LATVIJAS UNIVERSITĀTE

MAĢISTRA DARBS

RĪGA 2020

UNIVERSITY OF LATVIA
FACULTY OF HUMANITIES
DEPARTMENT OF ENGLISH STUDIES

COLLOCATION VARIATIONS IN THE WEB-BASED CORPORA OF THE ENGLISH LANGUAGE VARIETIES

VĀRDU SAVIENOJUMU VARIĀCIJAS ANGĻU VALODAS VARIANTU TĪMEKĻA KORPUSOS

MASTER THESIS

Katrīna Vēvere
Matriculation card No. kv18079
Adviser: assoc. prof. Zigrīda Vinčela

RIGA 2020

ACKNOWLEDGEMENTS

I would like to thank my advisor Zigrīda Vinčela for her guidance, encouragement and patience through this writing process. I would like to thank the Department of English Studies for the knowledge, challenges and opportunities presented in the study process. Finally, I must express my gratitude to my family and friends for their continuous support throughout the years of study.

ANOTĀCIJA

Kopš pasaulē internets ir kļuvis aizvien pieejamāks, palielinājusies arī saziņa starp cilvēkiem visā pasaulē. Šāda informācijas apmaiņa var izraisīt globalizācijas efektu valodās vai valodu variantos, it īpaši valstiskajā līmenī. Neraugoties uz to, dažādi pētnieki ir analizējuši dažādās pasaules valstīs lietotās angļu valodas variāciju atšķirības, nevis līdzības. Tādējādi šā, uz korpusu balstītā pētījuma mērķis ir analizēt kolokācijas, pamatojoties uz izvēlēto lietvārdu sarakstu, kas veidots, izmantojot četras nacionālās angļu valodas variantus – amerikāņu, austrāliešu, britu un kanādiešu. Analīze tika veikta divos korpusos – *Corpus of Global Web-Based English (GloWbE)* un *Corpus of Four National English Varieties (CoFNEV)*, kurš īpaši radīts šim pētījumam. Nepieciešamo datu iegūšanai izmantota biežuma, logaritmiskās varbūtības un savstarpējās informācijas analīze. Pētījuma rezultāti norāda, ka lietvārdu pielietojuma līmenī globalizācijas efekts nav novērojams, izņēmums – vārds *vacation* kanādiešu angļu valodā. Tomēr kolokācijas norāda uz līdzību starp valodu variācijām.

Atslēgvārdi: korpuss, nacionālās angļu valodas variācijas, kolokācija, biežums, savstarpējā informācija, logaritmiskā varbūtība

ABSTRACT

The communication among people from around the world has increased since the use of the Internet has become more available. This information exchange can lead to the globalization effects in languages or language variants, especially on the national level. Despite that, various researchers have performed studies on the differences but not similarities of national English varieties seen all around the world. Thus, the purpose of the current corpus-based research is to analyse collocations based on the selected noun list that represent four national English varieties – American, Australian, British, and Canadian. The analysis is performed within Corpus of Global Web-Based English (GloWbE) and Corpus of Four National English Varieties (CoFNEV), which is created for the purposes of this study. Frequency, log-likelihood and mutual information analysis is used to carry out the necessary data analysis. The results of the research indicate the on the level of noun use no globalization effects can be seen, except for the use of *vacation* in Canadian English. Yet the collocation use indicates similarities among varieties.

Keywords: corpus, national English varieties, collocation, frequency, mutual information, log-likelihood

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS 1
INTRODUCTION 2
1. CORPORA AND THEIR USE FOR LANGUAGE VARIETY ANALYSIS 5
1.1. Classification of corpora 5
1.1.1. Corpus relations to the Web 9
1.2. Varieties of English 11
2. COLLOCATIONS 15
2.1. Parts of speech as the basis for collocations 15
2.2. Definition of collocations 17
2.3. Collocation analysis 19
3. DIFFERENCE OF NOUNS AND THEIR COLLOCATION USE AMONG FOUR NATIONAL ENGLISH VARIETIES 23
3.1. Methodology 23
3.2. Procedure 24
3.2.1. Selection of nouns 24
3.2.2. Selection of GloWbE and creation of CoFNEV 25
3.3. Results 28
3.3.1. Noun use in Corpus of Global Web-Based English (GloWbE) 28
3.3.2. Noun use in Corpus of Four National English Varieties (CoFNEV) 32
3.3.3. Noun collocation in GloWbE 36
3.3.4. Noun collocation in CoFNEV 52
CONCLUSIONS 66
THESES 68
REFERENCES 69
Internet sources 71
APPENDIX 1 CORPUS DESCRIPTION 73

LIST OF ABBREVIATIONS AND ACRONYMS

AF – absolute frequency

BNC – British National Corpus

COCA - Corpus of Contemporary American English

CoFNEV – Corpus of Four National English Varieties

GloWbE – Corpus of Global Web-Based English

LL – log-likelihood

MI – mutual information

RF – relative frequency

INTRODUCTION

With the development of technology, various parts of our lives have changed and have become more technology based, whether it is the use of mechanics or the Internet. This increased technology usage presents a question of how or if there have been any changes seen in language usage. Due to the wide access to information, not only do people have the opportunity to learn new languages, keep records of extinct languages, and carry out different types of language research and analysis, but people also have the access to see how one language has manifested in different places and evolved into different variants.

Due to globalization, English has become one of the most used languages in the world, with a lot of information exchange happening specifically in English. Because there are various countries with English as an official language, the information exchanged by people can be in different variants, depending on which country they come from. Additionally, the quantity of texts, in particular written texts, is impressive and they could be used to analyse languages and their varieties. Thus, in order to carry out such analysis, it is useful to collect certain texts in one or more databases. That is why the creation of corpora and their analysis enables the investigation of the differences in such aspects of a language and its varieties.

In relation to this, several researchers have carried out corpus-based researches to analyse English varieties. For instance, corpora-based study was carried out by Xiaohui Xu (2017), who focused on the differences of English variants that are present in Africa, with the main focus on the comparison of Kenya and Tanzania corpus in comparison with Jamaica corpus. Cook and Brinton (2017) developed corpora based on national English varieties, specifically Canadian English instead of analysing the existing corpora. Corpus-based approach was also favoured by Larsson (2012) with the focus on researching which English variant was preferred by EFL learners in Bulgaria, Italy and Sweden. A different aspect was investigated by Simaki, Simakis, Paradis and Kerren (2017), who chose to analyse national English varieties in social media, with a corpus comprised of texts from Facebook and Twitter.

The focus of these prior researches was to find the differences in various national English variants and differences in language use. However, although news and social media is different in each country, due to the increased accesses of the Internet, language exchange among people from various regions has also increased. That is why the direction of the current paper is to investigate if there are common elements between some of national English varieties in blog and news articles. More specifically, to see if in a span of a decade, the language use tendencies among varieties have become more similar.

That is why **the goal** of this research is to create a noun list, which is used as a basis for collocation analysis across four national English varieties (American, Australian, British, and Canadian) in two corpora – Corpus of Global Web-Based English (GloWbE) and Corpus of Four National English Varieties (CoFNEV), which is created for the present study.

To fulfil the selected goal of the current paper, these **research questions** are proposed:

1. What similarities and differences of noun frequencies in American, Australian, British, and Canadian English in GloWbE corpus can be found in relation to the typical use of the selected nouns?
2. What similarities and differences of noun frequencies in American, Australian, British, and Canadian English in CoFNEV corpus can be found in relation to the typical use of the selected nouns?
3. What collocations are present in each corpus for the selected nouns?
4. How the obtained collocations support the noun use trends in the four English varieties in GloWbE and CoFNEV corpora?

To answer the proposed research questions, these are the suggested **enabling objectives**:

1. to collect and analyse theoretical material relating to corpus, web as corpus, English language varieties, nouns, and collocations;
2. to collect and create a corpus of articles from the USA, the UK, Australia and Canada;
3. to select a list of nouns from the four English language variants;
4. to collect and analyse the statistical results on each noun and its collocations provided in GloWbE;
5. to collect and analyse the statistical results on each noun and its collocations provided in Corpus of four National English Varieties;
6. to compare the results on noun use in GloWbE and Corpus of Four National English Varieties;
7. to compare the results on collocation use in GloWbE and Corpus of Four National English Varieties;
8. to draw conclusions based on the obtained material.

The current research is comprised of a theoretical and empirical part. To obtain the theoretical background on corpora, English language varieties, nouns and collocations, comparative analysis was used. This type of analysis is necessary to get a better understanding of the theory and how different researches view each of these aspects. In regards to the empirical part, both qualitative and quantitative research methods were used as the basis for corpus-based research. Quantitative analysis in the form of statistical analysis is used to

obtain data from each corpus, such as, descriptive statistics (absolute and relative frequencies) and inferential statistics (log-likelihood and mutual information). On the other hand, qualitative research method of description of data on noun and their collocates is used. The research is based on the GloWbE corpus; however, due to the corpus being created almost a decade ago, there is a necessity for more recent data, which can be achieve by the creation of another corpus. The current research is presented through the view of corpus-based analysis. A detailed description of corpus-based analysis is provided in the methodology chapter.

The first chapter is dedicated to theoretical background of corpus linguistics. The focus is on the various principles that corpus creation is based on, which include the purpose of corpora, the level of annotation, and data collection methods. Additionally, a subchapter dedicate do the discussion of language varieties is presented, due to the possibilities of researching them via the use of corpora.

The second chapter is dedicated to collocations. Firstly, there is a brief discussion of parts of speech, followed by phrases that are created with parts of speech, which are used as the basis for collocation creation. Additionally, there is also a discussion of some ways of analysing collocations, including frequency analysis, log-likelihood, and mutual information analysis.

The third chapter is focused on the methodology and the discussion of the results. Firstly, the aspects of corpus-based research are discussed and the differences it has from other corpus related researches. Further on, the information on the chosen noun word list is presented, followed by the criteria for corpus selection, which is needed for the creation of Corpus of Four National English Varieties (CoFNEV) that is specifically created for this research. The description of the research procedure is also included. Lastly, the discussion of the results is presented. The results section is divided in four subchapters, two of which are dedicated to the results of frequency statistical data in both corpora, and two subchapters dedicated to the statistical data of the obtained collocations in both corpora.

The present research is summarised with the conclusions of the current study. The research is finalised with the research thesis, references, as well as, an appendix in which the description of the CoFNEV corpus is provided.

1. CORPORA AND THEIR USE FOR LANGUAGE VARIETY ANALYSIS

This chapter is dedicated to providing information about types of corpora, the main fields where they are used and how corpora can be classified. Researchers have proposed various classifications, but the focus in this study will be on the theories suggested by Granger (2008), McCarthy and O'Keeffe (2010), McEnery and Hardie (2012), and Bianchi (2012). Additionally, the discussion of English language varieties is discussed. The main focus of the current study of on the theoretical material provided by Davies (2005), Bergs and Briton (2017), Peters (2017), Dollinger (2017), and Hundt (2017).

1.1. Classification of corpora

This chapter focuses on the classification of the principles necessary for corpus creation. It is vital to analyse these approaches, because many cover essential aspects of corpora, such as how corpus can be used. Additionally, some classifications focus on the level of annotation, while others outline where and how corpus can be used, and how corpus is related to the Web.

Before looking into the use of various corpora, it is important to state the basis in which fields can corpus be used. Such classification is provided by McCarthy and O'Keeffe (2010), highlights six types of Corpus Linguistics and, subsequently, corpus uses – language teaching and learning; discourse analysis; literary studies and translation studies; forensic linguistics; pragmatics; sociolinguistics, media discourse and political discourse (2010: 7-12). When it comes to this particular research, the analysis of national English varieties using corpus is more related to sociolinguistics; however, language variety analysis through corpus can be useful in various aforementioned fields.

Researchers normally range corpora in several types depending to their use. McEnery and Hardie (2012) suggest 'the principles underlying the use of corpora', which are 'mode of communication; corpus-based versus corpus-driven linguistics; data collection regimen; the use of annotated versus unannotated corpora; total accountability versus data selection; multilingual versus monolingual corpora' (2012: 3). There are variations of these categorisations and some researchers have provided a different outlook towards the typology, or have classified a part of these six types.

One such author is Granger (2008), who views corpora classification from a different angle, but having some commonalities with the aforementioned McEnery and Hardie (2012)

classification. Granger's (2008) main focus is on corpus in relation to learner corpora. However, this classification is comprised of various aspects which are applicable beyond learner corpora. Granger's (2008) list includes six types of corpora – commercial vs. academic, big vs. small, English vs. non-English, writing vs. speech, longitudinal vs. cross-sectional, and immediate vs. delayed pedagogical use (Granger, 2008: 261-263).

Thus, the first type offered by Granger (2008) is of commercial or academic corpora. Granger (2008) indicates, that 'commercial learner corpora, initiated by major publishing companies, and academic learner corpora, which are compiled in educational settings' (Granger, 2008: 261). Both of these types can be used by and for learners, although there is a distinctive difference in their size. There is the distinction that 'there are more academic than commercial corpora, commercial corpora tend to be much larger and have a wider range of mother tongue backgrounds' (Granger, 2008: 261). The classification of commercial and academic corpora indicates the second type – big vs. small corpora.

Next type of corpus discussed by Granger (2008) deals with the size (size and purpose, matching), which nowadays can be created quite big in a small amount of time (2008: 261-262). With the help of computers, larger texts can be added to a corpus in a small amount of time, making it easier to create big corpora. However, there are upsides to small corpora, for example, 'a detailed longitudinal study of one single learner is of great value if the focus is on individual interlanguage development', while large corpora can provide better data representation (Granger, 2008: 262).

The next type, proposed by Granger (2008) is English vs. non-English corpus. From the name of the terms it is also understandable that there is a possibility for each language to create a corpus, but English has been the leading language in corpus creation, as well as, there are many corpora in English with millions of words (Granger, 2008: 262). Other language corpora are on the rise, especially, French, Swedish, Norwegian, Dutch, Spanish, and German, that can be used by learners, which also means that there are a lot of available learner corpora, but they are not available to the general public (Granger, 2008: 262). On the other hand, McEnery and Hardie (2012) have a bit of a different view with the classification of monolingual and multilingual corpora (2012: 18). However, instead of focusing on which language is used in a corpus, the authors focus on the quantity of languages present in a corpus (McEnery and Hardie, 2012: 18).

Further on, Granger's (2008) classification offers writing vs. speech corpora. This is a useful distinction as often written language differs from spoken language due to differences in the preparation of the text. Additionally, it is much easier to record and store a written text than spoken text. This is something that could change 'as increased use of information and

communication technologies (ICT) in foreign language teaching allows for quick and easy compilation of a wide variety of computer-mediated communication between learners' (Granger, 2008: 262). Not only does technology help keeping records of written texts, but it also has made it more accessible and faster to record and store spoken texts. However, there is a new problem that has arisen, which is 'the difficulty is compounded in the case of multimedia learner corpora, which contain learners' texts linked to audio-video recordings' (Granger, 2008: 263). As well as, younger generations have started to communicate with images and emoticons, which also is difficult to record for compiling corpora and subsequently learn from that corpora.

The following, is Granger's (2008) classification of longitudinal vs. cross-sectional corpus. The difference between these two corpora is that one is used more as a learner corpus, which is 'cross-sectional, i.e. they contain data gathered from different categories of learners at a single point in time', whereas longitudinal corpora takes a longer period of time to gather data from learners (Granger, 2008: 263). However, the use of technology might be useful, as it may help to make up for the current dearth of good longitudinal data' (Granger, 2008: 263). Although, mentioned by Granger (2008) specifically in the context of learners, this type can apply to other corpora as well. The longitudinal corpus is comparable with McEnery and Hardie's (2012) offered classification of "monitor corpus", which also is developed over time (2012: 6-7). As a contrast to monitor corpus, McEnery and Hardie (2012) offer the type 'balanced corpora, also known as sample corpora, try to represent a particular type of language over a specific span of time' (2012: 8). Both of these classifications have similar notions but presented with a different focus.

The last but not least of the corpora types discussed by Granger (2008) is immediate vs. delayed pedagogical use. In this case, the title also clearly states the difference between them, because immediate indicates the ability to use such corpus instantly, while delayed signals to the fact it is created for other uses. Another difference is that immediate corpora 'are not used directly as teaching/learning materials by the learners who have produced the data', while delayed corpus means that 'the learners are at the same time producers and users of the corpus data' (Granger, 2008: 263). This last type of corpora is the only one in Granger's (2008) list, which is specifically focused on corpora use in teaching, because here the most important aspect is the attention to learners.

Bianchi (2012) outline a different corpus, which also covers aspects from the six types proposed by McEnery and Hardie (2012). Bianchi (2012) classifies corpus by how it is annotated, also known as, corpus markup, which is 'is the act of adding explicit (meta-) information to a corpus' (Bianchi, 2012: 40). A lot of corpora are annotated, but there are

some corpora that have stayed un-annotated. Depending on the focus of the corpora, there are specific aspects of the text that can be annotated. Bianchi (2012) provides some aspects that are annotated in a corpus, such as, 'textual, such as part of speech information (POS tagging), syntactic annotation (parsing), semantic annotation; and meta-textual, such as sociolinguistic information' (Bianchi, 2012: 40). These annotations cover a variety of language aspects from grammatical and syntactic division of language, to external aspects that could influence the information provided in the texts.

