##  dim    value      %   cum%   scree plot
##  1      0.263523  64.5  64.5  ********************
##  2      0.032115   7.9  72.4  ***
##  3      0.015098   3.7  76.1  *
##  4      00000000   0.0  76.1
```

The contribution of the variables to the formation of the first dimension provides information regarding the interpretation of the results. The output is shown below.

```
      ## $`Dim 1`
## $quali
##                       R2       p.value
## Type        0.69569647 2.823318e-149
## Intensifier 0.69569647 2.788443e-133
## Speaker     0.05328175  1.983216e-08
##
## $category
##                                 Estimate       p.value
## Type=limit                     0.83048146 2.223488e-103
## Intensifier=completely_EST     2.54594300  2.538208e-27
## Intensifier=completely_ENG     1.77478876  3.292115e-15
## Intensifier=very_EST           0.01736399  1.860993e-10
## Speaker=NNS                    0.19480130  1.983216e-08
## Intensifier=totally_ENG        2.54594300  2.188769e-07
## Intensifier=absolutely_ENG     1.07707778  5.682027e-07
## Type=extreme                   0.02561659  2.058235e-06
## Intensifier=totally_EST        2.54594300  2.726632e-04
## Intensifier=so_ENG            -0.19338411  3.824416e-02
```

```
## Intensifier=extremely_EST    -1.76322904  2.888707e-02
## Intensifier=particularly_ENG -1.76322904  3.666864e-05
## Speaker=NS                   -0.19480130  1.983216e-08
## Intensifier=really_ENG       -1.31105484  3.791418e-34
## Type=scalar                  -0.85609805 3.796856e-120
```

The first dimension is formed by contrasting limit and scalar adjective types. The output confirms that in terms of the first dimension *very* in LINDSEI-EST cooccurs with larger proportions of limit adjectives. However, the estimate number is quite low for the *very,* which means that the effect of this correlation is not significant. Similarly, the effect of correlation of *so* found in the LOCNEC data is minor. The only significant correlation in terms of the first dimension is between the booster *really* and the scalar type of adjectives. However, given that the first dimension has high explanatory power, accounting for 64.5% of the data variance, it is possible to confirm the tendencies discussed in the previous subchapter. It is clear that there is a division between boosters and maximizers in terms of types of adjectives. Thus, there is an overall preference for maximizers to cooccur with limit adjectives. Most of the boosters, which are found on the left in Figure 16, preferably cooccur with scalar adjectives. Additionally, there is a relatively small difference between Englis L1 and Estonian L1 speakers as the proportion of limit adjectives in LINDSEI-EST is slightly higher.

The maximizer *absolutely*, which is not plotted for the reasons of space as it would make it hard to read the biplot, correlates with extreme adjectives. However, since this adjective is not found in the LINDSEI-EST data, it is not possible to examine the differences between English L1 and Estonian L1 speakers. The dataset shows that *absolutely* tends to modify extreme adjectives in LOCNEC as there are only two instances of it cooccurring with limit adjectives.

The second dimension explains 8% out of a total 72% of the variance in the data. Even though this number is low, it provides valuable insides into the patterns in the data. The output of the contribution of variables for the second dimension are shown below.

```
## $`Dim 2`
## $quali
##                     R2       p.value
## Type        0.6231834 1.372104e-122
## Intensifier 0.6231834 1.410662e-107
##
## $category
##                                  Estimate       p.value
## Type=extreme                   1.16799700 3.090146e-113
## Intensifier=absolutely_ENG     3.65171471  3.952768e-35
## Intensifier=so_ENG             1.23071737  4.591588e-16
## Intensifier=particularly_EST   1.64852780  2.373655e-08
## Intensifier=completely_ENG     0.98746816  4.835512e-03
## Intensifier=very_ENG           0.05689972  5.659333e-03
## Intensifier=really_ENG         0.03736327  1.349257e-02
## Intensifier=very_EST           0.05279146  3.794327e-02
## Intensifier=really_EST         0.36994654  4.586170e-02
## Intensifier=particularly_ENG -0.47485368  2.138507e-02
## Intensifier=too_ENG          -0.31460124  2.058610e-02
## Intensifier=totally_EST      -1.95722541  6.003980e-03
## Intensifier=too_EST          -0.84544661  1.363558e-04
## Intensifier=totally_ENG      -1.95722541  9.541595e-05
## Type=scalar                  -0.40139493  1.536324e-11
## Intensifier=completely_EST   -1.95722541  1.132339e-15
## Type=limit                   -0.76660207  1.350766e-18
```

The second dimension contrasts such types of adjectives as limit and scalar with extreme adjectives. In this dimension, the effect of a small number of extreme adjectives cooccurring with the maximizer *completely* in the LONCEC contributes to plotting it in the same quadrant with the extreme type. The same maximizer in the LINDSEI-EST data is found correlating with limit type. These results confirm the observations made when examining the results of hierarchical cluster analysis. The difference between these maximizers is that *completely* modifies a small number of extreme adjectives in LOCNEC, while in LINDSEI-EST it is found with limit adjectives only.

