VILNIUS UNIVERSITY

FACULTY OF PHILOLOGY

Ieva Gulbinaitė

English Studies (Linguistics)

A Contrastive Analysis of Referential Cohesion in L2 Written Learner English

MA thesis

Academic Supervisor: Assoc. Prof. Dr. Rita Juknevičienė

Vilnius

2025

TABLE OF CONTENTS

ABSTRACT

This master's thesis investigates patterns of referential cohesion through the anaphoric pronoun *it* in written English, comparing native speakers with Lithuanian and Norwegian English as a Foreign language (EFL) learners. The research focuses on two types of antecedents—nominal phrases (NPs) denoting concrete to abstract entities, and verbal constructions representing highly abstract discourse elements, examining their distribution, distances, and discourse positioning across three corpora (LOCNESS, LICLE, and NICLE).

The findings reveal that while NP antecedents predominate across all datasets, EFL learners demonstrate a statistically significantly higher reliance on verbal antecedents compared to native speakers. This divergence aligns with previous psycholinguistic research suggesting that native English speakers prefer demonstratives for verbal antecedents (Wittenberg et al., 2021; Çokal et al., 2016). The difference appears influenced by first language transfer—Lithuanian “tai” ('it') overlaps with English “it” only in denoting general phenomena, while Norwegian learners may underuse “dette” ('this') due to its primarily formal register status in Norwegian.

Antecedent-anaphor distances varied considerably, with native speakers maintaining cohesion over longer stretches for NP antecedents (up to 125 tokens) through thematic continuity and lexical reiteration. In positioning, NP antecedents appeared in the same sentence as their anaphors more frequently (76% in LOCNESS, 68% in LICLE) than verbal antecedents (56% in LOCNESS, 36% in LICLE). Norwegian learners demonstrated statistically significant differences from native speakers by placing NP antecedents in sentences preceding “it” more frequently.

This master's thesis additionally identifies areas for future investigation: hypothetical antecedents, implicit antecedents, and number agreement violations using the singular pronoun “it”.

Keywords: referential cohesion, pronouns, EFL, anaphora, Contrastive Interlanguage Analysis (CIA), contrastive analysis

Chapter 1. INTRODUCTION

Ever since the focal work on cohesion was published by Halliday and Hasan (1976), cohesion analysis has been studied extensively with the focus on its different aspects across different genres. Halliday and Hasan (ibid.) categorised cohesion into five distinct types: reference, conjunction, substitution, ellipsis, and lexical cohesion. Despite considerable research attention to these categories, significant areas remain underexplored.

In the context of second language acquisition, cohesion in written production has been analysed from both learner and native speaker perspectives. However, corpus-based contrastive analyses are predominantly focused on conjunctions (Connor, 1984; Granger and Tyson, 1996; Altenberg and Tapper, 1998; Hinkel, 2001; Liu and Braine, 2005; Bikelienė, 2012; Tejada et al., 2015). As for the research on referential cohesion, it has been primarily analysed in the context of personal reference, that is, when the antecedent is in reference to a person (Ryan, 2015; Díaz-Negrillo & Rosillo, 2024)

Findings from the analysis of native speaker corpora suggest that among different genres, personal pronouns are mostly widespread in conversations and are least frequent in academic prose (Biber et al. 2021:239). While these discoveries might be a point of reference for understanding anaphoric expressions in different genres, the treatment of anaphoric expressions in the written production of students and, even more specifically, L2 learner English remains little investigated. The argumentative essays produced by EFL learners are a distinct genre involving a number of variables. Students are only developing their writing skills and are influenced by pedagogical methodologies which face challenges focusing their attention on the core of argumentation (McGee, 2019; Schneer, 2014).

The need to study cohesion, namely, grammatical cohesion and reference, stems from the unique nature of the argumentative essay genre and the scarce comparative analyses of native speakers (NS) and non-native speakers (NNS) employment of anaphors in academic discourse.

The selection of the anaphoric *it* for this thesis is motivated by its unique property to point back both to specific NPs as well as sentences and clauses in discourse (Quirk et al. 1972:700). *It* in the English language is a highly productive pronoun and also poses challenges due its abstract nature, where the same form appears as an empty subject/object, anticipatory subject/object, and as a subject in cleft constructions (Biber et al. 2021: 331-332). Considering this complexity, it would not be surprising that EFL learners may struggle with the acquisition of *it,* depending on the extent to which their L1 pronominal systems are similar to that of English. This study hypothesises that greater functional overlap between the L1 pronoun system and English would result in better pronoun acquisition, as measured against native speaker (NS) production. Therefore, this study examines learner English produced by EFL learners from two different L1 backgrounds, namely, Lithuanian and Norwegian.

The Lithuanian pronoun system, as opposed to English, displays more characteristics of a classical Indo-European language in which grammatical gender of an entity defines the selection of a pronoun, therefore, an exact counterpart of *it* does not exist in Lithuanian and it is rather a set of different Lithuanian pronouns that correspond to *it* in English. The most notable difference between the two languages is that the 3rd person pronoun paradigm in Lithuanian does not include the neuter pronoun *tai* 'it' which is instead treated as a demonstrative along with its counterparts marked for gender which include *tas* 'that, masculine', *ta* 'that, feminine gender', *šitas* 'this, masculine ', *šita* 'this, feminine' As for its function, the neuter demonstrative *tai* is resorted to solely for pointing back to generalized phenomena, previously introduced into discourse (Ramonienė et al., 2019:72).

As for the Norwegian 3rd person system, it displays more similarities to English than to Lithuanian as the distinction is made between animate *han* 'he' and *hun* 'she' and inanimate entities *den* 'it, masculine/feminine' and *det* 'it, neuter' which are, as opposed to English, marked for masculine/feminine and neuter genders while. *Den* 'it, masculine/feminine' and *det* 'it, neuter' are not solely used in reference to inanimate objects, but also points back to a phrase or clause (Holmes and Enger, 2018:145). It is thus possible to hypothesize that native languages of EFL learners may be reflected in the way Lithuanian and Norwegian EFL learner use anaphoras.

To fill the existing gap on referential cohesion in linguistic literature, the present thesis explores referential cohesion in L2 written learner English.

The aim of the present thesis is to compare the usage of the anaphoric marker *it* in native speaker (NS) group and two non-native speaker (NNS) groups.

In order to reach this aim, the following **research questions** are raised:

1. What types of antecedents are linked with the anaphor *it* in the written production of native speakers and EFL learners?
2. To what extent do native speakers and EFL learners prefer positioning the anaphor *it* within the sentence that contains their antecedents?
3. What differences can be identified in the written production of native speakers and EFL learners in terms of antecedent types and the distance between the anaphor and its antecedent?

The structure: This thesis consists of four chapters – Introduction, Literature Review, Data and Methods, and Results and Discussion. The Introduction identifies the traditional five cohesive devices and discusses research regarding cohesion, highlighting existing research gaps. Additionally, the Introduction examines the properties of the pronoun *it* in the English language and overviews its counterparts in Lithuanian and Norwegian. The Literature Review section introduces relevant scholarship on cohesion, discusses key concepts and terminology, focusing particularly on the concept of reference. Special attention is given to third-person pronouns, comparing the interchangeability of demonstratives with *it*. The Data and Methods chapter presents the methodological procedure followed to conduct the analysis, identifies data sources, and introduces the coding categories used. The Results and Discussion section includes both quantitative and qualitative analyses of the data, organised into relevant thematic categories. Finally, the Conclusions answer the research questions and provide interpretations regarding the differences between the learner groups. The thesis concludes with references and a summary in Lithuanian.

Chapter 2. LITERATURE REVIEW

2.1 Coherence and cohesion

Coherence and cohesion are two interdependent notions which are often discussed together. Having that said, linguistic literature is at times unclear regarding the distinction between the two, the terms at times being used interchangeably. Despite coherence being closely related to cohesion, there are differences. Morris and Hirst (1991:25) summarise that cohesion has to do with text hanging together, while coherence is a term related to text making sense.

Regarding the relation between text and cohesion, Halliday and Hasan (1976:298–299) emphasise that cohesion is a necessary though not a sufficient condition for the creation of text and what creates text is the textual, or text-forming, component of the linguistic system, of which cohesion is one part of. In other words, cohesion provides connectedness between sentences and ideas in a text, and the textual system consists not only of cohesion but also the elements of grammar and lexis.

According to Halliday and Hasan (1976:4) the concept of cohesion is fundamentally a semantic one; it refers to relations of meaning that exist within the text, and that define it as a text. Cohesion occurs when an interpretation of an element in discourse depends on interpreting another element.

While cohesion can be measured through specific linguistic features, coherence interacts with language proficiency and background knowledge, indicating a scope beyond purely linguistic features (Crossley et al., 2018a). Despite the presence of cohesive devices, coherence may not be established if a reader lacks relevant background knowledge. Thus, coherence relates more directly to text readability.

The difference between cohesion and coherence can be thus summarised as follows: cohesion is merely a facilitative means to contribute to coherence or increase the readability of the text but does not guarantee that the text will be perceived logically by the reader.

2.2 Types of cohesion

Cohesion is traditionally divided into grammatical, lexical, and lexicogrammatical types based on closed-ended and open-ended systems (Halliday and Hasan, 1976:301). Lexicogrammatical cohesion, expressed through conjunctions, represents a borderline case—often realised through closed-class elements but involving lexical choice in conjunction selection.

As a more empirical approach towards cohesion has started to emerge, the terminology around the topic has been expanding. For example, Crossley et al. (2016) explored how learners of a second language (L2) improve their use of cohesive devices in writing over a semester. The researchers focused on essays written by university students in English for Academic Purposes (EAP) courses, categorising cohesive devices based on the level of text they connect, namely sentence, paragraph and overall structure of text (local, global and textual cohesion respectively).

Grammatical cohesion encompasses reference, substitution and ellipsis. The following subsection discusses theoretical background of reference, which is a primary focus of the present master's thesis.

2.3 Classification of reference

Halliday and Hasan (1976) categorise reference into personals, demonstratives and comparatives. Empirical research has analysed these types to varying degrees. For example, Reid (1992), limits her analysis to personal and demonstrative pronouns, excluding the comparative type. The author was interested in how speakers whose native languages are Arabic, Chinese, Spanish and English differ in their written production in terms of the use of this cohesive device. It was found that the number of pronouns used by NS was significantly lower compared to Arabic, Chinese and Spanish NNS of English. What is more, the distribution of pronouns also depended upon the topic of the essay. For example, when given to write comparison/contrast topics, the percentage of the pronoun usage *I, you* was greater, but when the writing task concerned the description of graphs, more demonstratives and third-person pronouns *it* were used. While analysing the pronouns, Reid (ibid.) also included the instances of personal pronouns such as *I* and *you*, which are not cohesive as they are exophoric, that is, occurring outside the text (Halliday & Hasan, 1976, p. 48).