Part of speech tagging or POS is the first aspect mentioned by Bianchi (2012), which means that, for the purposes of corpus creation, parts of speech are marked. Providing this type of annotation helps researchers focus on specific aspects of language. That also refers to collocations, because by tagging POS it is easier to find these collocations and what parts of speech are put together. There are various aspects that can be tagged, for example, 'adjective, comparative; noun, countable, singular; verb, simple present, 3rd person,' as well as, punctuation and other aspects (Bianchi, 2012: 41). The examples provided by Bianchi (2012) indicate that words can be tagged by their word class and by the additional information that they carry. That includes specific indications for each word class, such as, the degree of comparison, tense, countable or uncountable tagging, and other aspects that define each of the word class (Bianchi, 2012: 41). However, Bianchi (2012) also highlights that POS tagging does come with certain constraints. The biggest challenges are choosing the tag types, figuring out what to do with contracted forms, and what to do 'grammatical units' that consist of more than word, for example, '*so that*, or *such as*' (2012: 41). This suggests that corpus programmes have difficulties automatically tagging sentence constituents, because although sometimes some words function individually, often enough there are some grammatical units that do function together. Bianchi (2012) suggest ways to make the tagging easier, which is carrying out post-editing, that can be done manually and recently automatically, although, that still will not 'guarantee and error-free tagged corpus' (2012: 42). There is a possibility for improvement, and hopefully in the future this will change.

Further on, next in Bianchi's list (2012) is lemmatisation. McEnery and Hardie (2012) provided definition explains that it is 'a form of corpus annotation were every token in the corpus is labelled to indicate its lemma' (2012: 245). This type of annotation does not focus on classifying words by their characteristics as POS does, but it focuses on their core. It focuses on the grammatical aspect and in corpus these lemma are found by every word they represent, which mean that corpus can recognize all inflections that in English refers to verbs, nouns, and adjectives, but in other languages there could be additional cases (Bianchi, 2012: 42). This type of annotation also has some problems with tagging. As mentioned by Bianchi

(2012), and often occurrence ‘represented by use of apostrophes, as in the case of Italian definite article *l’*, or English Saxon genitive *‘s*’ with programmes tagging these as belonging to words (2012: 43).

The next annotation type is semantic annotation. As the name suggests, this focuses on semantic aspects of each word and the tagging ‘indicates the semantic field to which the word belongs’ (Bianchi, 2012: 43). Semantic field provides an understanding of some concepts that are present in a text, such as synonyms, hypernyms, and hyponyms (Bianchi, 2012: 43). This annotation type is a bit more complicated due to it illustrating abstract aspects of a language rather than concrete grammatical aspects. Additionally, because these are abstract aspects, they can be open to interpretation. However, University of Lancaster’s tool USAS has a solution, which is ‘attaching several separable labels to the same word’ (Bianchi, 2012: 44).

The previously discussed corpus types and ways of classifying them are useful and provide an understanding of the various aspects of corpus and their creation. It is clear that there are different ways to look at a corpus and classify it, which is evident by the variations researchers have when providing the typology.

1.1.1. Corpus relations to the Web

In one of the aspects provided by McEnery and Hardie (2012), they mention ways of collecting data. In these data collection methods, the aspect of corpus and the Web is mentioned, which is relevant for the current study. According to McEnery and Hardie (2012), this relation ‘takes as its starting point a massive collection of data that is ever-growing, and uses it for the study of language’, and by creating a corpus it in a way puts a small stop point for the growing information volume (2012: 7).

There is some controversy regarding the third corpus association with the Web, which is considering the Web as corpus. Some researchers agree with statement, such as, Kilgarriff and Grefenstette (2003), who base their statement with the explanation, that ‘a corpus is a collection of texts when considered as an object of language or literary study’ (2003: 2). However, over time this view shifted towards a different view of corpus relationship with the Web. According to McEnery and Hardie (2012), the original view presented some constraints, mainly due to the Web being comprised of vast amounts of texts and the fact that errors might be present (2012: 7-8).

From these shifts in views, a classification of types of usage of The Web in relation to corpus needed to be presented. Bianchi (2012) has gathered four different ways that corpus is associated with the Web:

1. querying the Web via commercial search engines and using the retrieved data as concordance lines (i.e. using the web as corpus surrogate);
2. creating corpora from the Web (i.e. using the Web as a corpus shop);
3. considering the Web as a corpus proper;
4. creating a new object, a sort of mini-Web (or mega-corpus) adapted to language research (Bianchi, 2012: 36).

These are the four ways corpus use can be divided according to its relation to the Web. With this distinction, corpus relations with the Web could start from just a web search, such as, carrying out a Google search, to being invested to create a separate large corpus. However, the second and third classification could be considered more complex. Using the Web as a corpus shop might cause some challenges, especially how to show language aspect representation (Bianchi, 2012: 38). On the other hand, when it comes to the Web as corpus proper, other challenges arises, such as, the large size, various types of texts or some form of duplicates, fast changes, as well as, classifying texts (Bianchi, 2012: 37-38). However, these classifications of corpus still have to be research as the Web is developing and growing fast.

For the purposes of this research, the most important one of these usages is creating corpora from the Web or using the Web as a corpus shop. As it is stated by Gatto (2014), this type of Web use is 'researchers using the web as corpus shop select and download texts retrieved by search engines to create "disposable corpora" either manually or […] semi-automatic (e.g. using a toolkit which allows crawling and extensively downloading from the web, such as BootCat)' (2014:37). With the help of corpus collecting programmes, this corpus collection method has become easier for researches, because the process has become more automatic. Additionally, it has become easier to create large corpora because the capabilities of computers have also increased the ability to have bigger storage spaces capacities and gather a vast variety of texts from the Web. One of the programmes that provides this opportunity is *Sketch Engine*, which, as mentioned by Buendía-Castro and López-Rodríguez (2013) 'is a corpus query system' that has the ability to provide such functions as 'concordance, keywords, and wordlists' and 'it integrates grammatical relations, a distributional thesaurus and word sketches' (2013: 59). This programme can help researchers use web as shop and carry out various data collection methods provided in the system.

This subchapter focuses on the corpus relation to the Web, and specifically how the Web can be used to gather the corpus sample. Although there are four ways in which the Web can be used for corpus creation, the use of Web as corpus shop is the specific method used in this research.

1.2.Varieties of English

This chapter focuses on the different varieties of English with a special emphasis on the varieties that are in the focus of the present study. The main theoretical background is taken from authors, such as Davies (2005), Bergs and Briton (2017), Peters (2017), Dollinger (2017), and Hundt (2017).

Each language has a set of different varieties the origin and existence of which depend on various factors that can influence a language. According to Davies (2005) there are two main distinctions – 'variation and the individual' and 'variation and the group' (2005: 2-6). These are the two main differences and they can be divided further on into subgroups, because there are many aspects in each category. In the first category, these varieties are due to the fact that 'each of us has our own verbal repertoire or range of speech styles to draw on the particular situations' (Davies, 2005: 2). That signals to the fact that not only can different individuals have different language varieties, but one person can have a range of language varieties and be able to switch between them depending on the situation. This skill is important for communication with others, because it means that all people in a certain situation can recognize the context and know how to act and communicate. It means that 'the individual way in which each of us speaks is called our **idiolect**, and it is a combination of the ways in which we use the sounds, words and grammar of the language' (Davies, 2005: 3). Although the language that we use is the same, each person expresses themselves differently.

The other variation is related to language and certain group of people. According to Davies (2005), 'linguists use the word **code** to describe the particular language, dialect or variety we choose to use on any occasion', where switching between language, dialect and variety is called '**code-switching**' (2005: 5). This can be applied not only to variations in relation to groups, but also to individuals, because code-switching happens each time there is a necessity to switch between variations. However, when it comes to a group of people, they are called a '**speech community** to describe a group of speakers who share the same language varieties or speech repertoires' (Davies, 2005: 5). In speech communities individuals already have at least one common variety and code-switching is not required in this group, only when a situation requires to. This leads to the various types of varieties that a speech community can share. Davies (2005) indicates that English is one aspect that the communities share; however, English speakers can be divided 'regionally, ethnically, and socially, as well as through factors like their gender, jobs and interests' (2005: 5). It means that speech communities are not limited to the official languages of countries, but can be related to

different varieties that relate to personal interests. Language varieties are depended on the context not just different types of languages.

For the purposes of this research, the most important variety is regional because the main focus of this research in regards to language varieties is specifically the difference in regional varieties. As it is suggested by Vinčela (2017), there are several researchers, such as Cragg et al. (2000), McArthur (2003), and Swan (2005), who indicate that language varieties are differentiated by several linguistic aspects, such as 'grammatical, spelling, pronunciation and vocabulary' (Vinčela, 2017: 162). Bergs and Briton (2017) have edited volume five of the series *The History of English*, which specifically focuses on different English varieties based on the geographic location. Mainly, the focus is on four biggest geographic or national English varieties – British, American, Canadian, and Australian.

The first variety up for discussion is British English. According to Peters (2017), there are two contexts in which British English could be classified as, which are

> At the narrow end of the scale it refers to (i) the language of English, or (ii) the language of Great Britain, or (iii) the language of the British Isles. At its broadest (iv) it may refer to the variety of English used in British Commonwealth countries where there was a sufficiently large community of British settlers to establish it as the national language (e.g. Canada, Australia). (2017: 96-97)

Due to the fact that in this particular research Canadian and Australian English are looked at as two separate types of English, British English is considered at the narrow end, which refers to the British Isles. Peters (2017) adds that British English originated from Middle English dialects (2017: 98). It means that British English could be considered as one of the oldest English variants that exist nowadays. Additionally to being one of the oldest variants of English, it is also, 'with the processes of colonial expansion and postcolonial contraction', became an English standard together with American English (Peters, 2017: 115). These variants are also the ones that are taught, when teaching English as a second language. Peters (2017) also indicates about some of British English specifics, which 'is less homogenous and less standardized than American English in its orthography – as well as its morphology and aspects of syntax' (Peters, 2017: 110). That means that British English is more complex than other varieties of English, and it as a more complicated spelling and grammar.

Next type of variety is American English. This variety has officially been recognized, 'when the first American expression reached England, or less sensibly, when the present-day United States declared independence from the United Kingdom of England, Scotland and Ireland' (Bailey, 2017: 9). The independence of the United States made it easier for a different

English variant to develop because now as a separate country it has no outside influence. That means that British English and American English started to evolve separately, making it possible now have two very different variants. However, it also led to some dislike by the British, who 'were raising alarms that American English was corrupting the excellence of English at home' (Bailey, 2017: 16). It is understandable that the British would have this opinion towards American English, given that Modern English originated from Britain, which is why they could be thinking that their English is the original and subsequently correct English. As mentioned above, Peters (2017) argues that in the case of British English, American is not as complex as British English (2017: 110).

Next variation of English is Canadian English. According to Dollinger (2017), when it comes to Canadian English, the standard version is spoken by most Canadians who primarily speak English, but that does not mean that those are most of Canada's citizens (2017: 55). That is due to the fact that there are some parts of Canada that speak French. One of the main influences of Canadian English is that 'immigration from the British Isles is central in the Canadian context where pro-British sentiments have been felt' (Dollinger, 2017: 59). Based on a research presented by Dollinger (2017) it can be said that in regards to Canadian lexis, it is a mixture between British English and American English, as well as, of compounds that have originated in Canada (2017: 65-66). Lexis plays and important part in distinguishing linguistic varieties, as an example, Vinčela (2017) researches a word belonging specifically to the Canadian variety can be mentioned, which is 'the slang name *loonie* for the Canadian one dollar coin' (2017: 162).

Last but not least is the Australian English variant. Hundt's (2017) approach was to analyse Australian English together with New Zealand English, because of the 'similar input of English, Scottish, and Irish dialects, with a predominance of south-eastern English dialects' (2017: 291). Although, these are separate countries, both of them have a similar language development and have a history in connection to the British Commonwealth. Hundt (2017) indicates that there is a belief that Cockney English influenced Australian English, but it might be a different story, where a combination of dialects formed Australian English (2017: 293). However, there are different ideas about which dialect is the main influence for Australian English. Additionally, Hundt (2017**)** suggests that in regards to lexis there are also researches that show that Aboriginal languages, Maori and, also, nowadays, American English have influenced it (2017: 296-299).

This chapter focused on varieties of languages and English varieties. There are several types, such as geographic or national, ethnic, social, gender, job or interest related, but for the purposes of this research, the focus is on geographic variety, specifically, the four largest

English varieties – British, American, Canadian, and Australian. British is the oldest one of these varieties and its lexis is more complex than American. While Canadian and Australian lexis is combined from British and American English, with Australian English also having some influence from local tribe languages.

2. COLLOCATIONS

This chapter will focus on parts of speech, collocations, their creation and ways of analysing them. The main theoretical material will be gathered from works by Biber, Conrad and Leech (2003), Nugues (2006), Brown and Miller (2013), and Lobeck and Denham (2014).

2.1. Parts of speech as the basis for collocations

Parts of speech are the basis of any language, as well as, the sentence structure. Word class is a language aspect taught early on in schools in both native and foreign languages. However, it can cause confusion for non-native speakers, when studying a foreign language, especially if there is not enough of information provided and if there are significant differences in native and foreign language. Additionally, particularly for the present research, the main focus is on nouns as they are the basis of wordlist selection, and the creation of parts of speech phrases, because that is the bases of how the collocation will be searched in this research analyses.

It is important to know the division of syntactic categories of parts of speech because it shows the usage of them. Lobeck and Denham (2014) have provided a classification of syntactic categories, dividing them into lexical and functional categories. They provide a list for these categories, which are 'lexical categories (noun, verb, adjective, and adverb) and functional category (determiner, numeral, quantifier, pronoun, preposition, conjunction, degree word, auxiliary verb, and modal)' (Lobeck and Denham, 2014: 11). Both of these aspects work hand in hand, when it comes to creating a sentence. These categories complement each other and help provide the correct categorization in a sentence, when they are put together. As well as, these categories have different meanings, such as, 'lexical categories […] express the main content, or meaning in a sentence' and 'functional categories […] express grammatical information about definiteness, number, tense, gender, etc.' (ibid.). That is why it is important to know that these categories work together, because functional category is the one, which provides an in-depth understanding of lexical categories. Lobeck and Denham (2014) also provide another way of categorising parts of speech, which is according to word affixes, seen in Table 2.1. :

Table 2.1 **List of affixes (Lobeck and Denham, 2014: 13)**

Nouns	Verbs	Adjectives	Adverbs
-ity, -ment, -ion, ex-, -s	-ize, -ate, -ify, en-, dis-, -s, -ed, -ing, -en	-ly, -ish, -ful, non-, -er, -est	-ly, -wise, -like

This is just a brief list of the affixes available, but the morphological aspect is one way how to categorise some parts of speech. Word inflections are also influenced by whether the word is singular or plural.

Further on, in relation to creating collocations, it is important to analyse separately each part of speech as it gives information about how they are formed and these parts of speech interact in a clause, additionally, part of speech phrases provides the basis for collocations. To understand how these word classes function, it is necessary to know what types of phrases are there and how they are created. Part of speech phrases indicate how words function with each other, which is also the basis of collocations. All phrases are created with 'lexical word, there is a major phrase type with an example of that class as the head: noun phrase, verb phrase, adjective phrase, adverb phrase, and prepositional phrase' (Biber, Conrad and Leech, 2003: 41). Each phrase is named after the main word in it. It means that 'each phrase type can often consist of just one word: the head' (ibid.). Further on, each phrase type should be discussed in more detail.

Firstly, there is a noun phrase, where the head of the phrase is a noun. As mentioned by Lobeck and Denham (2014), there are various types of noun – abstract and concrete, common and proper, count and mass, collective and generic nouns (2014: 25-30). That is why this is a large part of speech that is represents and covers various aspects. When it comes to noun use in phrases, they can consist of other components, which are 'determiners**,** such as *the, a, her,* and can be accompanied by modifiers - elements which describe or classify whatever the head refers to; […] also be followed by complements, which complete the meaning of the noun, especially that-clauses or infinitive to-clauses' (Biber, Conrad and Leech, 2003: 41-42). That means a noun phrase can be just a noun by itself, or it can be a combination of various other modifiers or even clauses, which are there to support the noun. Some examples of noun phrases are 'a house; these houses' or 'the little girl next door' (ibid.: 41).

Next, there is the verb phrase, which head is 'lexical verb or primary verb' (Biber, Conrad and Leech, 2003: 42). Verb phrase, the same as noun phrases, can have modifiers to help support the meaning. However, unlike the noun phrases, the 'verb phrases are the essential part of a clause, referring to a type of state or action' (ibid.). That is why items that modify the verb are meant to specify state of the clause. Some examples of verb phrases are 'will always have' and 'has definitely started' (ibid.: 43).

Further on, there is the adjective phrase, where the head of the phrase is an adjective. Adjective phrases, the same as the two previously mentioned phrases, have modifiers, however, these 'modifiers typically answer a question about the degree of a quality',

additionally, ‘adjective heads can also take complements’ (ibid.: 43). Unlike, verb and noun phrases, adjective phrases work more as modifiers to them, not as main components of a clause. Some examples are ‘he’s totally crazy’ and ‘you couldn’t have a better name than that’ (ibid.: 44).

Next, there are the adverb phrases, which work similarly as adjective phrases. As well as in other phrases, adverb phrase can have modifiers, that can be before or after the head, and their purpose is ‘they typically express degree’ (ibid.: 44). Because these phrase work similarly as adjective phrases, they are also there to add meaning to the sentence, but they are not the main component. Some examples of adverb phrases include ‘much more quickly than envisaged’ and ‘they sang boomingly well’ (ibid.: 44).

Lastly, there is a prepositional phrase, where a preposition is the head of the phrase. This type of phrase cannot stand work by itself, because ‘consist of a preposition followed by a noun phrase, known as the prepositional complement’ (ibid.). It means that by adding a preposition to a noun phrase, a completely new type of phrase is created. That means that the modifiers and complements of a noun phrase can go together with a prepositional phrase, additionally, they ‘can be “extended” by an initial adverbial particle, which adds a meaning such as place, direction, or degree’ (ibid.: 45). Some examples of prepositional phrases are ‘in a street’ and ‘it was hard to live in Missouri after spending so much time in California’ (ibid.: 45).