Another observation that can be made by examining the second dimension is that the intensifier *so* in the LOCNEC data is has a stronger correlation with adjectives of the extreme

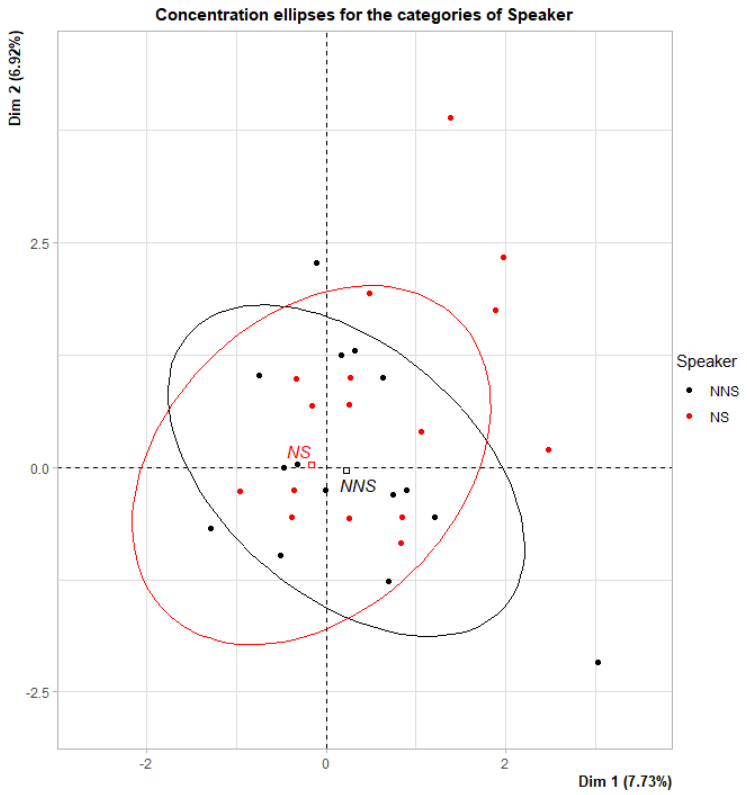


Figure 17. Confidence ellipses for the categories of speakers in the joint LOCNEC-LINDSEI-EST data subset

type. The fact that the second dimension has an overall lower explanatory power than the first dimension makes it possible to interpret this result as pointing to the differences in proportions of extreme adjectives compared to limit adjectives. The data also supports this, as *so* cooccurs with 11 extreme adjectives and 7 limit adjectives in LOCNEC. In the LINDSEI-EST, the distribution of extreme and limit adjectives that are found with the booster *so* is more balanced: 8 extreme and 9 limit adjectives.

The category of speakers is plotted in green colour in Figure 17 and shows that there are no significant differences as they are situated close to the zero points in the middle of the plot. They are plotted slightly apart in terms of the horizontal axes, meaning that there are differences in terms of proportion of scalar and limit adjectives. Figure 8, represents confidence ellipsis plotted around the examples of intensifiers and adjectives types that belong to two groups. The figure confirms that there is a substantial overlap between English L1 and Estonian L1 groups.

Finally, I present the results of validation of the exploratory power of the two dimensions examined previously. For this, logistic linear regression was used. The dimensions were applied as explanatory variables to the response variables English L1 and Estonian L1 speakers, which were coded as NS and NNS. The output of the logistic linear regression is presented below.

```
m <- lrm(df.data$Speaker ~ dim1 + dim2)
m

## Logistic Regression Model
##
##  lrm(formula = df.data$Speaker ~ dim1 + dim2)
##                        Model Likelihood      Discrimination    Rank Discrim.
##                              Ratio Test             Indexes          Indexes
##  Obs           578    LR chi2      32.38      R2       0.073    C       0.609
##   NNS          245    d.f.             2      R2(2,578)0.051    Dxy     0.217
##   NS           333    Pr(> chi2) <0.0001    R2(2,423.5)0.069    gamma   0.217
##  max |deriv| 9e-13                           Brier    0.231    tau-a   0.106
##
##            Coef    S.E.   Wald Z Pr(>|Z|)
##  Intercept  0.3107 0.0866  3.59  0.0003
##  dim1      -0.5955 0.1123 -5.30  <0.0001
##  dim2       0.1156 0.1118  1.03  0.3010
```

The results show that the dimensions do not have substantial explanatory power. It provides grounds for concluding that there are no significant differences between English L1 and Estonian L1 speakers in the use of intensifiers.

The result presented in previous subchapters and their implications for the research questions and future research will be discussed in-depth in the next subchapter.

2.5. Discussion

In this thesis, I aimed to answer the following research questions:

1. What is the distribution of the intensifiers in two data sets?

2. What are the differences and similarities between two groups of speakers (English L1 and Estonian L1) in terms of the use of intensifiers?

The distribution of intensifiers shows that both in LINDSEI-EST and LOCNEC the most frequent intensifiers are *very*, *really* and *so*. These intensifiers collocate with by far the largest numbers of adjectives and can be regarded as the primary items used for intensification by both English L1 and Estonian L1 speakers.

The results presented in the previous subchapter show that the overall frequencies of intensifiers in both data sets are similar when the frequencies are normalised. There are differences in terms of types of intensifers used by English L1 and Estonian L1 speaker. Thus, LOCNEC data shows a greater number of different intensifiers. Additionally, the distribution of intensifiers differs in terms of maximisers that are much less frequent in LINDSEI-EST than LOCNEC. The differences are also observed in terms of the use of intensifiers in different tasks. Similarly to the findings reported in Pérez-Paredes (2010) and Perez-Parades and Camino Bueno-Alastuey (2019), learners of English and native speakers had different approaches to picture description task. Thus, in LOCNEC there are overall less intensifiers in the picture description task when compared to LINDSEI-EST. The frequency of intensifiers in the set topic task also varies according to the data sets as there are overall less instances of intensification in this task in LOCNEC compared to LINDSEI-EST.

The main differences between the use of intensifiers in LINDSEI-EST and LOCNEC are the frequencies of top intensifiers. The most frequent intensifiers in LINDSEI-EST are

very, really, so and downtoner *quite*. These intensifiers are also the most frequent items in the LOCNEC dataset. This reflects the tendencies reported by Xiao and Tao (2007) that *very*, *really* and *so* tend to be associated with the spoken register. However, the difference between English L1 and Estonian L1 speakers are in terms of the use of maximizers. In the LOCNEC data, the maximizers *totally* is more frequent comapared to the LINDSEI-EST data. The maximizer *absolutely* is absent in the LINDSEI-EST data. This observation contributes to the fact that that these two datasets differ in terms of number of types of intensifiers. There are 28 diiferent intensifiers in LINDSEI-EST and 35 in LOCNEC.

