Vilnius University

Faculty of Philology

Department of English Philology

Edminas Šmergelis

Analysing collocations by association measures: the case of native and non-native written English

Thesis submitted in partial fulfilment of requirements for the degree of BA in English Philology

Supervisor: Assoc. Prof. Dr Rita Juknevičienė

Vilnius, 2024

Abstract

The present paper investigated the use of adjective-noun collocations in academic essays written by native and non-native speakers of English. Two association measures, Mutual Information (MI) and t-score, were used to determine the association strength of collocations. The findings revealed that non-native speakers utilize high-frequency collocations (attested by high t-scores) comparably or even more frequently than native speakers, but they significantly underuse strongly associated formulas (attested by high MI values). The present research was carried out in order to replicate the study conducted by Durrant & Schmitt (2009), using a different dataset. The results of the present study mostly align with Durrant and Schmitt's findings, however, it was observed in their study that the non-native speakers' reliance on high-frequency collocations was notably amplified by the repetition of favoured items.

Keywords: collocation, association measure, Mutual Information, t-score, native speakers, non-native speakers, formulaic language.

Table of Contents

1. Introduction

The inception of large corpora has had a profound impact on many subfields of linguistics. It has provided researchers with vast amounts of linguistic data, enabling them to conduct empirical studies on a scale that was not possible before. Systematic analysis of corpora has also proved to be useful in the identification of various recurring lexical patterns, most prominently sequences of formulaic language.

Many linguists have recognized the formulaic nature of language. It has been suggested that formulaic lexical sequences in large part determine language fluency (Schmitt 2004:4) and play a major role in language processing and use (Nesselhauf 2005:1). Kjellmer (1990:112) has noted that formulaic combinations of words constitute a large part of our mental lexicon. Some researchers have gone as far as to claim that at least one-third to one-half of language is comprised of formulaic lexical units (Erman and Warren 2000:37; Foster 2001:85). Considering how prevalent formulaic sequences are, it would only be natural to assume that the acquisition of such lexical units would be of crucial importance in second language learning. Some linguists have even argued that both first and second language learning essentially depends on the memorisation and analysis of pre-fabricated sequences of language (Ellis 1996:93). And while research has shown that formulaic units are quite abundant in the language of native speakers, some scholars have suggested that this may not be the case for non-native speakers. According to Kjellmer (1990:125–126), the mental lexicon of even the most advanced non-native speakers tends to lack pre-constructed word sequences. Instead, it would appear that non-native speakers tend to rely more on what Sinclair (1991:109) calls the ‘open-choice’ principle, i.e., they primarily construct their utterances word-by-word. Because of this, second language learners may face a series of potential challenges when it comes to fluent and effective language use.

Intuitively, the claim that second language learners might underuse formulaic sequences when compared to native speakers, seems to be correct. However, without empirical data it would be difficult to come to any sensible conclusion regarding this issue. Fortunately, considerable research has already been conducted to investigate the use of formulaic sequences in native and non-native speaker language. Some early research regarding this issue includes studies conducted by Granger (1998a), who focused on native and non-native speaker texts, and Foster's (2001) study, which investigated native and non-native speaker speech.

Granger’s (1998a) study aimed to substantiate Kjellmer’s (1990) assertion regarding non-native speakers' underutilization of pre-constructed lexical sequences. Granger compared

native speaker English writing and texts written by French-speaking learners of English in order to investigate whether learners would make less use of prefabricated sequences in their writing than their native speaker counterparts. The results did indeed confirm that advanced leaners use formulaic combinations to a much lesser extent than native speakers of English.

The study conducted by Foster (2001), analysed oral language productivity of native and non-native speakers in a task-based context. More specifically, the study aimed to investigate the extent to which native and non-native speakers utilized fully or partially pre-made sequences of formulaic language during an interactive task, and whether the provision of planning time before the task began had any significant impact on the degree to which such language was called upon. The results of the study showed that the frequency of formulaic language was overall much lower in non-native speaker data, when compared to native speaker data. Furthermore, it was shown that planning time had no discernible impact on the degree to which non-native speakers used lexicalized sequences.

These studies offer some valuable insight and provide empirical evidence to support the claim that native and non-native speakers make use of formulaic language to different extents. However, one potential shortcoming of these studies stems from the way formulaic language was identified in each case. For instance, Granger (1998a) analysed all combinations that had a particular grammatical form, regardless of whether those combinations could be considered formulaic or not. In the study conducted by Foster (2001), native speaker intuition was taken into account when determining which sequences were formulaic. What may potentially be problematic about these two approaches is the fact that they do not offer any concrete way of defining formulaic language. In general, many of the early studies that analysed formulaic sequences in native and non-native language have relied mostly on traditional linguistic criteria of semantic non-compositionality and syntactic fixedness to identify pre-fabricated lexical units. However, in more recent times researchers dealing with this particular phenomenon have relied on quantitative methods to identify pre-constructed phrases.

The study conducted by Durrant and Schmitt (2009) was one of the first to address this shortcoming by employing a more frequency-based approach to formulaic language analysis. The focus of their study were collocations. They were identified on the basis of their association scores in a large native reference corpus, which in this particular study was the *British National Corpus* (henceforth: the BNC). The two association measures that were calculated were Mutual Information (MI) and t-scores. MI tends to highlight word combinations that are less common, but whose constituent words are not often found apart, while t-scores tend to show very frequent collocations, usually composed of high-frequency words (Stubbs 1995:33-34). Durrant and

Schmitt (2009:160) note that many high-frequency collocations are neither semantically opaque nor frozen in form. This means that the analysis of high-frequency collocations would include not only phraseological sequences that meet the Traditional linguistic criteria of non-compositionality and full-fixedness, but also combinations that in the quantitative sense are still formulaic. The data for their study consisted of essays written in English by native speakers, which were compared against essays written by Turkish EAP students and essays from the Bulgarian sub-corpus of the *International Corpus of Learner English* (ICLE) (Granger et al. 2020). The paper centred around the analysis of premodifier-noun word pairs in terms of their t-scores and MI. The specific approach which was employed in Durrant and Schmitt's study allowed for the reveal of a more nuanced picture of formulaic language usage in native and non-native texts. Their results showed that non-native writers relied more heavily on high-frequency collocations (attested by high t-score), like *good example*, *long way*, and *hard work*, but that they underused less frequent, strongly associated combinations (attested by high MI), like *densely populated*, *bated breath*, *preconceived notions*, etc.