The information about phrases that create a clause is necessary, because that is also basis for collocations. Knowing the phrase types and their modifiers, it is clearer what word classes work together and which carry the meaning of the sentence. By knowing this, it is easier to understand that collocations can be part of one phrase or components combined from various phrases.

To sum up, this subchapter is about parts of speech and what phrases can be created with them. The main distinction is between lexical and functional because that determines which word classes are used for the meaning and which serve as the connectors and the support for lexical words. The understanding of how parts of speech phrases are created is the basis for collocations, because it shows how words function with each other.

2.2. Definition of collocations

In the previous chapter, word classes and phrases were discussed, which helps to understand how each of them function together and what word class has a main role or head position in a

phrase and what works more as a support for the head word. Further on, by knowing the possibilities of phrase creation, the next step is to understand collocations and the possible collocation types.

Most languages have the possibility to create collocations, and according to Brown and Miller (2013) they are the 'the relation between individual lexical words such that they frequently occur together or one requires the other: e.g. *brand* and *new* in *brand new*, *staple* and *diet* in *staple diet*' (Brown and Miller, 2013: 86). From the first part of the definition, it is revealed that there is little difference between word phrases and collocations, apart from the fact that collocations are between lexical words. However, it is indicated that these sets of words occur frequently side by side. Nugues (2006) highlights that these word sets are fully known and understood by native speakers, adding that 'collocations underlie word preferences that most of the time cannot easily be explained by a syntactic or semantic reasoning: they merely resort to usage' (Nugues, 2006: 106). Subsequently, it indicates that collocations can cause challenges for non-native speakers, because for each language only the native speakers are able to control collocation usage freely. This is also highlighted by Smadja (1993), who indicates the problems that are faced when collocations are translated, stating that 'in most cases, the learner cannot simply translate word-for-word what s/he would say in her/his native language' (Smadja, 1993: 146). If there cannot be word-for-word translations for many collocations, then that means that one way to translate collocations would be to find equivalents in one's own language, which indicates that only the translator would be able to find a collocation in his/her own native language.

McCarthy and O'Dell (2005) talk about the necessity of learning collocations and some of their typology. They indicate that 'some collocations are fixed, or very strong, for example *take a photo*' and on the other side, 'some collocations are more open, where several different words may be used to give similar meanings, for example keep to/ stick to the rules' (McCarthy and O'Dell, 2005: 6). Linking this together with what was discussed by Smadja (1993), it could be said, that open collocations might be easier for non-native speakers to learn, because of the possible variations available; however, that still can create difficulties in translating the collocations. This furthermore indicates the necessity to learn collocations for non-native speakers. That is why McCarthy and O'Dell (2005) highlight three reasons to learn collocations that can – 'give you the most natural way to say something; give you alternative ways of saying something, which may be more colourful/expressive or more precise; improve your style of writing' (McCarthy and O'Dell, 2005: 6). These reasons provide ways on how to improve a text and make a person seem more knowledgeable of a foreign language. However, there are other reasons why collocations are useful, such as, 'making students aware of low-

frequency collocates that native speakers have internalised' and 'for demonstrating the existence of bias or connotation in words' (Baker, Hardie and McEnery, 2006: 38).

Further on, Baker, Hardie and McEnery (2006) focus on the collocations in relation to their use in corpus linguistics. The authors mention the definition, which is similar to what has been established by McCarthy and O'Dell (2005), but additionally, Baker, Hardie and McEnery (2006) mention how corpus linguistics benefit the finding of collocations. The authors indicate that there are programmes that help finding collocations, such as, 'within WordSmith users can specify a window within which collocational frequencies can be calculated' (Baker, Hardie and McEnery, 2006: 37). Additionally, Baker, Hardie and McEnery (2006) add that there are some methods, which help indicated better information on how likely it is that some word combinations are an actual collocation. These corpus linguistic methods are mutual information (MI), Z-score, MI3, log-log and log-likelihood (Baker, Hardie and McEnery, 2006: 37). With the use of these methods, it is possible to get more concrete results, because by just looking at the frequency, it does not necessarily show the relation between words, because it focuses more on the quantity used in the text.

To sum up, this chapter is dedicated to the definition of collocations, with the main ideas coming from Smadja (1993), McCarthy and O'Dell (2005), Nugues (2006), Baker, Hardie and McEnery (2006), and Brown and Miller (2013). The main ideas that are put forth by them indicated that collocations come natural to native speakers and have to be learnt by the foreign language learner. They are hard to translate and are mostly created with lexical words.

2.3. Collocation analysis

In this chapter the ways of analysing collocations will be discussed. Some of these aspects include keywords, frequency, MI (mutual information), and log-likelihood calculation and analysis.

Keywords are the basis for corpus search and in a way it opens doors for further research. That is why it is important to discuss keywords in relation to collocations, because it is basis for finding collocations. In other words, Bianchi (2012) indicated that 'the notion of keyword includes the idea of statistical significance deriving from frequency comparisons' (2012: 47). This shows one aspect of how keywords function with other analysis ways which is frequency. At the basis of these analysis is the statistical evidence that provides concrete results. To gather results about keywords in a specific corpus, it is necessary to compare 'the

wordlist of the corpus under investigation with the wordlist of a suitable reference corpus; any word of the given corpus whose frequency is found to be outstanding with respect to the reference corpus is considered a keyword' (Bianchi, 2012: 47). This is useful when analysing corpora by language learners, because it gives the opportunity to compare keywords in their corpus to the ones in the reference corpus.

Keywords are also useful when analysing concordance lines, which 'are chunks of text that show the node word in context – hence the term KWIC (Key Word In Context) format' (Bianchi, 2012: 48). By seeing concordance lines, it is possible investigate how the chosen keyword functions in a text. It gives an idea what words or word classes precede and follow the keyword, as well as, the keyword's placement in a sentence. Concordance lines could be used for various reasons, such as, finding sentences with specific words, looking for the context of words, finding 'collocation, colligation, semantic preference and semantic prosody which are usually considered in corpus linguistics the four descriptive components of units of meaning' (Bianchi, 2012: 48). Apart from collocations, there are colligations, which are 'is the relation between general units in a construction: e.g. between adjective and noun in a noun phrase' (Brown and Miller, 2013: 86). Then there is semantic preference, which 'is understood as the semantic field a word's collocates predominantly belong to,' and there is semantic prosody which 'is restricted to a more general characterisation of these collocates, chiefly in terms of a positive or negative evaluation' (Oster, van Lawick, 2008: 335). That is why keywords are useful, because they provide wide opportunity to research various aspects of language and obtained detailed information about specific words and their meaning. Additionally, Bianchi (2012) mentions that KWIC provides the opportunity to read differently, because now people do not have to read all the compiled texts one by one, but gather information from all texts simultaneously (2012: 49). This change in reading allows finding the necessary information quicker and saving time on data collection.

In relation to keywords and, subsequently, collocations, the notion of frequency is also an essential part of research in corpus linguistics. Frequency provides the statistical information from the corpus, for example, as mentioned by Baker, Hardie and McEnery (2006), 'frequencies can be given as raw data, e.g. there are 58,860 occurrences of the word man in the British National Corpus (BNC); or (often more usefully) they can be given as percentages or proportions' (2006: 75). It relates back to keywords, as this information can be used to analyse them or statistical information about other items, which might not necessarily be keywords. Another aspect which is mentioned by Baker, Hardie and McEnery (2006) is that this device can be used to create a comparison between 'words in a corpus – for example *man* (602.91 per million) tends to occur more frequently than *woman* (225.43 per million),

suggesting that *man* is the marked or 'prototype' term' (Backer, Hardie and McEnery, 2006: 75). With just comparing these two words, the researchers are already able to provide some conclusions by just seeing the word frequency. If words in a corpus can be compared, then that means that in a similar manner, a comparison could be carried out between various corpora or types of corpora, for example, it could be one of the types suggested by Granger (2008), such as, written or spoken corpora. Other aspects suggested for analysis by Baker, Hardie and McEnery (2006) are about grammatical forms, word list, collocations, dispersion data, and token ratio (2006: 75-76). These are aspects that solely refer to corpus data analysis. However, frequency analysis results can be used for purposes outside of corpus data analysis. One such instance is discussed by Okamoto (2015), which is frequency use for selecting vocabulary for teaching purposes. As indicated in *Sketch Engine* programme descriptions, there are two types of frequencies used in corpus data collection – absolute and relative frequencies (Online 3and Online 8). The notion of absolute frequency 'refers to the number of occurrences or hits', while relative frequency means 'a number of occurrences (hits) of an item per million, also called i.p.m. (instances per million)' (Online 3 and Online 8).

Another way to analyse collocation is by using MI or mutual information. This is also a calculation that is available to use in various corpora and corpus creation programmes. This analysis provides the opportunity 'to ascertain one of the most important parameters, time delay, in reconstructing phase space from nonlinear time series' (Jiang, Huang, Zhang, Li, Zhang and Hua, 2010: 2948). It shows items work together taking away the differences in time. The authors in the article provide many ways in which the mutual information can be calculated, but these formulas are only needed if everything for the corpus is created from the beginning. However, if corpus creation programmes are used, these calculations are not necessary, because there are programmes, where various types of formulas are already programmed in. Additionally, this calculation helps to find out 'where two items share information, knowing one of the items reduces how much information has to be discovered about the other one' (Brown and Miller, 2013: 300). This is very helpful in collocation analysis, because it can provide the information on whether some words work together or not.

Last but not least is log-likelihood analysis. According to Jiang and Wong (2012), this calculation is used for large scale material (2012: 833). However, when it comes to linguistics, the use of this calculation has different purposes. According to Brown and Miller (2013), log-likelihood is 'a statistical test for comparing the frequency of a given word in one corpus with the frequency of the same word in another corpus and determining whether any difference in frequency is statistically significant' (Brown and Miller, 2013: 271). This is useful when comparing corpus by learners and corpus for learners, because then then the

language learner's progress can be tracked. This can also apply to other types of corpora, for example, comparing American English corpora to British English corpora. Not only will this calculation can help compare the frequency, but also show how significant the differences are.

To conclude, this subchapter is meant to show the ways in which collocations can be analysed. Works by various authors have been used to find out the information about these variations, such as, Baker, Hardie and McEnery (2006), Bianchi (2012), Brown and Miller (2013), and others. There provided options for analysis is keyword analysis, frequency analysis, mutual information analysis, and log-likelihood analysis that each have a specific focus and purpose for the analysis of texts.

3. DIFFERENCE OF NOUNS AND THEIR COLLOCATION USE AMONG FOUR NATIONAL ENGLISH VARIETIES

3.1. Methodology

This chapter focuses on the methodology used in this research paper and also discusses the specifics of two corpora used in this study.

Corpora have been used in many researches to analyse language, so there are also some researches that have been done specifically on language varieties. Additionally, researchers use either already existing corpora or, if there is a necessity, they create their own corpus. Corpus can be used in various ways to carry out a research. Bruckmaier (2017) mentions three types – corpus-informed, corpus-based, and corpus-driven research (2017: 16). These three uses can be applied to language research, but there are some differences between these approaches. As stated by Bruckmaier (2017) corpus-informed 'uses the data of corpora to illustrate and exemplify a linguistic issue', while corpus-based is where 'quantitative information is used to formulate new conclusions' in relation to theory (2017: 16). These two methods are very similar and both of them focus on the theory how it can be supported or improved. However, when it comes to corpus-driven, the situation is completely different. Instead of basing the research on theory or theories, 'corpus-driven research uses corpora to formulate new theories and to redefine established concepts' (Bruckmaier, 2017: 16-17). That is why Bruckmaier's (2017) research is corpus-based, as it is based on already exciting theories and notions that are applied in the research of analysing sociolinguistic aspects in Jamaica and Singapore.

There are numerous other studies, where researchers have used corpus-based methodology. For example, Xu (2017), who used corpus-based methodology to study English varieties present in Africa. Larsson (2012) who chose to analyse which English varieties are preferred by language learners, was also carried out as a corpus-based research. As well as, Hundt (1998), whose theory on Australian and New Zealand English was discussed in the previous chapter about English language varieties, has used a corpus-based approach to research grammar in New Zealand English.

Based on similar researches carried out on language varieties, the present research is also a corpus-based research. The goal of this research is find out whether among four English language varieties, some are present in the other languages. This research is based on already existing theoretical material, which is discussed in the chapters above about corpus, language

varieties and collocations, which are going to be discussed in order to compare the use of them among four English varieties – British, American, Canadian and Australian.

3.2. Procedure

The research procedure covers the steps that were taken to gather data and carry out further analysis. The first step was to select a word list containing noun pairs, which are used in the same meaning in different English varieties. Further on, the selection of English language varieties was carried out that was followed by the selection of the corpus for analysis (GloWbE) and then by the selection and creation of a second corpus (CoFNEV) for analysis. Therefore, the final steps were taken before the data collection began.

3.2.1. Selection of nouns

This subchapter is dedicated to the procedure of the research, and to discuss the steps that are taken to gather the data.

Firstly, it is necessary to state the wordlist, which is vital for the corpora analysis. The part-of-speech up for discussion is nouns, which were selected manually, based on their meaning and use in various varieties. The attention was paid to noun pairs that are used in the same meaning, but in different varieties. The nouns were chosen from four national English varieties – American, Australian, British, and Canadian. These varieties were chosen because they are the four most noticeable geographic English national varieties. The nouns were chosen from sources, which indicate the differences in word use among different varieties. These sources include - *IELTS Online Practice* (Online 5), *Ryerson University Student Learning Support* (Online 9), *Lexico* (Online 6), and *Australian English* (Online 1). Below in Table 3.1 there is list of noun list chosen for analysis.

Table 3.2 **Noun wordlist for four English national varieties**

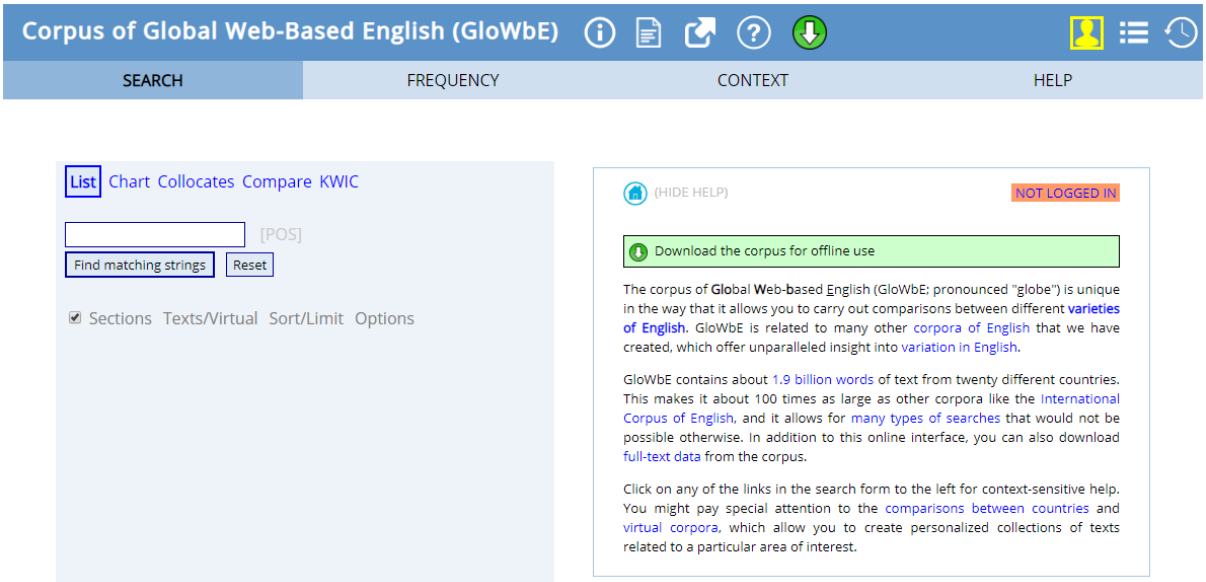
American English	Australian English	British English	Canadian English
candy	sweets	sweets	candy
elevator	lift/elevator	lift	elevator
gasoline	gasoline	petrol	gasoline
pants	pants/trousers	trousers	pants
subway	subway	underground	subway
vacation	holiday	holiday	holiday

The noun list provided in Table 3.2, comprises a collection of six noun pairs. The nouns are divided according to their use in each variety. American and British English varieties are opposites of each other, while Australian and Canadian English has a mix of American and British. The next step in the research procedure is to select the corpora for analysis.

3.2.2. Selection of GloWbE and creation of CoFNEV

The next steps deal with the selection and creation of corpora. Firstly, the search of corpora containing national English varieties had to be carried out. The selection was done taking in mind previous experiences in corpora use as well as with the help of Google Search. From that, several corpora were selected with English varieties. There were a few corpora, which cover one specific English variety, for example, *British National Corpus (BNC)* or *Corpus of Contemporary American English (COCA).* However, these corpora are of different sizes and created in different time periods, which indicated a need for one corpus containing various varieties in the same time frame. This lead to the selection of *Corpus of Global Web-Based English (GloWbE).*

Figure 3.1 **Main page of GloWbE**



The GloWbE is a large unparalleled corpus, containing around 1.9 billion words, which are collected from 20 countries that use a different national English variety. This corpus consists of articles from the web, which also includes news articles, a vast list of blogs posts and other websites. (Online 2) The GloWbE corpus is based in the Brigham Young University system. However, this corpus does not include current texts and is limited to articles from 2012 to

2013. (Online 2) Due to this limitation in the time period, there is a necessity to have a corpus, which includes recent texts, for the purposes of comparison of the data.