Taking into account the fact that previous research on the use of intensifiers in spoken register reported that boosters *very, really* and *so* (Tagliamonte 2003, Tagliamonte and Roberts 2005 ,Tagliamonte 2008) are particularly frequent in this register, it is possible to assume that the findings of this study reflect the tendencies of spoken language more generally. Additionally, the fact that *absolutely* is found in LOCNEC among the top 10 frequent intensifiers reflects the fact that it is associated with the spoken register in the British variety of English reported by Xiao and Tao (2007). The speakers in the LOCNEC dataset are native speaker of the British variety. Estonian L1 speakers on the other hand show a preference for the use of maximizer *completely*. Thus, the differences based on the raw and normalised frequencies show the collocational preferences between the two groups of speakers: English L1 and Estonian L1.

Contrary to the study conducted by Pérez-Paredes (2010), there was no significant underused of intensifiers by Estonian L1 speakers. The learner of English with Estonian L1 background used more intensifiers of the type booster compared to the English L1 speakers. However, the analysis shows similar results as in Perez-Parades and Camino Bueno-Alastuey (2019), where the authors reported the differences in frequencies of intensifiers in the picture description task. Thus, the overuse of intensifiers by the Estonian L1 speakers in

picture description task is driven by the fact that they showed greater involvement in this part of the interview compared to English L1 speakers. This fact suggests that the tasks themselves influence the use of intensifiers and this should be considered in the future studies on the use intensifiers based on the LINDSEI and LOCNEC data.

The question is whether the differences in terms of the overall frequencies are suggestive of significant over- or underuse reported in the previous studies (Granger 1998 , Lorenz 1999). Schweinberger (2020: 165) brings attention to the fact pointed out by Gries (2018: 303-304) that focusing on the frequencies alone disregards the complexity of the data and possible influence of multiple factors that influence the results. As a conclusion, Schweinberger (2020) and Gries (2018) suggest using methods that allow to validate significance of the frequencies according to multiple factors.

In this thesis, I made an attempt to examine the use of intensifiers using two types of methods: hierarchical cluster analysis (HCA) and multiple correspondence analysis (MCA) (subchapter 2.3.1 and 2.3.2 respectively). Even though, the methods are exploratory and are limited to the number of factors that can be taken into account, they provide valuable insights into patterns observed in the data used in this thesis. Thus, incorporating such factors as speaker (English L1, Estonian L1) and types of adjectives according to the classification proposed by Paradis (1997), allowed for identification of particular collocational preferences in terms of types of adjectives based on a set of the most frequent boosters and maximizers in LINDSEI-EST and LOCNEC.

Application of the HCA allowed identifying groups that intensifiers form based on the types of adjectives they modify. The main clusters identified in both LINDSEI-EST and LOCNEC are frequent boosters *very*, *really* and *so*, maximizers and less frequent boosters *too*, *particularly*, *terribly* and *extremely*. This allows to conclude that there is observable

difference between the intensifiers that belong to the class of amplifiers. The maximizers preferably collocate with extreme and limit types of adjectives. This fact reflects the findings of Paradis (1998) that indicate the role of semantics of adjectives when it comes to the use of intensifiers and this is especially true for the class of maximizers. However, the data used in this thesis show that the main difference between boosters and maximizers concerns the proportion of the types. Thus, the proportion of extreme and limit adjectives contributes to the identification of the clusters of maximizers. The boosters have a higher proportion of scalar adjectives. However, the data show that the most frequent boosters *very, really* and *so* coocure with all three types of adjectives, even though the proportion of extreme and limit adjectives is much lower compared to the adjectives of the scalar type.

The function of boosters is not limited to the modification of scalar adjectives. This observation reflects the point made by Paradis (1998: 59) that contextual modulation contributes to the fact that boosters are not limited to the modification of scalar adjectives only. I would argue that considering the results presented in this thesis, contextual modulation plays a very important role in contributing to the distribution of intensifiers of the booster category. It is highly unlikely that speakers consider such factors as scalar, limit or extreme types of adjectives when they choose to use particular intensifier in combination with adjectives. Considering, that the limit adjective *different* cooccurs with *very* 12 times in LOCNEC is suggestive that it is rather an example of an existing prefabricated patterns of collocation. The same observation is made in LINDSEI-EST data, where this adjective cooccurs with the maximizer *completely* 5 times and 4 times with the booster *so*. Additionally, despite the fact that this adjective belongs to the type of limit adjectives, it can receive scalar reading, which is why it is modified by boosters.

A similar situation can be observed in terms of the use of intensifiers with the adjective *amazing*. Thus, in LOCNEC it is found 5 times, 3 of which with the maximizer

absolutely, and 2 times with boosters *really* and *so*. The examples allows to assume that there is a tendency of *amazing* to coocure with *absolutely*, however any conclusion beyond this are not possible given the limited size of the data, especially since all three of the examples are found in one interview.

The observation on the bases of the distribution of intensifiers based on the types of adjectives modified, can be also analysed in the light of the results provided by Ito and Tagliamonte (2003), Tagliamonte and Roberts (2005) and Tagliamonte (2008).The authors point out that one of the most important factors contributing to the distribution of intensifiers is grammaticalization. Thus, the intensifiers that are more grammaticalized tend to cooccur with larger number of adjectives. The authors attribute this factor to the increase in the use of the intensifier *so*, which is used as frequently as *very* and *really*. This fact helps to explain why the intensifiers *very*, *really* and *so* cooccur with all types of adjectives in the LINDSEI-EST and LOCNEC data sets. Thus, the observed underuse in terms of different types of maximizers in LINDSEI-EST in comparison to LOCNEC, can reflect the overall tendency of *very*, *really* and *so* to modify adjectives of different types. Given the predominant frequency of these boosters they are the preferable strategy, that does not only characterise learners of English of Estonian L1 background. These tendencies are also observed in LOCNEC.