Many recent studies concerned with formulaic sequences in non-native speaker language seem to employ a similar strategy to the one devised by Durrant and Schmitt (2009). For instance, Granger and Bestgen's (2014) study also presents an analysis of collocations in terms of their association measures. However, instead of investigating the differences between native and non-native language, their focus shifts towards phraseological competence at different proficiency levels. Their data sample consisted of essays written in English by German, French and Spanish speakers. The essays were extracted from the International Corpus of Learner English and then assigned a rating based on the demonstrated level of proficiency. Results of the study showed that intermediate learner texts contained a smaller number of lower-frequency, but strongly-associated collocations than advanced learner texts and a higher proportion of high-frequency collocations.

Other studies have also analysed collocations that were produced or processed by Lithuanian learners of English, though such research remains relatively sparse compared to the abundance of studies focusing on learners of English from other linguistic backgrounds. Juknevičienė (2008) analysed collocational competence of Lithuanian learners of English. The study investigated learners' ability to produce collocations with high-frequency verbs, i.e. HAVE, DO, MAKE, TAKE and GIVE, to then compare it with data from a comparable corpus of native speakers of English. Another study, conducted by Vilkaitė and Schmitt (2017) approached the issue from a psycholinguistic point of view. Their research investigated whether advanced non-native speakers show processing advantages for non-adjacent collocations to the

same or similar degree as native speakers do. Participants came from a variety of linguistic backgrounds, including Lithuanian.

The studies that analysed this phenomenon seem to present consistent results, showing that non-native speakers generally tend to underuse formulaic language, when compared to native speakers. However, different approaches employed in these investigations seem to reveal this tendency in different capacities.

The present paper attempts to replicate the study conducted by Durrant and Schmitt (2009), using a different dataset. The purpose of this is to verify whether the results of the present analysis are consistent with the findings of their study. Additionally, Durrant and Schmitt's research focused on premodifier-noun combinations. This includes both non-noun and adjective-noun collocations, which were analysed together. Some researchers have claimed that, because of the phraseological difference between noun-noun and adjective-noun collocations, it may be worth analysing the two types of combinations separately (Granger and Bestgen 2014:234). Thus, the present study will only investigate adjective-noun collocations in native and non-native speaker texts.

The following research questions have been raised in this study:

RQ1: To what extent do texts produced by native and non-native users of English differ in terms of collocations?

RQ2: How do the results of the present analysis compare to the findings of Durrant and Schmitt's (2009) study?

RQ3: Which association measure, i.e. t-score or MI, is a better indicator of differences in the use of collocations by native and non-native speakers?

The remainder of the paper will be organized in four parts: Data and Methods, Results, Discussion and Conclusion. Data and Methods will present a detailed summary of the corpora used for the study, criteria set for the selection of native and non-native essays, description of the methodology employed for the identification and analysis of collocations, the type of collocations that will be analysed, and definitions of the relevant association measures that will be used in the analysis. The Results section will present the findings of the study, while the Discussion section will include the interpretation of those findings. The implications of the results will also be discussed and compared to the findings of previous research. Lastly, the Conclusion section will summarize the study and provide some ideas for future research.

2. Data and Methods

The purpose of the present study is to compare the use of collocations in native and non-native texts. More specifically, the study only focuses on directly adjacent adjective-noun collocations used in academic essays.

The methodology applied to this research project has been adapted from the study conducted by Durrant and Schmitt (2009). Their approach centred around the frequency-based analysis of collocations. In this approach, collocation can be defined as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991:7). In other words, collocations are comprised of words that appear together more frequently than their individual frequencies would predict.

2.1. Native and non-native speaker corpora

Two corpora have been complied for this study: a native speaker corpus and a non-native speaker corpus. The native speaker corpus is comprised of English essays extracted from the *Louvain Corpus of Native English Essays* (LOCNESS) (Granger 1998b). Native speaker essays represent only British English variety, essays written in US English were not included. The non-native speaker corpus is comprised of English essays extracted from the *Lithuanian component of the International Corpus of Learner English* (LICLE) (Granger et al. 2020). Each compiled corpus consists of 30 essays. The texts were selected mostly randomly, while only briefly overviewing the general subject matter of each essay in order to ensure diversity of topics. The number of words and extracted collocations in the native speaker corpus differs quite significantly from the non-native speaker corpus, with the former consisting of 30,542 words while the latter is comprised of 17,753 words. The average length of an essay in the native speaker corpus is approximately 1,018 words, while the average length of an essay in the non-native speaker corpus is approximately 591 words. As many authors have pointed out, it is quite difficult to identify native texts that would be equivalent in type and similar in length to non-native speaker texts (Durrant and Schmitt 2009:162; Granger et al. 2020:13; Lorenz 1999:14). To combat this, Durrant and Schmitt opted to distinguish between texts of different lengths. The complete dataset of their study was divided into shorter native and non-native texts that were analysed separately from extended native and non-native writing. However, this procedure yielded very similar results with the same tendencies ultimately emerging from both analyses. Still, the distinction between essays of different lengths would nevertheless contribute to the overall accuracy of the analysis. However, because of the time constraints and limited

scope of the present study, no attempts to differentiate between texts of different lengths were made.

2.2. Association measures

The two key metrics that were used to determine the collocational strength of retrieved word pairs are *Mutual Information* (MI) and *t-score*. These statistical measures tend to highlight different things. Mutual information compares the probability of two words occurring together with the probabilities of those words occurring independently (Church and Hanks 1990:23). High MI typically indicates a strong collocation whose constituent words are less frequently found apart. Phrases like *tectonic plates* and *immortal souls* are usually given as examples of strong collocations with a high MI score. T-score usually indicates high-frequency collocations whose constituent words are also highly frequent. Examples of collocations with high t-scores usually include pairs such as *hard work* and *good example*.

MI and t-scores of each collocation were retrieved from the *British National Corpus* (The BNC). The BNC is a 100-million-word corpus of British English that covers both spoken and written language. It has been a very popular choice in contemporary linguistic research ever since its release. The BNC has been utilized in many previous studies that deal with a similar subject matter, including the one conducted by Durrant and Schmitt. Because of this, it was selected as a reference corpus for this project.

2.3. Collocations analysed

As was mentioned earlier, only directly adjacent adjective-noun phrases were analysed. Phrases that contained proper nouns, acronyms, pronouns, possessives, semi-determiners, numbers/ordinals were removed from the final dataset, just as they were in Durrant and Schmitt's study. Additionally, many of the retrieved word pairs had extremely low association values. Some scholars have suggested that phrases with t-scores lower than 2 and/or MI values lower than 3, might not even qualify as collocations (Hunston 2002; Stubbs 1995), mostly due to the lack of semantic restriction, and in the study conducted by Durrant and Schmitt, such collocations were excluded from the analysis outright. Thus, in order to achieve more comparable results, only adjective-noun phrases with t-scores of 2 or higher and/or MI scores of 3 or higher were analysed in the present study.