That is why there is the need for a second corpus, which would contain recent articles to create a comparison between the corpora and to see how language has changed in almost a decade. In order for the corpora to be comparable, the selection of articles needed to be as random as possible, but similar to the way texts were selected in the GloWbE corpus. However, these texts needed to be blog and news articles because they are the most comparable to the GloWbE corpus. The texts for GloWbE were selected through an n-gram search in Google, to get a random selection of web pages (Online 2). The selection of the blogs and news websites for Corpus of Four National English Varieties was not possible to be completely randomised, because the programme *Sketch Engine*, which was used for the collection of the corpora, were not compatible with some websites. Despite that, the websites were selected based on Google search of the most popular blogs and news websites in each of the four countries representing each English variant (American, British, Canadian, and Australian). Through compiling period, it could be seen, which websites are compatible with *Sketch Engine* and were added to the corpus. Most of the blogs were related to lifestyle, travel and cooking, that is why, if possible, blogs that cover all of these aspects were chosen. With news websites, the situation was different, because with them the most important aspect was to find popular new sites, which have free access that for some countries was more difficult than others.

Taking all of that in mind, the criteria list for the research has to be specified. These criteria can be seen in the list below:

1. there has to be a mix of blogs and news websites,
2. the websites have to be from each country representing a national language variant,
3. there has to be an equal amount of website representing each language variant;
4. the websites have to be free access or mostly free access, without obligatory subscription;

The list of websites used in the creation of the Corpus of Four National English Varieties can be found in Appendix 1.

This programme *Sketch Engine* can be used to access already exciting corpora or create a new one. The programme provides the opportunity to upload documents from a computer or to use the Web to compile a corpus. There are possibilities of adding specific links or add a general website, from which up to 2000 links are taken automatically. For this research, the option of using Web, and specifically, adding websites was used to create this corpus. (Online 3 → 4) However, there are restrictions of storage space in the programme, which lead to each

variety having a sample of 1 million words. In the Table 3.3 the comparison of the sample size in GloWbE and CoFNEV corpora is presented.

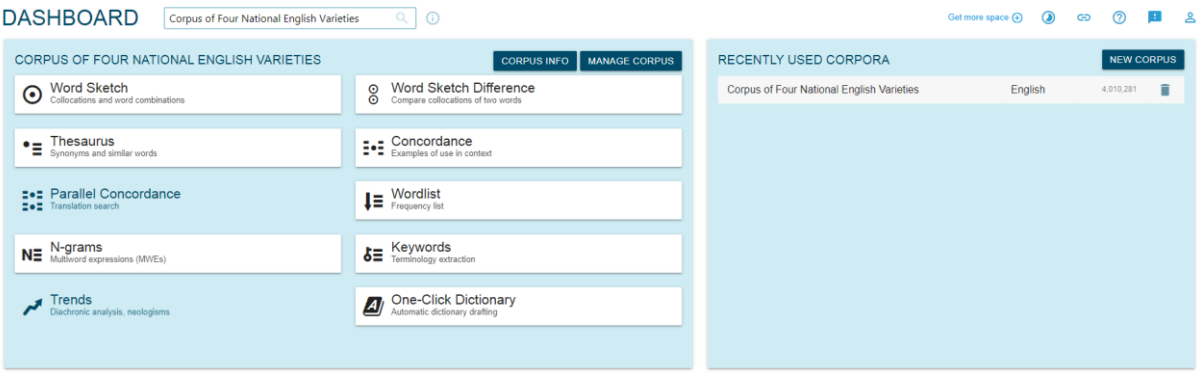
Table 3.3. **Sample size in GloWbE and CoFNEV**

Variety	Sample size in GloWbE (Online 2)	Sample size in CoFNEV
American	386,809,355	1,000,546
Australian	148,208,169	1,002,920
British	387,615,074	1,004,430
Canadian	134,765,381	1,001,388
Total	1,057,397,979	4,010,281

As seen in Table 3.3, the GloWbE corpus, has a considerably broader sample size than CoFNEV. Despite that, CoFNEV corpus can be used to get an idea of some tendencies in the varieties. It has to be mentioned that in GloWbE corpus, American and British English have a similar sample size, and Australian and Canadian also have a similar sample size, but half as much as available for American and British English, while in CoFNEV corpus all the varieties have a similar sized sample.

Additionally, *Sketch Engine* was chosen because there were similar data collection possibilities as GloWbE has. In Figure 3.2, the main page if *Sketch Engine* webpage can be seen.

Figure 3.2. **The main page of *Sketch Engine* website**



Both of these programmes provide frequency of words, covering absolute frequency found in the texts, and relative frequency that shows the ratio of word per one million. Both of these programmes also provided the option of searching for collocations, with the possibility to select the indication of mutual information (MI). In both corpora, it is possible to select the range of words taken for collocations; however, the in the default setting the range 4 tokens to the left and to the right are set. In order to obtain the necessary data, each query was done by writing the word in the search box with the added tag _*n**, which means that the search will

select ‘common noun, neutral for number (e.g. sheep, cod, headquarters)’, except in the cases where the noun is in plural form (Online 10).

3.3. Results

This chapter is dedicated to the discussion and results of the current research. This chapter is divided in five subchapters. Ahead of analysing collocations, it is vital to state the noun frequencies to understand if the chosen noun frequencies coincide with their typical use in each variety. Thus, the first subchapter covers the differences and similarities of nouns in the selected noun variations in corpus GloWbE, which includes discussing the frequency of nouns in each national English variety. Further on, the second subchapter is focused on discussing the same arguments, but in Corpus of Four National English Varieties, which is specifically created for the purposes of this research. Additionally, in the second subchapter, the log-likelihood analysis of the frequencies, where each noun appears the most, will be looked at. This is needed to ensure that the frequencies obtained have a significant connection. Due to possible mistakes that could occur in both corpora in part of speech annotation, the main attention will be put towards the analysis of collocations present for each noun. Therefore, the third and fourth subchapters are dedicated to the collocation analysis in English varieties in GloWbE and Corpus of Four National English Varieties accordingly. In these chapters to find if the collocations have a significant connection, the mutual information (MI) score will be analysed. The last subchapter is focused on the discussion of results of the research and discussing what is similar and different in the two corpora.

3.3.1. Noun use in Corpus of Global Web-Based English (GloWbE)

In this subchapter the use of nouns in the Corpus of Global Web-Based English or GloWbE are discussed. The main focus is on analysing the frequencies of the words in each variety. One instance is the data on absolute frequency or AF as seen in Table 3.4, which provides the information on how many times each node appears in the corpus. However, because GloWbE is not a parallel corpus and each variety has a different number of words included in the corpora, for that reason, relative frequency or RF is used to ensure the normalization basis of each word. Despite, the corpus not being parallel, all of the nouns selected for the current research were present in all English varieties.

The first noun pair is *candy* and *sweets*, which results are displayed in Table 3.4. With this pair the results are inconclusive, given that the frequencies are quite small in all the varieties, which might be due to the query being carried out in for the plural form.

Table 3.4 **Use of *candy* and *sweets* in GloWbE**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
candy	6867	17.8	945	6.4	2419	6.2	1953	14.5
sweets	1012	2.6	493	3.3	2000	5.2	375	2.8

The normalised data provided through the relative frequency in all varieties for the node *sweets* are very similar. When it comes to the word *candy*, the data indicated that in British English the results are higher than with the word *sweets.* In Australian English similar results to British English can be seen. In the case of American English, it is clear that this is the form, which is the norm in the variety. The same can be said about Canadian English, which usually opts for *candy*, which is as well evident in this linguistic data. In this particular case British English comes out as the most frequent, which coincides with the use of the noun *sweet*. However, in Australian English, where the use of *sweets* is also considered the norm, the relative frequency is lower than British English and closer to American and Canadian English.

The next pair up for analysis is *elevator* and *lift*. In Table 3.5 the linguistic statistical data from the GloWbE corpus is provided. In the first columns for each variety there is the absolute frequency (AF) of all the words in a variety, and in the second columns, the relative frequency (RF) is provided, which provides the normalisation bases for comparison

Table 3.5 **Use of *elevator* and *lift* in GloWbE**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
elevator	4510	11.7	442	3.0	1248	3.2	1221	9.1
lift	2160	5.6	1540	10.5	4287	11.1	1182	8.8

The results show that in the case of *elevator*, the most frequent use is seen in American English, while with the word *lift*, American English is the variety in which this noun appears the least. The opposite is in Australian and British English, where it is used frequently and also with a similar relative frequency. As for Canadian English, it is in the middle of the statistical results. An interesting indication is of Canadian English is that the relative frequency is almost the same between *lift* and *elevator*, although, elevator would be the more common choice. By these results, it can be stated that in this corpus sample Australians

choose *lift* over *elevator*, although both cases would be acceptable. Despite these results, American English firmly stays with the use of *elevator* over the use of *lift.*

The next noun pair is *gasoline* and *petrol*, with the results presented in Table 3.6. Concerning the output of this query the following clarifications are vital for the noun *gasoline.* Due to the fact that, although, *gas* is often used instead of *gasoline* as the shortened form; however, in the corpus the query results showed a noticeable number of cases for *gas* as a state of a substance. Due to that, the query search was limited to *gasoline.*

Table 3.6 **Use of *gasoline* and *petrol* in GloWbE**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
gasoline	5611	14.5	366	2.5	830	2.1	1347	10.0
petrol	484	1.3	2422	16.3	5204	13.4	120	0.9

As seen in Table 3.6, the results are quite surprising for Australian English. The noun *petrol* is not seen as the common choice from the pair; however, the results from GloWbE show the opposite. Additionally, the normalisation basis for Australian English is even bigger than in British English, were the use of *petrol* is the norm. While clearly in American and Canadian English this form is a rarity. That is the opposite with *gasoline*, where it is most frequently seen in American and Canadian English varieties. Although, Australian and British English rarely sees the cases of *gasoline*, it is not as little as *petrol* in Canadian and American.

Further on, the discussion of the results for the word pair *pants* and *trousers* can be carried out. These statistical results are presented in Table 3.7.

Table 3.7 **Use of *pants* and *trousers* in GloWbE**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
pants	7021	18.2	2718	18.3	3775	9.7	2539	18.8
trousers	790	2.0	658	4.4	4240	10.9	197	1.5

It comes as a surprise to find that in American, Australian, and Canadian English the normalisation basis data of *pants* is almost identical, with a few decimal differences. These results are curious not only because in Australian English both noun usage are acceptable, but also because of the large score of relative frequency. The relative frequency provides a normalisation basis, which indicated that the use of *pants* is used almost as often as in American English. In British English the relative frequency is half of what is seen in the other three varieties. Whereas in the case of *trousers*, the relative frequency is noticeable in British

English, while in the other varieties it is small, or in Canadian English a rare occurrence, as seen in Table 3.7.

The next noun pair up for discussion is *subway* and *underground*. The linguistics statistics in the Corpus of Global Web-Based English are presented in Table 3.8. The absolute frequency is marked as AF and relative frequency marked as RF in Table 3.8 and all the following tables in this chapter.

Table 3.8 **Use of *subway* and *underground* in GloWbE**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
subway	3778	9.8	579	3.9	1463	3.8	1881	14.0
underground	1136	2.9	302	2.0	2211	5.7	368	2.7

In the Table 3.8 it can be observed that *underground* is not a frequently used noun in the varieties in comparison with *subway*. Traditionally, *underground* is prevalent in British English, which is supported by the absolute frequency (AF) and subsequently the normalisation basis provided by relative frequencies (RF) presented in Table 3.8. More unusual results appear in the query for *subway*, where the word has a higher relative frequency in Canadian English than in American English. The noun *subway* is used in Canadian English; however, it is surprising that the word is more frequent in Canadian English over American English, taking in mind the differences in size of both varieties present on the corpus. This peculiarity could be explained with the fact that this is not a parallel corpus, which could indicate to Canadian English having a substantial representation of texts which in some way cover the topic of *subway.*

Last but not least is the case of is the case of *vacation* and *holiday,* with the results presented in Table 3.9.

Table 3.9 **Use of *vacation* and *holiday* in GloWbE**

Variant	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
vacation	9390	24.3	1565	10.6	3824	9.9	5292	39.3
holiday	15105	39.1	8778	59.2	28083	72.5	6635	49.2

With this noun pair, the statistical results are greater in size than in any other pair used for analysis in the current research. The statistics show, that both nouns are used often in all the variants; however, in the case of *vacation,* it is curious to see, that it appears the most in Canadian variant, although, the noun is specific to American English. When it comes to the noun *holiday*, it is the most frequent in British English. Despite it being used in Australian,

British, and Canadian English, the nouns is seen being frequently used also in American English even more frequently than the noun *vacation.*

In this subchapter, the discussion is directed towards the analysis of absolute frequency (AF), which provides the number of times a word was found in a corpus, and relative frequency (RF), which provides the normalisation basis in order to compare the subcorpora, of the nouns pairs in GloWbE corpus. All of the chosen nouns were present in all English national varieties.

3.3.2. Noun use in Corpus of Four National English Varieties (CoFNEV)

In this subchapter the linguistic statistics of absolute frequency (AF) and relative frequency (RF) in the Corpus of Four Nation English Varieties or CoFNEV are discussed. This corpus was created specifically for the use in the current research. CoFNEV consists of around four million words, with each variety having around a million words. That is a relatively small sample size, which is why it is expected that not all nouns from the chosen wordlist are present in the corpus. Despite this, some results can still be made, even if the word does not appear in a certain variety. Additionally, this corpus has one advantage, which is not present in GloWbE corpus, which is that all varieties have a similar size sample, unlike, GloWbE, which has quite a noticeable discrepancy among the varieties. To ensure that the obtained data from both corpora is significant and not accidental, the log-likelihood analysis is carried out for the most frequent words in CoFNEV corpora in relation to the same word in GloWbE corpus. For this calculation, the log-likelihood calculator, provided in Lancaster University is used (Online 7).

The first word pair is *sweets* and *candy,* which is presented in Table 3.10. The same as seen in the previous subchapter, the absolute frequency (AF) of the word is presented first, and then, to get the normalisation basis, the relative frequency (RF) is provided. Both of the nouns are present in three varieties.

Table 3.10 **Use of *candy* and *sweets* in CoFNEV**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
candy	112	24.27	0	0	6	1.30	10	2.16
sweets	24	5.20	5	1.09	25	5.42	0	0

There is an absence of the query *candy* in Australian English, while in other varieties it is present. In British English, this noun appears only six times, which is the lowest frequency

from all the varieties. Canadian English is not far behind British English, although typically Canadians opt for the form *candy.* Americans, who typically use the noun *candy*, have a high result for the noun *candy* as seen by the relative frequency (RF), which is seen in the Table 3.10.That is why for the noun *candy*, it is necessary to look at the log-likelihood between GloWbE and CoFNEV. The results provided by the log-likelihood calculator indicate the score of 222.72. In the case of log-likelihood, the higher the score the better the fit of the two results, which in this case is a high score and signifies that the data is not accidental.

Surprisingly, *sweets* had an almost identical frequency in American and British English, which is unusual for the American variant. However, Canadian English had no cases of the query *sweets* and Australian had only five instances. Due to *sweets* having a bit of a higher normalisation basis results in British English, the log-likelihood of this variety is looked at. Unlike in the case of *candy*, the log-likelihood score for *sweets* is 38.85, which is significantly smaller score. That could signal to *sweets* not having as high of a use or more sample texts would be needed.

The next noun pair up for discussion is *lift* and *elevator.* The statistical data is presented in Table 3.11. With this noun pair, the situation is a bit different, and the nouns have appeared two and three times respectively.

Table 3.11 **Use of *elevator* and *lift* in CoFNEV**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
elevator	9	1.95	0	0	0	0	9	1.95
lift	0	0	17	3.6	27	5.85	26	5.63

The noun *elevator* is absent in both Australian and British English, which indicates that in this corpus sample Australian English leans heavily towards the British use, although, typically in this variety it is acceptable to use both forms. Additionally, the results for *elevator* are identical in American and Canadian English. That is why log-likelihood analysis is necessary for both. For American English the score is 0.66, but for Canadian English it is 0.00, which indicate that there is a lack of evidence from the corpus, and also that the use of *elevator* is slightly more preferred in American English.

The noun *lift* is absent in American English, which can be explained by the fact that this is not the common choice used from this pair. However, it is unusual that in Canadian English this use is almost tied with the British English absolute and relative frequencies, especially because in this variety it the noun *elevator* is the common choice. This might be due to the peculiarities of the corpus sample. Due to British English having a slightly higher score

according to the normalisation basis, the log-likelihood needs to be determined for this variety. The score is 16.12, which, although, is a small number, is still higher than anything else seen with this noun pair.

The noun pair up for analysis is *gasoline* and *petrol.* The statistical data is indicated in Table 3.12. With this pair there also were some constraints, because the shortened form of *gasoline*, which is *gas,* could not have been chosen, due to the corpora given results not only for gasoline, but also for the state of a substance.

Table 3.12 **Use of *gasoline* and *petrol* in CoFNEV**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
gasoline	15	3.25	0	0	0	0	0	0
petrol	0	0	31	6.71	8	1.73	7	1.51

When it comes to *gasoline*, it is only present in American English. This is an unusual case, especially because the word *gas* was present in all of the varieties. However, because the shortened for of *gasoline* is *gas*, which also represents a state of a substance, the base form *gasoline* was chosen for analysis. Its log-likelihood score is 0.02, which is low, although understandable because of the difference amount of usage of the shortened form and the full noun form.

As seen in Table 3.12, there is a peculiar instance, where the most frequent cases are seen in Australian English, although in Australian English the chosen form usually is *gasoline*. Meanwhile, in British English, which is the one variant out of the four that typically uses the form *petrol*, has an absolute frequency of eight, and relative frequency (RF) of only 1.73. That is barely above the results seen with Canadian English. For Australian English, the log-likelihood score is 10.21, which indicates, given the sample size, shows to a significant result in the noun usage.

Further on, the next noun pair for analysis is *pants* and *trousers.* The linguistic statistics are presented in Table 3.13.