The result of MCA show that the differences between two groups of speakers in the use of intensifiers are not statistically significant. The dimensions compiled on the basis of the MCA do not have high explanatory power when logistic linear regression is applied in order to test how well the model would predict the categorisation into English L1 and Estonian L1 speakers. Moreover, the confidence ellipses show that the two categories of speakers show substantial overlap in terms of the use of boosters and maximizers. Given that

limit and extreme adjectives are not as frequent as scalar adjectives, the influence of the types of adjectives does not result in a very good predictability of the model.

Schweinberger (2020) uses a different classification of adjectives that is based on Dixon (1977, 2004, cited in Schweinberger 2020: 170). This classification includes such types as Difficulty, Dimension, Human Propensity, Physical Property and Value (ibid). Operationalisation of adjectives using this classification could have potentially improved the results of MCA, as it provides more fine-grained distinction between the adjectives. However, this classification is not based on the contribution of gradability on intensifier-adjective combinations. Thus, it would not be possible to attest whether there are differences in the use of maximizers and boosters in terms of the semantic properties of adjectives.

In terms of methodology this thesis show that studies that look at the over- or underuse of particular linguistic items can benefit from incorporating multiple factors into the analysis. Despite the differences in terms of frequencies of particular intensifiers in LINDSEI-EST and LOCNEC, the MCA shows that there are no substantial statistical differences that would allow concluding that Estonian L1 speakers deviate from English L1 speakers. It is likely that the fact that LINDSEI-EST is based on the data of Estonian L1 speakers who are advanced learners of English contribute to these insignificant statistical differences in the use of amplifiers.

However, the fact that the data shows different collocational patterns in terms of particular adjectives, suggests that further research would benefit from calculating the association between particular types of intensifiers and adjectives using distinctive collexeme analysis (Gries and. Stefanowitsch 2004). This would allow for testing more precisely the influence of collocational preferences on the distribution of intensifiers. Another limitation of this study is disregarding of such factor as priming mentioned by

Schweinbwerger (2020) . The data in LINDSEI-EST and LOCNEC show that speakers tend to use the same intensifiers if they have used it in the previous or the same utterance or sentence. Additionally, research on intensifiers would benefit from controlling the individual differences between speakers. As it was shown previously that certain example come from the same interview. Thus, it is hard to conclude whether such examples are representative of the general tendencies of the data sets or characterise preferences in the use of intensifiers of an individual speaker.

The current study shows that such exploratory methods as HCA and MCA provide valuable insights into discovering the patterns in the data that allow to formulate hypotheses for further research. The patterns or general tendencies obtained by these methods have to be validated by hypothesis-testing methods such as linear mixed-effect regression models (Winter 2020). This method allows to test the influence of multiple factors on the use of particular linguistic item, which would allow to incorporate all of the above-mentioned factors that can potentially have an effect on the use of intensifiers. However, this requires the collection of more data as the data sets presented in this thesis are limited.

3. CONCLUSION

In this thesis, I aimed to examine the intensifiers using two comparable data sets: the Estonian component of the Louvain International Database of Spoken English Interlanguage (LINDSEI Gilquin et al. 2010) and the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004). The data sets allow for comparing the linguistic behaviour of the learners of English and English L1 speakers as the data sets contain transcriptions of interviews that require both groups of speakers to complete identical tasks. Thus, this type of data presents a perfect opportunity to investigate the use of intensifiers and identify similarities and differences between the advanced learners of English and native speakers.

Previous research on the use of intensifiers among the learners of English and the native speakers of English has focused largely on the written register (Granger 1998, Lorenz 1999,). The authors conclude that the learners of English deviate from the native speakers of English both in terms of frequencies of intensifiers and types. More recent studies have examined the use of intensifiers in the spoken register (Pérez-Paredes 2010, Perez-Parades and Camino Bueno-Alastuey 2019). The authors came to the conclusion that intensifiers are not a part of learners' active vocabulary. Additionally, the authors pointed out that the tasks influenced the distribution of the intensifiers. The learners of English of such L1 backgrounds as Chinese, German and Spanish showed a greater level of involvement compared to the English L1 speakers when presented with the picture description task. The previous research on the differences between the learners of English and English L1 speakers presented results suggesting that the main differences in the use of intensifiers concern the types of intensifiers. The study conducted by Schweinberger (2020) examined the use of *very* by the speakers of various L1 backgrounds and focused on the types of adjectives modified. The author used statistical methods to approach the frequencies of intensifiers

from the point of view of statistical significance. The results show that there are no significant deviations between the native speakers and the learners of English.

The various studies based on the use of intensifiers among English L1 speakers only showed that such factors as grammaticalization play a role in the frequency of the use of such boosters as *very*, *really* and *so* (Tagliamonte 2003, Tagliamonte and Roberts 2005, Tagliamonte 2008). Several authors approached the classification of intensifiers from different points of view. The most prominent classification that considers the correlation between the types of adjectives and the intensifiers that modify them was proposed by Paradis (1997), who divided adjectives into scalar, extreme and limit. This classification was adopted in this study to identify the differences in the use of particular types of intensifiers.

In order to answer the research questions raised in this thesis, the interviews in LINDSEI-EST and LOCNEC were analysed using the language processing tools available in the programming language Python. The instances of intensification of adjectives were extracted and manually annotated for the types of adjectives. The additional information about the examples (speakers code, part of the interview) was added in order to allow for a more fine-grained analysis. The primary methods used for identifying the differences and similarities in the use of intensifiers are hierarchical cluster analysis and multiple correspondence analysis. These methods allow examining the influence of the types of adjectives on the distribution of intensifiers in the LINDSEI-EST and LOCNEC data. Additionally, the results of MCA were validated using logistic linear regression in order to identify the significance of the found differences.