556 collocation tokens (i.e. both unique and repeated collocations) that met the aforementioned criteria were extracted from the native speaker corpus and 310 collocation tokens were retrieved from the non-native speaker corpus. The native speaker dataset contained 116 collocations that were repeated items, while the non-native speaker dataset contained 63

repeated collocations. These identical combinations were not completely excluded from the analysis, in order to investigate the extent to which favoured collocations are repeated in native and non-native texts. When identical combinations are removed, the number of collocations extracted from the native speaker corpus is reduced to 440 types (i.e. exclusively unique collocations), while the number of collocations extracted from the non-native speaker corpus is reduced to 247 types.

2.4. Procedure

As Durrant and Schmitt (2009:168) have pointed out in their study, a simple binary distinction between word pairs that qualify as collocations and those that do not would hardly be satisfactory, as different collocations tend to have vastly different association scores. Durrant and Schmitt's solution was to rank collocations across a scale of association strength. In their study, collocations were distributed across 7 bands of t-scores, in the following way:

$t = 2–3.99$; $t = 4–5.99$; $t = 6–7.99$; $t = 8–9.99$; $t = 10–14.99$; $t = 15–19.99$; $t \geq 20$

Collocations were also distributed across 8 bands of MI, as follows:

$MI = 3–3.99$; $MI = 4–4.99$; $MI = 5–5.99$; $MI = 6–6.99$; $MI = 7–7.99$; $MI = 8–8.99$; $MI = 9–9.99$; $MI \geq 10$

One thing to note regarding this model of categorization, is that a significant number of collocations that comprise the dataset have either a high enough t-score value ($t \geq 2$) or a high enough MI ($MI \geq 3$) but not both. This is important, because the currently defined bandings clearly neglect those collocations that have a high enough value in only one of the two association measures. The simplest solution to this problem would be to only analyse collocations that have both a high enough t-score and a high enough MI, but this would further reduce the size of the dataset and would neglect combinations that according to the pre-set minimal association requirements should be considered collocations. An alternative solution that was chosen for the present study was to introduce two additional bands of association scores: $t < 2$ and $MI < 3$ to account for collocations that had a low score in one or the other association measure. This was also done to ensure that the distribution across all bands sums up to a 100%.

It is also noteworthy that in Durrant and Schmitt's study the word combinations were extracted manually from a corpus. In the present study, the free corpus toolbox *LancsBox X* (Brezina and Platt 2023) was used to significantly streamline this process. The native and non-native speaker texts were uploaded onto LancsBox X and compiled into separate corpora, which

were then tagged for parts-of-speech by one of the available grammatical taggers. From there, a list of relevant word combinations was automatically compiled and extracted. This process is arguably faster than manual labour, however, raw human input could potentially ensure more accurate results, provided that human error can be avoided. Nevertheless, most modern grammatical tagging algorithms provide sufficient precision in the identification of parts-of-speech. Additionally, automatization of this process opens the possibility of analysing much greater volumes of data and has already been applied in a number of recent studies.

3. Results

This section presents a detailed description of the findings of the present analysis, including the calculated association score averages, summary of collocation repetition in native and non-native essays and the distribution of collocations across different levels of association strength. Collocation types and collocation tokens were investigated separately.

3.1. Association score averages

First, the association score averages were calculated. Though this procedure only has the capacity to represent the differences that exist between native and non-native speaker essays in a fairly general way, it is nevertheless an essential preliminary step in the analysis. When repeated items are excluded, the average t-score of collocations found in native speaker essays is lower than the average t-score of collocations found in non-native speaker texts, however, when average MI scores from the two datasets are compared, an opposite trend emerges, with the average MI of collocations in the native speaker corpus being higher than the average MI of collocation in the non-native speaker corpus. The average t-score of collocations found in the native speaker corpus is 8.07, while the average MI is 5.56. In contrast, the average t-score of collocations found in the non-native speaker corpus is 8.59, while the average MI is 4.96. For better reference, examples from the present study of collocations with low t-score would include items such as *religious theme* (t = 2.13), *good will* (t = 2.24), *human kind* (t = 2.55), *expensive car* (t = 3.84) while collocations like *human nature* (t = 20.97), *common law* (t = 33.56), *long time* (t = 57.39), *last year* (t = 94.6) could be considered as having a high t-score. As for MI, *great feeling* (MI = 3.04), *common good* (MI = 3.17), *established order* (MI = 3.39), *perfect world* (MI = 3.54) would be examples of collocations with relatively low MI, while *integral part* (MI = 10), *vivid illustration* (MI = 10.32), *instant gratification* (MI = 12.07), *divine providence* (MI = 13.55) are examples of collocations with high MI.

When repeated items are included in the calculation, both the average t-score and MI of collocations found in native speaker essays are higher than the average t-score and MI of

collocations found in non-native speaker essays. The average t-score of collocations found in the native speaker corpus increases to 9.02, while the average MI increases to 5.69. The average t-score of collocations found in the non-native speaker corpus increases to 8.9, while the average MI increases to 4.99. These results are summarized in Table 1.

Table 1. *Average MI and t-score of collocations in native (NS) and non-native (NNS) speaker texts*

	Excluding repeated items		Including repeated items	
	NS Corpus	NNS Corpus	NS Corpus	NNS Corpus
Average t-score	8.07	8.59	9.02	8.9
Average MI	5.56	4.96	5.69	4.99

3.2. Repetition of collocations

Concerning the repetition of collocations, 116 out of the 556 collocations in the native speaker dataset are repeated items, which makes up 20.86% of all collocations found in the dataset. In contrast, 63 out of the 310 or 20.32% of collocations in the non-native speaker dataset are repeated items. This shows that there is no significant difference between native and non-native speaker essays in terms of frequency of identical collocation repetition. However, considering that the inclusion of identical combinations versus their exclusion yielded different association score averages, it can be assumed that the repetition of collocations is not purely random, but rather that certain collocations are more preferred than others.

3.3. Distribution of collocations across different levels of association strength

As mentioned in the Data and Methods section, the focal point of the present study is the analysis of collocation distribution across different bands of t-score and MI. The aim is to reveal the extent to which native and non-native speakers utilize collocations with different association strength.