Not all types of reference function cohesively. A classification by Halliday and Hasan (1976) provided in Figure 1 illustrates a dichotomy between cohesive and non-cohesive types of reference. According to the authors (ibid.), "only the third person is inherently cohesive, in that a third person form typically refers anaphorically to a preceding item in the text" and "first and second person forms do not normally refer to the text at all" (p.48).

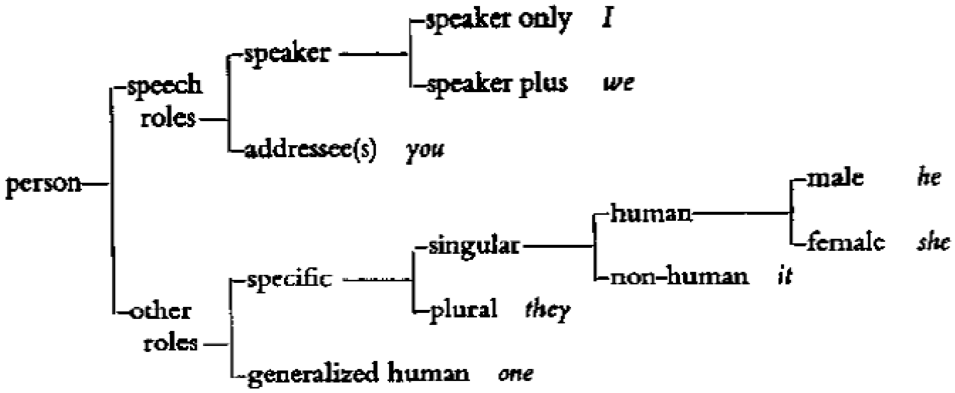

Figure 1. Semantic distinctions in the personal system (Halliday and Hasan 1976:44)

Although Halliday and Hasan (1976) classify the pronoun *it* among personals, this classification has been questioned on semantic grounds. Wittenberg et al. (2021) distinguish between personal and non-personal pronouns, noting that the non-personal pronoun *it* refers to events or non-human entities.

2.4 The concept of anaphora

Anaphora is defined as "the relation between an anaphor and an antecedent, where the interpretation of the anaphor is determined via that of the antecedent" (Huddleston and Pullum, 2016:1453). The term of anaphora corresponds to a *tie* in Halliday and Hasan's (1976) typology, while the term *anaphora* is there understood as a "presupposition, pointing back to some previous item" (ibid., 14). To avoid confusion, this master's thesis adheres to the more recent terminology by Huddleston and Pullum (2016), where anaphor is understood as a linguistic cue that signals a presupposition, as in the following sentence, where *he* presupposes *Max*:

(1) ***Max** claims **he** wasn't told about it.* (ibid.1453)

Consequently, the antecedent is an element that precedes the anaphor, in (X), that corresponds to *Max*. While the relation between *Max* and *he* is called anaphora.

Anaphora enjoys widespread interest across various branches of linguistics; anaphora resolution has been examined by computational linguists (Mitkov, 2016; Poesio et al., 2023). The cognitive mechanisms involved in the processing of anaphora have been a focus of cognitive linguists (Ariel, 1990; Holler & Suckow, 2016; Wittenberg et al., 2021), pragmatists (Liu, 2023; Geluykens, 1994; Huang, 2000), as well as grammarians (Safir, 2004; Chierchia, 1995).

2.4.1 The concept of anaphor

Traditionally, anaphors were conceptualized as reference signals and were distinguished between sentence or clause and noun phrase reference signals (Quirk et al., 1972:700). In (2), the anaphor *it* signals a sentence, in particular, *it* points back to the two preceding sentences, while in (3), the demonstrative *this* signals the noun phrase *his brown raincoat.*

(2) *Many students never improve. They get no advice and therefore they keep repeating the same mistakes. **It**'s a terrible shame.* (ibid.,701).

(3) *He asked for* **his brown raincoat.** *He insisted that* **this** *was his usual coat during the cold winter months.* (ibid. 704)

In subsequent grammars of English, knowledge about anaphors has been expanded beyond their properties of signalling, identifying the types of anaphoric expressions and their interaction with the antecedent. Biber et al. (2021:241) identified six types of anaphoric expressions, which are as follows:

1) demonstrative pronoun

2) personal pronoun
3) demonstrative with synonym
4) demonstrative with repeated noun
5) the with synonym
6) the with repeated noun

2.4.2 The concept of antecedent

2.4.2.1 Semantics of antecedents

Antecedents function as discourse referents that provide the necessary semantic content for anaphoric expressions to establish reference (Quirk et al., 1972:700). The attempts to classify both discoursal and extralinguistic entities, based on the level of their perceptual properties, are discussed in linguistic literature.

Lyons (1977) proposed a distinction between first-order, second-order, and third-order entities. First-order entities are physical objects with relatively stable perceptual properties existing in three-dimensional space (e.g., persons, animals, artifacts). Second-order entities are events, processes, and states-of-affairs that occur rather than exist, being located in time rather than space. Third-order entities represent abstract propositions outside spatiotemporal dimensions, such as concepts, ideas, and propositional content.

Building on Lyons' work, Asher (1993) developed a framework for understanding abstract objects in discourse, particularly focusing on how abstract entities serve as antecedents for anaphoric reference. Asher's theory of abstract objects distinguishes between eventualities (events, processes, states), facts, and purely abstract objects (propositions, questions, etc.).

Although an anaphor is often said to refer to both concrete entity antecedents expressed by noun phrases and to non-NP antecedents, such treatment is inaccurate. The term *reference* is inherently connected to external world entities, while non-NP antecedents occur exclusively in discourse and are co-textual.

It is thus more sensible to state that an anaphor, or an anaphoric pronoun, is "anaphoric to, or linked anaphorically with, its antecedent" (Huddleston and Pullum 2016:1457-1458). This means that the term *reference* can only apply to the most prototypical case of anaphora in which the anaphor is a pronoun and the antecedent is a noun phrase denoting an entity in extralinguistic reality. To avoid confusion, some scholars make use of the term *co-reference* to denote anaphors with NP antecedents (Loáiciga et al., 2017:1325). On the other hand, when an antecedent is a clause or sentence, it cannot be treated as an entity in the real world. Rather, clauses and sentences are discoursal elements.

2.4.3 Referential distance

It has been proposed that the selection of an anaphor and its antecedent is related to the distance separating them. This implies that the choice of an anaphor is not arbitrary but rather dependent on cognitive and psychological factors. Givón (1983) suggests that in cases of short referential distance, an anaphor tends to take the simplest linguistic form, that is, a pronoun, as the antecedent remains highly activated in the speaker's and hearer's working memory.

In terms of information structure, anaphors mark givenness. Anaphoric pronouns denote old information, while the antecedent of the anaphor occupying a fuller linguistic expression is understood as new information. The tendency for shorter linguistic forms to be associated with given information and longer forms with new information is explained by Ariel's Accessibility Theory (1990). The theory distinguishes between:

1) Low-accessible elements: These require a low degree of mental activation and can be expressed through reduced forms like pronouns.
2) High-accessible elements: These exhibit a high degree of activation and often require fuller linguistic forms for clarification.

The Givenness Hierarchy by Gundel et al. (1993) complements this theory by categorizing cognitive statuses into six levels: in focus, activated, familiar, uniquely identifiable, referential, and type identifiable. Each status corresponds to a specific type of referring expression, reflecting the speaker's assumptions about the hearer's cognitive status of the referent.

It is suggested that the specificity of the anaphor correlates with the larger anaphoric distance, that is, the most specific anaphors, *the* with repeated noun, tend to occur the most remotely from the antecedent (refer to Figure 2). The tendency is related to the cognitive processing of anaphora and discourse comprehension.

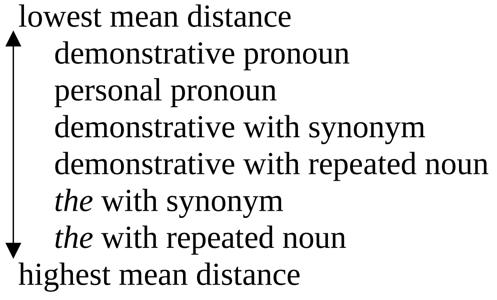

Figure 2. The relation between specificity of an anaphor and anaphoric distance (Biber et al., 2019:241–242)

The hierarchy presented in Figure 2 is also complemented with empirical research on the subject. For example, the limited anaphoric distance of demonstrative pronouns is associated with their fundamental meaning; they indicate entities that are recognised in relation to the immediate context, which may include either the situational context or the surrounding co-text. Noun phrases that incorporate demonstrative determiners exhibit greater specificity than demonstrative pronouns, resulting in a broader anaphoric distance. Similarly, the distinction between repeated nouns and synonyms aligns with the pattern observed in the noun phrases marked by demonstratives (Dietrich et al., 2024).

However, the principles of accessibility theory can evidently deviate from the expected patterns. A corpus-based study by Demol (2007), which used Ariel's (1990) accessibility theory as a framework and compared the French third-person clitic pronoun *il* 'he' with the demonstrative pronoun *celui-ci* 'this one', the results partially deviated from the expected pattern. The clitic pronoun *il* 'he' in terms of its accessibility is regarded as a high accessibility marker, while the demonstrative pronoun *celui-ci* 'this one' is an intermediate accessibility marker (Demol 2007:7). One of the unexpected results in the study (ibid.) was the smaller mean distance between antecedent and the anaphor *celui-ci* 'this one' than *il* 'he' when given the higher accessibility of the latter the vice-versa scenario was to be expected. However, in terms of topicality, *il* 'he' is found to occur more as subjects in sentence-initial position which suggests the low-accessibility of the pronoun.

2.5 Referential cohesion in L2 Discourse

2.5.1 Personal pronouns

This subsection reviews several experimental studies that focus on the use of third-person pronouns as animate entities, which are commonly explored within the context of second language acquisition.