The analysis identified that the main differences between the Estonian component of LINDSEI and LOCNEC can be observed in the number of intensifiers of category maximizers. The Estonian L1 speakers use a smaller number of maximizers compared to

English L1 speakers. However, the large proportion of the adjectives of all three types (scalar, extreme and limit) cooccur with the most frequent boosters in both data sets. Thus, the function of underused maximizers in the Estonian component of LINDSEI is fulfilled by the use of boosters. The results of the Multiple correspondence analysis show that the observed differences are not statistically significant. Thus, it is possible to conclude that the advanced learners of English do not differ from their English native speakers in the use of the most frequent amplifiers. The results of the analysis potentially contribute to the studies aiming at identifying under- and overuse among the learners of English.

REFERENCES

Altenberg, Bengt. 1991. Amplifier collocations in spoken English. In Stig Johannson, Anna-Brita Stenström (eds). *English Computer Corpora: selected papers and Research guide*, 127–148. Berlin, New York: Mouton.

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

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Bolinger, Dwight. 1972. *Degree words*. The Hague: Mouton.

Bäcklund, Ulf. 1973. *The Collocation of Adverbs of Degree in English*. Uppsala: Almqvist and Wiksell.

Croft, William, and D. Alan Cruse. 2004. *Cognitive Linguistics* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie (ed), *Phraseology: Theory, analysis, and applications*, 145–160. Oxford, UK: Clarendon.

Harrell Jr , Frank E. 2022. rms: Regression Modeling Strategies. R package version 6.3-0. Available at https://CRAN.R-project.org/package=rms , accessed May 2022

Harris, Charles R., Jarrod K. Millman, Stéfan J. van der Walt,and Ralf Gommers, and Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor,and Sebastian Berg,

Nathaniel J.Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585, 357–362.

Holmes, Janet. 1984. Modifying illocutionary force. *Journal of Pragmatics*, 8:3, 345–365.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ito, Rika and Sali Tagliamonte. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, 32:2, 257–279.

Kennedy, Graeme. 2003. Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly*, 37:3, 467-487.

Kennedy, Christopher, and Louise McNally. 2005. The Syntax and Semantics of Multiple Degree Modification in English. In Stefan Müller (ed). *Proceedingsof the 12th International Conference on Head-Driven Phrase Structure Grammar,Department of Informatics, University of Lisbon*, 178–191. Stanford, CA: CSLI Publications.

Le , Sebastien, Julie Josse, Francois Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25:1, 1-18.

Levshina, Natalia. 2015. *How to Do Linguistics with r: Data Exploration and Statistical Analysis.* Amsterdam: John Benjamins Publishing Company.

Lorenz, Gunter R. 1999. *Adjective intensification – Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

Peters, Hans. 1992. English boosters: Some synchronic and diachronic aspects. In Günter Kellermann and Michael D. Morrissey (eds). *Diachrony within synchrony: Language history and cognition*, 529–545. Frankfurt: Peter Lang.

Macaulay, Ronald. 2006. Pure grammaticalisation: The development of a teenage intensifier. *Language Variation and Change*, 18:3, 267-283.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert and Kurt Hornik. 2022. *cluster: Cluster Analysis Basics and Extensions.* Available at https://CRAN.R-project.org/package=cluster, accessed May 2022.

Mustanoja, Tauno Frans. 1960. *A Middle English syntax*. Helsinki: Société Néophilologique.

Méndez-Naya, Belén. 2008. Special issue on English intensifiers. *English Language and Linguistics*, 12:2, 213–219.

Nenadic, Oleg, Michael Greenacre. 2007. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software* , 20:3, 1-13.

Paradis, Carita. 1997. *Degree modifiers of adjectives in spoken British English*. Lund: Lund University Press.

Paradis, Carita. 2001. Adjectives and boundedness. *Cognitive Linguistics*, 12:1, 47–65.

Pérez-Paredes, Pascual. 2010. The death of the adverb revisited: Attested uses of adverbs in native and non-native comparable corpora of spoken English. In Jaén, María Moreno, Fernando Serrano Valverde and María Calzada Pérez (eds). *Exploring*

new paths in language pedagogy lexis and corpus based language teaching, 157-172. London: Equinox Publishing.

Pérez-Paredes, Pascual and María Belén Díez-Bedmar. 2012. The use of intensifying adverbs in learner writing. In Ykio Tono, Yuji Kawaguchi and Makoto Minegishi (eds). *Developmental and Crossslinguistic Perspectives in Learner corpus research*, 105–124. Amsterdam: John Benjamins.

Pérez-Paredes, Pascual and María Sánchez-Tornel. 2015 A multidimensional analysis of learner language during story reconstruction in interviews. In Marcus Callies, Sandra Götz (eds). *Studies in Corpus Linguistics Learner Corpora in Language Testing and Assessment*, 141–162. Amsterdam: John Benjamins.

Suzuki, Ryota, Yoshikazu Terada and Hidetoshi Shimodaira. 2019. *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. Available at https://CRAN.R-project.org/package=pvclust , accessed May 2022

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Rickford, John R., Isabelle Buchstaller, Thomas Wasow and Arnold Zwicky. 2007. Intensive and quotative all: Something old, something new. *American Speech*, 82:1, 3-31.

Recski, Leonardo Juliano. 2004. "...It's really ultimately very cruel...": Contrasting English intensifier collocations across EFL writing and academic spoken discourse. *DELTA: Documentação De Estudos Em Linguística Teórica e Aplicada*, 20:2, 211-234.

RStudio Team (2020). RStudio: Integrated Development for R. Boston, MA: RStudio, PBC.

Stoffel, Cornelis. 1901. *Intensives and Down-toners*. Heidelberg: Carl Winter.

Schweinberger, Martin. 2020. A corpus-based analysis of differences in the use of very for adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research*. 6:2, 163-192.

Tagliamonte, Sali and Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series Friends. *American Speech*, 80:3, 280–300.

Tagliamonte, Sali. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics*, 12:2, 361-394.