3.3.1. Distribution of collocations across different levels of t-score

First, the distribution of collocations was calculated across 8 bands of t-score, with repeated items excluded. This distribution is represented in Figure 1. Figure 1 shows that most collocations in both native and non-native speaker corpora fall into the t = 2–3.99 range (e.g., *final state* (t = 2.09), *terrible life* (t = 2.76), *negative aspect* (t = 3.83), etc.), however, the non-native speaker corpus has 1.37% more collocations within this band than the native speaker corpus. The lowest percentage of collocations from both corpora fall into the t = 15–19.99 band (e.g., *political union* (t = 15.42), *common ground* (t = 16.46), *free market* (t = 19.65), etc.), with the non-native speaker corpus having 1.04% more collocations within this band than the native

speaker corpus. The biggest difference between the two corpora in terms of collocation usage can be observed within the t < 2 band, with the native speaker corpus having 4.23% more collocations within this band than the non-native speaker corpus. Across all other bands (t ≥ 2) the difference in collocation use between the two corpora is relatively small. The second highest percentile difference can be observed within the t = 10–14.99 band (e.g., *physical contact* (t = 10.45), *free will* (t = 11.15), *monetary system* (t = 14.09), etc.), with the non-native speaker corpus having 2.05% more collocations that fall into this range than the native speaker corpus. The smallest difference in terms of collocation use between the two corpora can be observed within the t ≥ 20 band (e.g., *good job* (t = 20.84), *public opinion* (t = 27.89), *great deal* (t = 59.97), etc.). The native speaker corpus has 0,13% more collocations that fall into this range than the non-native speaker corpus. When the bands are collapsed into broader 'low' (t < 8) and 'high' (t ≥ 8) t-score bandings, some additional observations can be made, specifically in regard to the latter half of t-score ranges. The higher half of t-score bandings contain 33.42% of collocations from the native speaker corpus and 34.85% from the non-native speaker corpus, meaning that the non-native speaker corpus has 1.43% more collocations within this range than the native speaker corpus. Though the difference is very slight, it does explain why the average t-score of collocations in the non-native speaker corpus was slightly higher than the average t-score of collocations from the native speaker corpus. The reason for the difference in t-score averages is further made apparent when only the three highest bandings are taken into consideration. In the range t ≥ 10, there are 2.96% more collocations from the non-native speaker corpus than there are from the native speaker corpus. All of this suggests that non-native speakers take a slightly higher proportion of collocations from the highest bands of t-score than their native speaker counterparts.

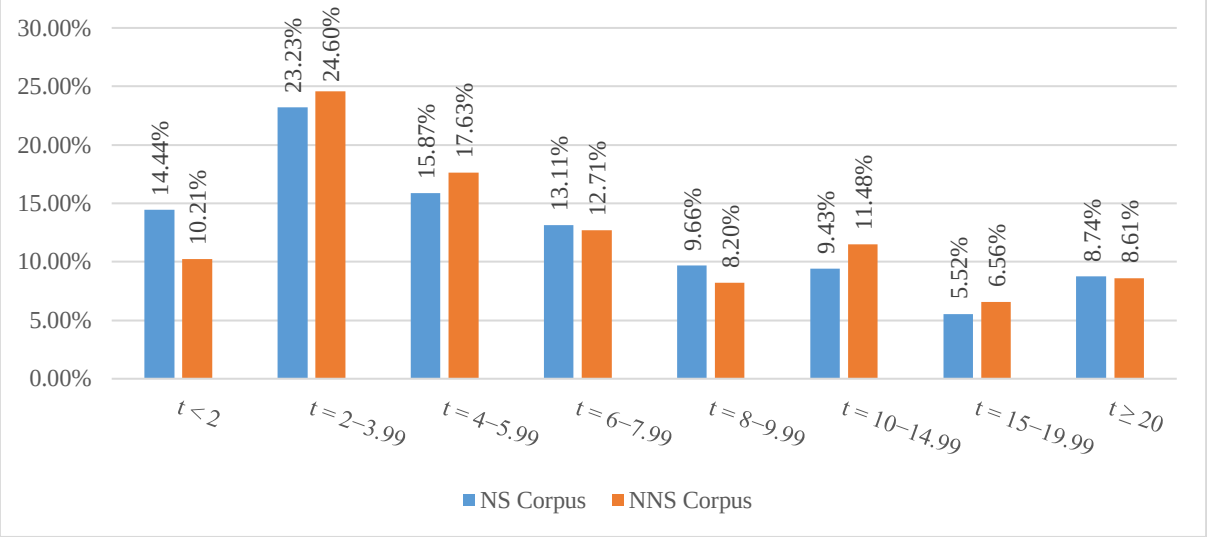

Figure 1. *Distribution of collocations across different levels of t-score bands (repeated items excluded)*

When the data is recalculated with repeated collocations, some new patterns emerge. From Figure 2, it becomes immediately evident that the differences between native and non-native speaker corpora are far more pronounced in almost every single t-score band after the inclusion of repeated items. The t = 2–3.99 range has retained the highest percentage of collocations from both corpora. With repeated items included in the analysis, this band exhibits the smallest difference between the two corpora in terms of collocation use, with the non-native speaker corpus having only 0,56% more collocations within this range than the native speaker corpus. The t = 15–19.99 band retains the lowest percentage of collocations from both corpora. The biggest difference between the two corpora can be seen within the t = 10–14.99 range, with the non-native speaker corpus having 4.66% more collocations within this band than the native speaker corpus. Perhaps the most interesting change can be observed within the highest t-score band ($t \geq 20$). Before the inclusion of repeated items, the proportion of collocations from both corpora within this band was nearly the same, but after the inclusion of repeated items, there are 2.5% more collocations from the native speaker corpus within this range than there are from the non-native speaker corpus. If the ranges are once again collapsed into 'low' ($t < 8$) and 'high' ($t \geq 8$) bandings, the same pattern that was deduced during the analysis of exclusively unique collocations persists when repeated items are included. The higher half of t-score bandings contains 36.72% of collocations from the native speaker corpus and 37.76% from the non-native speaker corpus, with the percentile difference being 1.04%. If only the three highest bands are considered ($t \geq 10$), the difference is once again amplified, with the non-native speaker corpus having 3.34% more collocations within this range than the native speaker corpus. From this pattern of results, it is evident that even after the inclusion of repeated items, within the higher end of t-score ranges there still remains a higher percentage of collocations from the non-native speaker corpus. This, on the surface, fails to explain why the average t-score of collocations from the native speaker corpus increased and became marginally higher than the average t-score of non-native speaker collocations after the inclusion of repeated items. However, if the average t-score of collocations from the highest bands is calculated, it becomes obvious why the increase occurred. When these three bands are isolated, the average t-score of native speaker collocations is 21.14, while the average t-score of non-native speaker collocations is 19.22. For comparison, if the three highest bands are isolated and the average t-score of only unique collocations is calculated, the average t-score of native speaker collocations becomes 19.89, while the average t-score of non-native collocations becomes 20.02. So, despite the fact that within the highest t-score ranges there are proportionally less native speaker collocations than there are non-native speaker collocations, those native speaker collocations on average have a slightly higher t-score.