2.5.1.1 Pronominal Reference Patterns Among EFL Learners

Research examining how Chinese EFL learners maintain reference for story characters throughout discourse found that NNS tend to use significantly fewer high accessibility markers in contexts where NS prefer them, resulting in overexplicitness (Ryan, 2015:839), as exemplified in (4):

(4) *The girl was running away, and she kn-ah knock into **Charlie Chaplin**, and **Charlie Chaplin** actually was very helpful.* (ibid. 847)

Besides NP redundancy, non-native speakers were also underexplicit when anaphorically referring to minor story characters (low accessibility referents) expected to be expressed in longer linguistic forms like *the* + NP (Ryan, 2015:845). This underexplicitness likely stems from L1 influence, as Chinese lacks overt definiteness expressions (articles) found in English.

The non-native like usage of anaphor by EFL learners has also been evident in the usage of syntactic contexts, namely, in coordinating sentences in which zero anaphor is licensed in the English language (Díaz-Negrillo and Rosillo, 2024).

The study by Díaz-Negrillo and Rosillo (2024) analysed the referring expressions (RE), namely, 3rd person singular grammatical subject, selection in lower intermediate (B1) and lower advanced (C1) Spanish EFL learner language, comparing it with the reference group of L1 English in both spoken and written modes. The findings of the study suggest that intermediate and advanced Spanish EFL learners are similar in their use of REs in spoken mode; however, as far as written production is concerned, advanced-level users display more similarities with L1 English speakers.

Another corpus-based study by Quesada and Lozano (2020) examines comparable findings by analysing Spanish EFL learners with varying levels of English proficiency (beginner, intermediate, advanced), comparing the results with the reference group of L1 English. It was found that for learners, even at an advanced level, creating cohesion in discourse is problematic. One finding has to do with overexplicitess arising from the overuse of explicit referential expressions. Learners rarely resort to zero anaphora in the contexts that license zero anaphora in English, even though zero anaphora is also allowed in Spanish. Another challenge that the learner group is found to encounter is the redundancy in an attempt to avoid possible ambiguity; that is, fuller NPs are used instead of pronouns in contexts where they would be more preferable.

There is also evidence that EFL learners, regardless of their proficiency level, tend to underuse pronouns and instead rely on referential expressions associated with lower accessibility. Studies have shown that the use of highly accessible forms, namely, pronouns, as opposed to more explicit forms like full noun phrases—is positively linked to evaluations of discourse cohesion, as measured by the pronoun-noun ratio (Crossley et al., 2016). More specifically, given information is found to be positively associated with the use of high accessibility forms, suggesting that speakers who maintain higher levels of referential accessibility produce more cohesive discourse (Crossley et al., 2016).

2.5.1.2 Pronominal Reference Patterns Among English CFL Learners

Difficulties in establishing referential cohesion in L2 discourse are not confined to native speakers of Chinese; inverse patterns have also been observed. For example, NNS of Chinese often process implicit referential expressions—that is, expressions lacking explicit pronouns or noun phrases—which is a common strategy in Chinese but not in English (Saner and Hefright, 2015). In their study, Saner and Hefright (2015) employed an eye-tracking experiment to measure participants' gaze movements while they listened to narrative descriptions containing instances of zero-anaphora in Chinese. The authors found that non-native speakers demonstrated longer fixation times and more regressions upon encountering omitted referents, indicating an increased cognitive demand in resolving references. This finding suggests that discourse practices from the native language may transfer to non-native language usage.

2.5.2 Demonstratives

2.5.2.1 Demonstratives vs *it*

The findings in cognitive linguistics suggest that demonstratives are slower to process for readers in comparison to pronouns (Gundel et al., 1993; Wittenberg et al., 2021). On the other hand, when the pronoun *it* refers to a non-NP antecedent but to a proposition, demonstratives are easier to process (Çokal et al., 2016). The native speaker spoken corpus data also seem to be in line with these findings, as only 5% of the analysed non-NP antecedents have *it* as an anaphoric marker, instead giving preference to demonstrative pronouns (Gundel et al. 2008: 357).

According to Wittenberg et al. (2021:5), *it* tends to quickly connect to the first noun phrase that fulfils both its relational and classificatory criteria, while the demonstrative pronoun *that* serves as a ‘universal bundler’, facilitating access to a segment of conceptual structure, regardless of whether it is linguistically encoded. This suggests that not only *that* could occur exophorically, but it is also more conceptually complex compared to *it.*

2.5.2.2 Demonstratives in academic discourse

The use of demonstratives is also identified as being related to genre, suggesting that the distribution and frequency of demonstratives varies across different discourse contexts. For example, in their study Gray and Cortes (2011) analysed how demonstratives are deployed in academic prose across fields of Applied Linguistics (AL) and Materials and Civil Engineering (MCE). The study (ibid.) found that in AL articles, demonstratives used as determiners accounted for 79% of the analysed cases, while 21% were used as pronouns. In contrast, MCE articles showed results of 83% for demonstratives used as determiners and 17% used as pronouns. These findings, however, to an extent challenge the prescriptivist view on bare demonstratives, as American Psychological Association states that "Simple pronouns are the most troublesome, especially *this, that, these,* and *those* when they refer to a previous sentence. Eliminate ambiguity by writing, for example, this test, that trial, these participants, and those reports" (APA cited in Gray and Cortes, 2011:32)

However, the mere fact that deviations from prescriptive guidelines can be observed in academic writing does not inherently imply that writers struggle to achieve cohesion. In fact, several psycholinguistic studies previously mentioned in this subsection (Gundel et al., 1993; Wittenberg et al., 2021) have demonstrated the contrary.

2.5.2.3 Demonstratives in EFL learners’ production

Demonstratives have also been analysed in the context of L2 production. Lee et al. (2021) explored the usage of (un-)attended demonstratives in high and low-rated essays of Chinese EFL learners. The authors (ibid.) identified that the usage of unattended, or bare demonstratives, is more frequently observed in low-rated written production, compared to high-rated essays in which attended demonstratives are preferred. In terms of their syntactic environment unattended *this/these* precede copular verbs more often in low-rated essays, while high-rated essays are characterised by a wider range of verb types. Nevertheless, it is worth noting that in native speaker academic prose, the distribution of demonstrative-adjacent verbs (copular vs. lexical) has been shown to be field and topic-related rather than indicative of cohesiveness in written production (Gray and Cortes, 2011).

2.6 Summary of Literature Review

The literature review section introduced the notion of cohesion and its distinction from coherence, examining the main categories within cohesion. It focused primarily on referential cohesion and interrelated notions such as anaphora, anaphor, and antecedent, which constitute the central focus of this thesis.

The reviewed literature reveals a predominance of psycholinguistic studies on pronominal anaphora, though corpus-based studies examining cohesion across different genres have increasingly emerged in recent years.

For EFL learners, the acquisition of referential cohesion appears to be influenced by native language background, often manifesting in patterns of overexplicitness or underexplicitness. However, evidence suggests that inadequate usage of cohesive elements may also stem from insufficient institutional instruction on cohesive expression. This inadequate institutional coverage is exemplified by Quesada and Lozano's (2020) study, which found that zero anaphora is underused by non-native speakers of English, even though their native languages permit pronoun omission. While some cohesion recommendations appear in APA writing manuals, these tend to adopt a prescriptivist orientation that may not align with empirical findings.

Additionally, this section explored demonstratives and their interchangeability with the personal pronoun *it*, analysing how demonstratives function in written academic discourse and examining the relationship between demonstrative usage and EFL proficiency levels.

Chapter 3. DATA AND METHODS

The data of written L2 production is obtained from International Corpus of Learner English (ICLE, Granger et al., 2020) and its subcorpora LICLE (Lithuanian learner English) and NICLE (Norwegian learner English) As a reference tool, a corpus of NS essays, namely, Louvain Corpus of Native English Students (CECL, 2005) was used. The corpus of LOCNESS is selected due to its purpose of being used as a reference corpus in contrastive studies of learner English.

The predefined number of 300 cases of anaphoric *it* per each group was selected, resulting in total of 330 essays (LOCNESS 136; LICLE 130; 138 NICLE).

The essays were extracted from the corpora and uploaded into Sketch Engine (Kilgarriff et al., 2014), which facilitated the extraction of concordance lines. However, the extraction was partially conducted manually, as the tool does not differentiate between non-referential and referential instances of *it*. From 5280 concordance lines automatically generated by the tool, 900 were manually selected for the study sample.

This thesis employs an exploratory corpus-driven approach based on Contrastive Interlanguage Analysis (CIA); a methodology proposed by Granger (1996) designed specifically for contrastive research of learner corpus data.

The samples were analysed by extracting concordances with the Key Words in Context (KWIC) *it* and *its*. Although *its* is not the focus of this analysis, it was included because learners might mistakenly use *its* (possessive determiner) instead of *it's* (contraction of *it is*).

Prior to manually selecting concordance lines containing anaphoric *it*, a predefined list of instances not relevant to anaphoric expressions was selected. This list includes:

a) Extrapositional and impersonal *it*, as in: ***It's*** *ridiculous that they've given the job to Pat*.

b) The *it*-cleft construction, as in: ***It*** *was precisely for that reason that the rules were changed*.

c) Weather, time, place, condition, as in: ***It*** *is raining;* ***It*** *is very noisy in this room*; *I don't like* **it** *when you behave like this*.

d) *It* as subject with other predicative NPs, as in: a. ***It*** *was a perfect day. b. The day was perfect*.

e) *It* in idioms, as in: *Hold* **it**!; *We'll play* **it** *safe* (Huddleston and Pullum, 2016:1481–1483)

Additionally, cases of direct quotations were also excluded from the analysis.

To ensure a balanced sample, the concordance lines were retrieved from the middle sections of the essays, excluding the introductory and concluding paragraphs. Additionally, to minimise the potential influence of writers' idiolect, the number of concordance lines containing it was limited to a maximum of three instances per individual essay.