Van Rossum, Guido, Fred L.Drake Jr. 1995. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace

Wagner, Susanne. 2017. Totally new and pretty awesome: Amplifier-adjective bigrams in GloWbE. *Lingua*, 200, 63–83.

Winter, Bodo 2020. Statistics for Linguists:An Introduction Using R. New York:Routledge

Xiao, Richard, and Hongyin Tao. 2007. A Corpus-Based Sociolinguistic Study of Amplifiers in British English. *Sociolinguistic Studies*, 1:2, 241-273

Appendix 1: The scripts for the Rstudio

The scripts of HCA and MCA are available at:

https://github.com/Deniss182/MA_thesis_analysis.git, accessed on 17 of May, 2022

Appendix 2: The sample used hierarchical cluster analysis (HCA) and multiple correspondence analysis (MCA)

Estonian L1 speakers:

Intensifier	Adjective	Type
particularly	impressive	extreme
particularly	impressive	extreme
particularly	impressive	extreme
really	lovely	extreme
really	lovely	extreme
really	lovely	extreme
really	crowded	extreme
really	tiny	extreme
really	topical	extreme
really	inspiring	extreme
really	intense	extreme
really	great	extreme
really	amazing	extreme
really	worn_out	extreme
really	impressive	extreme
really	devastating	extreme
really	great	extreme
so	disturbing	extreme
so	impressive	extreme
so	awful	extreme
so	vague	extreme
so	impressive	extreme
so	bizarre	extreme
so	impressive	extreme
so	stubborn	extreme
very	grand	extreme
very	weird	extreme
very	topical	extreme
very	impressive	extreme
very	intensive	extreme
very	exhausting	extreme
very	stressful	extreme
very	cringey	extreme
very	mean	extreme
very	lovely	extreme
very	basic	extreme
very	passionate	extreme
very	satisfying	extreme
very	delicious	extreme

very	eye_opening	extreme
completely	different	limit
completely	normal	limit
completely	different	limit
completely	shifted	limit
completely	different	limit
completely	isolated	limit
completely	different	limit
completely	different	limit
really	modern	limit
really	dumbed_down	limit
really	realistic	limit
really	relatable	limit
really	different	limit
really	repetitive	limit
really	mechanical	limit
really	mechanical	limit
really	humorous	limit
really	relaxing	limit
really	bored	limit
really	outgoing	limit
really	impressed	limit
really	toxic	limit
so	different	limit
so	different	limit
so	different	limit
so	different	limit
so	satisfied	limit
so	clear	limit
so	used	limit
so	impressed	limit
so	humid	limit
too	critical	limit
too	stressed	limit
very	controversial	limit
very	free	limit
very	sceptical	limit
very	emotional	limit
very	safe	limit
very	focused	limit
very	specific	limit
very	humorous	limit
very	TRUE	limit
very	thought_provoking	limit
very	unfortunate	limit
very	little	limit
very	shocked	limit
very	realistic	limit

very	impressed	limit
very	different	limit
very	foreign	limit
very	uncommon	limit
very	sciency	limit
very	different	limit
very	monoracial	limit
very	colonial	limit
very	rare	limit
very	helpful	limit
very	delusioning	limit
very	fulfilling	limit
very	fond	limit
very	TRUE	limit
very	likely	limit
very	visible	limit
very	likely	limit
extremely	uncomfortable	scalar
extremely	uncomfortable	scalar
particularly	good	scalar
particularly	good	scalar
particularly	good	scalar
particularly	good	scalar
particularly	old	scalar
particularly	good	scalar
really	cute	scalar
really	posh	scalar
really	cool	scalar
really	pretty	scalar
really	good	scalar
really	good	scalar
really	long	scalar
really	interesting	scalar
really	big	scalar
really	difficult	scalar
really	long	scalar
really	sad	scalar
really	nice	scalar
really	funny	scalar
really	interesting	scalar
really	good	scalar
really	cool	scalar
really	warm	scalar
really	scared	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	interesting	scalar

really	good	scalar
really	interested	scalar
really	quick	scalar
really	good	scalar
really	happy	scalar
really	good	scalar
really	rich	scalar
really	deep	scalar
really	tired	scalar
really	strange	scalar
really	rich	scalar
really	beautiful	scalar
really	good	scalar
really	interesting	scalar
really	pretty	scalar
really	cool	scalar
really	cool	scalar
really	fun	scalar
really	cool	scalar
really	grateful	scalar
really	tired	scalar
really	nice	scalar
really	clean	scalar
really	clean	scalar
really	nice	scalar
really	nice	scalar
really	scary	scalar
really	good	scalar
really	loud	scalar
really	happy	scalar
so	big	scalar
so	sorry	scalar
so	pretty	scalar
so	similar	scalar
so	ugly	scalar
so	kind	scalar
so	funny	scalar
so	expensive	scalar
so	friendly	scalar
so	cheery	scalar
so	friendly	scalar
so	friendly	scalar
so	welcoming	scalar
so	beautiful	scalar
so	long	scalar
so	strange	scalar
so	good	scalar
so	expensive	scalar

so	beautiful	scalar
so	warm	scalar
so	friendly	scalar
so	small	scalar
so	poor	scalar
so	nice	scalar
so	familiar	scalar
too	beautiful	scalar
too	happy	scalar
too	long	scalar
too	large	scalar
too	young	scalar
too	few	scalar
very	formal	scalar
very	informal	scalar
very	happy	scalar
very	interesting	scalar
very	happy	scalar
very	similar	scalar
very	high	scalar
very	strict	scalar
very	high	scalar
very	happy	scalar
very	productive	scalar
very	interesting	scalar
very	successful	scalar
very	interested	scalar
very	interesting	scalar
very	friendly	scalar
very	uncomfortable	scalar
very	depressing	scalar
very	enjoyable	scalar
very	enjoyable	scalar
very	violent	scalar
very	good	scalar
very	strange	scalar