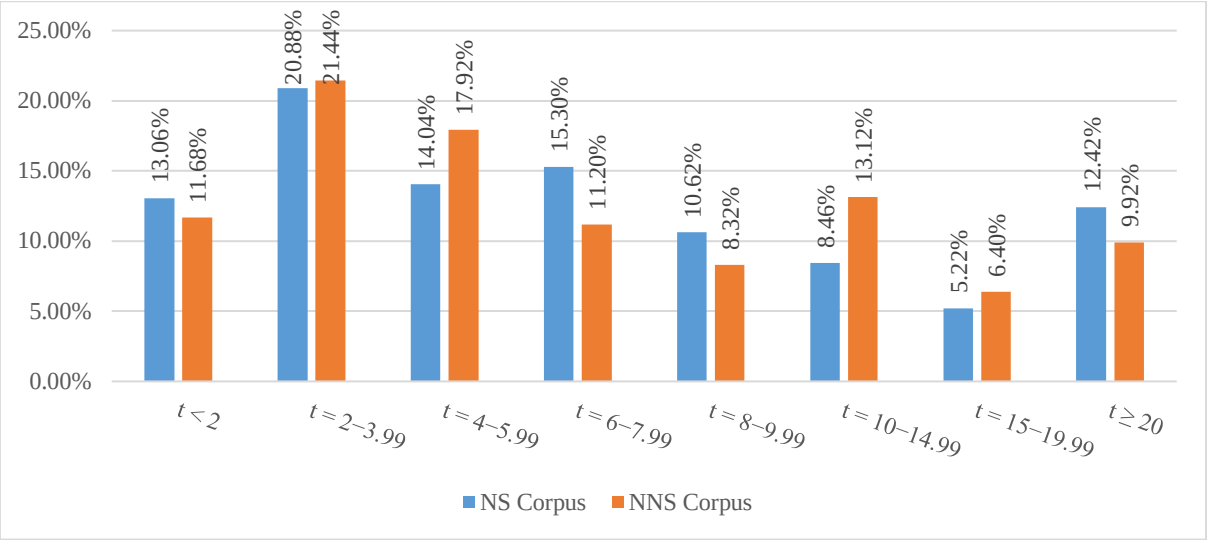

Figure 2. *Distribution of collocations across different levels of t-score (repeated items included)*

Table 2 shows the degree to which the percentages across all 8 bands increased or decreased after the inclusion of repeated collocations. These findings show that native speakers most extensively repeat collocations that fall into the t ≥ 20 band, while non-native speakers mostly repeat collocations that fall into the t = 10–14.99 range.

Table 2. *Percentile change across bands of t-score when repeated collocations are included*

	t < 2	2–3.99	4–5.99	6–7.99	8–9.99	10–14.99	15–19.99	t ≥ 20
NS Corpus	-1.38%	-2.35%	-1.83%	2.19%	0.96%	-0.97%	-0.30%	3.68%
NNS Corpus	1.47%	-3.16%	0.29%	-1.51%	0.12%	1.64%	-0.16%	1.31%

The 3.68% increase in the highest t-score band (t ≥ 20) and the notable percentile decrease in the three lowest categories (t<6), especially in the t = 2–3.99 ranges (-2.35%), offer a satisfying explanation for the notable increase of the average t-score of collocations in the native speaker corpus after the inclusion of repeated items. Concerning the percentile changes across t-score bands with collocations from the non-native speaker corpus, it is evident that no drastic increase (at least not to the same degree as in the native speaker data) had occurred. However, a significant decrease did occur in the t = 2–3.99 band (-3.16%). Furthermore, a 1.47% increase in the lowest t < 2 range can be observed. This would explain why the average t-score of collocations in the non-native speaker corpus did not increase as drastically as it did with collocations from the native speaker corpus. So, while the repetition of collocations in native and non-native speaker texts happens to the same extent, those repeated items are clearly distributed differently in the two corpora. All of this shows that native speakers are more likely

to repeat strong collocations that would fall into the highest t-score range, while non-native speakers do not seem to depend as heavily on repetition of collocations from any particular t-score band.

3.3.2. Distribution of collocations across different levels of MI

The same procedure that was used to calculate the distribution of collocations across different ranges of t-score was also carried out when dealing with the MI of collocations. Figure 3 summarizes the distribution of collocations across different bands of MI. It becomes immediately evident that collocations from the native speaker corpus are overall distributed far more evenly across all bands of MI than collocations from the non-native speaker corpus. It can be seen that the use of collocations in the non-native speaker corpus spikes in two categories: MI = 3–3.99 (22.14%) and MI = 5–5.99 (25.01%). There also exists the biggest percentile difference between the two corpora within these two bands. The non-native speaker corpus has 7.19% more collocations that fall into the MI = 3–3.99 range (e.g., *national law* (MI = 3.22), *natural process* (MI = 3.68), *principal character* (MI = 3.98), etc.) than the non-native speaker corpus, and 7.3% more collocations within the MI = 5–5.99 range (e.g. *terrible disaster* (MI = 5.03), *general public* (MI = 5.62), *free movement* (MI = 5.97), etc.) than the native speaker corpus. The highest percentage of collocations (18.40%) in the native speaker corpus fall into the MI = 4–4.99 range (e.g., *important task* (MI = 4.14), *academic discussion* (MI = 4.26), *political action* (MI = 4.8), etc.). There are 3.05% more collocations from the native speaker corpus within this band than there are from the non-native speaker corpus. Another overarching trend emerges when the bands are collapsed into broader 'low' (MI < 6) and 'high' (MI ≥ 6) categories. The lower half of MI ranges (MI < 6) are primarily 'dominated' by collocations from the non-native speaker corpus (MI = 4–4.99 is an exception). 73.76% of all collocations in the non-native speaker corpus fall into this broader range of MI scores. The majority of collocations in the native speaker corpus also fall into the MI < 6 range (60.9%). However, it is evident that native speaker essays contain significantly more collocations that fall into the higher MI ≥ 6 range. For reference, 39.1% of all collocations in the native speaker corpus fall into this range versus 26.24% that were found in non-native essays (12.86% difference). Even more importantly, the MI ≥10 range (e.g., *own volition* (MI = 10.14), *instant gratification* (MI = 12.07), *proportional representation* (MI = 13.45), etc.) stands out, due to the fact that not a single collocation from the non-native speaker corpus has an MI value within this range. Conversely, 4.60% of all collocations in the native speaker corpus fall into the MI ≥10 range.