The data samples are coded for six categories as follows: **1. Antecedent** the specific discourse entity to which the anaphor is linked to. For example, the antecedent may be expressed by NPs such as *social prestige, literature, welfare system*, or by verbal elements encompassing both finite and non-finite clauses as in *[that] cultures become more and more similar to each other; [that] such a rapid migration decreased the number of jobless people; to have children* (infinitival phrase); *knowing the English language* (gerund phrase). The identification of the antecedent is an initial step in the analysis, as the specific linguistic expression linked to *it* is interrelated with subsequent categories, namely, the type of antecedent (nominal or verbal) and the distance between antecedent and anaphor, with the specific antecedent acting as a point of reference for measuring this distance. **2. The type of antecedent** is distinguished between the nominal type, which encompasses referring expressions that vary in terms of their degree of concreteness (e.g., *the house; the idea of Optimism*), denoting extra-linguistic entities, and the verbal type, which includes discourse-specific entities such as thoughts, facts, events, or propositions (e.g., *to have a free, independent, individualized nation*; *[whether] it is possible to have censorship without "harming" anyone*). Considering that the antecedent type linked to the anaphoric *it* signals native-like acquisition of referential cohesion, namely employing *it* mostly to point back to NP antecedents (Wittenberg, 2021; Gundel et al., 2008), it is expected that the distribution of antecedent types might yield different results in NS and NNS written production. An additional coding category is included for verbal antecedents, with subcategories being clause, gerund phrase, or infinitive phrase. **3. The distance between antecedent *it*** refers to the number of tokens between the antecedent and its anaphor. This category is coded considering the findings from NS production, namely that pronouns tend to occur relatively close to their antecedents to facilitate cognitive processing of the anaphora in question (Givón, 1983; Ariel, 1990; Biber et al., 2021). In cases where *it* is separated by another anaphor, as in (5), the anaphoric distance is counted from the immediately preceding anaphor, not from the initial antecedent. This approach is taken because reiteration of the same referent distributes the antecedent across the discourse, maintaining its salience.

> (5) *He shockingly writes about the sexually transmitted disease, syphillis. The docter who got* ***this disease*** *got it from a friar, who, after a series of events, received it directly from Christopher Columbus.* (LOCNESS: usprb1023)

4. Several *it* in succession marks whether the sample under analysis is a part of a chain where there have already been the same preceding pronoun in a sentence i.e. the pronoun *it, as in* (6)*:*

> (6) *She said that if that was the case then by right she owned* ***the house*** *because she cleaned* ***it*** *and kept* ***it*** *up.* (LOCNESS: usscu2014)

This category is interrelated with the preceding category *the distance between antecedent and anaphor* as the presence of several *it* in succession may contribute to an increased distance: "There can be a very large distance between the first antecedent in a chain and the final anaphor, greater than would typically be permitted for a direct link: it is the intermediate links that keep the referent salient in the context of discourse so that reference to it can be made employing a personal pronoun or other anaphor with little intrinsic content." (Huddleston and Pullum, 2016:1457) **5. Anaphora occurs in the same sentence** marks whether the anaphor occurs in the same sentence as its antecedent. This category is important because "pronouns favour a position where the antecedent occurs in the previous sentence" (Ariel 1990:18). **6. Subject of the following sentence** marks cases where the antecedent occurs in the preceding sentence; this category marks whether the anaphor occurs in the subject slot of the following sentence, that is, the subject of the main clause.

It should be noted that the spelling and grammar errors in the examples of learner writing are intentionally preserved to maintain the authenticity of the data. Each example is followed by a reference to the corpus and code of the text from which it is taken.

Chapter 4. RESULTS AND DISCUSSION

The present chapter overviews the empirical findings, organised into four categories as follows: the types of antecedents, the distance between antecedent, anaphor position in discourse and deviated use of anaphoras. Each section presents both quantitative and qualitative analyses of the respective category.

4.1 Types of Antecedents

In terms of their types, antecedents are divided into noun phrases (henceforth: NPs) and verbal types. The results are summarised in Table 1 below:

Table 1. Distribution of types of antecedents in raw frequency and percentage

	LOCNESS	LICLE	NICLE
NP	231(78%)	206 (68%)	206 (68%)
Verbal	69 (22%)	94 (32%)	94 (32%)

A chi-square test of independence was conducted to examine the relationship between the categories NP and verbal antecedents across the corpora of LOCNESS and LICLE and NICLE. The results indicated a significant association between the variables, χ^2 (1, *N* = 600) = 6.63, *p* = .010 which means that differences among the corpora are statistically significant, and the choice of antecedent is related to the variety of English represented by each corpus.

The following subsection of this thesis discusses the verbal antecedents in more detail, identifying the subtypes of the verbal antecedents and providing examples from empirical data.

4.1.1 Verbal antecedents

Verbal antecedents are found to be expressed by both finite and non-finite forms, the latter is further divided into infinitive and gerund phrases.

Table 2. Distribution of verbal antecedents in raw frequency and percentage

	LOCNESS	LICLE	NICLE
Finite	40 (58%)	59 (63%)	54 (57%)
Non-finite (gerund phrase)	21 (30%)	26 (27%)	23 (24%)
Non-finite (infinitive phrase)	8 (12%)	9 (10%)	17 (19%)

The distribution among verbal antecedents does not show statistically significant differences between NS and NNS corpora. The test comparing the data from corpora LOCNESS and NICLE yielded a chi-square value of *χ2(2,N*=163*)=1.62, p=*.445 while the comparison between LOCNESS and LICLE resulted in the following: χ2(2,*N*=163)=1.62, *p*=.445.

All three groups most often express verbal antecedents by finite clauses, as in (7), which are the most syntactically complex, followed by gerund phrases, as in (8), and infinitive phrases, illustrated in (9):

(7) ***We have airplanes and submarines****. This is practical, of course, and* ***it*** *makes it easier to get where you want to go* <…> (NICLE:NOOS1018)

(8) *Furthermore,* ***being a member of the EU*** *means changing some economic aspects. First of all,* ***it*** *concerns the shift from Lithuanian currency (litas) to a common currency (Euro).* (LICLE:LTVI2072)

(9) *Can we expert* ***a scientist to bear this additional burden for the whole world****? In truth no,* ***it*** is *unreasonable.* (LOCNESS:alev8021)

This subsection introduced to the subtypes of verbal antecedents, indicating that the distribution in the three corpora is not statistically significant and presented examples of each of the subtype. The verbal antecedents are also discussed in more detail in the section **4.2.2**.

4.2 Distance between antecedent and anaphor

Table 3. Distance between antecedent and its anaphor in tokens

NP antecedents			
	LOCNESS	**LICLE**	**NICLE**
Average distance	9	8.22	9.19
Max. distance	125	36	37
Min. distance	0	0	0
Verbal antecedents			
	LOCNESS	**LICLE**	**NICLE**
Average distance	7.22	5.27	7.91
Max. distance	33	28	123
Min. distance	0	0	0

Table 4. The ranges in antecedent-anaphor distance and number of instances in raw frequency and percentage

NP antecedents			
	LOCNESS	**LICLE**	**NICLE**
Low distance (0-10 tokens)	169 (73%)	146 (72%)	128 (64%)
Medium distance (11-51 tokens)	57 (25%)	57 (28%)	71 (35%)
High distance (51-125)	4 (2%)	N/A	2 (1%)
Verbal antecedents			
	LOCNESS	**LICLE**	**NICLE**
Low distance (0-7 tokens)	41 (61%)	65 (69%)	64 (68%)
Medium distance (8-15 tokens)	17 (25%)	25 (27%)	21 (22%)
High distance (16-45)	9 (14%)	4 (4%)	9 (10%)

In this thesis, the distance between anaphors and their antecedents is measured separately for NP and verbal antecedents. This distinction is made in order to find out whether the type of antecedent may influence the measured distance.

The results in Table 3 indicate that the average NP antecedent distance is similar across all groups (around 9 tokens for native speakers and Norwegian EFL learners, and slightly lower—8.22—for Lithuanian EFL learners). This result suggests that, on average, all writers locate NP antecedents relatively close to their anaphors, which is likely to facilitate anaphora resolution. The preference by all groups to locate anaphor close to its antecedent is also evident in the results provided in Table 4, indicating that the vast majority (64-73%) of NP antecedents occur separated from their anaphor by low distance of up to ten tokens. It is also notable that the minimum distance is 0 tokens in each corpus, indicating instances where the anaphor immediately follows the antecedent.

The maximum distance for native speakers (125 tokens) is substantially higher than for both Lithuanian (36 tokens) and Norwegian (37 tokens) EFL learners. This marked difference in the

LOCNESS corpus may suggest that the learners adopt different strategies for maintaining referential cohesion in discourse.

For verbal antecedents, native speakers average 7.22 tokens and Lithuanian EFL learners average 5.27 tokens. The slightly shorter average distance in Lithuanian EFL texts again suggests a preference for keeping antecedents close to their anaphors. Norwegian EFL learners, on the other hand, average 7.91 tokens, which may be indicative of NS like anaphor usage.

In terms of maximum and minimum distance, all three groups employ immediate links between the antecedent and its anaphor—as evidenced by a minimum distance of 0 tokens. However, the maximum distance varies notably among the groups. While native speakers and Lithuanian EFL learners keep the distance between the antecedent and its anaphor within the range of 28 to 33 tokens, Norwegian EFL learners demonstrate a significantly longer span, reaching up to 123 tokens.

The examples in the following subsection illustrate how referential cohesion is maintained at both maximum and minimum distances when the antecedents are NPs and when they are of a verbal type.

4.2.1 Maximum distance between an NP antecedent and an anaphor

Example (10) illustrates the most salient case in terms of antecedent-anaphor distance found in the analysed data set, demonstrating an antecedent-anaphor distance of 125 tokens.

> (10) ***"In Vitro fertilisation"*** *is the* ***fertilisation*** *of an egg in the laboratory ie. in a testube. The egg is taken from the mother and placed in an environment which will optimise the chances of* ***fertilisation*** *by the sperm from the father. Once* ***fertilisation*** *has occured the* ***fertilised*** *egg is implanted back into the mothers womb and from there on the pregnancy will be normal. Normally more than 1 egg is taken from the mother so that the eggs can be stored and used later if the pregnancy is unsuccesful or so that more than one can be* ***fertilised*** *at the same time to increase the chance of a succesful pregnancy. This usually leads to multiple births ie. twins, triplets etc. There are people who are agains this, saying* ***it*** *is not natural and is it fair to the child having started life in a test tube, as they believe life starts from the moment of conception.* (LOCNESS:alev8003)

Although such a large distance between antecedent and anaphor distance in (10) is considered outstandingly marginal in the data set among both in native speaker data and in learner English, the distance between the NP antecedent *"In Vitro fertilisation"* and *it* could be explained by the explicatory nature of the discourse segment the anaphora occurs in and the fact that the thematic continuity of fertilisation is retained throughout the entire discourse segment also being cohesively expressed by the usage of the noun *fertilisation* and the past participle counterpart *fertilised*.

Given that cohesive relation in a long antecedent-anaphor distance is possible does not inherently suggest that in cases where distances are lower, an anaphoric link is easier to be created.