Table 3.13 **Use of *pants* and *trousers* in CoFNEV**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
pants	63	13.65	5	1.08	12	2.60	57	12.36
trousers	0	0	0	0	23	4.99	0	0

With this pair it is hard to make clear conclusions, because for *pants* there are cases for each variety, while for *trousers* there is only one. It can be confidently said that *pants* is

mostly used in American and Canadian English, because the relative frequency is quite big, especially in comparison to the amount of words present for each English variety. These two varieties have also very similar frequencies for the word *pants.* Due to American English having a slightly higher relative frequency, the log-likelihood score is calculated for this variety between the two corpora. The score is 66.76, which is a high score, and signals that the number of frequencies is not accidental.

Similarly, it can be seen that this noun is not commonly used in British or Australian English. In regards to the noun *trousers,* in this corpus it is found only in the British English variety and the frequency is more substantial than for the word *pants* in this variety. In comparison to the log-likelihood score for *pants,* in the case of *trousers* the score is 12.84, which could be interpreted also as to signalling to the usage being similar, but due to the sample size, the score is smaller.

The next noun pair displayed in Table 3.914 is *underground* and *subway.* This was on one of the least frequent noun pairs found in Corpus of Four National English Varieties, although it still proved some data for analysis.

Table 3.14 **Use of *subway* and *underground* in CoFNEV**

Variety	**American**		**Australian**		**British**		**Canadian**	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
subway	93	20.15	0	0	9	1.95	10	2.16
underground	0	0	0	0	6	1.30	0	0

As seen in Table 3.14, in the case of *subway,* the word is absent from Australian English. There is no clear idea of why that would be the case; however, most likely it is due to the size of the corpus. Meanwhile, understandably, the highest relative frequency is present in American English, while in Canadian and British English, the relative frequency is comparatively low. This is supported by the log-likelihood score, which is 250.79.

The noun *underground* is found only in British English. That is also the language variant, where it is used the most commonly. However, given that there a travel blog is included in each language variety as well as news articles; it is interesting to find that *underground* is not present in other varieties, which are the ones that could have had mentioned *underground* at least in the context of British mode of transportation. When performing the log-likelihood analysis, the result of 0.01 was obtained, this also indicated to the lack of word present in the sample.

The last noun pair up for analysis is *vacation* and *holiday.* The linguistic data is presented in Table 3.15.

Table 3.15 **Use of *vacation* and *holiday* in CoFNEV**

Variant	American		Australian		British		Canadian	
Frequency	AF	RF	AF	RF	AF	RF	AF	RF
vacation	138	29.91	7	1.5	49	10.62	336	72.83
holiday	275	59.61	250	54.19	308	66.76	59	12.78

This is the only noun pair in Corpus of Four National English Varieties, which was present in all the four varieties. The data shows similar results to GloWbE corpus, due to *vacation* being the most frequent in Canadian English variant. It is also supported by log-likelihood score 832.08, which is the largest score from all of the words analysed in the present research.

Further on, *holiday* is more frequent in American English than *vacation.* However, unlike in GloWbE, in this corpus *vacation* is very rare in Australian English. The most frequent use of the noun *holiday* is seen in British English, which is supported by the log-likelihood analysis score 416.33, when comparing between the two corpora.

To conclude, this subchapter focused on the absolute and relative frequency analysis in the Corpus of Four National English Varieties (CoFNEV), which was created for the use in the current paper. The statistical data obtained from this corpus had some similarities and differences with GloWbE, which were observed due to the contrasting sample sizes. However, in many cases, when analysing the log-likelihood of the most frequent words of CoFNEV corpus in comparison to the same nouns in GloWbE corpus, the results indicated that number of frequencies is not accidental and is comparable between the corpora. For the analysis it is revealed, that no globalization effect can be observed in almost all of the national English varieties. The only distinct exception observed from both corpora is the case of *vacation* in Canadian English, where the typically used noun should be *holiday.* This observation indicated to an American variant influence over Canadian English, which is due to their proximity.

3.3.3. Noun collocation in GloWbE

In this subchapter, the collocations of each noun chosen for the current research in the GloWbE corpus are discussed. The noun analysis served as basis for the collocation analysis, given that the collocation are analysed based on the part of speech relationships. The focus is on noun collocation with lexical parts of speech – nouns with nouns, nouns with verbs, and nouns with adjectives. The only lexical part of speech not included in the research is adverbs,

due to programmes often having challenges in recognising or categorising adverbs automatically. Additionally, the statistical data representing mutual information (MI) are presented. Each collocation was selected based on the most frequent one in each part of speech, which also has an MI score 3 or above, and with a range of four words to left and right. Then for that collocation the examples were manually looked through. If there were mistakes either in the placement of the collocation, or the word was annotated as the wrong part of speech, the next most frequent collocation in the same parameters was manually selected.

In Table 3.16 the collocation and mutual information score for the noun *candy* is presented. Unlike, the other half of the pair - *sweets*, with *candy* the mutual information score is lower.

Table 3.16. ***Candy*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	Halloween	8.28	eat	4.42	favorite	3.43
Australian	eye	7.24	stick	4.02	sweet	5.13
British	eye	6.94	eat	3.43	pink	5.49
Canadian	eye	6.63	eat	4.57	hard	3.14

The first examples are of noun and noun collocations for the word candy:

1. *American: I've still got gobs of Halloween candy leftover, and the skeleton is still on my front door.*
2. *Australian: I was assumed to be eye candy, the pretty face of a publication whose content was provided by people with actual talent.*
3. *British: Aspinal of London, Brook Street: For some serious eye candy, this is the place to be.*
4. *Canadian: The piggy bank is cute and adorable eye candy that's more decorative than functional.*

The strongest collocation is with the American collocation of *Halloween*, which is understandable due to the popularity of the celebration in the country. While in the other three varieties the word collocating the most with *candy* is *eye*, which forms an expression that can be used to describe something or someone.

The next examples are with noun and verb collocations for the noun *candy*:

1. *American: Today is the day that parents indulge their children's wishes to eat as much candy as they can hold.*

2. *Australian: Stick the candy canes around the circumference of the candle, so they fit together snugly.*
3. *British: The idea is that beginning on 1 December the children may eat one piece of candy each day till Christmas Eve.*
4. *Canadian: It took us a while to eat all that candy!*

In this case, three out of four varieties have the same verb as the most collocating verb, which is the verb *eat*. This is the case collocation as with the noun *sweets.* However, in this case there is also Australian English, who has the collocation *stick* as the most frequent collocation with *candy.*

Lastly, these are the examples with noun and adjective collocations for *candy*:

1. *American: What's your least favorite candy?*
2. *Australian: Sweet as candy, yet sometimes rebellious.*
3. *British: My favorite set-up was against the gigantic wall of pink cotton candy.*
4. *Canadian: Some (usually thin) individuals report that they need to eat every three hours or they experience hypoglycemic symptoms -- they might even carry snacks or hard candy for "emergencies".*

With the noun *candy,* the adjective collocations are very different. Each variety has their own most frequent collocating adjective. Additionally, the mutual information is also relatively low in comparison to other collocations.

The collocations and mutual information is analysed for the noun *sweets*. Firstly, the collocations and mutual information for *sweets* is displayed in Table 3.17. Unlike the other nouns analysed previously, the noun *sweet* has larger mutual information score in noun and noun collocations and noun and adjective collocations.

Table 3.17. ***Sweets*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	fats	9.29	eat	5.53	sugary	9.82
Australian	coffee	5.61	eat	4.79	traditional	4.32
British	chocolate	7.45	eat	5.34	boiled	9.33
Canadian	craving	10.69	eating	5.79	sugary	9.73

The first examples are for noun and noun collocations with the noun *sweets*:

1. *American: Retail price increases during that time were much lower for sweets and fats than for vegetables and fruit.*

2. *Australian: Offering an array of coffee and sweets including; extra-large cinnamon buns, giant cupcakes and red velvet cake all designed for sharing, the? JavaBlue Cafe? will be located adjacent to the? Shake Spot? on Promenade Deck.*
3. *British: And you know, I prefer it when it's topped with chocolate sweets rather than jelly type sweets.*
4. *Canadian: I've got my cravings for sweets under control now and things are starting to balance out.*

All the noun collocations seen with the noun *sweets* are very different. The lowest mutual information score is for Australian collocation *coffee*. However, the rest of the noun collocations have a high mutual information score, with Canadian collocation *craving* having the highest mutual information score out of all collocations with the noun *sweets.*

The next examples are of noun and verb collocations with the noun *sweets*:

1. *American: Myth: People with diabetes can't eat sweets or chocolate.*
2. *Australian: I can even eat sweets again with absolutely no problems!*
3. *British: Was the agreement that you wouldn't eat sweets, or that you would lose weight.*
4. *Canadian: He also enjoys playing soccer, strumming the guitar, and eating sweets.*

With these verb collocations, the verb that collocates the most with the noun *sweets* is *eat.* Additionally, this verb is a collocation in all four English varieties. The only difference is that for American, Australian, and British English the form which collocates the most with *sweets* is *eat*, while in Canadian English it is the form *eating.*

The last examples are with noun and adjective collocations for the noun *sweets*:

1. *American: This is caused by all the processed foods and sugary sweets that is cheap.*
2. *Australian: Imagine this: you wake up at 5am to cook a whole batch of traditional sweets to offer guests at the party you're hosting that night.*
3. *British: It is fashionable to serve boiled sweets after a dinner party...' After Eights' are really OUT.*
4. *Canadian: My cravings for sugary sweets, bread - all gone.*

In the case of adjectives, the adjective *sugary* is used in both American and Canadian English variants. While for Australian English the collocate is *traditional*, but in British English, the collocate is *boiled.* There is a significant difference in the mutual information scores, where American, British, and Canadian English the score is above 9, while in Australian English the mutual information score is above 4.

The next is the analysis of the noun pair *elevator* and *lift*. The statistical data and the collocations are seen in Table 3.18.

Table 3.18. ***Elevator*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	guy	4.42	ride	4.63	express	3.46
Australian	pitch	8.27	took	3.50	private	3.33
British	pitch	7.34	stuck	4.88	stuck	3.88
Canadian	pitch	8.06	take	3.08	co-operative	7.40

Firstly, these are the examples with noun and noun collocations for *elevator*:

1. *American: I've been in elevator guy's position, I've been in plenty of situations where I've wished desperately to be cool and suave and charming and just ended up making a young woman uncomfortable.*
2. *Australian: Many experts will tell you to hone your 60 second elevator pitch.*
3. *British: Develop a description of yourself and what you do as a written statement, and as a verbal statement (an' elevator speech' or' elevator pitch' - so called because it makes a successful impact in the time you share an elevator with someone who asks: "What do you do? ").*
4. *Canadian: The elevator pitch initially grabbed my attention and I love the value proposition.*

Three out of four varieties have the same most frequent collocation – *pitch.* It is necessary to also point out, that same as with *ski lift* and *hydraulic lift*, *elevator pitch* is a set phrase, which is not replaced by *lift.* When it comes to American English, they have a different collocation, which is also explained in the example.

Further on, the next examples is with noun and verb collocations for the noun *elevator*:

1. *American: The park offers guaranteed fun whether you ride the elevator or take the stairs.*
2. *Australian: We took the elevator down to the third floor.*
3. *British: It could also be useful if you find yourself stuck in a broken elevator with four other people and one of them is a psychopath that wants to mess you about and make you believe that one of the other people in the elevator is Satan.*
4. *Canadian: If you're not afraid of heights, take the transparent elevator to the top.*

In verb collocations, Australian and Canadian English have the same verb as the most frequent one used, but they are used in different forms, while American and British English used different collocations.

Last of the examples for collocations with *elevator* are for noun and adjective collocations:

1. *American: It was an express elevator, with no stops below the thirty-ninth floor, and the building was deserted.*
2. *Australian: A private elevator leads to the suite, or more precisely four-floor, four-bedroom apartment.*
3. *British: Me and my mum got stuck in the glass elevator for 15 mins.*
4. *Canadian: There exists no newspaper, no nurse, no veterinary surgeon, no doctor, no Protestant church, no co-operative elevator, no practicing lawyer, no homemakers' club, no hospital, no creamery, no printing establishment. and practically no telephones.*

In the case of noun and adjective collocations for *elevator,* each variant has a different most frequent collocation, but what is noteworthy, is that all adjectives precede the noun without any insertions.

The next analysis is of the noun *lift.* Firstly, the statistical data of *lift* and its examples is presented in Table 3.19.

Table 3.19. ***Lift*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	ski	8.21	dropped	3.01	hydraulic	7.49
Australian	stairs	7.29	ride	3.03	crowded	2.84
British	floor	4.49	stuck	4.10	hydraulic	7.53
Canadian	chair	7.78	accessed	5.85	hydraulic	7.67

The first examples are with noun and noun collocations for *lift*:

1. *American: I had a group on which included the ski lift and all you can eat buffet.*
2. *Australian: Take the stairs instead of the lift and walk up escalators.*
3. *British: They took the lift to the second floor.*
4. *Canadian: Most of the hill was serviced by just one double chair lift line.*

Lift collocations with nouns are very different in all varieties. One peculiar instance is of *ski* and *chair* because these words are related to skiing. What is worth noting is that in skiing *lift* is used in all variants, because *elevator* is not used to describe this lifting device.

The next examples are with noun and verb collocations for *lift*:

1. *American: I stopped for a minute and looked around for something else to do, or something else to occupy myself with for the two or three minute wait as the lift dropped someone off up high and then came down to collect us.*
2. *Australian: Do I have to buy a ticket if I want to use the terrain park and not ride the lift?*
3. *British: It's a book about campaigners, veterans, enthusiasts and computer geeks, as well as Twitter, trees, and Stephen Fry stuck in a lift.*

4. *Canadian: In the summer months, our hostel is right next to some of the best lift accessed downhill mountain biking trails in British Columbia and Canada.*

The same as with noun and noun collocation, with verbs each variant also has its own most frequent collocations, which are completely different from each other. In all the cases with these collocations, the words are side by side or separated only by the functional parts of speeches.

Last examples for the noun *lift* are with noun and adjective collocations:

1. *American: A decade after his graduation from Gregory's International Health Institute, a team of city firemen needed a hydraulic lift remove the ailing high from the Brooklyn apartment where he'd spent the last five years.*
2. *Australian: Aromatherapy - have you ever dropped a quiet one in a crowded lift, and asked loudly' can you smell petrol'?*
3. *British: Each 16-person two-stop hydraulic lift was designed by Stannah.*
4. *Canadian: The hydraulic lift on the plow had a hand pump.*

The collocations with adjectives indicate that the most frequent collate in three varieties is *hydraulic*, while in Australian the most frequent case is *crowded.* It is also important to mention, that similarly with lifts used for skiing, hydraulic lift are also something that tend to appear in all varieties and not with the noun *elevator*.

Further on, the next noun pair up for analysis is *gasoline* and *petrol.* The collocations and mutual information score is presented in Table 3.20.

Table 3.20. ***Gasoline*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	prices	7.50	reduce	3.22	higher	3.38
Australian	prices	6.56	fell	4.67	conventional	7.07
British	prices	6.61	drinking	6.32	higher	3.68
Canadian	prices	6.87	rising	4.82	cheaper	7.40

The first examples of noun and noun collocations for the noun *gasoline* are:

1. *American: Gasoline prices increased due to the poor economy or something.*
2. *Australian: However, survey data indicated that retail gasoline prices were about flat in early October.*
3. *British: High gasoline prices make for angry constituents.*
4. *Canadian: The oil industry doesn't like high gasoline prices any more than you do.*

All the collocating nouns for *gasoline* are the same in all four English varieties. Additionally, the mutual information score is also almost identical, with American English being slightly in front of the other English variants.

Further on, the examples with noun and verb collocations for *gasoline* are seen below:

1. *American: The best way to reduce your gasoline costs is simply to drive less.*
2. *Australian: Gasoline stockpiles nationwide fell 59,000 barrels, the data showed.*
3. *British: Also bear in mind that these vehicles have a habit of drinking more gasoline than compact or saloon style cars.*
4. *Canadian: The discussions came amid rising gasoline prices.*

Unlike with nouns, the verb collocations are very different. Their mutual information scores are also quite low, except for British English with the collocation of *drinking*, which has the MI score over 6, while the rest are around 4.

Following are the examples with noun and adjective collocations for the word *gasoline*:

1. *American: We think US households can absorb the demand shock of higher gasoline prices as long as there is a corresponding expansion of employment and income.*
2. *Australian: The only source of power is conventional gasoline.*
3. *British: Higher gasoline tax would encourage people to use oil more conservatively.*
4. *Canadian: Cheaper oil exports would mean cheaper gasoline for the Americans (though not for Canadians).*

For American and British English *higher* is the one adjective which has the highest frequency of collocating adjectives. While Australian and Canadian English have a different set of collocating adjectives. The more peculiar aspect is that Australian and Canadian English collocation are also the ones with a higher MI score, which is above 7.

The other half of the noun pair up for discussion for the GloWbE corpus is *petrol*. Statistical data about *petrol* can be seen in Table 3.21.

Table 3.21. ***Petrol*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	station	7.39	thrown	5.47	cheap	4.34
Australian	station	6.91	sniffing	10.39	unleaded	11.33
British	station	7.23	throwing	4.85	unleaded	11.36
Canadian	stations	8.43	ran	6.30	leaded	13.55

The first examples are of noun and noun collocations for the word *petrol*:

1. *American: Elsewhere in the country, scattered protests went off peacefully but a petrol station was burnt by angry youths in the country's second largest northern city of Irbid.*

2. *Australian: I loved this little old petrol station in Speers Point.*
3. *British: His father had a petrol station in the village.*
4. *Canadian: Privatisation of new Petrol Stations is also under consideration.*

All the collocations with petrol are the same for all the four varieties. The only difference is that American, Australian, and British English has the singular form of the noun as the most frequent collocation, while Canadian English has the plural for *stations* as the most frequent collocation. At the same time, the mutual information is relatively close for each variant.