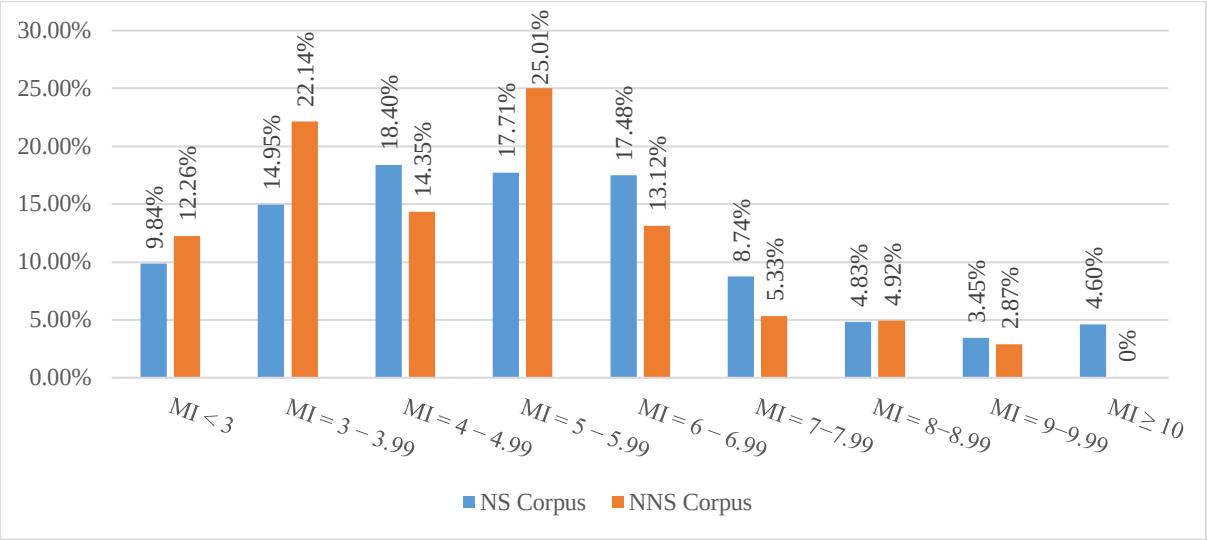

Figure 3. *Distribution of collocations across different levels of MI (repeated items excluded)*

When the distribution of collocations across MI bands is recalculated with repeated items, the general picture remains essentially the same. The data spikes that were recorded previously in ranges MI = 3–3.99 and MI = 5–5.99 remain after the data is recalculated, as seen in Figure 4. After the recalculation, the highest percentage of collocations from the native corpus fall into the MI = 6–6.99 band (e.g., *constant awareness* (MI = 6.07), *national referendum* (MI = 6.42), *collective guilt* (MI = 6.96), etc.).

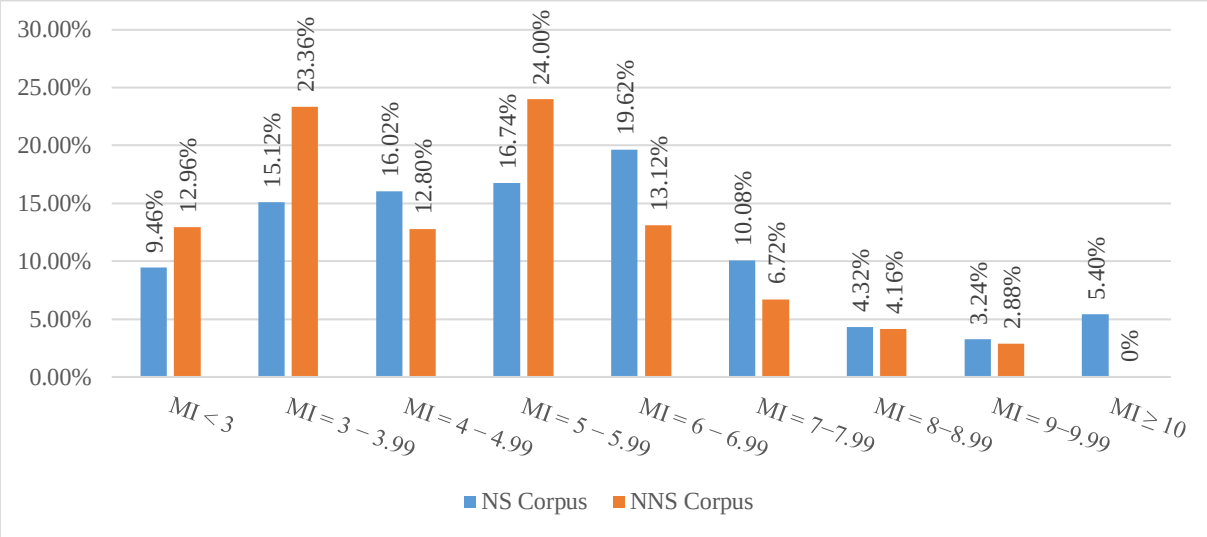

Figure 4. *Distribution of collocations across different levels of MI (repeated items included)*

Table 3 shows the degree to which the percentages across all 9 bands of MI either increased or decreased after the inclusion of repeated collocations. It would appear that native speakers most extensively repeat collocations that fall into the MI = 6 − 6.99 band, while non-native speakers mostly repeat collocations that fall into the MI = 7−7.99 range.

Table 3. *Percentile change across bands of MI after the inclusion of repeated collocations*

	MI < 3	3–3.99	4–4.99	5–5.99	6–6.99	7–7.99	8–8.99	9–9.99	MI ≥ 10
NS Corpus	-0.38%	0.17%	-2.38%	-0.97%	2.14%	1.34%	-0.51%	-0.21%	0.80%
NNS Corpus	0.70%	1.22%	-1.55%	-1.01%	0.00%	1.39%	-0.76%	0.01%	0.00%

Here, the percentile changes in both corpora have occurred relatively evenly across all ranges. This stands in stark contrast to the more drastic changes that occurred across the different bands of t-scores after the inclusion of repeated items, especially with collocations from the native speaker corpus. Additionally, during the t-score analysis, the most prominent percentile changes occurred at the opposite ends of the outlined t-score ranges (2–3.99 and t ≥ 20). Conversely, during the MI analysis, the more notable changes that did occur after the inclusion of repeated collocations, happened in the middle MI bands (4 – 4.99; 6 – 6.99; 7–7.99). This would explain why the calculation of MI averages did not reveal a significant increase after the inclusion of repeated collocations in either of the corpora, and the calculation of t-score averages did.