> (11) *Secondly, politicians also claim that by making students pay for* ***their studies*** *they will increase the quality of education.* (1) *However, students disagree with* ***it****.* (2) *According to the association of Lithuanian students, first the quality should be offered to them and them they should be asked to pay for* ***it****.* (3) (LICLE: LTVI3005)

In (11) the distance between the antecedent and anaphor is 36 tokens, which is deemed medium in the sample, as indicated in Table 4. However, the attempts of the writer to keep the referent salient are very different compared to (10) For one thing, in (11) the NP antecedent overlaps with the clausal antecedent *[that] by making students pay for their studies they [politicians] will increase the quality of education*, which is anaphorically linked with *it* in the second sentence of the excerpt. Because the two anaphors are linked to different types of antecedents, possibly, the cohesive anaphor sequence is not created. Another aspect potentially decreasing the cohesion in (11) is the intervening discourse referent *quality* in the third sentence, which could be regarded as a potential antecedent for two reasons. First, *quality* is closer to *it,* although at the same time it could be argued that *pay for quality* is not an adequate collocation here, particularly because the verb pays for collocates with *their studies* in the preceding co-text. Nonetheless, the second reason why *quality* could be presupposed by *it* is the fact that the author mistakenly uses a plural antecedent for the singular anaphor *it*, resulting in a violation of number agreement in anaphora.

Similarly to (10), the thematic continuity is also evident in (12).

> (12) *Most of us have seen at least one or two* ***soap operas****, and consequently know what happens in every soap opera in the whole world.* (1) *The plot is predictable, all the characters are deceptive and hide terrible family secrets, and you can jump in in any episode and still understand everything without difficulty*. (2) *So why do people watch* ***it****?* (3) (NICLE: NOBE1025)

In (12), the NP antecedent-anaphor distance of 37 tokens is illustrated. The first sentence proposes that many people have seen soap operas in their lifetime and know what to expect from them. The second sentence, then, continues the topic of the predictability of the series. Finally, in the third sentence, the author questions why soap operas, despite their predictable nature, are watched by people. The anaphor *it* in (12) presupposes *soap operas* because the NP *soap operas* has been explicitly introduced and its characteristics discussed in the preceding sentences. The anaphoric link is further enhanced by the near synonyms *seen* in the first sentence and *watch* in the third sentence.

Considering the cases of maximum distance in each corpus, it could be stated that the greater number of tokens intervening between antecedent and anaphor does not inherently imply a violation in referential clarity, and this is evident in the data of NS. As illustrated in (10) and (12), the long and medium distances may be justified in case the topic throughout the discourse is maintained. However, as far as the case in LICLE is concerned, although the maximum distance between antecedent and anaphor is significantly lower than encountered in the LOCNESS corpus, the identification of a discourse referent might be hindered by such factors as overlapping referents contributing to potential referential ambiguity.

4.2.2 Maximum distance between a verbal antecedent and an anaphor

Having overviewed the anaphoric behaviour in maximum distances when the antecedent is of verbal type, this subsection provides with analysis of the verbal antecedents in high-distance discourse environments.

Referential cohesion is also observed to be maintained through contrasting concepts, as exemplified in (13).

> (13) *While* ***reading a piece of literature****, the audience is given more time to accept the knowledge with which it has been confronted.* (1) *Each reader continues at his own pace and may handle the passages in his own way.* (2) *Because* ***it*** *is a slower, more thought-provoking process than hearing the spoken word for instance* (3) <*...*> (LOCNESS:usprb1001)

The first sentence of (13) introduces the idea that reading provides more time to process the information, while in the second sentence, the author supports this idea by proposing that each reader can choose their own pace and manner in which they read a literary work. The third sentence then concludes by stating that *reading a piece of literature*, is a more cognitively challenging activity as opposed to *hearing the spoken word*. This contrasting idea is also expressed by resorting to the same verbal antecedent subtype, that is, a gerund phrase. The similarity in linguistic form may be a facilitating factor in handling the presupposition.

The referential cohesion may be reinforced by an explicit expression of an antecedent in combination with the lexical co-text the anaphor occurs in, as illustrated in (14).

> (14) *It can be stated* ***that the language of politics is very often the language of lies and illusions****.* (1) *The history of the world wars that devastated the world in the 20th century and continues to do that till today can serve as a good proof of* ***it****.* (2) (LICLE:LTVI1053)

In the first sentence of (14), the writer makes an explicit statement about language of politics being a language of lies and illusions while the lexical bundle a *good proof of* that immediately precedes *it* presupposes a verbal antecedent and the only possible candidate in the discourse segment is *that the language of politics is very often the language of lies and illusions.*

In contrast, referential cohesion may be hindered by the occurrence of multiple referents in a discourse segment and topical shifts, as in (15).

> (15) ***To treat those who already are experienced in the field of crime****; it takes more work to get them to start a different life. Maybe it is more to it than just to make them to quit criminal*

> *activities. Maybe their situation makes them criminals, but they really do not want to steal or do other kinds of criminal activities. They have to be given a new start, and they have to be given the necessary guide lines away from what they are used to do. And these guides should not just drop by once a week, they should be some sort of a person who gives support and look after them as much as possible. This of course costs money, and it is all up to the governments if they want to do anything about* ***it****.* (NICLE: NOAG1020)

In (15) a long-distance antecedent-anaphor of 123 tokens is illustrated. The theme of this discourse segment is the rehabilitation of criminals, which is expressed in discourse by the verb phrase '*To treat those who already are experienced in the field of crime*.' The segment then goes on to describe the potential feelings of criminals, and finally, the author expresses their own opinion on the rehabilitation matter; thus, a shift in topics is evident, and consequently, a cognitive shift for the reader may be expected. While the author provides a possible solution for criminal rehabilitation expressed by the clause *they should be some sort of a person who gives support and look after them as much as possible,* the latter becomes a potential presupposed item and the ambiguity is also created as it is no longer clear to which proposition the author expresses his opinion towards. The infinitival bundle *to do anything about* preceding the anaphor it however, presupposes that it is the overall situation that should be in question.

The examples discussed in this subsection demonstrate that non-immediate links between the verbal antecedents and anaphor may be established by contrasting ideas that are expressed by the same type of antecedent. Alternatively, the example from the LICLE subcorpus demonstrates that referential cohesion may be reinforced by the co-text immediately preceding the anaphoric *it.* However, when the distance between anaphor and antecedent is as long as 123 tokens, as in (15), the presupposed item and an anaphor may be intervened by several sentences that discuss a number of topics and include multiple discourse referents.

4.3 The position of anaphor in discourse

4.3.1 LOCNESS vs LICLE

In this subsection, the position of anaphor in discourse across the corpora LOCNESS and LICLE is compared. These findings, presented in Table 5, are interpreted considering other categories such as average distance between the antecedent and anaphor.

Table 5. Distribution of anaphor occurrences: Same vs separate sentence position

NP antecedents			
	LOCNESS	**LICLE**	**NICLE**
Yes	177 (77%)	139 (68%)	126 (87%)
No	54 (23%)	65 (32%)	76 (13%)
Verbal antecedents			
	LOCNESS	**LICLE**	**NICLE**
Yes	39 (57%)	34 (36%)	42 (45%)
No	30 (43%)	60 (64%)	52 (55%)

The results in the anaphor in are found to be statistically significant between LOCNESS and LICLE corpora, with the result, $\chi^2(1, N = 435) = 3.93$, $p = .048$.

However, the relation between anaphor position in discourse and antecedent-anaphor distance in cases where an NP antecedent occurs in the same sentence does not form a basis for the presumption that there are substantial differences in how NS and NNS locate anaphors in their texts, results indicating average distance 7.13 in LOCNESS and 7.47 in LICLE.

In cases where a verbal antecedent is positioned in a preceding sentence in reference to *it*, the average distance is found to be 16.90 for LOCNESS and 9.81 for LICLE. Such differences could be accounted for by the NS tendency to use more syntactically complex sentences compared to NNS, and consequently increasing the antecedent-anaphor distance, although this would require an additional analysis.

The separation of anaphora on a sentential level may a hindering factor if no explicit linguistic links, such as conjunctions are introduced, as in (16).

(16) ***The V-chip*** *is an electrical device that blocks out violent television shows.* (1) *Some people, such as the Senate Commerce Committee, are not satisfied with* ***it.*** (2) (LOCNESS; usscu3007)

In (16), an NP antecedent and an anaphor occur in different sentences, presenting two distinct topics. The first sentence involves the explication of the term the *V-chip,* which introduces the first topic into the discourse. While the second sentence introduces the counterpoint of the referent in question and does not provide an explicit link, such as a conjunction, to mark the contrast between the two topics.

It is also evident that referential cohesion may be enhanced even in the case of sentential separation of anaphor and antecedent, as in (17).

> (17) *In many ways* ***the money*** *creates an attitude of greediness among many sports players* .(1) *With this attitude it changes their whole outlook on life.* ***It*** *makes them think that they are better than other people and* ***it*** *also makes them think that they can get away with other things.* (2) (LOCNESS: usscu2008)

In (17) the NP antecedent *the money* is linked with the same anaphor *it* twice, with one intervening sentence separating the sentences in which non-referential *it* appears. However, due to the fact that in the first and third sentence of the segment, the topic of money in sports culture is retained and the first anaphoric *it* appears in the initial subject position, minimising the antecedent-anaphor distance, referential cohesion is maintained. As far as the second referential *it* is concerned, referential cohesion is enhanced by the same syntactic pattern, also employing the same verb *makes,* anaphorically linking *it* to the same referent in discourse. What is more, although the antecedent-anaphor distance is extended in relation to the second anaphoric *it,* with the distance at 31 tokens, which is deemed to be medium in the data set, but does not result in referential ambiguity as there are no other potential NP candidates expressed between the antecedent and anaphor.

On the other hand, separating an antecedent and an anaphor on a sentence level, may result in referential ambiguity, as evident in (18).

> (18) *So the statement that* ***education*** *is important is a fact* (1)*. But how to treat* ***it*** *is more complicated question which is dangerously tend to become a rethorical one, because of the big amount of opinions about* ***it****.* (2) (LICLE: LTVI3021)

In (18), an ambiguity between two NP antecedents can be observed. The first sentence of the excerpt introduces the extended NP, *the statement that education is important,* which is a potential candidate for the antecedent linked with both of the anaphors in the second sentence. Although the antecedent is more thematically salient than the NP antecedent *education,* which is embedded in the extended NP*,* yet *education* is more likely to be presupposed by *it.* The NP antecedent in the first sentence may be discarded based on the verb *treat* which is more likely to collocate with *education* than it is with *the statement.* As for the second *it,* the preference is also given to the antecedent *education* due to the limited potential propositions available in the context; people tend to form opinions about the concept itself rather than the statements about the concept.