very	good	scalar
very	happy	scalar
very	funny	scalar
very	similar	scalar
very	big	scalar
very	important	scalar
very	sensitive	scalar
very	good	scalar
very	flat	scalar
very	good	scalar
very	far	scalar
very	nice	scalar

very	difficult	scalar
very	difficult	scalar
very	nice	scalar
very	interested	scalar
very	small	scalar
very	poor	scalar
very	rich	scalar
very	famous	scalar
very	posh	scalar
very	long	scalar
very	convenient	scalar
very	convenient	scalar
very	nice	scalar

English L1 speakers:

Intensifier	Adjective	Type
absolutely	amazing	extreme
absolutely	fantastic	extreme
absolutely	amazing	extreme
absolutely	brilliant	extreme
absolutely	amazing	extreme
absolutely	sure	limit
absolutely	flabbergasted	limit
completely	horrible	extreme
completely	gorgeous	extreme
completely	wild	extreme
completely	new	limit
completely	new	limit
completely	different	limit
completely	foreign	limit
completely	foreign	limit
extremely	good	scalar
particularly	good	scalar
particularly	good	scalar
particularly	good	scalar
particularly	bad	scalar
particularly	good	scalar
particularly	enjoyable	scalar
particularly	nice	scalar
really	amazing	extreme
really	amusing	extreme
really	frustrating	extreme
really	hectic	extreme
really	impressive	extreme
really	lovely	extreme

really	lovely	extreme
really	tense	extreme
really	versatile	extreme
really	weird	extreme
really	bilingual	limit
really	black	limit
really	fed_up	limit
really	grey	limit
really	impressed	limit
really	solid	limit
really	attractive	scalar
really	bad	scalar
really	bad	scalar
really	bad	scalar
really	beautiful	scalar
really	beautiful	scalar
really	beautiful	scalar
really	big	scalar
really	bright	scalar
really	close	scalar
really	confident	scalar
really	cool	scalar
really	difficult	scalar
really	easy	scalar
really	effective	scalar
really	effective	scalar
really	friendly	scalar
really	friendly	scalar
really	fun	scalar
really	fun	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar

really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	good	scalar
really	handy	scalar
really	happy	scalar
really	heavy	scalar
really	hot	scalar
really	hot	scalar
really	important	scalar
really	interesting	scalar
really	interesting	scalar
really	interesting	scalar
really	interesting	scalar
really	interesting	scalar
really	interesting	scalar
really	lucky	scalar
really	nervous	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	nice	scalar
really	noisy	scalar
really	noisy	scalar
really	organised	scalar
really	posh	scalar
really	rich	scalar
really	sad	scalar
really	slow	scalar
really	small	scalar

really	sunny	scalar
really	sweet	scalar
really	ugly	scalar
really	warm	scalar
really	young	scalar
so	versatile	extreme
so	confusing	extreme
so	daunting	extreme
so	great	extreme
so	amazing	extreme
so	miserable	extreme
so	impressive	extreme
so	professional	extreme
so	frustrating	extreme
so	fascinating	extreme
so	generous	extreme
so	defining	limit
so	realistic	limit
so	different	limit
so	little	limit
so	little	limit
so	cutoff	limit
so	overcast	limit
so	cool	scalar
so	bad	scalar
so	unfriendly	scalar
so	frustrated	scalar
so	old	scalar
so	nervous	scalar
so	cold	scalar
so	nervous	scalar
so	hot	scalar
so	beautiful	scalar
so	expensive	scalar
so	cheap	scalar
so	interested	scalar
so	interesting	scalar
so	friendly	scalar
so	kind	scalar
so	aggressive	scalar
terribly	tragic	extreme
terribly	modern	limit
too	knackered	extreme
too	realistic	limit
too	worthwhile	limit
too	bored	limit
too	hot	scalar
too	nice	scalar

too	long	scalar
too	complicated	scalar
too	difficult	scalar
too	scared	scalar
too	bad	scalar
too	hot	scalar
totally	FALSE	limit
totally	different	limit
very	demanding	extreme
very	mad	extreme
very	comical	extreme
very	impressive	extreme
very	versatile	extreme
very	exciting	extreme
very	smug	extreme
very	valuable	extreme
very	upsetting	extreme
very	miserable	extreme
very	painful	extreme
very	accurate	extreme
very	stressful	extreme
very	inadequate	extreme
very	weary	extreme
very	sympathetic	extreme
very	impressive	extreme
very	significant	extreme
very	striking	extreme
very	tiny	extreme
very	lush	extreme
very	grey	limit
very	rare	limit
very	affected	limit
very	ill	limit
very	different	limit
very	different	limit
very	Spanish	limit
very	Spanish	limit
very	different	limit
very	north	limit
very	reserved	limit
very	different	limit
very	TRUE	limit
very	impressed	limit
very	different	limit
very	different	limit
very	Spanish	limit
very	pleased	limit
very	different	limit

very	different	limit
very	different	limit
very	separate	limit
very	different	limit
very	Danish	limit
very	rare	limit
very	claustrophobic	limit
very	different	limit
very	different	limit
very	impressed	limit
very	green	limit
very	relaxing	limit
very	traumatic	limit
very	complete	limit
very	touristy	limit
very	reserved	limit
very	good	scalar
very	nice	scalar
very	interested	scalar
very	hard	scalar
very	good	scalar
very	big	scalar
very	big	scalar
very	good	scalar
very	happy	scalar
very	long	scalar
very	hard	scalar
very	nice	scalar
very	nice	scalar
very	careful	scalar
very	good	scalar
very	good	scalar
very	high	scalar
very	good	scalar
very	traditional	scalar
very	nice	scalar
very	interesting	scalar
very	strange	scalar
very	lucky	scalar
very	happy	scalar
very	difficult	scalar
very	good	scalar
very	happy	scalar
very	small	scalar
very	beautiful	scalar