These are the examples with noun and verb collocations for the word *petrol*:

1. *American: While they are packing, a petrol bomb is thrown through the front door of their flat, highlighting the extreme danger they face.*
2. *Australian: I once went to a conference on petrol sniffing representing the Commonwealth of Australia.*
3. *British: Meanwhile, up to 7,000 took to the streets, throwing petrol bombs and rocks at riot police who responded with tear gas.*
4. *Canadian: The last time I was in a car that ran out of petrol was 10 years ago or so, on a snowy winter night back in Ottawa.*

This is an interesting case, because for American and British English, the both have the verb *throw* as the main collocation, but in different forms, the highest mutual information score belongs to Australian English. Their collocation with *petrol* is the verb *sniffing.*

Lastly, the examples with noun and adjective collocations for *petrol* are seen below:

1. *American: You can't get labour anywhere as cheap as the petrol we buy.*
2. *Australian: They simply shifted the auto industry in a new direction, away from leaded petrol and towards unleaded petrol.*
3. *British: All petrol stations provide unleaded petrol and diesel.*
4. *Canadian: Leaded petrol (70fils a litre) is only available at certain stations.*

This is an interesting case, because there is the collocation with the smallest mutual information score and the highest mutual information score. American English has the collocation *cheap*, which has a low MI score, while Canadian English has the highest MI score for the adjective *leaded.*

Further on, the analysis of the pair *pants* and *trousers* and their collocation is carried out. The linguistic data for *pants* is seen in Table 3.22. The case of *pants* is very interesting, because every collocation is the same, except for one instance, which will further on be discussed with the examples.

Table 3.22. ***Pants*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	pair	6.86	wear	6.70	black	3.93
Australian	pair	7.01	wear	6.13	black	4.32
British	pair	6.27	wear	5.92	hot	5.03
Canadian	pair	6.74	wear	6.39	black	4.87

The first examples are with noun and noun collocations for *pants*:

1. *American: She was wearing a loose white shirt tucked into a pair of tight leather pants.*
2. *Australian: I'm 150cm tall and haven't had a pair of pants that I haven't had to take up.*
3. *British: Now we have a hole in this pair of pants.*
4. *Canadian: You have at least 1-2 dresses and 1 pair of pants too many.*

All varieties have the same noun as a collocation for the word *pair.* Additionally, all the collocations are the same form *pair of pants*, with the exception of American English, where collocation is extended with additional adjectives.

Further on, the examples with noun and verb collocations for *pants* are displayed below:

1. *American: I am so excited to actually be able to wear these pants!*
2. *Australian: Unfortunately I can not for the life of me find pants to wear at the office, it's like they're made for men!*
3. *British: That's right... I wear the pants in this family!*
4. *Canadian: Wear long pants and a long sleeved shirt.*

When it comes to collocations with verbs in all four varieties they are exactly the same, the only difference is the word placement in the sentence.

Last examples are with noun and adjective collocations for the word *pants*:

1. *American: Problem is, I need more black pants, no polyester, and all of them are those stupid hip huggers that were awful the last time they were in style too.*
2. *Australian: I did not pick a red jumper and black pants by some accident or wardrobe malfunction.*
3. *British: Their broad design makes them similar to mini hot pants and ideal for sports.*
4. *Canadian: In black pants and a maple leaf-red shirt, she dressed like Tiger Woods on Sunday.*

Three out of four varieties have the same collocation with *pants*, with the exception of British English, who have the collocation of *hot* used with the noun *pants.*

The next statistics are of the noun *trousers* that are presented in Table 3.23. Similarly to *pants*, there are only slight differences in these collocations.

Table 3.23. ***Trousers*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	shirt	8.25	wear	6.76	white	4.28
Australian	pair	7.46	wearing	6.94	black	5.69
British	pair	7.02	wear	6.49	black	4.85
Canadian	pair	8.10	wearing	6.58	white	5.39

In the place are the examples with noun and noun collocations for the word *trousers*:

1. *American: You can wear cheaper trousers and shirt with a brand name jacket.*
2. *Australian: The only thing missing is a pair of golf trousers and reference to a frat party.*
3. *British: Beyoncé looked fabulous in a pair of leather trousers tucked into some knee-high boots.*
4. *Canadian: I tried on a pair of high-wasted trousers last week and loved them!*

Three out of four varieties have the collocation *pair* for the word *trousers.* In American English, the most frequent collocation is with *shirt.*

The next examples are with noun and verb collocations for the noun *trousers*:

1. *American: I shall wear white flannel trousers, and walk upon the beach.*
2. *Australian: I tried wearing trousers and proper shoes on previous trips.*
3. *British: Might be advisable to wear brown trousers and a shirt the colour of blood.*
4. *Canadian: He is wearing plaid trousers two sizes too small and he appears to be dancing.*

Unlike with noun and noun collocations, in this case, the verb is the same in all the varieties. However, the most frequent form for American and British English collocations is *wear*, while for Australian and Canadian English, the most frequent form of the verb is *wearing.*

The last case for the noun trousers are the examples with noun and adjectives collocations:

1. *American: They wear large blue turbans, scarlet coats and white trousers.*
2. *Australian: The elf was clothed in black trousers and a sky blue shirt of fine tailoring.*
3. *British: A tall man in black uniform trousers and blue shirt.*
4. *Canadian: Dark-blue trousers with a double white stripe were issued for dismounted duties.*

In this collocation case, American and Canadian English has the most frequent collocation of *white.* While in Australian and British English the most frequent collocation is with *black.* However, these collocations are not too different, because both are created with an adjective signalling towards the colour of the item, it just the matter if which colour is used the most often in terms of collocating with the noun in question.

The next collocation analysis is of the noun pair *subway* and *underground.* The most frequent collocations, which are above the score of 3 in mutual information for *subway* are presented in Table 3.24.

Table 3.24. ***Subway*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	system	4.75	ride	4.85	crowded	6.57
Australian	station	6.98	closed	4.83	underground	6.62
British	station	7.10	riding	5.63	central	3.91
Canadian	station	7.16	built	3.26	downtown	3.55

To better understand the collocations that are seen in Table 3.24, the examples for each collocations are necessary. Firstly, these are the examples for noun and noun collocations with *subway*:

1. *American: Our subway system is small but cool.*
2. *Australian: The South Ferry subway station, near Battery Park, is flooded with seawater.*
3. *British: There is a lack of good transportation options for the city. There is no subway station, the buses are rare and uncomfortable, and the taxis are quite expensive.*
4. *Canadian: Kipling station and numerous other subway stations have elevators.*

Similarly to noun collocations with *underground*, in three out of four noun collocations the most frequent one is with *station.* However, here the outlier is the American variant, in which the most frequent collocation is with the noun *system.*

The next examples are with noun and verb collocations for the word *subway*:

1. *American: I live in New York City, I ride the subway pretty much everywhere.*
2. *Australian: The New York Times says the city's subway may be closed for four or five days due to flooded tunnels.*
3. *British: I leave my hair and make-up how it is because I love riding the subway looking like a freak.*
4. *Canadian: Imagine if the subway had built in 10-minute layovers for every train?*

With these examples some similarities can be seen. In the case of American and British English variants, both of them have the verb collocation of *to ride,* but in different verb forms. While inn Australian and Canadian English variants the collocations indicate different contexts.

The last examples for the word *subway* are the noun and adjectives collocations:

1. *American: Just try opening a laptop on a crowded subway train or bus.*

2. *Australian: Trams ran until 1968, only to disappear almost overnight. They were quickly replaced by an underground subway system.*
3. *British: The Central Subway is an essential addition to our local transit network.*
4. *Canadian: That being said, many downtown subway exits still involve stair.*

Similar to the adjective collocations for the noun *underground,* the collocations with *subway* differ in all four variants. Additionally, there is a difference between British English and the other variants, because unlike American, Australian, and Canadian English, in British English the collocation is an adjective modifying a noun, and it is also a part of a proper name.

In Table 3.25 the collocations for the word *underground* are presented. The collocations in each category with its mutual information (MI) can be seen side by side for each of the English language variants.

Table 3.25. ***Underground*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	station	4.76	confirmed	3.89	famous	3.68
Australian	rail	7.16	referred	4.02	ornate	9.38
British	station	6.99	travelling	4.22	nearest	5.80
Canadian	station	4.34	headed	4.96	Biotronic	14.48

First, the analysis of noun and noun collocations can be analysed. Although the mutual information for these collocations differ, in three out of four national English variants, the most frequent collocation with *station,* but Australian English has the collocation rail, which also has the highest mutual information ration out of all four variants. To better understand the collocation, it is fundamental to establish examples from each English variant. These are the examples of noun and noun collocations in the same order that is presented in Table 3.25:

1. *American: The plan was supposed to be quite simple: take the Underground to the appointed station.*
2. *Australian: Building the Melbourne Underground Rail Loop was an extraordinary undertaking.*
3. *British: From the Underground station turn left along Victoria Embankment.*
4. *Canadian: I made that familiar approach from the Underground station, down the dark alley, past tourists posing for photos in front of the cathedral.*

From the examples above, it can be seen that in American, British, and Canadian English the context is very similar, where someone is going towards or from a certain underground station, with the Australian example indicating towards building a part of an underground.

The next example are or noun and verb collocations:

1. *American: I also spent about 30 minutes on the phone with JCP &L yesterday and they confirmed the Underground Cable Fault.*
2. *Australian: The other iconic mode of public transport is the London Underground, referred to as "the tube" by locals.*
3. *British: Travelling on the Underground is the quickest way to get around London.*
4. *Canadian: I still had promise of a warm flat waiting for me, so I wandered toward the Underground and headed home.*

In the case of noun and verb relationship with the noun *underground*, the collocations are very different. This is a completely different situation from what was seen with noun and noun relationships.

The last examples for the noun *underground* are its adjective collocations. These are the examples with each of the four English varieties:

1. *American: He was the most famous Underground "Stationmaster" in history.*
2. *Australian: From the famous domes of St Basil's Cathedral and the ornate underground of the Russian Metro, you will see the diversity and contrast of this city and its grand past to what it is evolving to be today.*
3. *British: The nearest Underground Station is Uxbridge.*
4. *Canadian: Their base - known as "the lab" - is a disused station in the London Underground that includes their biotronic computer TIM, voiced by Philip Gilbert.*

All the adjective collocations are very different and touch upon various contexts. However, there is a difference in the placement of the collocations. The programme has recognised a collocation in Canadian variant that stands further apart, unlike in the other three variants, where the adjective precedes the noun.

The last pair up for discussion is *vacation* and *holiday.* In Table 3.26 the linguistic data for *vacation* is presented.

Table 3.26. ***Vacation*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	family	3.75	taking	3.06	paid	3.76
Australian	family	3.43	planning	4.07	relaxing	6.48
British	summer	5.39	enjoy	3.82	paid	3.13
Canadian	pay	4.43	taking	3.02	annual	4.43

The first examples are for noun and noun collocations with the word *vacation*:

1. *American: For the past three years I have taken my family to vacation exclusively in Franconia during Christmas break and to ski Cannon.*

2. *Australian: THE best lake in the country to plan a family vacation is Pickwick Lake!*
3. *British: You can work on-campus 40 hours a week during breaks and the summer vacation.*
4. *Canadian: The vacation pay is given at the start of the vacation or when employment is terminated.*

For American and Australian English variant users, the most frequent collocation with *vacation* is the noun *family.* In British English, the most frequent collocation is with *summer* and for Canadian English it is *pay.* Because all of these noun collocations are frequently used words outside of the collocations with *vacation*, the mutual information appear to be so low. If these words were rare and exclusively used with *vacation* the mutual information would be much higher.

The second case of examples is with the noun and verb collocations for the noun *vacation*:

1. *American: They might be taking an expensive vacation which is giving them spending guilt deep down.*
2. *Australian: When planning a vacation, try finding seminar or conferences offered in your field to attend.*
3. *British: Thanks for a wonderful place to enjoy our vacation.*
4. *Canadian: Debating over taking a vacation this summer?*

In the case of verbs, American and Canadian English has the same verb used in the same form that is the most frequent collocation used. While for Australians it is *planning* and for British English it is *enjoy*. Despite that, the mutual information with these collocations is also no higher than 4.

Last but not least, are the examples with noun and adjective collocations for *vacation*:

1. *American: They will receive one additional week of paid vacation time going forward.*
2. *Australian: Adelaide can also provide a more relaxing vacation.*
3. *British: My husband got 2 weeks paid vacation, but I did not have any paid vacation.*
4. *Canadian: To avoid burnout, you also need an annual vacation (4 weeks).*

In this case, American and British English have the same most frequent collocation – *paid.* Canadian English has *annual* as the most frequent collocation. When it comes to Australian English, the most frequent collocation is with the adjective *relaxing;* however, it also has the highest mutual information score, which is 6.48.

Further on the collocation analysis of *holiday* and *vacation* is discussed. In Table 3.27 the collocation and mutual information for *holiday* is presented.

Table 3.27. ***Holiday*** **collocations in GloWbE**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	season	6.54	celebrate	5.57	annual	4.53
Australian	family	3.30	planning	3.77	public	3.68
British	bank	4.86	booked	5.67	lovely	3.13
Canadian	season	5.76	shopping	4.29	public	3.87

The first instance is of the examples with noun and noun collocations for *holiday*:

1. *American: It's peak holiday season in NYC so be sure to plan ahead.*
2. *Australian: A family holiday overseas can be a memorable experience for both parents and kids.*
3. *British: As it was bank holiday, the phone could not be picked up until Tuesday 28th.*
4. *Canadian: I hope that you all have a wonderful holiday season!*

In American and Canadian English the collocations are the same, with the combination of *season* and *holiday*. While for Australian and British English, the collocations are different.

The next instance is the examples with noun and verb collocations for *holiday*:

1. *American: What a way to celebrate the holiday season!*
2. *Australian: Where do you even start with planning a holiday?*
3. *British: My partner, Kevin, proposed and we booked a holiday to Dubai for next summer.*
4. *Canadian: Maybe they're all out getting their holiday shopping done together.*

With verb collocations, the verbs are different in each collocation. Additionally, the mutual information is also similar for the collocations, which means that none of these collocations are stronger than others in this list.

Lastly, the collocation examples with nouns and adjectives for the word *holiday* are seen below:

1. *American: This past weekend was an annual holiday bazaar and yard sale at a local church.*
2. *Australian: Australia Day is a designated public holiday and is a time for Australians of all backgrounds to celebrate national unity.*
3. *British: We're having a lovely holiday in Tuscany.*
4. *Canadian: Good Friday is a public holiday at a national level in Canada.*

In adjective collocations, Australian and Canadian collocations are the same, with *public* collocating with *holiday*. For American and British collocations differ. However, here also mutual information is similar, that is why it is hard to determine which the strongest collocation is.

This subchapter is dedicated to the analysis of the noun collocations in the GloWbE corpus. The main aspects up for analysis were collocation of each noun with other nouns, verbs, and adjectives, which all are lexical parts of speech. All the collocation chosen had a minimum mutual information score of 3.

3.3.4. Noun collocation in CoFNEV

In this subchapter, the analysis of noun collocation in the Corpus of Four National English Varieties (CoFNEV) is carried out. The procedure followed in the chapter is the same as the one carried out in the previous subchapter, where first the table with the collocating words and the mutual information is presented, with exampled displayed after it. This corpus posed some challenges due to it having a relatively smaller size in comparison to GloWbE. That is why in most cases some varieties did not have any examples or collocations for the nouns, because the nouns do not exist in this particular corpus. Often, for those nouns that can be found, one collocating part of speech was not present. However, the information found can still offer some insight and provide the opportunity to compare these corpora. Additionally, it has to be mentioned that due to the size of the corpus, the mutual information scores will appear more substantial. The same collocation selection methods were applied when selecting the collocations from CoFNEV corpus.

The first noun pair for analysis is *candy* and *sweet*. The word *candy* has no collocations in the Australian variant, while it is present in other variants. These collocations and statistical data is presented in Table 3.28.

Table 3.28. ***Candy*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	cane	11.75	eating	7.41	dang	10.91
Australian	0	0	0	0	0	0
British	cane	11.54	0	0	burnt	11.3
Canadian	store	6.29	sell	6.46	Suger-free	10.42

These are the examples for the noun collocations:

1. *American: Draw out a simple holiday shape like an ornament, candle, tree, or candy cane that's about twice as big as your gift.*
2. *British: Hide a selection of candy canes around your home – the more obscure the place the better – and watch as your family roam around with excitement.*

3. *Canadian: For fresh, fantastic fudge, stop into the Café Arômes et Saveurs, which not only offers great coffee in the back but is attached to a huge candy store.*

Interestingly, there is a case where in two variants there is the same collocating noun. Those are American and British English variants, where *cane* collocates with *candy.* This might be due to the fact that that is a name of a specific type of candy. This collocation also has a high MI score present in both varieties.

The following are examples for verb collocations with *candy*:

1. *American: I almost cried laughing when I saw this photo because I've never seen three people look more miserable while eating candy.*
2. *Canadian: They look like little red Bus shelters, but they sell candy, newspapers, cigarettes and other newsstand type things.*

Surprisingly, the American collocation is also similar to what can be seen in GloWbE corpus. Additionally, the mutual information score is similar in the CoFNEV corpus for both verb collocations, indicating to the words having a strong collocation relationship. When comparing the results obtained from CoFNEV to the data present in GloWbE corpus, it is revealed that the verb collocations are similar with the verb collocation not only with *candy* but also with *sweets.* This observation shows that collocation use is not dependent on the varieties.

Lastly, these are the examples for adjective collocations:

1. *American: While I love the spirit of Halloween and the trick or treating and the pre parties, the kids usually come away with a sugar high that just won't quit and so much dang candy!*
2. *British: A visual feast for all the senses; smoke from open campfires billows into the frosty darkness, the aroma of burnt candy and caramelised almonds permeates the air and the regal sounds of bugles signal the start and end of the festivities perfection in a Christmas Market!*
3. *Canadian: These sugar-free candies typically have less than 5 grams of sugar in a 50-gram bag – great news for everyone with a sweet tooth.*

These adjective collocations are not similar among the varieties. As well as, they are not similar to GloWbE collocations for *candy*. However, these do have a high mutual information score, which is above 10 for all the adjectives, showing a strong link between the words.