Another relation is found between an antecedent not occurring in the same sentence as its anaphor and the placement of the anaphor in the following sentence. The cases where an anaphor is in a subject position of the main clause, usually occurring in a sentence initial position, account for 50% of the cases in LOCNESS, while the percentage in LICLE for the same conditions is 60%. This kind of anaphor placement in the subject slot suggests the attempt to keep the anaphor in the most adjacent position to the antecedent possible, and in this way ensuring cohesion by minimising antecedent-anaphor distance. This is especially prominent in cases where the antecedent-anaphor distance is 0 tokens. Within LICLE, such cases account for 11% of the occurrences in which an NP antecedent and the anaphor *it* appears in separate sentences; in LOCNESS, the corresponding figure is 7%. The immediate anaphoric links are illustrated in (19)-(22).

> (19) *The question seems to be profoundly complicated; nevertheless, it has already been answered in the recent issue of* ***Lietuvos Rytas****.* ***It*** *presents a survey with the cultural artifacts that Lithuanians consider to be the most representative of the Lithuanian national identity.* (LICLE: LTVI2065)
>
> (20) *Ronald Barther even introduced the term "readerly" and "writerly" which refer to the already mentioned roles of the reader. In addition, a very important phenomenon in modern poetry was* ***imagism****.* ***It*** *was started by Ezra Pound and later approved by ee. Cummings, Doolittle, etc.* (LICLE: LTVI1022)
>
> (21) *The flag that is currently over the state house is* ***the battle flag****.* ***It*** *is the flag that is red with blue bars, with stars in them, crossing through the center of it.* (LOCNESS: usscu2001)
>
> (22) *Through the first part of the play Oreste shows signs of innocent* ***bad faith****.* ***It*** *is innocent as he has no past experience.* (LOCNESS: brsur1007)

In cases where the anaphor occurs in the subject position of the following sentence, and the antecedent-anaphor distance is medium, cohesion is maintained by means other than short distance.

> (23) ***Television and magazines*** *have implanted in most peoples' minds that if a woman is not beautiful and thin, then in same way she doesn't measure up, therefore a lot of young girls are left feeling that they have to look or act a certain way if they want to fit in.* (1) ***It*** *also puts into the minds of young men that this is the way a young man should be, and that's who they should want.* (2) (LOCNESS:usscu2015)

Example (23) demonstrates the presupposition enhanced by the reiteration of lexical items. The first sentence of the excerpt discusses the negative impact television and magazines have on the population's perception of women and mentions that this type of media has embedded itself in the minds of most people. The second sentence then continues the topic of media influencing the mind by explicitly repeating the noun. While the repeated noun could facilitate anaphora resolution, establishing the anaphoric link might also be hindered by the violation of agreement in number as *it* is used to refer to a plural discourse entity.

In a similar manner to (23), lexical reiteration is employed in (24).

> (24) *Stories like Cinderella and Sleeping Beauty, show females waiting to be rescued by a man.* (1) ***This type of story*** *feeds to little girls that its okay to be dependent on a man, you don't have to be independent, especially if you are nice and sweet.* (2) ***It*** *also shows little boys that their role is not played in the house doing "girly" things.* (3) (LOCNESS: usscu2015)

In (24), the verb *show* in the first sentence reoccurs in the third sentence. However, the first and third sentences are not anaphorically linked. In the first sentence, the NP *Stories like Cinderella and Sleeping Beauty* are subsumed under the NP *this type of story* in the second sentence. Although the repetition of the verb *show* in the first and third sentences may create the impression of an anaphoric connection, such an interpretation would suggest a grammatical error, as there is no agreement in number between the NP in the first sentence and *it* in the third sentence.

In some ways, the lexical reiteration strategy in (23)–(24) is similar to that used in (25).

> (25) ***The agreement which was signed by politicians*** *claims that from the year of 2008 students should pay for their education.* (1) *The sum of money may vary from 2000 to 5000 litas.* (2) *To get such sums students will be offered to take loans.* (3) ***It*** *also says that 30% of the best students from every study programm will not have to pay*. (4) (LICLE: LTVI3005)

In (25), rather than repeating the exact verb or noun present in the preceding co-text, the author employs synonyms, namely, *claims* in the first sentence and *says* in the fourth sentence. Regarding antecedent–anaphor distance, they are separated by 35 tokens, which is considered a medium range. Moreover, the antecedent and anaphor are divided by two thematically distinct sentences: the second sentence discusses the price of studies, while the third sentence introduces the ways in which students can obtain funds for their tuition fees. Finally, the fourth sentence returns to the topic of the agreement's content. This combination of verbal synonyms in the first and fourth sentences, along with the broad topical range intervening between the antecedent and the anaphor, could potentially result in difficulties in anaphor resolution.

Referential cohesion may also be reinforced by positioning both the antecedent and its anaphor in subject positions in the separate clauses, as exemplified in (26).

> (26) *What has* **colonialism** *done to the countries whose linguistic and cultural existance would today be obscure without the intervenience of Great Britain?* (1) ***It*** *has robbed the nations of the opportunity to be unique.* (2) (LICLE: LTVI2037)

In (26), the interrogative form of the first sentence presupposes an answer, which is provided in the second sentence. The alignment of the antecedent and anaphor in the subject position, thus, reinforces the cohesion of the discourse.

A tendency to primarily use finite clauses as antecedents of *it* could be accounted for by writers attempting to navigate through difficult conceptual material within an essay and language economy principles. One example of such navigation is observed in the context of explication of ideas previously introduced into discourse. In case of LICLE, this explicatory function co-occurring with verbal finite antecedents is observed, each of the cases being explicitly marked with lexical bundles such as *it means, it may rather explained, it is due to the fact* such cases are illustrated in (27) – (29).

> (27) *These career women stay absolutely alone and have no one to take care of them when they become old. Furthermore, those feminists who have children usually bring up them alone.* ***It*** *means that they make their lifes really exhausting: they have to earn money and give education without any help.* (LICLE: LTVI2081)
>
> (28) *Males took more and longer turns in comparison to females, however, the amount of words did not determine one's power.* ***It*** *may be rather explained by individual features of a person's character.* (LICLE: LTVI2018)
>
> (29) *There are abstractions in Hopi that do not have equivalent terms in English.* ***It*** *is due to the fact that the these abstractions is a part of the Hopi speakers' vitalistic and animistic beliefs.* (LICLE:LTVI1063)

It is also noticeable that in all the examples above (27)-(28), the anaphor *it* occurs in a prominent syntactic position of subject, antecedent-anaphor distance kept at 0 tokens. By locating the anaphor at the beginning of the following sentence, the thematic focus is maintained and, thus, the reader may be directed back to the preceding proposition in case the latter is not completely retained.

Interestingly, the occurrence of *it* in explicatory contexts is found exclusively in the LICLE corpus, suggesting that the recurring phraseology might be influenced by culture-specific discourse organising strategies.

This subchapter has overviewed the strategies writers employ to keep their texts cohesive in cases where the antecedents and anaphors occur in different sentences. It was identified that writers resort to syntactic and lexical reiterations, the use of synonyms and immediately preceding antecedents, where the distance is 0 tokens. In addition, some problematic aspects of keeping the referent salient in discourse were identified, namely, embedded antecedents, as in (18) and the violations of agreement in number, as in (23).

4.3.2 LOCNESS vs NICLE

In terms of NP antecedent-anaphor position in discourse, the differences between NICLE and LOCNESS are statistically significant. The results of the chi-square test indicate a significant relationship between LOCNESS and NICLE in terms of anaphor placement, $\chi^2(1, N = 433) = 10.41$, $p = .001$ with the NNS group placing NP antecedents in preceding sentences in relation to *it* more frequently. The separation of anaphora on a sentential level also results in a larger average of antecedent-anaphor distance, with 13.42 tokens. In contrast, when the antecedent and *it* occurs in the same sentence, the average antecedent-anaphor distance is at 7,18 tokens. This suggests that in the case of antecedent-anaphor separation on a sentential level, the anaphor is not necessarily placed in the position most adjacent to the antecedent in the preceding sentence, which may increase the distance between antecedent and anaphor and therefore diminish the saliency of a referent. One of the ways to diminish the referential saliency is through thematic shifts, as illustrated in (30).

> (30) *But I want to stress one thing, in matters like these, abortion should no be used as contraceptives!* (1) *One serious matter, which once in a while, some poor girl or woman suffer from, is rape.* (2) *In this case I see only one solution, which is* ***abortion*** *and should not an issue at all.* (3) *A woman should not have to look into a rapist's eyes again...* (4) *Of course* ***it's*** *entirely her own decision, but you have to think about yourself in a situation like this, and also imagine what it would be like in the future, telling your child that the father is a rapist…* (5) (NICLE: NOOS1029)

In the first sentence of (30), the writer expresses their opinion towards abortion, arguing that it should not be a method of contraception. The second sentence discusses the potential sexual abuse one might encounter, in the third sentence then the solution after having conceived as a result of rape is suggested. In the fourth sentence, the topic switches and now pertains to the author's opinion that a woman should not encounter the sexual abuser again. Finally, in the fifth, the topic of abortion is brought back, and it should be presupposed that the decision in question is abortion.

However, it is also evident that anaphora separation on a sentential level may not always result in difficulty in anaphora resolution, as other cohesive elements might facilitate the understanding of what is being presupposed. For example, writers may resort to lexical reiteration, as in (31)

> (31) ***Censorship in Western society*** *today isn't so much about political restrictions, as it is in for example countries with dictatorship.*(1) ***It*** *is more about increasing quality of life for everyone.* (2) (NICLE: NOHE1001)

In (31), the preposition *about* is repeated in the first and second sentences. In addition, both the antecedent and the anaphor appear in the subject positions, resulting in syntactically comparable sentences.

The writers may also resort to syntactic reiteration to maintain referential cohesion, as demonstrated in (32).

> (32) *If **the real world** means everything that comes after studying, then there is no education singlehanded that is good enough.* (1) *If **it** means that the education have to prepare the work-situation, it still lack plenty.* (2) (NICLE: NOBE1003)

In (32), a syntactic pattern of a conditional clause is reiterated along with the verb *means* in the first and second sentences

Naturally, anaphoric expression is expected to be preceded by the NP, and not vice versa. Having that said, such a deviation is encountered in the study sample used for this thesis (1 case across all the corpora), and exemplified in (33).