4. Discussion

In the beginning of the paper, three research questions were raised, all of which are now going to be addressed.

RQ1: To what extent do texts produced by native and non-native users of English differ in terms of collocations? The analysis of collocation distribution across bands of t-score revealed that non-native speakers utilize collocations with very high t-scores slightly more extensively than native speaker do. However, this difference is very marginal, especially when only unique collocations are analysed. The analysis of collocation distribution across different bands of MI, on the other hand, revealed some noteworthy differences between native and non-native speaker texts. The results showed that non-native speakers, in comparison to native speakers, strongly underuse collocations with very high MI values. In fact, not a single collocation from the non-native speaker corpus fell into the highest MI range. Furthermore, it was discovered that non-native speakers significantly over-rely on collocations with very low MI values (e.g., *reasonable solution* (MI = 3.04), *social position* (MI = 3.35), *happy life* (MI = 3.69), etc.).

RQ2: How do the results of the present analysis compare to the findings of Durrant and Schmitt's (2009) study? The results of the present study in essence match Durrant and Schmitt's

findings. However, there are some notable irregularities that should be addressed. Firstly, their analysis of collocation tokens (i.e. all unique and repeated collocations) showed that non-native speakers heavily over-relied on collocations with very high t-scores. Though the results of the present analysis also showed that non-native speakers utilized collocations with high t-scores to a greater extent than native speakers did, this tendency was not as prominent as it was in Durrant and Schmitt's study. Most importantly, it was revealed in their study that within the highest t-score band ($t \geq 20$) the percentage of collocations from the non-native speaker corpus was considerably higher than the percentage of collocations from the native speaker corpus. The analysis of tokens in the present study pointed to the complete opposite trend. As mentioned in the Results section, the usage of collocations within the highest t-score band ($t \geq 20$) was 2.5% higher in the native speaker corpus than it was in the non-native speaker corpus. Secondly, even the analysis of collocation types (i.e. exclusively unique collocations) in Durrant and Schmitt's study showed that non-native speakers were at least marginally more reliant on collocations that fell into the highest t-score band than native speakers were. In the present analysis of collocations types, it was revealed that native speakers were at least slightly more reliant on collocations that fell into the $t \geq 20$ range than non-native speakers were. Whether the calculation of collocation tokens or types is the more accurate approach to this issue is up for debate. However, even Durrant and Schmitt put emphasis on the fact that the assumption regarding non-native speakers' overreliance on strong collocations could only be confirmed if repeated collocations are taken into account. Otherwise, the difference between native and non-native speakers' use of strong collocations is far more marginal. So, even though the results of the present t-score analysis point to the opposite trend, it is evident that the difference between native and non-native texts in terms of strong collocation usage is likewise minimized when only unique collocations are calculated.

The findings of the MI analysis are largely consistent with Durrant and Schmitt's findings and essentially point to the same tendencies. Both their analysis and the present analysis revealed that non-native speakers heavily underuse collocations with high MI values. Furthermore, their analysis of short texts also revealed that not a single collocation from the non-native speaker texts had an MI within the highest defined range ($MI \geq 10$). Their findings have also exhibited a data spike in the 5–5.99 band, which confirms that non-native speakers overuse collocations with lower MI values.

RQ3: Which association measure, i.e. t-score or MI, is a better indicator of differences in the use of collocations by native and non-native speakers? As already discussed, the t-score analysis has revealed only marginal differences between the native and non-native speakers'

usage of collocations. When compounded with the fact that the minor differences that were detected are is some aspects inconsistent with the findings of Durrant and Schmitt's analysis, some uncertainty arises in relation to the reliability of t-score as an indicator of differences between native and non-native speakers' use of collocations. The MI analysis, by comparison, has proven to be quite reliable, as it revealed some very salient patterns that differentiate between native and non-native speakers on the basis of the use of collocations. Not only that, but the patterns that were revealed by the analysis of exclusively unique collocations remained consistent even when repetition of favoured items was accounted for. Furthermore, the results of the MI analysis were, as already mentioned, consistent with the results of Durrant and Schmitt's MI analysis. All of this would suggest that MI is overall a more reliable indicator of differences that exist between native and non-native speakers' use of collocations. With all that said, the role of the t-score analysis in this type of study should not be entirely discredited. After all, the minor inconsistencies between the current study and the one conducted by Durrant and Schmitt could have been influenced by a number of other variables. For instance, the sizes of the datasets are quite different in these two studies; the present study was limited to the analysis of exclusively adjective-noun collocations, while in the study conducted by Durrant and Schmitt both adjective-noun and noun-noun combinations were analysed; the level of proficiency of non-native speakers was not measured in the present study. Additionally, repetition has proven to be a highly influential factor, as it had a profound impact on the results of the t-score analysis in both the present study and the one conducted by Durrant and Schmitt. Furthermore, other studies have yielded more comparable results to Durrant and Schmitt's findings. For instance, Granger and Bestgen's (2014) analysis of collocations in intermediate and advanced learners' texts detected a higher proportion of high-frequency collocations (as attested by higher t-score values) in texts written by intermediate learners. Granger and Bestgen (2014:248) even concluded in their study that "it is useful to use two measures of collocational strength rather than just one" and that "the concurrent use of MI and t-score makes it possible to highlight two aspects of phraseology".

So, how can the present findings be interpreted? The fact that the difference between native and non-native speakers' usage of collocations with high t-scores was quite marginal is not entirely surprising. As mentioned in the Data and Methods section, high t-scores usually indicate formulas that are generally quite frequent in language and are composed of frequently occurring lexical items. It is thus unsurprising that frequently occurring collocations would become a prominent feature of any text, whether it was written by a native speaker or a non-native speaker. This shows that non-native speakers are quite capable of acquiring high-frequency phraseology and can utilize it just as effectively in their texts as native speaker can.

On the other hand, the results of the MI analysis seem to suggest that non-native speakers have a harder time acquiring strongly-associated formulas (marked by high MI values). It is also possible that non-native speakers simply lack confidence and or certainty when it comes to the production of correct collocations. Thus, their strategy when completing written assignments is to opt for safer and, consequently, higher-frequency phraseology, in order to avoid making mistakes. It would then appear that the intuitive claim made by certain linguists (Kjellmer 1990; Sinclair 1991), regarding the lack of pre-constructed phraseology in the language of non-native speakers is not entirely incorrect, but the formulation of this claim can be improved upon. A better description of non-native speakers' phraseology is proposed by Durrant and Schmitt (2009:175) who suggest that "non-native phraseology differs from that of natives not because it avoids formulaic language altogether but because it overuses high-frequency collocations and underuses the lower-frequency, but strongly-associated, pairs characterised by high mutual information scores".