Following is the other half of noun pair *sweets.* The collocation data is presented in Table 3.29.

Table 3.29. ***Sweets*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	notes	9.07	cramming	11.68	rare	11.09
Australian	0	0	love	5.2	minimal	12.42
British	foods	6.03	loves	5.2	retro	11.44
Canadian	0	0	0	0	0	0

These are the examples for American and British English noun collocations:

1. *American: We would wrap them up with notes and sweets, and leave them on the porch.*
2. *British: Sweets and junk foods provide calories, but not the needed nutrients.*

These collocations differ and indicate a sequence of items rather than nouns appearing with the noun *sweets.* Despite that, the mutual information still is shown as 6 and 9, which is a high result. Thus, in this corpus, these words collocate due to the context in which they are used.

The next are examples of verb collocations with *sweet*:

1. *American: I figured I'd spend today doing nothing but cramming sweets into my mouth, but right now I feel fine.*
2. *Australian: I love sweets and love to offer them to my boys, but I try to keep is as low as possible*
3. *British: For example, my Dad loves cheese so I would get him a chess hamper but my husband loves American sweets so I would get him a hamper of American chocolates and sweets.*

This is one of the rare cases, where there are identical collocations in two varieties. In this case those are Australian and British, who have the collocation *love* for the noun *sweets.* Interestingly, they also have identical MI scores, although the verb form differs, they are used equally frequently in both varieties.

The last examples are of adjective collocations with the noun *sweets*:

1. *American: It's an impulse control issue dating back to my childhood, when sweets were rare (no one in my family has a sweet tooth, and they are not bakers).*
2. *Australian: Sweets are very minimal around here but as you mention, need to be there so there's not that underlying fascination to want to.*
3. *British: From makeup and candle filled calendars to nail polishes, alcoholic miniatures, retro sweets and bath bombs, there is something out there for everyone.*

The adjective collocations are quite different from each other; however, all of them have a high mutual information score. Not only are they different in CoFNEV, they also differ from GloWbE collocations.

The next noun pair is *elevator* and *lift.* Firstly, in Table 3.30 the collocations for *elevator* are presented. There are no collocations from Australian and British English.

Table 3.30. ***Elevator*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	Brooklyn	7.27	blocked	8.45	Senators-only	13
Australian	0	0	0	0	0	0
British	0	0	0	0	0	0
Canadian	0	0	getting	10.09	copper	10.91

Below is an example of the American noun collocation:

1. *American: A woman was doused with a flammable liquid and burned alive in a Brooklyn elevator this afternoon – and police are searching for her killer, law enforcement sources said.*

This is a proper noun collocation, which also has the most noticeable mutual information score. Thus, in this corpus there are several cases where the context is about *elevator* in *Brooklyn.* It is no surprise that there are no collocations in British and Australian English, because *elevator* is not the chosen noun in these varieties.

For verb collocations, these are the examples available for American and Canadian English:

1. *American: On Friday morning, two women raced past reporters and security officers and blocked a senators-only elevator in the US Capitol.*
2. *Canadian: Just getting off the elevator there is an immediate sense of calm as you get sight of the pool through the window.*

Both of these collocations have a high mutual information score. The collocations differ in this corpus and differ from what can be seen in the GloWbE corpus.

The last examples are of adjective collocations for the noun *elevator*:

1. *American: On Friday morning, two women raced past reporters and security officers and blocked a senators-only elevator in the US Capitol.*
2. *Canadian: It was designed to in still fear and show power with an impressive copper elevator cut through the mountain to take people to the top.*

In American English variant, the collocation has a high MI score; however, that is due to this word combination being exclusively used together in this corpus. There is a chance that if the corpus was considerably larger size, the score would be a lot different. The same applies also to Canadian collocation.

Secondly, the noun *lift* is displayed in the Table 3.31. For this noun there are no cases for American English, as well as, only a noun collocation for Australian English.

Table 3.31. ***Lift*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	0	0	0	0	0	0
Australian	panels	8.37	0	0	0	0
British	system	5.89	features	5.85	rapid	8.96
Canadian	platform	10.33	taking	4.26	High-speed	9.87

Below, the examples of noun collocations with *lift* are displayed:

1. *Australian: Door handles, pedestrian crossings, lift panels, handrails, bin lids, kitchen taps, vending machines – pulling, pushing, pressing, turning.*
2. *British: It boasts the most modern lift system you could wish for.*
3. *Canadian: Climb the steps for a tantalizing peek of the Colosseum and Roman Forum just beyond, or see the whole historic center by taking the platform lift to the top.*

In this case, all the noun collocations are different. The collocation with the lowest MI score is *system* which is a British collocation, while the highest is Canadian *platform.* These collocations also differ from the GloWbE collocations.

For verb collocations, there are only cases for British and Canadian English:

1. *British: Neustift is the closest sizable village to the Stubai glacier, features 28 lifts and 7382 feet of vertical descent in the winter.*
2. *Canadian: Dome €6 taking steps, €8 taking lift to terrace then steps.*

There are also different verb collocations in the varieties; however, the Canadian collocation is similar to what can be seen in GloWbE corpus for the noun *elevator*. That indicates to similar collocation use between the nouns and beyond one variety.

Lastly, the same two varieties have examples for adjective collocations:

1. *British: Within a few hours you can reach the summit of the Trois Vallées massif, the Pointe de Thorens (3266 m) at Val Thorens, by means of the efficient and rapid ski lifts built in the area.*
2. *Canadian: It has high-speed chair lifts and is a great place for beginners.*

Although, here the adjective are also different, it has to be mentioned that both of these adjectives have a similar meaning. As well as, the mutual information for both collocations is quite noticeable, which could mean a similar usage of the collocations.

The data obtained from CoFNEV on *lift* and *elevator* has only one similarity with the GloWbE corpus. This similarity is the Canadian verb collocation, which coincides with data from other varieties seen in GloWbE corpus.

The next pair is *gasoline* and *petrol.* With the noun *gasoline,* there are only two collocations found in the CoFNEV corpus. The statistical data can be seen in Table 3.32.

Table 3.32. ***Gasoline*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	prices	5.46	adding	5.87	0	0
Australian	0	0	0	0	0	0
British	0	0	0	0	0	0
Canadian	0	0	0	0	0	0

Below, the American English noun and verb collocation examples can be seen:

1. *Noun: As gasoline prices continue to soar, city drivers yesterday began dreading what may be the inevitable – the $5 gallon.*
2. *Verb: And I shouldn't be adding gasoline to that already overrun fire.*

The noun collocation follows the same pattern as seen in GloWbE, because the collocation also is *price.* While the verb collocation differs from what has been seen previously.

The next is the case of *petrol*, with an exception for Canadian English, who use the spelling *petroleum* is used instead of *petrol.* The collocations and mutual information score is presented in Table 3.33.

Table 3.33. ***Petrol*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	0	0	0	0	0	0
Australian	bomb	11.36	ration	11.54	horrible	12.19
British	station	7.79	pour	8.81	0	0
Canadian	sale	7.73	exclude	9.83	0	0

There are no examples for American English. However, there are noun examples seen in other varieties:

1. *Australian: 'Kill their families,' rioters yelled as they hurled petrol bombs as a policeman and his family cowered.*
2. *British: Vigil at petrol station where police tasered dad in front of his 'screaming' son.*
3. *Canadian: Canadian Tire also took a hit from declining petroleum sales, as more people stayed home and off the roads during COVID-19.*

Australian English is the only one in this set, which has above 11 mutual information score; however, the other MI scores are also high and show a high probability of the usage of this collocation. While British English is the one, whose collocation is the same as in GloWbE corpus for the word *petrol.*

A different situation is with verbs, which examples can be seen below examples:

1. *Australian: Researchers suggest an age-old plan used to ration petrol and water may be a safe and quick way Australia can exit COVID-19 restrictions*
2. *British: Ms Dugdale has previously criticised David Cameron's inflammatory decision to announce he was pushing ahead with so-called "English votes for English laws" hours after Scotland voted No in 2014 and has accused Ms Davidson of wanting to "pour petrol" on the issue for electoral gain.*
3. *Canadian: Revenue in its retail segment fell 2.4 per cent to $2.5-billion; excluding petroleum, revenue declined 1.8 per cent.*

None of these collocations are similar to GloWbE corpus. Additionally, none of these collocations coincide with each other. Despite that, all of the verb collocations have a high MI score.

Lastly, this is the example for Australian English of the adjective collocations with *petrol*:

1. *Australian: And wow petrol is horrible there, I'm paying around $1.55 back in the country town where I'm from and I thought that was bad.*

This is the only adjective collocation in all the varieties with *petrol.* Additionally, it does not coincide with what could be seen in GloWbE corpus, despite that, it has a high MI score.

The analysis of *gasoline* and *petrol* in the CoFNEV corpus reveal two collocations that coincide with the data in GloWbE. These collocations are not restricted to their use in varieties. The noun collocation *prices* seen in CoFNEV corpus as a collocate for the noun *gasoline* is seen in the two corpora, alongside the noun collocation *station* which is used in relation to *petrol* in CoFNEV corpus.

The next noun pair is *pants* and *trousers.* Luckily, the case of *pants* has only one collocation type missing, which is noun collocation in Australian English that can be seen in Table 3.34.

Table 3.34. ***Pants*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	pocket	9.14	wear	8.19	white	8.39
Australian	0	0	wear	7	proper	8.71
British	T-shirt	9.64	wearing	5.42	patterned	12.68
Canadian	tops	10.18	include	6.39	long	7.29

The first examples are of the noun collocations that are displayed below:

1. *American: The hubby is originally from New York, and as such, he keeps his wallet in his front pants pocket.*
2. *British: The side-splitting video starts off with Ant preparing to shoot the golf ball off the tee as he dons a navy pair of golfing pants and a bright pink T-shirt.*
3. *Canadian: Layers – pack the basic tropical clothes and make sure to include long pants, long-sleeve tops and some warm clothes for those chilly nights.*

These collocations also have a high MI score. Additionally, they are quite different from what was seen in the previous subchapter of the collocations in GloWbE corpus.

A different situation can be seen with the verb collocations, which examples can be seen below:

1. *American: I probably won't do video chat because I don't like to wear pants or a bra when I'm home.*
2. *Australian: Boys will be able to wear skirts and girls will be able to wear pants under a new government initiative to offer cheaper, gender-neutral uniform options.*
3. *British: So would the mouth-watering cookies he baked before he went on to inform us all he wasn't wearing any pants.*
4. *Canadian: Gear rentals include pants, jackets, boots, mittens, and helmet.*

These collocations very much coincide with what was seen in GloWbE corpus. In three out of four varieties, the most frequent collocation is with the verb *wear* in different forms. The only different collocation is in Canadian English. This verb usage indicates to a strong argument of these words being collocates, due the connection among varieties and between the corpora.

These are the examples for the adjective collocations with *pants*:

1. *American: I like to wear these white wide-legged pants paired with a band tee or a vintage tee for a no-brainer outfit.*
2. *Australian: In two weeks you will scrape those PJs off your body and swear to wear proper pants.*
3. *British: Patterned pants are vivid, so you should be careful while combining them with the jacket.*

4. *Canadian: Avoid packing skirts or shorts – instead invest in a good pair of long pants made from a breathable fabric like linen.*

These collocations differ from each other. However, the American collocation is similar to what was seen in GloWbE corpus. The adjective collocation *white* is observed in various cases independently of the usage in varieties in the GloWbE corpus. This similarity indicated to the fact that collocations are created irrespective from the noun use in national varieties.

In comparison, a direr situation can be seen in Table 3.35, where the collocations for *trousers* are presented only in the British English variant.

Table 3.35. ***Trousers*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	0	0	0	0	0	0
Australian	0	0	0	0	0	0
British	waistcoats	13	wear	6.41	leather	9.33
Canadian	0	0	0	0	0	0

Because all of the collocation examples are from British English, they are displayed in one list below:

1. *Noun: The modern business suit can date as far back as the 17th century when those attending European royal courts would wear trousers, long waistcoats, and cravats.*
2. *Verb: The modern business suit can date as far back as the 17th century when those attending European royal courts would wear trousers, long waistcoats, and cravats.*
3. *Adjective: Grey knitted jumper, faux fur gilet and black leather trousers. This is such a cosy outfit.*

Although the mutual information scores are high, they still are rare in the corpus. Despite that, the noun only being present in the British varieties, there is a significant importance to the verb collocation *wear*. The verb collocation *wear* and the derived form *wearing* are present in almost every variety in both corpora. This observation leads to the conclusion that even if the mutual information in some varieties has a lower score, the united use in all the varieties supports the claim that this is a strong collocation with the nouns *pants* and *trousers.*

The next noun pair is *subway* and *underground.* The nouns have only a few instances of collocations. As seen in Table 3.36, there are no collocations for Australian English, British English lacks verb and adjective collocations, and Canadian lacks verb and adjective collocations.

Table 3.36. ***Subway*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	station	9.35	take	4.67	noisy	11.68
Australian	0	0	0	0	0	0
British	system	5.89	0	0	0	0
Canadian	cars	6.71	0	0	0	0

The examples presented below are of the noun and noun collocations for the word *subway*:

1. *American: A public service announcement related to the coronavirus is seen on screen in the Times Square subway station in New York City as the number of cases in New York state rise, March 5, 2020.*
2. *British: I bet their subway system is ten times better than NYC's.*
3. *Canadian: Vancouver city planner Brent Toderian – are predicting that continued public fears of physical proximity on buses and subway cars will help boost car use.*

All the noun collocations are completely different in all varieties where the noun is present. The most frequent collocation with the highest mutual information score is *station* used in American English. Both of the noun collocations with *subway* appear in GloWbE corpus; however, the collocation *station* is the one which appears the most in GloWbE for *subway* and *underground.*

There is only one example for American English for collocation with noun and verb sets for the word *subway*:

1. *American: You can also take the subway, but there isn't a stop terribly close by, so be ready to walk.*

This collocation also has a relatively low score for mutual information. Due to the small size of the corpus, this might indicate to the verb being used in different contexts more often than with *subway.*

There is also only one example for adjective collocation with *subway*, which is also present in American English. This is the example for the adjective collocation:

1. *American: I'd heard the subways were noisy.*

However, positively, this collocation has a high mutual information score.

The second noun up for discussion is *underground.* The statistical data for *underground* is presented in Table 3.37.

Table 3.37. ***Underground*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	0	0	0	0	0	0
Australian	0	0	0	0	0	0
British	London	7.89	0	0	0	0
Canadian	0	0	0	0	0	0

The challenge with *underground* is that it is only present in British English, and there is no verb and adjective collocation. That is why only an example for noun collocations is presented below:

1. *Noun: Building projects worth £1.5bn which depend on loans from the European Union, including signalling work on the London Underground and the construction of an offshore wind scheme in Merseyside, are in jeopardy as Brexit approaches.*

Unlike in the GloWbE corpus, here the most frequent collocation is *London*, which is understandable, due to it being the name of the underground system in London. The difference is also seen in that this is a proper noun collocation, while in GloWbE in the most frequent collocation were not with proper nouns.

The last pair up for discussion is *vacation* and *holiday*. This is the only pair which is represented in almost all instances except for one place. In Table 3.38, the statistical data for *vacation* is displayed.

Table 3.38. ***Vacation*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	spot	7.56	getting	5.06	great	7.11
Australian	backpack	10.61	planning	6.18	0	0
British	home	6.31	book	8.62	best	4.29
Canadian	time	6.02	plan	10.68	next	8.54

The first examples are of noun and noun collocations:

1. *American: Favorite vacation spot: Santa Cruz on the beach.*
2. *Australian: Me and my husband live in Melbourne and we are planning a 3 week backpack vacation to Europe this December, a cheap and best one, but I'm still struggling with where to start from!*
3. *British: A couple head to their vacation home out in the country but are quickly terrorized by masked psychopaths in the night.*
4. *Canadian: Can my employer force me to use my vacation time during a business closing?*

These collocations are very different from what was seen in the previous chapter about the GloWbE corpus. Additionally, they all have a high MI score, with Australian *backpack* having the highest MI score out of all varieties, meaning that all of these collocations are strong collocations in this corpus.

The next examples are of noun and verb collocations:

1. *American: The hard-hit hospital's front-line workers will be getting a three-night complimentary vacation for them and a loved one.*
2. *Australian: Me and my husband live in Melbourne and we are planning a 3 week backpack vacation to Europe this December, a cheap and best one, but I'm still struggling with where to start from!*
3. *British: If you want to have your kids all to yourself, this is the best time to book a vacation and it is also sensible because the kids are off school and have free time.*
4. *Canadian: This Antarctica travel guide will help you plan your next vacation.*

In the case of verb collocations, Australian and Canadian English have the same verb as the most frequent collocation in GloWbE, except having different verb forms – *plan* and *planning.* The rest of the collocation are different but convey similar meanings.

Next examples are of noun and adjective collocations, where there is no example for Australian English:

1. *American: And remember – part of the reasons vacations are great is that they're vacations.*
2. *British: For the discerning traveller wanting to escape the Christmas crowds this winter here are the best winter vacations in Europe for travel snobs...*
3. *Canadian: This Botswana travel guide will help you plan your next vacation.*

For American and British English, although the adjectives used are different, the idea is similar. While in Canadian variant it is very different. Additionally, the British collocation has a low MI score for the size of the corpus.

The last noun for analysis is *holiday*. This is the only noun in this corpus, which has all the collocations present that are displayed in Table 3.39.