> (33) *How could a girl/woman live on for the rest of her life knowing that her child is the result of a rape? The man she hates the most she has to face every day in her child's eyes. What is she going to tell her child when it comes to the fatherhood? This is something that we all can understand. What about the mother to be? Usually when it comes to abortion the girl is very young. **It** would certainly turn her life around and in some cases it would ruin the girl's life.When you're still in your teens there's a lot of things to try out, mistakes to be made and you're not fully developped. **A baby** needs full attention when it's born, and the girl of sixteen is not able to give her all that. Maybe in some cases the child is better off, because it's not wanted, and the mother can't support the child.* (NICLE: NOOS1027)

Although (33) illustrates a marginal case of anaphora, the example is meaningful in terms of presupposition. In (33), the author begins by considering a hypothetical situation of a pregnancy as a result of rape and the potential distress that a forced childbirth is associated with. Next, in relation to the latter topic the author introduces a new sentence in which they turn the attention to the NP *abortion* that becomes a salient candidate for an antecedent of *it*, which appears in a subject position of the following sentence. However, based on the author's attitude expressed in the preceding co-text, *abortion* would hardly *ruin the girl's life*, consequently, it is the NP *a baby* which is separated from the anaphoric marker by a whole sentence and antecedent-anaphor distance of 39 tokens.

This subchapter discussed how Norwegian EFL learners navigate through discourse when anaphoras occur in different sentences. It was found that to maintain a cohesive link between an antecedent and *it* resorted to lexical and syntactic reiteration. In addition, it was identified that thematic shifts may decrease the saliency of a referent.

4.4. Deviant usage of anaphoras

During the data analysis, three more aspects of anaphora usage were identified. One involves hypothetical antecedents, another pertains to inferred antecedents, and finally, the use of *it* when referring to plural entities. These categories are concisely explored in this subsection.

4.4.1 Hypothetical antecedents

A number of anaphoras in the analysed data set are found to be connected to hypothetical and vaguely indicated antecedents expressed by indefinite pronouns such as *something*, *everything,* the modifying adjective *certain,* or hypothetical clauses, for example:

> (34) *However, one conclusion might be met at this stage for writing to be amusing: a person writing a text on* ***a certain topic*** *must be interested in* ***it*** *and appropriate for* ***it*** *(for example, the topic of plumbers would not suit fashion editor).* (LICLE: LTVI1058)

> (35) *Universeties don't seem to be as serious as we think it should be.There are proffesors that are bad and that not really know what they are talking about or you can get those who think they know* ***everything*** *but does't know how to teach* ***it*** *to you.* (NICLE: NOAC1015)

It could be asserted that in (34), the antecedent *a certain topic* is not explicit as replacing the anaphor *it* with this NP would not result in a sensible proposition. The antecedent here therefore could be considered partially inferable; the noun *topic* is one part of the antecedent but instead of being modified by the adjective *certain,* it is inferred to be preceded by a determiner such as *that* or *the*.

In (35), the formal antecedent *everything* does not satisfy the presupposition triggered by the anaphor due to the indefiniteness of the noun.

It could be asserted that in Example 34, the antecedent *a certain topic* is not explicit as replacing the anaphor *it* with this NP would not result in a sensible proposition. The antecedent here therefore could be considered partially inferable; the noun *topic* is one part of the antecedent but instead of being modified by the adjective *certain,* it is inferred to be preceded by a determiner such as *that* or *the*. Such antecedents are relatively scarce - in LOCNESS they make up 1% of the analysed data, while in LICLE and NICLE - 5% and 4% respectively.

Another minor category will be presented in the following subsection of the thesis.

4.4.2 Inferred antecedents

A small proportion across all the corpora are found not to have an explicitly expressed antecedent, which conflicts with the idea that cohesive elements are explicit linguistic cues. One such example is given below:

> (36) *Opposing to himself Tomas Niurka explains that he has noticed the influence of working late hours on his quality of studies.* (1) *The focus from studies switched to work.* (2) *In this*

> *situation the job is in the first place because this is the main source of revenue to pay for studies.* (3) ***It** is a good example of forgetting the alternative to study good enough to avoid the possibility to be one of the 36 percent tuition fee paying ones.* (4) (LICLE: LTVY1010)

In (36) is only possible to infer the antecedent type; the phrase *a good example* in the fourth sentence presupposes that it is the entirety of the situation in the preceding co-text that is being discussed. Given that situations are discussed employing either finite or non-finite verb forms rather than NPs, the antecedent could be encapsulated based on the information provided in the sentences preceding the anaphor as follows: *Tomas Niurka's job taking precedence over his studies*

Similarly, in Example 26 the reader is trusted to derive the NP antecedent themselves.

> (37) *Go skiing into the beautiful nature with our little rucksack and a packed lunch to eat in the free. We still want this pure and simple things. Just to take your dog for a walk in the forest on a sunny afternoon. Sitting on the top of the mountain looking out over the sea, and just imagination how your life will be. Lying in the grass, looking up in the sky, just dreaming. We have not become robots, and I don't think we ever will.* ***It** will not be like it is in science fiction movies.* (NICLE: NOHO1042)

In (37),based on the fact that sentences preceding *it* discuss a variety of pleasant human experiences, the antecedent could be encapsulated by the nouns *life* or *the world.*

4.4.3 *It* in reference to plural entities

Even though cases in which violations in the number agreement of antecedent-anaphor are relatively scarce, considering the advanced language level of the learners, the ungrammatical usage of *it* might signal a certain degree of students having little awareness of the grammatical properties of *it.* Interestingly, *it* in reference to plural entities is also encountered in the NS data, and produced by seven individual writers, such cases are illustrated in (38) and (39) below:

> (38) *We learn about **the ways of African-American culture**; we denounce **it** at first, then we learn to accept **it** and finally we understand **it**.* (LOCNESS: usprb1009)

> (39) ***Children's drawings** are a good example of this. In these, verbal skills are minimized and focused mainly on the communication of the basic idea. These are very basic, but **it** is interpretable by nearly everyone in a precise, pictorial form.* (LOCNESS: usscu4007)

All the categories presented in this chapter may be a starting point for further research, specifically combining it with psycholinguistic methodologies to identify how such anaphoras are perceived by readers.

CONCLUSIONS

The present master thesis distinguished between two types of antecedents: NPs, which denote entities (varying from the most concrete to the most abstract), and verbal antecedents that are understood as highly abstract discourse elements. Verbal antecedents are further subdivided into finite and non-finite types, the former involving a verb that shows tense and the latter encompassing gerund and infinitive phrases.

While the vast majority of occurrences across all three data sets are NP antecedents, a statistically significant result is observed in the distribution of verbal antecedents, with both EFL learner groups more frequently relying on verbal antecedents for referential cohesion in comparison to native speakers.

The less frequent occurrence of verbal antecedents in native speaker data aligns with previous psycholinguistic and cognitive research on referential cohesion (Wittenberg et al., 2021; Çokal et al., 2016), suggesting that native speakers favor demonstratives when antecedents are of verbal type.

The distribution of antecedents in EFL learner writing, however, might be accounted for by native language background. In Lithuanian, the most prototypical equivalent for the English pronoun *it* is *tai* ('it'), which overlaps in only one function—denoting general phenomena previously introduced into discourse. As for Norwegian EFL learners, the reason for more frequent use of anaphoric *it* to point back to verbal antecedents is less prominent because the translational equivalent of *it* in Norwegian, namely, the pronoun *det* (‘it’), in Norwegian shares a strong resemblance to *it*, and the set of demonstrative pronouns in both languages is comparable. However, in terms of register, the Norwegian *dette* ('this') is mostly used in formal contexts (Holmes and Enger 2018:160). Given that the demonstrative pronoun *this* in English would be preferable when pointing back to verbal antecedents, the unexpected usage of *it* by Norwegian EFL learners in written production might suggest that *dette* ('this') is less established in Norwegian.

The maximal distance for NP antecedents in LOCNESS, LICLE, and NICLE was found to be 125, 36, and 37 tokens, respectively. The native speaker data showing a long distance of 125 tokens suggests that referent saliency can be maintained through thematic continuity and lexical reiteration despite considerable distance. In the case of the medium distance of 37 tokens presented in Norwegian EFL learners' data, the strategy for maintaining referential cohesion also pertains to lexical choices, namely the employment of near synonyms. The example from LICLE with the medium distance of 36 tokens demonstrates that referential cohesion may be hindered by an intervening anaphoric *it* that appears between anaphors in question, where the two instances of *it* are not co-referential.

The maximal distance for verbal antecedents in LOCNESS, LICLE, and NICLE was found to be 28, 33, and 123 tokens, respectively. In the example from the LOCNESS corpus, referential cohesion is maintained through contrasting ideas, both expressed by gerund forms. In LICLE, the anaphoric link is signaled by linguistic cues occurring immediately preceding the anaphor, namely a lexical bundle presupposing a verbal antecedent and the absence of other verbal candidates in discourse. The example from NICLE demonstrates that a long distance between an antecedent and its anaphor is a potentially hindering factor if the intervening text contains multiple propositions and the author expresses their opinion toward one of them.

The results regarding anaphor position in discourse were not found to be statistically significant between LOCNESS and LICLE corpora. It is evident that NP antecedents tend to occur in the same sentence as their anaphors (76% in LOCNESS and 68% in LICLE), while verbal antecedents appear in the same sentence as their anaphors in 56% of cases in LOCNESS and 36% in LICLE. Moreover, in cases where the anaphor is positioned in the following sentence in reference to its antecedent, the anaphor tends to occupy the subject slot of that sentence. Such positioning suggests an attempt to keep the anaphor in the most adjacent position to the antecedent possible, thereby ensuring cohesion by minimising antecedent-anaphor distance. It is also observed in a number of instances that Lithuanian EFL learners collocate the anaphor *it* with lexical bundles such as *it means, it may rather be explained*, and *it is due to the fact* when the antecedent is of finite subtype and occurs in the preceding sentence.

In terms of NP antecedent-anaphor position in discourse, the differences between NICLE and LOCNESS are statistically significant, with Norwegian EFL learners placing NP antecedents in preceding sentences in relation to *it* more frequently.

The present analysis also highlighted three additional aspects that could be subjects of further research. First, the fact that some antecedents are hypothetical and therefore could be presupposed only to an extent. Second, that implicit antecedents are possible, although this does not align with the idea that cohesive elements are explicitly expressed in discourse. Third, that singular pronoun anaphora is used in some cases, which violates agreement in number.