very	romantic	scalar
very	nice	scalar
very	famous	scalar

very	famous	scalar
very	short	scalar
very	organised	scalar
very	loud	scalar
very	proud	scalar
very	high	scalar
very	strict	scalar
very	big	scalar
very	important	scalar
very	quiet	scalar
very	hot	scalar
very	interesting	scalar
very	warm	scalar
very	interested	scalar
very	sad	scalar
very	popular	scalar
very	rich	scalar
very	good	scalar
very	interesting	scalar
very	interesting	scalar
very	interesting	scalar
very	good	scalar
very	cheap	scalar
very	good	scalar
very	busy	scalar
very	difficult	scalar
very	good	scalar
very	good	scalar
very	good	scalar
very	good	scalar
very	strange	scalar
very	interesting	scalar
very	good	scalar
very	good	scalar
very	wet	scalar
very	long	scalar
very	good	scalar
very	rich	scalar
very	pretty	scalar
very	high	scalar
very	convenient	scalar
very	young	scalar
very	small	scalar
very	loud	scalar
very	proud	scalar
very	expensive	scalar
very	gentle	scalar
very	hot	scalar

very	nice	scalar
very	flat	scalar
very	lucky	scalar
very	bumpy	scalar
very	nice	scalar
very	friendly	scalar
very	interesting	scalar
very	long	scalar
very	depressing	scalar
very	nice	scalar
very	rich	scalar
very	poor	scalar
very	nice	scalar
very	hard	scalar
very	happy	scalar
very	heavy	scalar
very	cold	scalar
very	big	scalar

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Denys Savchenko

A corpus-based study of adjective intensification among native speakers and learners of English

Korpuspõhine uurimus omadussõnade intensiivsuse väljendamisest inglise keelt emakeelena kõnelejate ja võõrkeelena õppijate hulgas

Magistritöö
2022
Lehekülgede arv:93

Magistritöö uurib omadussõnade intensiivsuse väljendamist inglise keelt emakeelena kõnelejate ja võõrkeelena õppijate hulgas. Uurimuse andmed pärinevad kahest korpusest – Louvain International Database of Spoken English Interlanguage (LINDSEI) eesti keele all-korpusest (LINDSEI-EST) ja Louvain Corpus of Native English Conversation (LOCNEC) korpusest. Kokku uuritakse töös 1 317 omadussõna ja määrsõna kombinatsiooni intensiivsuse väljendamisel. Andmete kättesaamiseks tuli kirjutada Pythoni kood ja andmetes sõnaliigid automaatselt määrata. Vajalik oli andmete käsitsi puhastamine ja annoteerimine. Magistritöös otsitakse vastuseid järgmistele uurimisküsimustele: Kuidas jagunevad intensiivsust väljendavad määrsõnad kahe andmestiku vahel? Milliseid erinevusi võib märgata kahe rühma vahel (inglise keelt emakeelena kõnelejad ja võõrkeelena õppijad) nende väljendite kasutamisel? Varasemate uurimuste põhjal võib eeldada, et erinevad rühmad kasutavad erinevat tüüpi määrsõnu omadussõnade intensiivsuse väljendamiseks ja et erinevat liiki omadussõnad käivad kokku erinevate määrsõnadega.

Töö koosneb ühest teoreetilisest ja ühest empiirilisest peatükist. Magistritöö sissejuhatuses selgitatakse, miks on vajalik uurida omadussõnade intensiivsuse väljendamist eesti keelt emakeelena kõnelevate inglise keele õppijate hulgas. Lisaks sellele antake sissejuahtuses ülevaade peamistest töös kasutatud terminitest ja kirjeldatakse töö uurimisküsimusi. Esimene peatükk kirjeldab varasemaid uurimusi, mis on vaadelnud inglise keele omadussõnade intensiivsuse väljendamist nii inglise keelt emakeelena kõnelejate kui inglise keele õppijate hulgas. Peamiselt on uuritud omadussõnade intensiivsust väljendavaid määrsõnu erinevates inglise keele variantides. Töö teises peatükis antakse ülevaade töös kasutatud korpustest – LINDSEI-EST ja LOCNEC. Põhiline rõhk on kahe olulise uurimismeetodi selgitamisel – hierarhiline klasteranalüüs ja mitmene korrespondentsanalüüs. Mõlema meetodi kasutamine lubab töös minna kaugemale kui lihtsalt sagedusandmete analüüsimine. Hierarhilise klasteranalüüsi ja mitmese korrespondentsanalüüsi tulemusena on selge, et üldiselt on Eesti inglise keelt võõrkeelena edasijõudnud õppijate määrsõnade kasutus omadussõnade intensiivsuse väljendamisel sarnane inglise keelt emakeelena kõnelejate kasutusega. Kõige sagedasemateks määrsõnadeks on mõlemas all-korpuses *very, really, so, quite*. Erinev on määrsõnade loend, mida kumbki rühm kasutab. Emakeele kõnelejad kasutavad suuremal hulgal erinevaid määrsõnu omadussõnade intensiivuse kirjeldamiseks. Statistilise andmeanalüüsi tulemusena selgus, et jagades omadussõnad kolme liiki – skalaarsed, ekstreemsed, piiratud omadussõnad (*scalar, extreme, limit*) – on võimalik välja tuua erinevate määrsõnade klastrid, mille käitumismustrid materjalis on erinevad. Määrsõnade klastrite teket mõjutab omadussõna liik ja mitte keelekõnelejate taust.

Märksõnad: Inglise keel, õppijakeel, korpuslingvistika, määrsõnad, omadussõnad, klasteranalüüs, korrespondentsanalüüs

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Denys Savchenko

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

A corpus-based study of adjective intensification among native speakers and learners of English

mille juhendaja on Jane Klavan

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

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

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

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

Denys Svachenko

17.05.2022

Autorsuse kinnitus

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

Denys Savchenko

17.05.2022

Lõputöö on lubatud kaitsmisele.

Jane Klavan

17.05.2022