5. Conclusion

The present study investigated the use of collocations in texts written by native and non-native speakers of English, focusing specifically on adjective-noun collocations in academic essays. The analysis utilized two association measures, Mutual Information (MI) and t-score, to determine the collocational strength of word pairs. The collocations were extracted from manually compiled native and non-native speaker corpora and then ranked across a scale of association strength.

The findings revealed a nuanced picture of the differences that exist between native and non-native speaker essays. While non-native speakers tended to utilize collocations with higher t-scores to the same if not slightly greater extent than native speakers, they significantly underused strongly associated collocations marked by high MI value. These results are mostly consistent with Durrant and Schmitt's (2009) findings, however, their study showed that non-native speakers relied on high-frequency collocations significantly more heavily than native speaker counterparts, though this trend was quite notably amplified by the repetition of favored items.

Regarding the reliability of association measures, MI emerged as a more consistent indicator of differences between native and non-native speakers' use of collocation. The differences that were revealed by t-score were not very salient and were not entirely consistent with the results of other studies.

These findings contribute to a better understanding of non-native speakers' phraseology, suggesting that while they are capable of acquiring and using high-frequency collocations, they struggle with incorporating strongly associated, lower-frequency combinations. This highlights the importance of considering both frequency and association strength in studying collocational patterns in second language acquisition and developing resources for learners of English.

The present study had a fairly limited scope. Further research could employ a much larger dataset and explore additional factors influencing collocation use. The comparison of the present results with the findings of other studies seems to also indicate that more precise measurement of non-native speakers' proficiency in a target language can lead to more accurate representation of the different tendencies that determine collocation use. The separate analysis of different types of combinations (i.e. noun-noun, adjective-noun, verb-noun, etc.) also appears to be potentially resourceful. Thus, future research could compare the usage of a wider array of syntactic types of collocations.

References

Primary sources

Corpora

1. British National Corpus Consortium. (2007). *British National Corpus*, version 3 (BNC XML Edition). Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/
2. Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) *Learner English on Computer*. Addison Wesley Longman : London and New York, 3–18.
3. Granger, S., Dupont, M., Meunier, F., Naets, H. and Paquot, M. (2020) *The International Corpus of Learner English*. Version 3. Louvainla-Neuve: Presses universitaires de Louvain.

Software

4. Brezina, V. and Platt, W. (2023) *#LancsBox X* [software], Lancaster University, http://lancsbox.lancs.ac.uk.

Secondary sources

5. Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16 (1): 22–29.
6. Durrant, P. and Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, 47 (2): 157–177
7. Ellis, N. C. (1996). Sequencing in SLA – Phonological Memory, Chunking, and Points of Order. *Studies in Second Language Acquisition*, 18(1), 91–126.
8. Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. *Third Text*, 20, 29–62.
9. Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*, Martin Bygate, Peter Skehan and Merrill Swain (eds.), 75–94. London: Longman
10. Granger, S. (1998a). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology: Theory, Analysis, and Applications*, Anthony P. Cowie (ed.), 145–160. Oxford: Oxford University Press.

11. Granger, S. and Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics*, 52(3): 229–252.
12. Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
13. Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
14. Juknevičienė, R. (2008). "Collocations with High Frequency Verbs in Learner English: Lithuanian Learners vs Native Speakers", *Kalbotyra*, 59: 119–127.
15. Kjellmer, G. (1990). A mint of phrases. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Karin Aijmer and Bengt Altenberg (eds.), 111–127. London: Longman.
16. Lorenz, G. (1999). *Adjective Intensification – Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.
17. Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins
18. Schmitt, N. (ed.) (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins
19. Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
20. Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative methods. *Functions of language*, 2 (1): 1–33.
21. Vilkaitė, L. and Schmitt, N. (2017). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, 40(2), 329–354.

Santrauka

Kolokacijų analizė pagal traukos įverčius: gimtakalbių ir svetimkalbių vartotojų rašytinės anglų kalbos atvejis

Šiame darbe buvo tiriamas kolokacijų, sudarytų iš būdvardžių ir daiktavardžių, vartojimas gimtakalbių ir svetimkalbių anglų kalbos vartotojų akademiniuose rašiniuose. Gimtakalbių tekstai buvo išrinkti iš gimtakalbių anglų kalbos vartotojų rašinių tekstyno (LOCNESS). Panašaus pobūdžio svetimkalbių tekstai buvo išrinkti iš tarptautinio negimtakalbių anglų kalbos studentų tekstyno lietuviškojo komponento (LICLE). Tyrimui buvo surinkta 30 gimtakalbių ir 30 svetimkalbių anglų kalbos vartotojų rašinių. Iš šių tekstų, kolokacijos buvo išrinktos pasitelkiant tekstynų analizės įrankį LancsBox X. Gimtakalbių tekstyne buvo rastos 556 kolokacijos. Svetimkalbių tekstyne buvo rasta 310 kolokacijų.

Siekiant nustatyti išrinktų kolokacijų traukos stiprumą, buvo pasitelkti du traukos įverčiai – Mutual Information (MI) ir t-score. Aukštas MI dažniausiai žymi stipriai susijusias kolokacijas, kurias sudarantys žodžiai kalboje yra retai vartojami atskirai. Aukštas t-score dažniausiai žymi kolokacijas, kurias sudaro dažnai kalboje pasikartojantys žodžiai. Šie įverčiai buvo gauti iš Britų nacionalinio tekstyno (BNC). Pagal t-score vertes, kolokacijos buvo suskirstytos į 8 traukos stiprumo lygmenis. Pagal MI vertes, kolokacijos buvo suskirstytos į 9 traukos stiprumo lygmenis.

Tyrimo rezultatai atskleidė, jog svetimkalbių vartotojų kalboje dažnesnės kolokacijos su aukštomis t-score vertėmis, kurias jie vartoja taip pat dažnai ar net dažniau nei gimtakalbiai, tačiau jie žymiai rečiau produkuoja stipriai susijusias kolokacijas, turinčias aukštas MI vertes. Šis tyrimas buvo atliktas siekiant pakartoti Durrant ir Schmitt (2009) tyrimą naudojant kitokį duomenų rinkinį.

Šio darbo rezultatai iš esmės sutampa su Durrant ir Schmitt išvadomis, tačiau, jų darbe buvo nustatyta, jog svetimkalbiai vartotojai žymiai dažniau kartojo dažnai anglų kalboje pasikartojančias kolokacijas, todėl tokio pobūdžio kolokacijų svetimkalbių tekstuose iš viso buvo rasta daug daugiau.