SUMMARY (in Lithuanian)

Šiame magistro darbe nagrinėjami referencinės kohezijos modeliai, pasireiškiantys per angliškojo anaforinio įvardžio „it" vartojimą rašytinėje anglų kalboje. Tyrimas skirtas palyginti gimtakalbius anglų kalbos vartotojus su lietuvių ir norvegų anglų kalbos kaip antros kalbos (K2) besimokančiaisiais. Tyrime išskirtos dvi antecedentų rūšys – daiktavardinės frazės (DF), žyminčios konkrečius ir abstrakčius objektus, bei veiksmažodinės konstrukcijos, atspindinčios labai abstrakčius diskurso elementus. Analizę sudaro jų pasiskirstymo, nutolimo nuo įvardžio ir diskursinio pozicionavimo tyrimas trijuose tekstynuose (LOCNESS, LICLE ir NICLE).

Rezultatai atskleidžia, kad nors DF antecedentai vyrauja visose imtyse, K2 besimokantieji statistiškai reikšmingai dažniau remiasi veiksmažodiniais antecedentais, palyginti su gimtakalbiais. Šis skirtumas atitinka ankstesnius psicholingvistinius tyrimus, rodančius, kad gimtakalbiai anglų kalbos vartotojai nurodydami veiksmažodinius antecedentus teikia pirmenybę parodomiesiems įvardžiams (Wittenberg et al., 2021; Çokal et al., 2016). Šiam skirtumui įtakos turi gimtoji kalba – lietuvių kalbos „tai" sutampa su anglų kalbos įvardžiu „it" tik žymint bendruosius reiškinius, o norvegų gimtakalbiai gali nepakankamai dažnai remtis „dette" („this"/„tai") dėl šio įvardžio pirminio formalaus registro statuso norvegų kalboje.

Antecedento-anaforos atstumai ženkliai skyrėsi – gimtakalbiai išlaikė teksto rišlumą netgi esant ilgesniems DF antecedentų intervalams (iki 125 žodžių) per teminį tęstinumą ir leksinį kartojimą. Pozicionavimo atžvilgiu, NP antecedentai dažniau pasirodydavo tame pačiame sakinyje kaip ir jų anaforiniai įvardžiai (76% LOCNESS, 68% LICLE) nei veiksmažodiniai antecedentai (56% LOCNESS, 36% LICLE). Anglų K2 vartotojų _norvegų duomenys parodė statistiškai reikšmingus skirtumus lyginant su gimtakalbiais, dažniau rašydami DF antecedentus sakiniuose, einančiuose prieš „it" („tai").

Magistro darbe taip pat nustatomos sritys tolimesniems tyrimams: hipotetiniai antecedentai, numanomi antecedentai ir skaičiaus derinimo neatitikimai vartojant vienaskaitinį anaforinį įvardį.

Raktiniai žodžiai: referencinė kohezija, įvardžiai, anglų K2, anafora, gretinamoji tarpukalbės analizė, gretinamoji analizė

DATA SOURCES

ICLE

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020) The International Corpus of Learner English. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

LOCNESS

Centre for English Corpus Linguistics. (2005). Louvain Corpus of Native English Essays (LOCNESS). Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/locness.html

Sketch Engine

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: ten years on. *Lexicography*, 1: 7-36, 2014.

REFERENCES

Altenberg, B. & M. Tapper. 1998. The use of adverbial connectors in advanced Swedish learners' written English. In Learner English on computer, ed. S. Granger, 80-93. London: Longman.

American Psychological Association. 2001. Publication manual of the American Psychological Association. Washington, DC: American Psychological Association.

Ariel, M. 1990 *Accessing noun-phrase antecedents*. London: Routledge.

Asher, N. 1993. *Reference to abstract objects in discourse*. Dordrecht: Kluwer Academic.

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. 2021. *Grammar of spoken and written English*. John Benjamins Publishing Company.

Bikelienė, L. 2012. *Connector usage in native and non-native learners' English writing: Contrastive Analysis: Summary of doctoral dissertation: Humanities, Philologie (04 H)* (dissertation). *Connector usage in native and non-native learners' English writing: contrastive analysis: summary of doctoral dissertation: humanities, philologie (04 H).* Vilnius University, Vilnius.

Chierchia, G. 1995. *Dynamics of meaning: Anaphora, presupposition, and the theory of grammar*. Chicago: University of Chicago Press.

Connor, U. 1984. A study of cohesion and coherence in English as a Second language students' writing. *Paper in Linguistics*, *17*(3), 301–316. doi:10.1080/08351818409389208

Çokal, D., Sturt, P., & Ferreira, F. 2016. Processing of*it*and*this*in written narrative discourse. *Discourse Processes*, *55*(3), 272–289. doi:10.1080/0163853x.2016.1236231

Crossley, S. A., Kyle, K., & Dascalu, M. (2018a). The tool for the automatic analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*(1), 14–27. doi:10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16. doi:10.1016/j.jslw.2016.01.003

Demol, A. (2007). Accessibility theory applied to French: The case of il and Celui-Ci. *Folia Linguistica*, *41*(1–2), 1–35. doi:10.1515/flin.41.1-2.1

Dietrich, S., Seibold, V. C., & Rolke, B. (2024). Discourse comprehension and referential processing: Effects of contextual distance and semantic plausibility on presupposition processing. *Language and Cognition*, *16*(4), 2032–2054. doi:10.1017/langcog.2024.45

Díaz-Negrillo, A., & Espinola Rosillo, M. C. (2024). Mode of production and (referential) cohesion: An L2 english corpus-based study of Syntactic Coordination. *System*, *121*, 103247. doi:10.1016/j.system.2024.103247

Geluykens, R. 1994. *The pragmatics of Discourse Anaphora in English evidence from conversational repair by Ronald Geluykens*. Berlin: Mouton de Gruyter.

Givón, T. 1983. *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam u.a: Benjamins.

Granger, Sylviane. 1996. From CA to CIA and Back: An Integrated Contrastive Approach to Computerized Bilingual and Learner Corpora. In: Karin Aijmer, Bengt Altenberg & Stig Johansson (eds.), *Lund Studies in English 88*: *Languages in Contrast. Text-based crosslinguistic studies, 37–51*. Lund: Lund University Press.

Granger, S., & Tyson S. 1996. Connector usage in the English essay writing of Native and Non-native EFL speakers of English. *World Englishes*, *15*(1), 17–27. doi:10.1111/j.1467-971x.1996.tb00089.x

Gray, B., & Cortes, V. 2011. Perception vs. evidence: An analysis of this and these in academic prose. *English for Specific Purposes*, *30*(1), 31–43. doi:10.1016/j.esp.2010.06.004

Gundel, J. K., Hedberg, N., & Zacharski, R. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274. doi:10.2307/416535

Halliday, M. A. K., & Hasan, R. 1976. *Cohesion in English. M.A.K. Halliday and Ruqaiya Hasan*. London: Longman.

Holler, A., & Suckow, K. 2016. *Empirical perspectives on anaphora resolution*. Berlin/Boston: De Gruyter.

Holmes, P., & Enger, H.O. 2018. *Norwegian: A comprehensive grammar*. Abingdon, Oxon: Routledge.

Hinkel, E., 2001. Matters of cohesion in L2 academic texts. *Applied language learning, 12*(2), 111-132.

Huang, Y. 2000. *Anaphora: A cross-linguistic study*. Oxford: Oxford University Press.

Huddleston, R., & Pullum, G. K. 2016. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Lee, J. J., Tytko, T., & Larkin, R. 2021. (un)attended this/these in undergraduate student writing: A corpus analysis of high- and low-rated L2 writers. *Journal of English for Academic Purposes*, *50*, 100967. doi:10.1016/j.jeap.2021.100967

Liu, D. 2023. *Abstract entity anaphora in argumentative texts: Pragmatic features and Referent Interpretation*. Singapore: Springer.

Liu, M., & Braine, G. 2005. Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*(4), 623–636. doi:10.1016/j.system.2005.02.002

Loáiciga, S., Guillou, L., & Hardmeier, C. 2017. What is it? Disambiguating the different readings of the pronoun ‘it.’ *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/d17-1137

Lyons, J. 1977. *Semantics* (Vol. 2). Cambrigde: Cambridge University Press.

McGee, I. 2019. Macro- and micro-linguistic management of the argumentative essay: Implications for teaching. *Educational Studies*, *46*(6), 640–657. doi:10.1080/03055698.2019.1627661

Mitkov, R. 2016. *Anaphora resolution*. London: Routledge.

Morris, J., & Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, *17*(1), 21-48.

Poesio, M., Yu, J., Paun, S., Aloraini, A., Lu, P., Haber, J., & Cokal, D. 2023. Computational models of Anaphora. *Annual Review of Linguistics, 9*(1), 561–587. doi:10.1146/annurev-linguistics-031120-111653

Quesada, T., & Lozano, C. 2020. Which factors determine the choice of referential expressions in L2 English discourse? *Studies in Second Language Acquisition*, *42*(5), 959–986. doi:10.1017/s0272263120000224

Quirk, R., Greenbaum, S., Leech, G.N. and Svartvik, J., 1972. A grammar of contemporary English.

Ramonienė, M., Pribušauskaitė, J., Ramonaitė, J. T., & Vilkienė, L. (2020). *Lithuanian: A comprehensive grammar*. London: Routledge, Taylor & Francis Group.

Reid, J. 1992. A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing*, *1*(2), 79–107. doi:10.1016/1060-3743(92)90010-m

Ryan, J. 2015. Overexplicit referent tracking in L2 English: Strategy, avoidance, or myth? *Language Learning*, *65*(4), 824–859. doi:10.1111/lang.12139

Safir, K. 2004. *The syntax of anaphora Ken Safir*. Oxford: Oxford University Press.

Saner, L. D., & Hefright, B. 2015. Cross-cultural differences in linguistic reference tracking. *Procedia Manufacturing, 3,* 4022–4027. doi:10.1016/j.promfg.2015.07.969

Schneer, D. 2014. Rethinking the argumentative essay. *TESOL Journal*, *5*(4), 619–653. doi:10.1002/tesj.123

Tejada, M. Á., Gallardo, C. N., Ferradá, M. C., & López, M. I. 2015. 2L English texts and cohesion in upper CEFR levels: A corpus-based approach. *Procedia - Social and Behavioral Sciences*, *212*, 192–197. doi:10.1016/j.sbspro.2015.11.319