Table 3.39 . ***Holiday*** **collocations in CoFNEV**

Variant	Noun collocations	MI	Verb collocations	MI	Adjective collocations	MI
American	season	10.51	celebrate	9.64	special	8.87
Australian	visa	11.84	take	4.99	feral	12.68
British	destination	8.55	book	9.65	good	7.52
Canadian	season	7.85	need	6.04	next	5.37

The first examples are of noun and noun collocations:

1. *American: I love having a few simple hairstyles up my sleeve for the busy holiday season.*
2. *Australian: Over 3 years ago I posted about how to apply for a Tier 5 UK Youth Mobility Visa (previously called the working holiday visa) as an Australian.*
3. *British: I read at many places about its various holiday destination, sandy beaches, spectacular sand dunes, windsurfing and unbelievable year round sunshine.*
4. *Canadian: Those are our favourite travel hacks to help you navigate the holiday season.*

In the case of American and Canadian English, they have the same noun used as a collocation, and they are also similar to GloWbE collocations. Additionally, all the collocations have a high mutual information score.

The next examples are of noun and adjective collocation set with the noun *holiday*:

1. *American: Do any of you also celebrate the holidays with an advent calendar of activities?*
2. *Australian: A new campaign is encouraging Aussies to take their holidays Down Under.*
3. *British: When you book a holiday in Europe there is always the concern that the beaches will be crowded.*
4. *Canadian: I can understand why you need a holiday after this!*

All of the collocations, except for Australian have a high MI score, which indicate to the verb *take* being used more often in other contexts. On the other hand, the Australian collocation is similar to what was seen in the previous analysis of the GloWbE corpus.

The last examples for the noun *holiday* are the adjective collocations:

1. *American: He will always remember that you made his costume and made this holiday so special.*
2. *Australian: If your family is anything like mine, your school routine will be out the window, you will all be a little holiday feral and the thought of school being just 2 weeks away will make your stomach tighten.*
3. *British: Some holidays are so good that we may even start to dream about living there - but which country would you choose?*
4. *Canadian: Looking for a little excitement on your next holiday but not too extreme?*

These collocations are also quite different from what has been seen in previous chapters. However, it is interesting to see, that the most unusual collocation – Australian adjective *feral,* has also the highest MI score.

The collocations present for the noun pair *vacation* and *holiday* have the most similarities with the GloWbE corpus out of all the noun pairs. These collocation connections

between the two words are seen among various varieties and are not restricted one specific variety.

In conclusion, this subchapter focuses on the discussion of collocations found in the Corpus of Four National English Varieties. Due to the size of the corpus, there was an absence of various collocation types for the nouns or in some cases the noun itself was not present in a variety. Despite that, some comparison still can be made between the two corpora. One observation is that the size of the CoFNEV corpus influenced the abilities to do a full comparison of every word with the GloWbE due to various noun absences in the CoFNEV corpus. Despite this absence being expected, it is surprising to find more than a half of the nouns in the CoFNEV corpus. Additionally, it is noteworthy, that corpus creation has evolved so far that the creation of a four million word corpus can be done in a span of a week. The second conclusion that can be observed is the substantial amount of noun and noun collocations and noun and verb collocations that are present in both corpora irrespective of the variety. This shows that only the noun differences are divided among the varieties, but their collocations do not have the restriction of a variety.

CONCLUSIONS

The purpose of this thesis was to investigate whether the increased access to the Internet and communication among people around the world have caused a change in national English varieties. Thus, the goal of this research was to create a noun list, which would be used as a basis for collocation analysis in four national English varieties (American, Australian, British, and Canadian) within Corpus of Global Web-Based English (GloWbE) and Corpus of Four National English Varieties (CoFNEV).

Based on the theoretical material investigation, it is concluded that corpora can be categorised by several principles. These principles are the usage of corpora, the level of annotation, and data collection methods, that relate to the corpora and the Web. The theories concerning the Web in the context of corpus linguistics are considerably controversial due to the distinction of two approaches; firstly, the theories that view Web as corpus, and, secondly, the theories that view Web as a resource for corpora.

The investigation of the English language varieties has revealed that distinctions between geographic or national varieties are based on the variations within their grammar, pronunciation, spelling, and vocabulary. The four largest geographic or national varieties of the English language are American, Australian, British, and Canadian English. British is the oldest one out these varieties, with American English having developed a less complex version of it, whereas Canadian and Australian varieties over time have had influences from British and American English.

The study of parts of speech indicates that parts of speech phrases are comprised of lexical words and prepositions. The same approach is applied to the creation of collocations. The investigation of the analysis methodology used for collocations analysis indicate the relevance of such descriptive statistics measures as absolute and relative frequency analysis, as well as such measures of inferential statistics as log-likelihood and mutual information.

Based on the theoretical principles of corpus creation and the Web relation to corpus enable the design of various corpora types, such as, CoFNEV by selecting of recent texts with the help of the programme *Sketch Engine*. The corpus creation process reveals data collection challenges, such as, incompatibility of websites with *Sketch Engine* programme. The corpus creation process resulted in a corpus sample comprised of four million. The other corpus used in the present study is a GloWbE, which has a broad corpus comprised of 20 English varieties containing data from 2012 to 2013.

The empirical research results confirm that the use of the selected nouns has not changed in the selected varieties irrespective of the increased access to the Internet and

communication among people around the world. The only exception is the case of *vacation* use in Canadian English, where typically *holiday* would be used. The explanation for this changed could be due to the influence of the neighbouring American variety.

The findings also reveal various collocations among the four English varieties. The challenges faced in this part of analysis were the absence of data in some varieties depending on the chosen noun. Despite that, the investigation of collocations indicates three phenomena. One finding indicated that between GloWbE and CoFNEV there are instances of the same noun having the same collocations in irrespective of the variety. The second finding indicates that between the two corpora a noun pair can have some identical collocations among the varieties. The last conclusion is that in most cases the collocations that are identical are either noun and noun collocations or noun and verb collocations. Furthermore, the noun pair *holiday* and *vacation* had the widest representation of collocations observed across the GloWbE and CoFNEV corpora. The data obtained from this noun pair indicated the highest similar collocation cases within the two corpora

The limitations faced during the research process were in regards to the differences in corpora sizes. The created corpus can be considered broad in diachronic aspect; although according to modern standards and technical possibilities it would be advisable to enlarge the corpus for the further research exceeding one million words per variety.

The current research could be useful for language learners, who could gain the understanding of the collocation use with the nouns that vary across the four national varieties. Moreover, it can be useful for language teachers to see that in many cases the collocations are not dependent on national English variety but on the meaning of the words, as seen in the current research.

There is potential for further research. Some of the proposed investigations are: (1) noun and their collocation analysis between two corpora of a similar sample size; (2) all lexical part of speech investigation in one or more English variety corpora; (3) to investigate which language variety is preferred in Latvian media articles.

THESES

1. Corpus is a language resource that is classified on bases of the theories on corpus purpose, size, the level of annotations used in a corpus, the data collection methodology, corpus application in research, and the usage of the Web in corpus creation.
2. Researchers' approaches to the Web in the context of Corpus Linguistics are contradictory. Their approaches have shifted from viewing the Web as corpus towards viewing the Web as a resource for corpus.
3. Language varieties range from their division depending on region (geographic or national) and ethnicity to various social aspects, which are differentiated through the use of grammar, spelling, pronunciation, and vocabulary. The four broadest English varieties from the national point of view are American, Australian, British, and Canadian English.
4. Noun is a lexical part of speech that covers a wide array of types, such as, abstract, concrete, common, proper, count, mass, collective, and generic noun.
5. Parts of speech phrases underlie collocations creation as collocations are defined as a co-occurrence of lexical words that frequently interact with each other.
6. Frequency of the selected nouns (collocation nodes) is determined by descriptive statistics (absolute and relative frequency analysis), with log-likelihood providing inferential statistics data of the selected nouns and mutual information providing the statistical data of collocations.
7. Corpus of Global Web-Based English (GloWbE) encompasses various language varieties including American, Australian, British, and Canadian English, and Corpus of Four National English Varieties (CoFNEV) is a comparable corpus created for the present study according to the corpora design theories.
8. Noun frequency analysis according to the varieties shows that American English nouns are more frequent in Canadian English than British English nouns, for example, the American English noun *vacation.*
9. Variety specific noun frequency analysis indicates to no significant noun usage changes within American, Australian, and British English.
10. The collocate analysis of the selected variety specific nouns in the GloWbE and CoFNEV corpora indicate to a similar usage of collocates irrespective of language variety, predominantly in noun and noun, and noun and verb collocations.

REFERENCES

1. Bailey, R. W. (2017) Standard American English. In A. Bergs and L. J. Brinton (eds.) *The History of English: Varieties of English, Volume 5.* (pp. 1-9). Berlin and Boston: Walter de Gruyter.
2. Baker, P., Hardie, A. and McEnery, T. (2006) *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press Ltd.
3. Bianchi, F. (2012) *Culture, Corpora and Semantics.* Università del Salento. Available from http://siba-ese.unisalento.it/index.php/culturecorpora/article/view/12427/11066
4. Biber, D., Conrad, S., and Leech, G. (2003) *Longman Student Grammar of Spoken and Written English.* Edinburgh: Pearson Education Limited.
5. Brown, K., and Miller, J. (2013) *The Cambridge Dictionary of Linguistics.* Cambridge: Cambridge University Press.
6. Bruckmaier, E. (2017) *Getting at GET in World Englishes: A corpus-based semasiological-syntactic analysis.* Berlin and Boston: Walter de Gruyter.
7. Buendía-Castro, M. and López-Rodríguez, C., I. (2013) The Web for Corpus and the Web as Corpus in Translator Training. *New Voices in Translation Studies*, 10: 54-71.
8. Cook, P. and Briton, L. J. (2017) Building and evaluating web corpora representing national varieties of English. *Lang Resources & Evaluation.* 51: 643-662.
9. Cragg, C., Czarnecki, B., Philips, H. I. and Vanderlinden, K. (2003) *Editing Canadian English*. Toronto: McClelland & Stewart.
10. Davies, D. (2005) *Varieties of Modern English: an Introduction.* London and New York: Routledge.
11. Dollinger, S. (2017) Canadian English in real-time perspective. In A. Bergs and L. J. Brinton (eds.) *The History of English: Varieties of English, Volume 5.* (pp. 53-79). Berlin and Boston: Walter de Gruyter.
12. Gatto, M. (2014) *Web as Corpus: Theory and Practice.* London and New York: Bloomsbury.
13. Granger S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) *Corpus Linguistics. An International Handbook, Volume 1* (pp. 259-275). Berlin & New York: Walter de Gruyter.
14. Hundt, M. (1998) *New Zealand English Grammar: Fact or Fiction?* Amsterdam and Philadelphia: John Benjamins Publishing Company.

15. Hundt, M. (2017) Australian/New Zealand English. In A. Bergs and L. J. Brinton (eds.) *The History of English: Varieties of English, Volume 5.* (pp. 1-9). Berlin and Boston: Walter de Gruyter.
16. Jiang A., Huang, X., Zhang, Zhen., Li, J., Zhang, Zhi., and Hua, H. (2010) Mutual information algorithms. *Mechanical systems and signal processing,* 24: 2947-2960.
17. Jiang, L. and Wong, A. C. M. (2012) On standardizing the signed root log likelihood ratio statistic. *Statistic and Probability Letters*, 82: 833-839.
18. Kilgarriff, A. and Grefenstette, G. (2003) Web as Corpus. *Computational Linguistics.* 29 (3): 1-15.
19. Larsson, T. (2012) On spelling behavio(u)r: A corpus-based study of advanced EFL learners' preferred variety of English. *Nordic Journal of English Studies.*11 (3): 127-154.
20. Lobeck, A. and Denham, K. (2014) *Navigating English Grammar: A Guide to Analyzing Real Language.* Chichester: Wiley-Blackwell.
21. McArthur, T. (2003) *Oxford Guide to World English.* Oxford: Oxford University Press.
22. McCarthy, M. and O'Dell, F. (2005) *English Collocations in Use: How Words Work Together for Fluent and Natural English: Self-study and Classroom Use.* Cambridge: Cambridge University Press.
23. McCarthy, M. and O'Keeffe, A. (2010) Historical perspective: what are corpora and how have they evolved? In A. O'Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). London and New York: Routledge.
24. McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice.* New York: Cambridge University Press.
25. Nugues, P. M. (2006) *An Introduction to Language Processing with Perl and Prolog*. Leipzig: Springer.
26. Okamoto, M. (2015) Is corpus word frequency a good yardstick for selecting words to teach? Threshold levels for vocabulary selection. *System*, 51: 1-10. Elsevier. Available from https://www.sciencedirect.com/science/article/pii/S0346251X1500038X [Accessed on 26 November 2019].
27. Oster, U. and van Lawick (2008) Semantic preference and semantic prosody: a corpus-based analysis of translation-relevant aspects of the meaning of phraseological units. *Translation and Meaning, Part 8,* (pp. 333-344).

28. Peters, P. (2017) Standard British English. In A. Bergs and L. J. Brinton (eds.) *The History of English: Varieties of English, Volume 5.* (pp. 96-120). Berlin and Boston: Walter de Gruyter.
29. Simaki, V., Simakis, P., Paradis, C. and Kerren, A. (2017) Identifying the authors' national variety of English in social media texts. *Proceedings of Recent Advances in Natural Language Processing,* (pp. 671-678).
30. Smadja, F. (1993) Retrieving collocations from Text: Xtract. *Computational Linguistics,* 19 (1): 143-177.
31. Swan, M. (2005) Practical English Usage. Oxford. Oxford University Press.
32. Vinčela, Z. (2017) Canadian Dollar in the English Language Varieties: Corpus-Based Study. *Baltic Journal of English Language, Literature and Culture.* 7: 161-171.
33. Xu, X. (2017) Corpus-based study on African English varieties. *Journal of Language Teaching and Research.* 8 (3): 615-623.

Internet sources

1. [Online 1] Australian English. Available from https://www.cs.mcgill.ca/~rwest/wikispeedia/wpcd/wp/a/Australian_English.htm [Accessed on 23 April 2020].
2. [Online 2] Corpus of Global Web-Based English (GloWbE). Available from https://www.english-corpora.org/glowbe/ [Accessed on 24 April 2020].
3. [Online 3] Frequency. Available form https://www.sketchengine.eu/my_keywords/frequency/ [Accessed on 20 May 2020].
4. [Online 4] Sketch Engine. Available from https://www.sketchengine.eu/ [Accessed on 24 April 2020].
5. [Online 5] IEALTS Online Practice. Available from https://www.ieltsonlinepractice.com/australian-english-vs-american-english-vs-british-english-vs-canadian-english/ [Accessed on 23 April 2020].
6. [Online 6] Lexico. Available from https://www.lexico.com/grammar [Accessed on 23 April 2020].
7. [Online 7] Log-likelihood and effect size calculator. Available from http://ucrel.lancs.ac.uk/llwizard.html [Accessed on 20 May 2020].
8. [Online 8] Relative frequency. Available from https://www.sketchengine.eu/my_keywords/freqmill/ [Accessed on 20 May 2020].

9. [Online 9] Ryerson University Student Learning Support. *Canadian English*. Available from https://www.ryerson.ca/content/dam/studentlearningsupport/resources/grammar-handouts/Canadian_English.pdf [Accessed on 23 April 2020].
10. [Online 10] UCREL CLAWS7 Tagset. Available from http://ucrel.lancs.ac.uk/claws7tags.html [Accessed on 26 April 2020].

APPENDIX 1 CORPUS DESCRIPTION

Nr.	Variety	Folder Title	Words	Link
1.	American	US (news) abcnews.com	250,090	www.abcnews.com
2.	American	US (news) nypost.com	250,197	https://nypost.com/
3.	American	US (blog) everywhereist.com	251,111	https://everywhereist.com/
4.	American	US (blog) sayyes.com	250,148	https://sayyes.com/
5.	Australian	AUS (news) news.com.aus	250,169	https://www.news.com.au/
6.	Australian	AUS (news) theaustralian.com.au	250,267	https://www.theaustralian.com.au/
7.	Australian	AUS (blog) theaussienomad.com	251,925	https://theaussienomad.com/
8.	Australian	AUS (blog) wellnourished.com.au	250,281	https://wellnourished.com.au/
9.	British	UK (news) inews.co.uk	250,741	https://inews.co.uk/news
10.	British	UK (news) mirror.co.uk	251,155	https://www.mirror.co.uk/
11.	British	UK (blog) globalgrasshopper.com	253,464	https://globalgrasshopper.com/
12.	British	UK (blog) thelifestylebloggeruk.com	249,070	https://thelifestylebloggeruk.com/
13.	Canadian	CAN (news) globeandmail.com	250,545	https://www.theglobeandmail.com/
14.	Canadian	CAN (news) nationalpost.com	250,270	https://nationalpost.com/
15.	Canadian	CAN (blog) dailydream360.com	250,219	https://www.dailydream360.com/
16.	Canadian	CAN (blog) theplanetd.com	250,629	https://theplanetd.com/
Total: 4,010,281 words				

Dokumentārā lapa

Maģistra darbs „Collocation Variations in the Web-Based Corpora of the English Language Varieties" (Vārdu savienojumu variācijas angļu valodas variantu tīmekļa korpusos) izstrādāts LU Humanitāro zinātņu fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Katrīna Vēvere _____ 28.05.2020.

Rekomendēju darbu aizstāvēšanai

Vadītāja: asociētā profesore Dr.Philol. Zigrīda Vinčela _____ 28.05.2020

Recenzents: asociētā profesore Dr.Philol. Jana Kuzmina _____ 28.05.2020

Studiju metodiķe: Ieva Melbārde _____ 28.05.2020

Darbs iesniegts Anglistikas nodaļā 28. 05. 2020.

Darbu pieņēma: docētājs Dr. Philol. Aleksejs Taube _____ 28.05.2020

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

2020. gada…… jūnijā, prot. Nr. ….., vērtējums ……..

Komisijas sekretārs/-e: