Pedagogická fakulta
Faculty of Education
Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích
Pedagogická fakulta
Katedra anglistiky

Diplomová práce

The naturalness of the language of English students in written texts

Přirozenost jazyka studujících angličtiny v psaných textech

Vypracoval: Bc. Martin Svoboda
Vedoucí práce: Mgr. Jaroslav Emmer, Ph.D.

České Budějovice 2023

Prohlášení

Prohlašuji, že svoji diplomovou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své diplomové práce, a to v nezkrácené podobě – v úpravě vzniklé vypuštěním vyznačených částí archivovaných fakultou elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 28. prosince 2023

Martin Svoboda

Poděkování

Na tomto místě bych rád poděkoval mému vedoucímu práce Mgr. Jaroslavu Emmerovi, Ph.D. za jeho užitečné rady a připomínky, bez nichž by tato práce nemohla vzniknout. Dále bych také rád poděkoval PhDr. Christopheru Koyovi, M.A., Ph.D. za poskytnutí studentských esejí pro analýzu, bez kterých by se taktéž tato práce neobešla.

Anotace

Tato studie se zaměřuje na analýzu přirozenosti jazyka studentů anglistiky na Pedagogické fakultě Jihočeské univerzity (PF JČU) prostřednictvím psaných textů. Úvodní část se věnuje především představení problematiky přirozenosti jazyka mluveného i psaného, co je pro přirozeně znějící jazyk potřeba a jaké výhody může zaměření výuky na přirozenost přivést.

Teoretická sekce se zprvu věnuje klíčovým konceptům, jako jsou "nativelike selection" a "idiom principle", a poskytuje jejich podrobný popis. Dále pak také zkoumá oblasti korpusové a textové lingvistiky, konkrétně se zaměřuje na koncepty "keyness" a "aboutness". Zahrnuje také obecný pohled na kolokace a frazémy. Závěr teoretické části identifikuje časté problémy, s nimiž se nerodilí mluvčí mohou setkat při psaní anglických textů.

Praktická sekce analyzuje autentické filmové recenze z internetu, sloužící jako referenční korpus, a porovnává je s esejemi studentů anglistiky na PF JČU na téže téma. Pro tvorbu korpusů, kolokačních profilů a analýzu je využit program #LancsBox, z jehož výstupu jsou patrné rozdíly mezi jazykem studentů a rodilých mluvčích. Jednou z nejčastěji objevujících se odlišností je tendence studentů často opakovat pro ně již zažité kolokace a fráze, čímž se sice mohou vyhnout případným chybám, avšak ubírají tak textu na pestrosti. To může mít za následek, že se text na první pohled tváří amatérsky napsaný, méně zajímavý, či například méně přehledný.

V závěru praktické sekce jsou taktéž diskutovány výsledky analýzy. Zde jsou zvolena ta nejčastěji používaná lemmata, která se vyskytují v obou korpusech, a jejich kolokační profily mezi sebou porovnány. Rozdíly mezi nimi, ač už v obecné frekvenci používání, či používání modálních sloves a zájmen jakožto kolokátů, jsou zde znázorněny pomocí grafů a popsány. Ke každému rozdílu jsou mimo jiné také doplněny možné příčiny vzniku.

Abstract

This study focuses on analysing the naturalness of language of English students at the Faculty of Education of the University of South Bohemia (PF JČU) through written texts. The introductory part primarily addresses the introduction of the issues related to the naturalness of both spoken and written language, discussing the requirements for natural-sounding language and the advantages that focusing on naturalness in teaching can bring.

The theoretical section initially delves into key concepts such as "nativelike selection" and the "idiom principle", providing a detailed description of these concepts. It further explores areas of corpus and text linguistics, specifically focusing on the concepts of "keyness" and "aboutness". The section also provides a general overview of collocations and phrasemes. The conclusion of the theoretical part identifies some common problems that non-native speakers may encounter when writing English texts.

The practical section analyses authentic film reviews from the internet, serving as a reference corpus, and compares them with essays written by English students at PF JČU on the same topic. The #LancsBox program is utilized for the creation of corpora, collocational profiles, and analysis, revealing differences between the language of students and native speakers. One of the most frequently observed differences is the tendency of students to often repeat collocations and phrases they are already familiar with, which, while helping them avoid potential errors, reduces the variety of the text. As a result, the text may appear amateurishly written, less engaging, or, for instance, less clear at first glance.

In the conclusion of the practical section, the results of the analysis are also discussed. The most frequently used lemmas shared by both corpora are selected, and their collocational profiles are compared. Differences between them, whether in general frequency of usage or in the usage of modal verbs and pronouns as collocates, are illustrated using graphs and further described. Possible causes and reasons for each difference are also provided, among other considerations.

Table of contents

I. INTRODUCTION

Speaking or writing like native speakers is something that most if not all non-native speakers who are invested in a given language strive for, however it is more often than not easier said than done. One can be fluent in a language and use it on a daily basis and yet be immediately identified by a native speaker as a non-native for using phrases and collocations that a native speaker would find inappropriate or non-standard.

While using said non-standard phrases and collocations may be grammatically and semantically correct, in the eyes of a native speaker these phrases might just sound awkward, unnatural, or even hide some ulterior motive behind them based on current situation. This however is something for non-native speakers to find out by themselves as it is typically not taught in schools.

This all fits well with students of languages, in this case students of English on PF JČU, as they use English almost every day either in spoken form in class or written form in assignments and essays. Their outputs, in this case their essays, can be used for analysis in comparison with a reference corpus, consisting of authentic texts on a similar topic, to identify any said inappropriate and non-standard phrases, collocations and possible problematic areas, to perhaps help streamline the teaching of vocabulary and writing English texts.

Focusing on this aspect of English in teaching may not immediately provide meaningful results but it may help the future generations of teachers to adjust and better streamline the teaching of vocabulary and writing English texts and learners to better understand what phrases and collocations to use based on situation to not sound or look inappropriate or awkward.

II. THEORETICAL PART

1. Nativelike selection

The term "nativelike selection", as described by Pawley and Syder (1983), can be roughly understood as the ability of the native speaker to routinely convey his meaning by expressions that are not only grammatical but also nativelike, i.e. the ability to resemble a native of that given language in terms of expected grammar and choice of vocabulary. This is intriguing because native speakers effortlessly choose sentences that are idiomatic and natural from a range of grammatically correct alternatives. While some may think it easy and that all it takes to achieve a nativelike resemblance of language is to observe native speakers and their usage of language in given situations and try to mimic it, however this is not the case as there are many aspects that need to be considered and taken note of. It is not only important to know what sentence or expression (that is natural and idiomatic) to select but also the reasons behind it, as for each given situation, there can be numerous grammatically correct phrases, many of which may be non-nativelike or highly marked usages (Pawley & Syder, 1983).

Although the general nature and practical importance of nativelike selection is recognized, at least tacitly, by all second language teachers, this linguistic ability presents specific problems of formal description and explanation that have generally been overlooked. Pawley and Syder suggest that, to describe and explain it, it is necessary to look at how native speakers understand grammar in a way that is somewhat different from what most grammar experts currently believe. They also argue that, based on research into how people express themselves in everyday English conversations and situations, being able to speak a language fluently and naturally as well as write it depends a lot on knowing commonly used word patterns or "sentence stems" that are firmly established or lexicalized. These patterns are not true idioms but represent regular form-meaning associations and are known to mature speakers in the language (Pawley & Syder, 1983).

This all also applies to the term "nativelike fluency", which being closely connected to nativelike selection, is the ability of the native speaker to produce fluent stretches of spontaneous connected discourse, even though their ability to plan and encode novel speech in advance seems limited. It is however not exclusive only to speech, as nativelike fluency can also be observed in written texts, specifically in texts that are easy to read and understand (Pawley & Syder, 1983).

Overall, Pawley and Syder's theory departs from the traditional view of separating grammar into productive rules (syntax) and fixed usages (dictionary). It suggests that many regular sequences can be known both as whole units and as products of syntactic rules, leading to some redundancy in the grammar. This perspective has implications for how we understand and describe the native speaker's linguistic competence (Pawley & Syder, 1983).

1.1. The "puzzle of nativelike selection"

Another topic closely connected to nativelike selection that Pawley and Syder (1983) touch on is the idea of a "generative grammar" and the connection between it and "linguistic competence". Primarily credited to Chomsky (1957), this concept has been widely accepted since the 1960s, suggesting that part of learning a language involves understanding a system of rules that generates an infinite number of sentences in that language, assigns correct structures to them, and identifies incorrect ones (Pawley & Syder, 1983).

Chomsky's approach emphasizes the creative potential of grammar rules, and most linguists agree that natural languages have an extensive variety of possible sentences. While there are debates about the specifics of generative grammar, it is generally accepted that knowing these rules is crucial for language proficiency (Pawley & Syder, 1983).

Pawley and Syder also address a less-explored issue: native speakers do not use the full creative potential of these grammar rules. In fact, only a small percentage of grammatically correct sentences sound natural to native speakers. Many

grammatical sentences are considered unidiomatic, odd, or foreign-sounding. This observation remains true even when considering sentences that make sense and are relatively short. For example, the sentences "I had four uncles." / "The brothers of my parents were four." or "That was one Christmas that I'll always remember…" / "There is not a time when my remembering that Christmas will not take place…". While the first and third sentences look like something an ordinary person would say, their paraphrased versions seem completely unnatural, even though they are grammatically correct. If a language learner is to achieve nativelike control, then they do more than just learn the usual generative grammar rules that define all the sentences of the language. They also need to learn how to recognize which well-formed sentences are considered natural and normal by native speakers as opposed to those that sound strange or unusual. How this distinction is made is what Pawley and Syder call the "puzzle of nativelike selection" (Pawley & Syder, 1983).

This all can be quite difficult for people who learn a new language primarily from a grammar book, especially if they have not had much exposure to how the language is actually used in everyday life. When learners try using their "book knowledge" in real conversations, even if they have studied hard and their sentences are technically correct, they may not sound quite right to native speakers. That is because native speakers do not usually talk the way grammar books teach. On the other hand, if they have learned a language by being part of a community where it is spoken from the beginning, they tend to pick up both natural-sounding speech and correct grammar at the same time without even needing to know the reason something is written or pronounced the way it is. Members of such groups or communities may not necessarily even find this to be an obvious problem, as it is natural for them (Pawley & Syder, 1983).

Pawley and Syder also state that grammarians might be tempted to dismiss nativelike selection as just a matter of style and not grammar, as if this would let them avoid trying to understand it. However, this does not really solve the problem. It merely gives it a name without explaining it properly. On the other hand, some might suggest that what is being touched on here is ungrammatical discourse, going against subtle grammar rules that have not been fully spelled out in grammatical

analysis. Whilst the idea deserves consideration, one should not rush to a solution by just labelling it. It may not be helpful to stretch the term "grammar rule" to include things that are quite different from what is usually classified under that label. Calling something by a familiar name does not automatically make it clear, especially if it is unfamiliar (Pawley & Syder, 1983).

It should be acknowledged that the problem's nature may not be well understood right now, and as Pawley and Syder state, there is no sharp boundary between the classes of nativelike and non-nativelike sentences, in much the same way as there is no sharp boundary between the categories of grammatical and ungrammatical sentences in English (Pawley & Syder, 1983).

2. The idiom principle

John Sinclair, in his book "Corpus, Concordance, Collocations" (1991), advocated for the use of corpus research in developing his concept of idioms. He argued that multiword expressions are not just random in language; they function as partially pre-formed phrases, essentially single choices. This concept is called the "idiom principle", which opposes the "open-choice principle", also described by Sinclair. The open-choice principle suggests that in grammatical language, users have the freedom to select from a range of word choices (Sinclair, 1991).

Sinclair's idea of the idiom principle has been widely accepted by linguists studying idioms and scholars like Grant & Nation (2006) and Levorato, Roch & Nesi (2007) have explored how often language users can rely on an identified idiom being used in an idiomatic sense rather than literally.

2.1. What are idioms?

An idiom can be considered a "fixed expression" where the overall meaning does not correspond to the meanings of its individual components. Čermák (2007) uses the term *compositionality*. For instance, "to kill two birds with one stone" means achieving two things with a single action, and "break a leg" means wishing someone

good luck. These idioms do not literally involve harming birds or breaking a leg (Benson et al., 1993). Identifying and understanding idioms can be challenging, particularly for non-native speakers of a language that lacks comparable idiomatic expressions for reference.

As outlined in "Collocations in a Learner Corpus" (Nesselhauf, 2005), word combinations can also be categorized into four distinct groups:

1. Free combinations – the elements of combination are used in the literal sense, e.g. "*drink tea*" and substitution can happen within a semantic field.
2. Restricted collocations – at least one element is used in its literal meaning, the other one has non-literal meaning, e.g. "*perform a task*", and substitution is limited.
3. Figurative idioms – they have figurative meaning but have literal interpretation, e.g. "*U-turn*" – to change one's behaviour. Substitution is rarely possible.
4. Pure idioms – they have figurative meaning and do not have literal interpretation, e.g. "*blow the gaff*". It is not possible to substitute the elements at all.

(in Nesselhauf, 2005)

In all languages, idioms and phrasemes are frequently observed, most of which initially had a literal meaning. Over time, we will likely come across newly coined idioms that have evolved from their original literal sense and are now associated with something entirely different.

2.2. Idiom principle vs. open-choice principle

As the differentiation between idiom and open-choice is central to the current topic, it is important to delve further into this topic. In his work, Sinclair (1991)describes the open-choice principle as follows:

> *"This is a way of seeing language text as a result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness. This is probably the normal way of seeing and describing language. It is often called a "slot-and-filler" model, envisaging texts as a series of slots which have to be filled from a lexicon which satisfies restraints. At each slot, virtually any word can occur. Since language is believed to operate simultaneously on several levels, there is a very complex pattern of choices in progress at any moment, but the underlying principle is simple enough."* (p. 109)

To accompany it, he offers the following examples of open-choice language use contrasted with idiom: *run a mile* (idiom: "Any normal Londoner would *run a mile* rather than lunch in the Westminster pub." / open-choice: "How fast can he *run a mile*?"), *kick up* (idiom: "Taste it, and, if desired, *kick up* its taste a little more by whisking a bit more of the flavourings… in."; open choice: "Slade's brave and brilliantly-judged penalty *kick up* the touchline."), and *stick out* (idiom: "… to find the activity and users that *stick out* as abnormal."; open choice: "… Klitschoko pulled a USB *stick out* of his pocket.") (Sinclair, 1991).

The idiom principle suggests that words in language do not appear as haphazardly as the open-choice principle suggests. Instead, words often occur together, and typical text or speech does not usually rely solely on the open-choice principle:

> *"The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single-choices, even though they might appear to be analyzable into segments…. At its simplest, the principle of idiom can be seen in the apparently simultaneous choice of two words for*

example, of course. This phrase operates effectively as a single word, and the word space, which is structurally bogus, may disappear in time, as we see in maybe, anyway, and another." (p. 110)

The idiom principle places constraints on both written and spoken language, establishing a sense of predictability based on the topic, situation, and context. A significant feature of the idiom principle, in contrast to the open-choice principle, is the idea of restricted exchangeability, meaning that at least one part of a preconstructed phrase cannot be substituted with a synonymous term without altering the meaning, function, or idiomatic nature of the phrase (Erman & Warren, 2000).

Similarly, Liu (2008) distinguishes between pre-established phrases, which have a fixed structure, and semi-pre-established phrases, which allow some structural variation. However, both of these categories fall under the idiom principle because they represent a single choice at the phrase level for language users. Overall, the idiom principle encompasses various aspects such as collocations, binomials, phrasal verbs, stock phrases, proverbs, and idioms (Liu, 2008).

3. Corpus linguistics and corpus

Some define corpus linguistics as "an area that focuses on a set of procedures of methods for studying language" (McEnery, T. & Hardie, A., 2011). Although it is not considered and independent branch of linguistics or a theory of language, it serves as a methodology for acquiring and analysing language data, either quantitatively or qualitatively. Corpus linguistics can be applied to nearly any area of language research, utilizing authentic, naturally occurring language as its primary subject (University of Helsinki, 2016).

Another term closely linked with corpus linguistic is the term "corpus" itself. It is defined as "in linguistics and lexicography, a collection of texts, spoken language, or other examples regarded as somewhat representative of a language, typically stored as an electronic database" (McArthur, 1992). A key function of a corpus is to validate

a language-related hypothesis, such as identifying the possible variations when employing a specific sound, word, or syntactic structure. Corpora can also serve as a starting point for linguistic description (Crystal, 1991).

To those unaccustomed to corpora, virtually any text might serve as a corpus or be transformed into one, but the truth is somewhat different. The text of a corpus must align with the hypothesis, be of a specified size, and be electronically stored because gathering data on frequencies, grammatical structures, and collocations is more efficiently accomplished with a computer rather than manually. Additionally, it should be accessible without restrictions, enabling research results to be cross-referenced, compared, and possibly replicated (University of Helsinki, 2016).

3.1. Text linguistics

The term "text linguistics", as described by Sarah Al-Otaibi from King Saud University (2014), refers to a branch of linguistics that deals with texts as systems of communication. Initially, its primary goal was to reveal and describe the grammatical structures within texts. However, the application of text linguistics has since expanded, moving beyond a narrow focus on traditional grammar to encompass the entire text (King Saud University, 2014).

The emergence of text linguistics as a branch of linguistics began in the early 1970s, coinciding with a shift in linguistic research away from the sentence as the primary unit of analysis. It was recognized that there was a need to explore units larger than the sentence and relationships within sentences. Central concerns include defining what makes a text a text (textuality) and categorizing texts based on their genre characteristics. With influences from pragmatics and psychology, there is a growing emphasis on the production, processing, reception, and social function of texts in society (King Saud University, 2014).

Text linguistics can be understood in two ways: as the study of the text itself as a product (text grammar), focusing on aspects like cohesion, coherence, organization, speech acts, and communicative functions, or as an examination of the text's

creation (theory of text), reception, and interpretation (Wikipedia, 2023) . In its examination of the text itself, text linguistics intersects with various other fields such a discourse analysis, stylistics, pragmatics, sociolinguistics, and narratology (King Saud University, 2014).

3.2. Keyness

In corpus linguistics, keyness stands for the quality a word or phrase has of being "key" or a "key word" in its context. A key word is a term that appears in a text more frequently than we would anticipate based on random chance alone. To identify key words, a statistical test (such as log-linear or chi-squared) is employed. These tests are able to compare the word frequencies in the text to the expected frequencies, which are determined from a significantly larger corpus serving as a reference for typical language usage. (Scott, M. & Tribble, C., 2006) The concept of keyness and key words is closely related to the concept of aboutness, which refers to comprehending the primary ideas, topics, or attitudes addressed in a text or corpus and will be explained further in its own chapter (Gabrielatos, C., 2018).

In contrast to collocation, which denotes the inherent connection between two words or phrases usually found within a specific range of each other, keyness is a characteristic of the text, not the language itself. This means that a word can possess keyness in a particular textual context, but it may lack keyness in different contexts. On the other hand, a node and collocate are frequently found together in texts of the same genre, so collocation can be considered primarily a linguistic phenomenon. When identifying a set of keywords within a given text that share keyness, they can be considered "co-keys". Words that are commonly found in the same texts as a key word are referred to as "associates" (Wikipedia, 2023).

3.2.1. Keyness analysis

According to Gabrielatos (2018), to analyse the keyness value of a corpus, to put it simply, one essentially has to compare frequencies. Presently, this analysis primarily seeks to identify significant differences in the frequency of word forms between two corpora, typically referred to as the "study" a "reference" corpus.

However, Gabrielatos claims there is a growing interest in using keyness analysis to establish both similarity and absence, which can be seen as instances of extreme frequency differences (Gabrielatos, C., 2018).

Unfortunately, the influence of practices from other quantitative disciplines and varying definitions of keyness have led to the adoption of inappropriate metrics. Gabrielatos claims that this, in turn, has given rise to several misconceptions related to the following:

a) The nature of keyness and keyness analysis
b) The types of linguistic units suitable for keyness analysis
c) The metrics appropriate for measuring keyness
d) The characteristics of the corpora being compared

(in Gabrielatos, C., 2018)

Lastly, he also argues that a study employing keyness analysis does not stop at identifying key items; this is just the initial step. A manual analysis is necessary to determine how these items are used in context. The precise and well-founded identification of key items is critical, as it significantly impacts the study's findings. Even when the manual analysis is thorough and contextually informed, flawed key item selection can lead to erroneous results and conclusions. Identifying key items and selecting those for the manual analysis is a multifaceted process, influenced by several misconceptions and thus should warrant a detailed examination (Gabrielatos, C., 2018).

Gabrielatos also presents examples of exploratory and focused approaches to keyness analysis that, although not entirely discreet, can be combined:

- **Example 1**: "The research starts with an exploratory approach, by deriving a list of key items ranked according to the value of the keyness metric used in the study. At this point, the researcher may switch to a targeted approach and select particular types of items for concordance analysis according to

explicit criteria, such as their normalised or raw frequency, part of speech, core sense, or relation to a particular topic."

- **Example 2**: "The research starts with a targeted approach, by specifying items to be included in, or excluded from, the analysis (as in the second stage in example one above). Members of the resulting key item list are then selected according to explicit criteria."
  (in Gabrielatos, C., 2018)

3.3. Aboutness

As mentioned before, the term aboutness can be roughly understood as the comprehension of the primary ideas, topics, or attitudes addressed within a text or collection of texts. Phillips (1989) argues that "aboutness stems from the reader's appreciation of the large-scale organisation of text". The concept of aboutness also plays a role in studies related to keyness and key words, and it could have had an impact on the evolution of keyness analysis, as this type of analysis is a means of establishing the aboutness of a text (Scott, 2001). Nonetheless, Phillips also states that the concept of aboutness was not determined by comparing frequency differences between (sub-)corpora. Instead, it relied on examining patterns of collocation within a (sub-)corpus. Despite this distinction, both methods have a common feature: the automated analysis typically does not consider the meaning of the linguistic forms being examined. The interpretation of results is where considerations of meaning come into play (Gabrielatos, C., 2018).

4. Collocations

According to information from Futurelearn.com (2021) provided in collaboration with Macquarie University in Sydney, Australia, the term collocation refers to a group of two or more words that are typically used together to convey a specific meaning. When different word combinations are employed, they often sound unnatural or awkward (Future Learn, 2021). These pairings are considered natural and appropriate by native English speakers, who use them regularly. For example, the phrase "a fast train" compared to "a quick train". Native English speakers

associate the word “fast” with movement and the word “quick” with the passage of time, enabling them to distinguish which collocation is more natural. In contrast, non-native English speakers might have difficulty discerning the difference. This does not necessarily imply that non-native speakers will not be understood, but it could require listeners to pay closer attention to the speech, potentially resulting in communication problems or difficulties. Utilizing appropriate collocations can also be advantageous if a speaker wishes to convey more information within a shorter context (Barfield & Gyllstad, 2009).

Combinations of words like these are highly significant and widely used by native speakers. Unfortunately, there is no straightforward rule for learning them. However, a helpful aspect of learning is that people tend to recall collocations more readily than individual words. Learning and retaining a collocation can be particularly advantageous for learners, as it can aid their ongoing language acquisition. When learners can recognize a familiar collocation in a text, it not only assists in comprehension but also boosts their confidence in their language skills (Nesselhauf, 2005).

According to Čermák (2006), collocations hold significant importance in the realm of education, where educators can leverage textbooks and materials grounded in collocation studies to assist their students in sounding more fluent. Moreover, Čermák suggests that translators might also gain advantages from collocations. By referring to a dictionary, they can identify more natural-sounding expressions and enhance the quality of their translations. It is worth noting that until a few decades ago, English textbooks emphasized individual vocabulary as the primary component of language, often overlooking the significance of collocations and their diverse variations (Barfield & Gyllstad, 2009).

4.1. Examining collocations

Theoretically, collocations can be described as lexical relations between two or more words that have a tendency to appear and co-occur within close proximity to each other. It is important to note that collocations can manifest in various ways,

shapes, or forms. To understand the various levels at which word co-occurrence can be categorized, we can consider the four types identified by Sinclair (1991): *collocation*, *colligation*, *semantic preference*, and *semantic prosody* (Geeraerts, 2010). However, for the purpose of this chapter, the primary focus will be only on collocations since they are the central subject of discussion.

Geeraerts (2010) explains that in a collocation, the word of interest is typically referred to as the *node*, while the accompanying word is known as the *collocate*. One common method of analysis involves creating a concordance for a specific text or group of texts. This concordance is essentially an alphabetical list of words in those texts, along with their immediate context. The typical way of presenting a concordance is through the Key Word in Context index (KWIC). This approach is frequently employed as an optional means of investigating the collocates of chosen nodes, including their position in relation to the node (either on the right or left), the distance between the collocates and nodes, and whether the collocates are found within the same sentence as the node or not (Geeraerts, 2010).

The node of a collocation analysis can either be a specific word form or a word itself, provided that lemmatization is applicable. Lemmatization involves treating all the inflected forms of a word as instances of a single lexical unit. Nodes within collocations can also encompass more complex expressions or phrases. It is worth noting that certain words, often referred to as *stop words*, such as *a*, *the*, *is*, *are*, *by*, *from*, and so on, which have limited explanatory power and carry less semantic significance, may potentially have a detrimental effect on the outcomes of collocational analyses. However, there are methods to address this issue, such as using stop lists as filters or employing various association measures designed to mostly exclude such words (Geeraerts, 2010).

4.2. Types of collocations

As stated by Kaplan International Languages (2021), the process of categorizing collocations can facilitate the learning of these word combinations.

The initial category they outline is the distinction between Strong and Weak (or Lexical) Collocations. In the case of strong collocations, the words involved do not easily combine with a wide array of other words. The connections within strong collocations are robust because there are few alternative and acceptable options to express the same idea. For instance, the phrase "*turn on a light*" is a strong collocation since most synonymous alternatives would sound peculiar and unnatural, like "*start a light*" or "*activate a light*". In contrast, weak collocations represent the opposite scenario. They encompass words that can be combined with numerous alternatives. For example, the phrase "*very interesting*" is frequently used, but the collocation itself is weak, as substitutes like "*extremely interesting*" or "*really interesting*" are also considered acceptable (Kaplan International, 2021).

The second category they describe is Grammatical Collocations. This is then further categorized into: Adverb collocations (adverb + adjective), Adjective collocations (adjective + noun), Noun collocations (noun + noun/verb) and Verb collocations (verb + noun/adverb) (Kaplan International, 2021).

Although Wei (1999) goes more into detail with Grammatical Collocations, he also describes a third collocational category in his work "Teaching Collocations for Productive Vocabulary Development". Concerning Grammatical Collocations, he divides them into two sub-categories, one being "Grammatical collocations that contain a preposition" and the other being "Grammatical collocations that involve a grammatical Structure". He then goes into more detail, showing contrasting examples. As the third category, he decided to include idiomatic expressions, saying that idiomatic expressions are the most fixed word combinations, where substitution of any of their components is virtually impossible, for example, "*kick the bucket*", "*play it by ear*" or "*let one's hair down*" (Wei, 1999).

The second category they outline is Grammatical Collocations. This category can be further broken down into Adverb collocations (combining an adverb with an adjective), Adjective collocations (combining an adjective with a noun), Noun collocations (combining a noun with another noun or a verb), and Verb collocations (combining a verb with a noun or adverb) (Kaplan International, 2021).

However, Wei (1999) provides a deeper exploration of grammatical collocations and introduces a third category in his study titled "Teaching Collocations for Productive Vocabulary Development". Within the realm of grammatical collocations, he further divides them into two subcategories: one being "Grammatical collocations containing a preposition," and the other being "Grammatical collocations that incorporate a grammatical structure." Wei goes on to provide detailed explanations with contrasting examples. For his third category, he includes idiomatic expressions, noting that these are the most fixed word combinations where it is virtually impossible to substitute any of their components. Examples of idiomatic expressions include phrases like "*kick the bucket*", "*play it by ear*", "*let one's hair down*", and so on (Wei, 1999).

4.3. Collocability

As per Čermák's definition (2007), collocability refers to the individual, formal, and semantic compatibility of language elements. This can be understood as the capacity of each language element to join with one or more others. Collocability is influenced by the collocational paradigms of the element and, in regular combinations, is determined by how well it pairs with them. When combined with valency, collocability plays a central role in the syntagmaticity of any language element. The specific realisation of collocability leads to the creation of a collocation (Čermák, 2007).

In his work Collocations, Collocability and Dictionary, he also claims that the whole collocational range (or collocability) of most words is and seems to be so large and unlimited that it is never given in full. Despite that, Čermák states that there is a select group of words that is evidently and strictly in its collocational capacity. This group has a very small list of collocates, which reverts the view adopted so far and suggests the possibility of viewing both the head and collocate as a single unit, identical, in many ways to idioms, compared to "*afraid*" (be afraid) or "*afoul*" (run afoul)(Čermák, 2006).

5. Phrasemes

According to Čermák (2006), a phraseme is a unique combination of at least two words, where each word does not function in the same way when combined with other words or appears exclusively in that particular combination. Phrasemes are fixed expressions carrying a specific meaning as a whole, with no room for inserting or substituting other elements (Čermák & Šulc, 2006).

The elements within a phraseme can be either compatible or incompatible. Phrasemes with compatible elements can convey both idiomatic and literal meanings. Čermák illustrates this with Czech examples, like "*bledá tvář*", which can mean both a white person in films about Native Americans (idiomatic) and a face that is literally white (literal), or "*dutá hlava*", which has only one idiomatic meaning, "a fool." Changing any element in a phraseme would render its meaning unrecognizable, for example, "*dutá ruka*" (Čermák & Šulc, 2006).

Phrasemes can be categorized into various groups based on two key factors: compositionality (whether their meaning results from a direct combination of the meanings of their individual components) and the type of restrictions imposed on the elements that can be freely chosen within them (Wikipedia, 2023). Non-compositional phrasemes are typically referred to as idioms, whereas compositional phrasemes can be further subdivided into collocations, clichés, and pragmatemes (Mel'čuk, 2012).

Lastly, while much of the conversation about phrasemes mainly focuses on multi-word expressions like the ones demonstrated earlier, it is important to recognize that phrasemes can also exist on the morphological level. Morphological phrasemes are established pairings of morphemes, and they include at least one component with selectional restrictions or in short, as described by Beck & Mel'čuk (2011), "phraseologized combinations of morphs inside a wordform". Similar to lexical phrasemes, morphological phrasemes can be either compositional or non-compositional. Two examples from English are the nominalizers used with particular verbal bases (e.g., *establishment* / **establishation*; *infestation* /

**infestment*; etc.), and the inhabitant suffixes required for particular place names (*Winnipeger* / **Winnipegian*; *Calgarian* / **Calgarier*; etc.); in both cases, the choice of derivational affix is restricted by the base, but the derivation is compositional (Wikipedia, 2023).

6. Problems non-native speakers experience when writing English texts

Writing in any language that is not the writer's native one can be a challenging endeavour; however, speakers of some languages may have it easier than others when trying to accommodate to the style of written English, especially when their native language is a part of the same language family as the one, they are trying to learn. Although every learner is different and even this advantage does not stop learners from making some common mistakes. The Mayfield Handbook of Technical & Scientific Writing (1997) describes the ten most common writing problems for non-native speakers of English:

1. Article and Noun Problems
2. Verb Problems
3. Word From
4. Word Order and Sentence Structure
5. Word Choice
6. Wordiness
7. Punctuation and Mechanics
8. Sentence and Paragraph Coherence
9. Organization and Stylistic Approach
10. Documentation and Use of Source language

(in The Mayfield Handbook of Technical & Scientific Writing, 1997)

Due to the sheer breadth of other sub-problems the categories above encompass, only a select few of them will be touched upon and described further described with

a bigger focus on Czech learners of English where possible, as their essays will be later analysed within the practical part of this thesis.

6.1. Articles and nouns

To start, one of the frequent challenges for Czech learners are articles and nouns. They often misuse articles or omit them altogether. The reason for this is that Czech does not have articles unlike English. Czech learners frequently apply the indefinite article to singular uncountable nouns, even though it should only be used with singular countable nouns. Singular invariable nouns generally maintain singular form although some also have a plural form (Poslušná, 2009). As for the nouns, there are often problems with countability, plurality and regularity. For example, Czech learners often tend to use the noun "informations" as in "Do you have any new informations?", which when translated to Czech being a completely normal sentence is incorrect in English. This and similar examples can most likely be attributed to grammatical interference between those two languages. Lastly, in English, it is not possible to create plurals by simply adding an "-s" ending to nouns with irregular plural forms. For example, "man" cannot become "mans" but rather "men". These forms have specific rules that need to be memorized, which may prove challenging (Poslušná, 2009).

6.2. Prepositions

Another quite common problem appears when Czech learners try to use prepositions in English the same way they use prepositions in Czech or translate them as if they were lexically independent units. The reason for this being that Czech prepositions tend to lack a direct equivalent in English, like in the case of "v". While in some cases, Czech "v" can be translated to "in" (*v krabici –> in the box*), in other cases preposition like "on" (*v neděli –> on Sunday*), or "at" (*v poledne –> at noon*) are correct equivalents. Last but not least, Czech learners often mix up prepositions of time like *before* and *after* with prepositions of place like *in front of* and *behind*, e.g. *before the meal / in front of the meal* or *it's behind him / it's after him* (Poslušná, 2009).

6.3. Word order and sentence structure

As per Poslušná (2009), the most frequent challenge lies in proper word order and sentence structure. Unlike Czech, English typically arranges declarative and imperative sentences in the following sequence: subject, verb, object, and then adverbials related to manner, place, and time. Hence, a sentence like "In England is spoken English", even though grammatically correct, may sound strange in English, although it can be used in Czech without any issue (Poslušná, 2009). The rules of correct word order, such as placing adjectives before nouns and adverbs after verbs, are not explicitly instructed but are instead acquired through years of practice. It is believed that native speakers intuitively adhere to a specific subjective-objective sequence/scale for adjectives. While there might be some ongoing discussion about these "rules," learners need not be discouraged. Typically, they can work around this in the beginning by constructing shorter sentences (Academic Language Experts, 2023).

6.4. Spelling variations

Another problem, although much less severe, is caused by the differences in spelling between British and American English, given how minor and easy-to-overlook the differences can be. These spelling mistakes most frequently occur with words ending in *-ise* and *-ize* (e.g. *realise*, *realize*) and *-or* and *-our* (e.g. *armor*, *armour*). Furthermore, British English often considers both spelling variants correct, but only one of them is predominantly employed in written works due to established conventions (Academic Language Experts, 2023).

6.5. Idiomatic and non-committal phrasing

Lastly, the problem of idiomatic and non-committal phrasing. Similar to employing first-person language in academic texts, excessive use of idiomatic expressions in writing can create an informal tone. Moreover, using idioms incorrectly may lead to confusion among readers. To avoid these problems, it is advisable to use idioms in moderation to ensure conciseness and readability of the text or in other cases avoid using idioms altogether (Academic Language Experts, 2023).

ESL writers also tend to avoid making definitive statements. However, there is nothing wrong with making a strong, well-supported statement when the evidence proves it. Indecisive writing tends to add unnecessary words to the text without adding substantive content. Given the principle that “less is more” in academic writing, learners should strive to deliver clear, concise statements that effectively convey your point. For example, the use “In conclusion, the effects of…” rather than “As a result of the analysis, it can be concluded that the effects of…” (Academic Language Experts, 2023).

III. PRACTICAL PART

7. Method of research and data collection

Before any analysis or research could take place, it was necessary to address a fundamental question: which essays to analyse? During their studies at the University of South Bohemia, students are tasked in writing numerous essays on various topics, during which they are taught the fundamentals of proper academic writing. The essay topics range from "The greatest Czech hero" to diverse ones like book analyses and descriptions of even comparisons of different teaching methods. Among these varied topics, one seemed particularly fitting: "Film reviews."

While a popular and seemingly easy topic among many students, its popularity was not the only reason why it was chosen. A substantial contributing factor in choosing this topic was also the fact that the University of South Bohemia offers dedicated film classes. In these classes, students first watch selected foreign films and then discuss the plot, background, themes, and other nuances of the film afterward. This process could then assist the students with writing their own film reviews, which were necessary to pass the class.

To get such essays that could be used for analysis, Dr Koy, one of the teachers of the film classes, was asked for assistance. He was of immense help and provided close to fifty students' film review essays for analysis. Only downside of this being that since these reviews were written for the film classes, the films they were based on were only the ones discussed in class and not entirely ones of the students' own choosing. The students could however choose between any of the discussed films so there was at least some space for variety. In the end, even though the topics of the reviews may not be as varied, it should not be a detriment to the analysis, as it is not important what the reviews are about but how they are written. After their collection, said reviews were used to create the target corpus.

Lastly, to contrast the film reviews written by students, authentic ones written by native English speakers, preferably those written by "professionals" on internet

websites specialising in film and other media reviews such as *Rogerebert.com* or *Polygon.com*. One hundred of such reviews, regarding new and popular films and shows at the time, were collected and subsequently compiled into a reference corpus with a combined total of approximately one hundred and four thousand words to provide variety and wide coverage of contextual language. This reference corpus then served as a basis for language comparison between itself and the target corpus in aim of determining if there are any similarities in the usage of idioms, collocations, etc (see Table 1 for both corpora).

Corpora	***Tokens***	***Words***	***Number of texts***
Target (non-native)	96844	96957	53
Reference (native)	104171	104315	100

Table 1: Corpora used in analysis

7.1. Chosen websites

To get a wide sample of reviews a total number of eight websites was chosen. Each of these websites were verified on websites such as *Transparencyreport.google.com* and *Similarweb.com* to determine their trustworthiness, the amount of internet traffic they experience and popularity compared to similar websites. However, not all of the chosen websites are the most popular as some of the lesser known and popular ones were also chosen to provide a varied sample and see if there are any substantial language differences between reviews from popular and not so popular sites.

7.1.1. RogerEbert.com

Launched in 2002 by a the late Rogert Joseph Ebert, a famous American film critic, journalist, and screenwriter, *RogerEbert.com* holds itself to a very high standard, posting very well structured and detailed reviews of films from all around the world. Managed by a group of professional critics personally selected by Ebert himself before his passing, the site boasts very high numbers of total monthly visits and overall user engagement and retention. One can find here a wide variety of film reviews ranging from all the popular ones currently being played in cinemas to more

indie, artistic, and experimental film projects. And with streaming platforms on the rise, no even platforms Netflix, Hulu or Peacock are ignored as films and even TV series are featured on the site. Lastly, the site also features frequent blogs with director or actor interviews, deep dives into filmmaking, and even overall coverage of film news (see Figure 1).

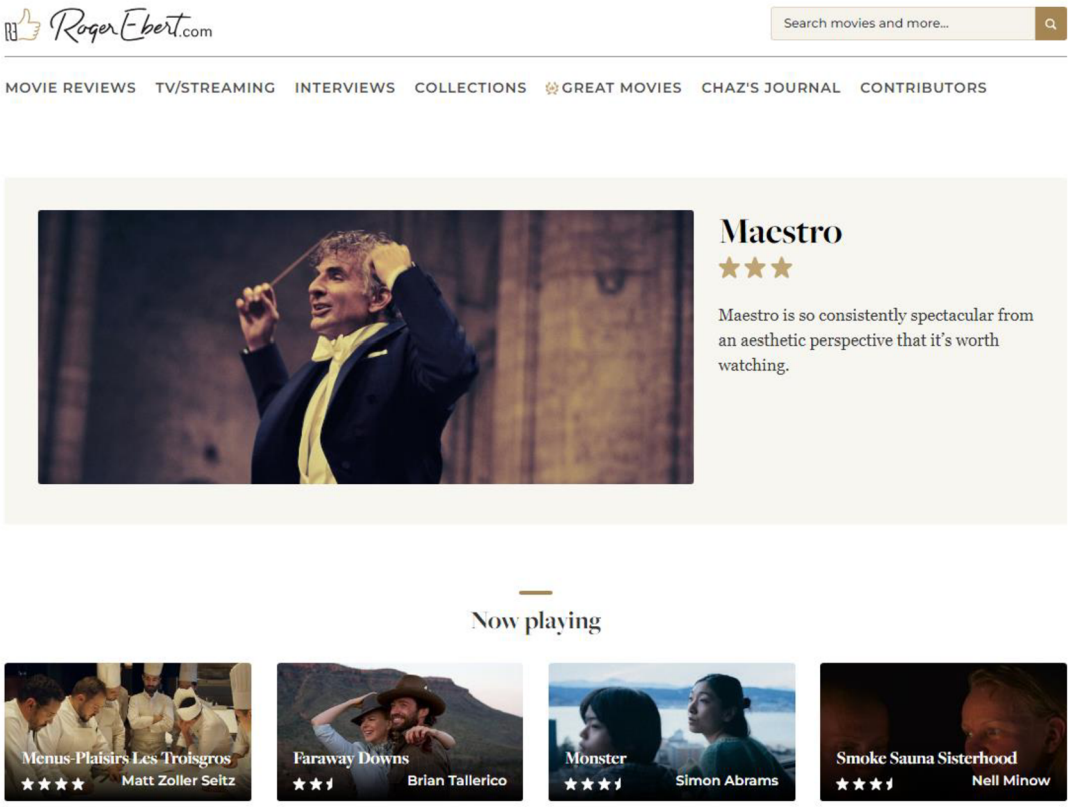

Figure 1: RogerEbert.com

7.1.2. Polygon.com

Polygon.com, another very popular entertainment website, was first launched in 2012 as a purely gaming blog. However, over the years as the website got increasingly popular it evolved and expanded into more of a general pop culture sphere and now covers everything from gaming news and reviews, to film and series reviews, recommendations on what is popular right now and even news or guides about tech and electronics. With around twenty-six million monthly visitors, *Polygon.com* currently as of writing this, ranks as the thirty second most popular pop culture and entertainment media website on the internet (see Figure 2).

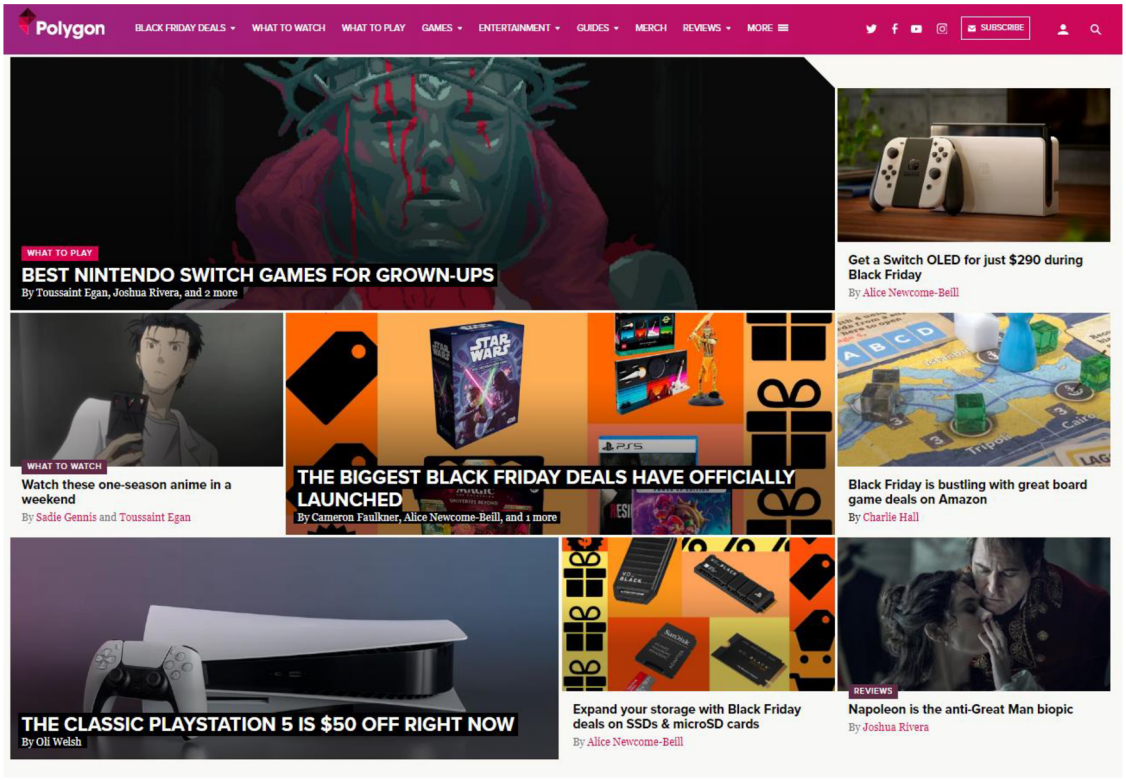

Figure 2: Polygon.com

7.1.3. IndieWire.com

Established in 1996, *IndieWire.com* is a film industry and review website whose main focused used to be predominantly independent film, although with the rising popularity of streaming platforms, the site's focus shifted to a broader one and now includes all mainstream film, television, and streaming media. What used to be a free daily mail newsletter service for independent film is now a sprawling film news and review website boasting around six million monthly visitors and growing. Lastly, the site is also host to many discussions regarding awards, award predictions, interviews, and overall happenings in Hollywood (see Figure 3).

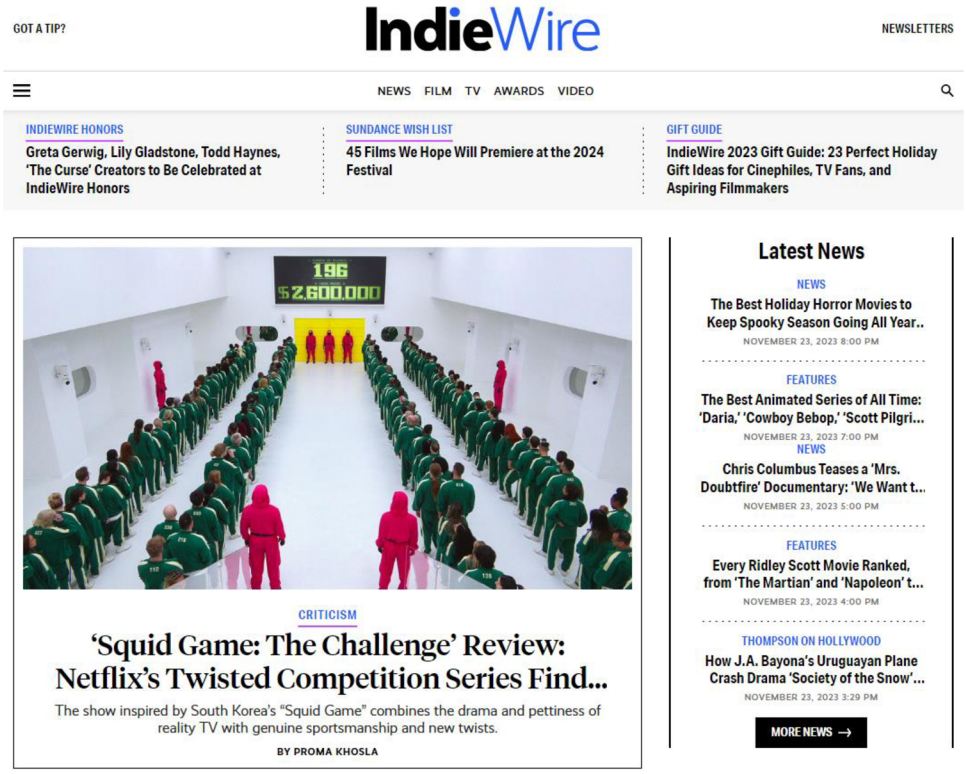

Figure 3: IndieWire.com

7.1.4. ScreenCrush.com

Ran by Townsquare Media, a radio network and media company based in New York, *ScreenCrush.com* is host not only to reviews but also to longform essays about films and film industry in general, trailers, top X lists, and even weekly podcasts discussing film news. While not as popular as previously mentioned websites, *ScreenCrush.com* is still visited by roughly half a million people every month (see Figure 4).

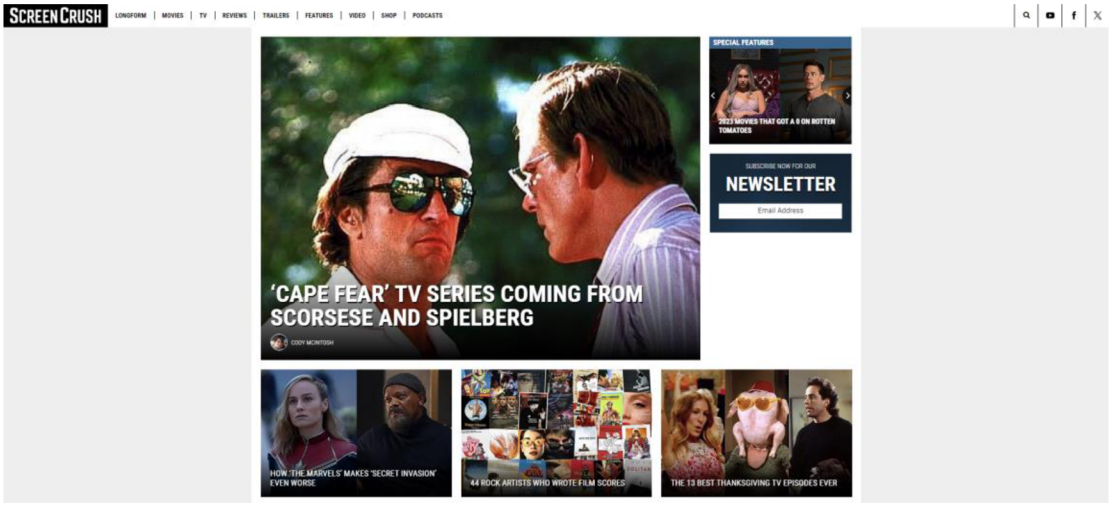

Figure 4. ScreenCrush.com

7.1.5. ReelViews.net

Not to be confused by *ReelReviews.com*, *ReelViews.net* serves as a personal blog for James Berardinelli, an approved film critic and fantasy novelist. Here Berardinelli shares his personal takes on recent films while also reviewing past years of film as a whole. One can also find numerous links to his other platforms like his social media accounts, RottenTomatoes film critic page, or even his Patreon page, where users can pay a monthly fee to get exclusive film news related content or early access to his normal content. *ReelViews.com* is visited by roughly two hundred thousand people every month, which is quite impressive for a personal blog (see Figure 5).

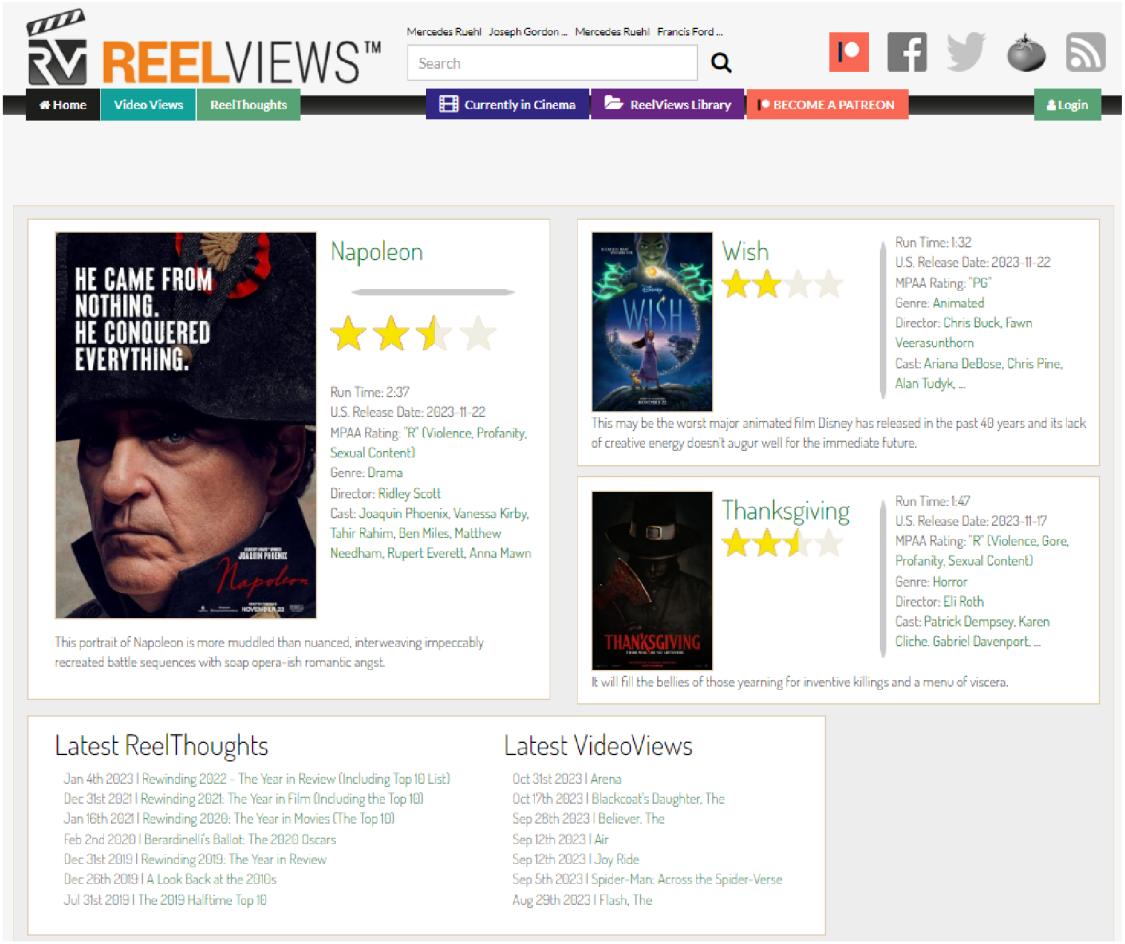

Figure 5: ReelViews.com

7.1.6. ScreenDaily.com

Managed by Screen International, a British film magazine covering international film business, *ScreenDaily.com* provides its viewers a real-time view of the film industry, include all matter of film news, interviews, and reviews. The site also provides information about box office sales from films, annual film festivals and awards. One very interesting feature, that other previously mentioned websites do not have is the option to sort reviews either based on festivals that the films were first screened on or even by their country of origin or if the country somehow participated on making of the film. *ScreenDaily.com* is visited by roughly seven hundred thousand people every month (see Figure 6).

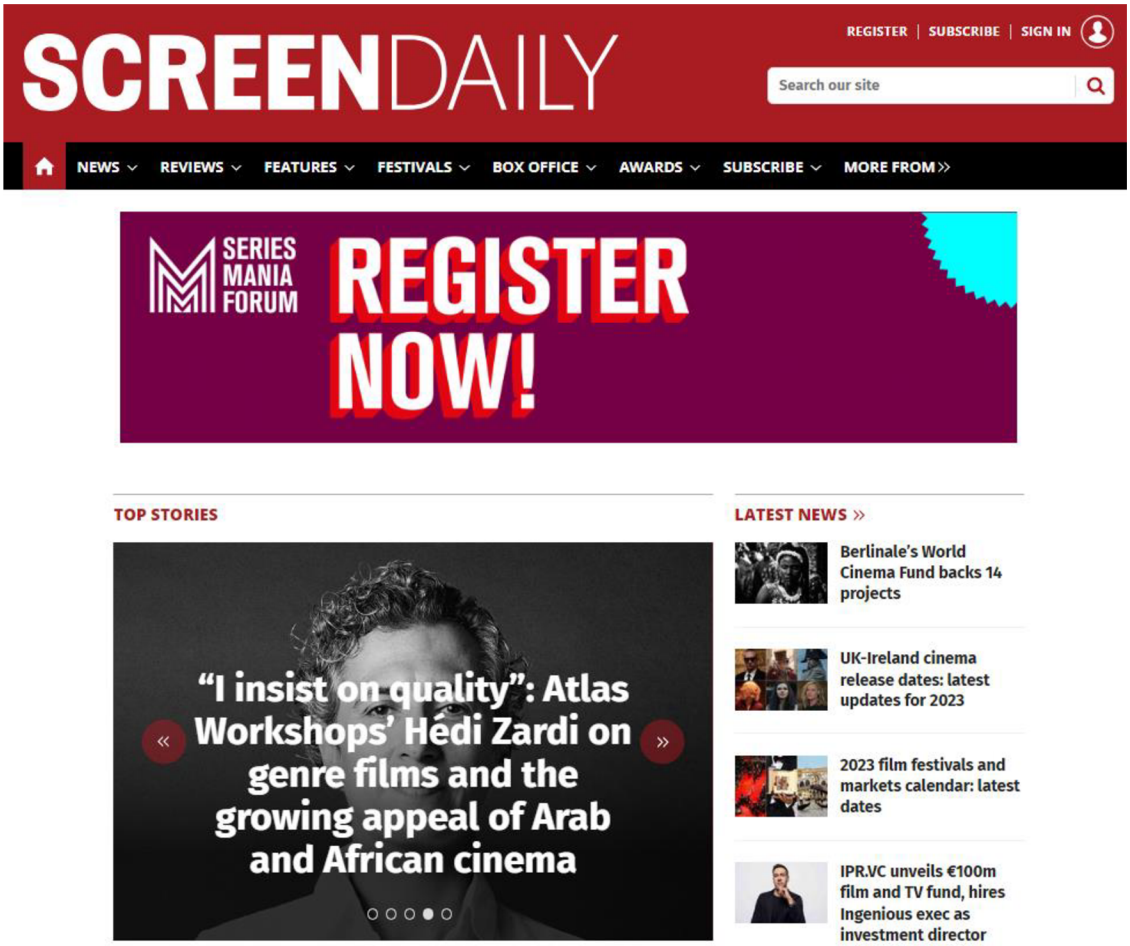

Figure 6: ScreenDaily.com

7.1.7. PlotAndTheme.com

Another personal blog, albeit smaller than *ReelViews.net*, is *PlotAndTheme.com*. Made by an amateur novelist a film critic Derek Jacobs, *PlotAndTheme.com* was used mainly for film reviews however as of 2023 has shifted more to discussing the overall aesthetics of film and writing. This resulted to the website not being updated as often as it used to be, as Jacobs is not writing any new reviews. His old reviews are however still free accessible. Due to its lack of new coverage and niche focus, *PlotAndTheme.com* sees only about forty thousand monthly visitors, which although impressive by itself is quite a small number compared to other mentioned websites (see Figure 7).

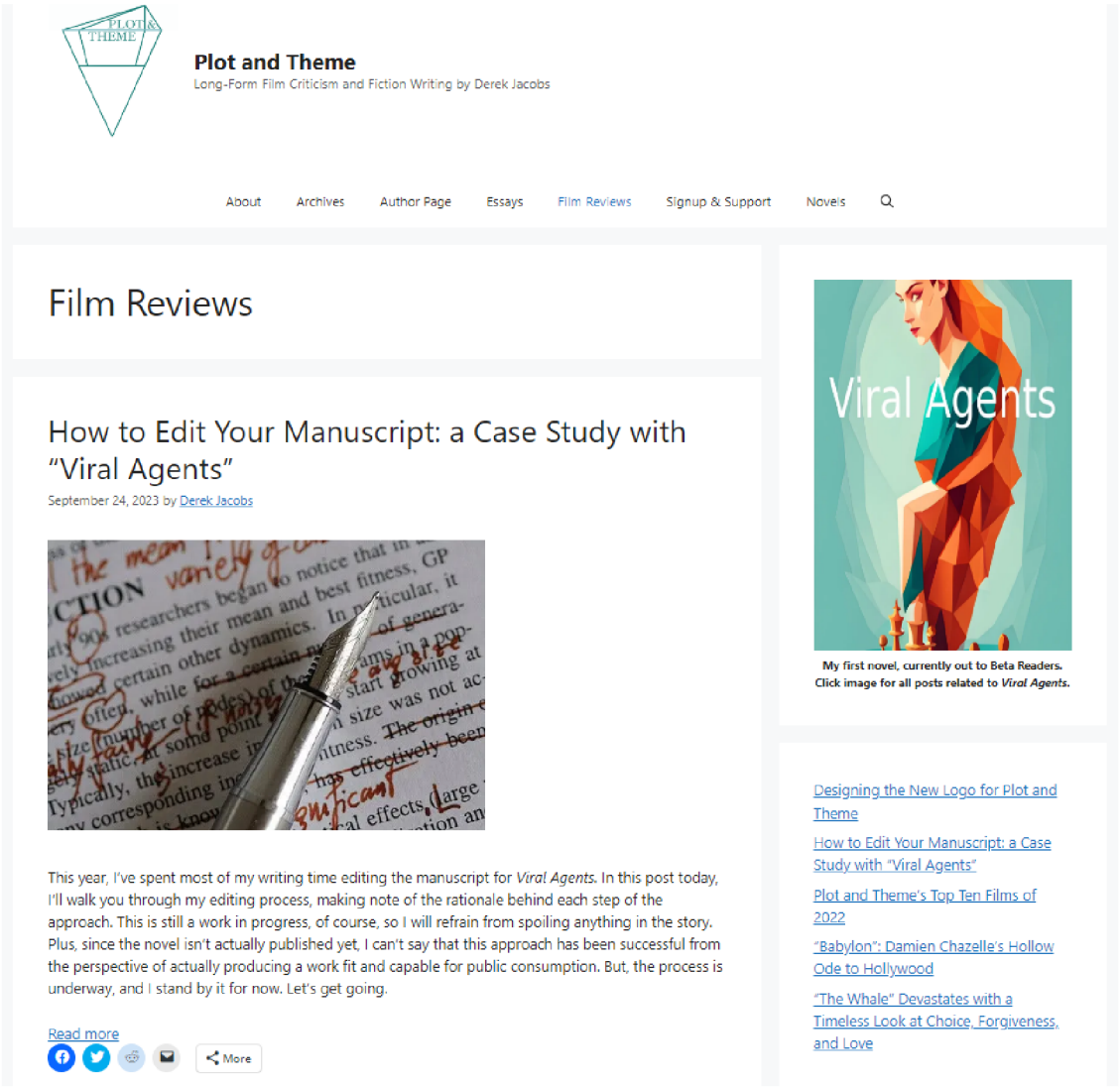

Figure 7: PlotAndTheme.com

7.1.8. LaTimes.com

While predominantly a news website based in Los Angeles, *LaTimes.com* not only include news but also a dedicated "Entertainment & Arts" section, which includes music, art and even film news and reviews. Articles in this section not only discuss all the recent film news but also reminisce about the "good old times" of film and how things have changed. Overall, *LaTimes.com* boast a very high popularity, being visited monthly by around fifty-three million people, although it is unclear, how many of those people visit the website purely to look at film reviews and read through discussion about upcoming blockbusters (see Figure 8).

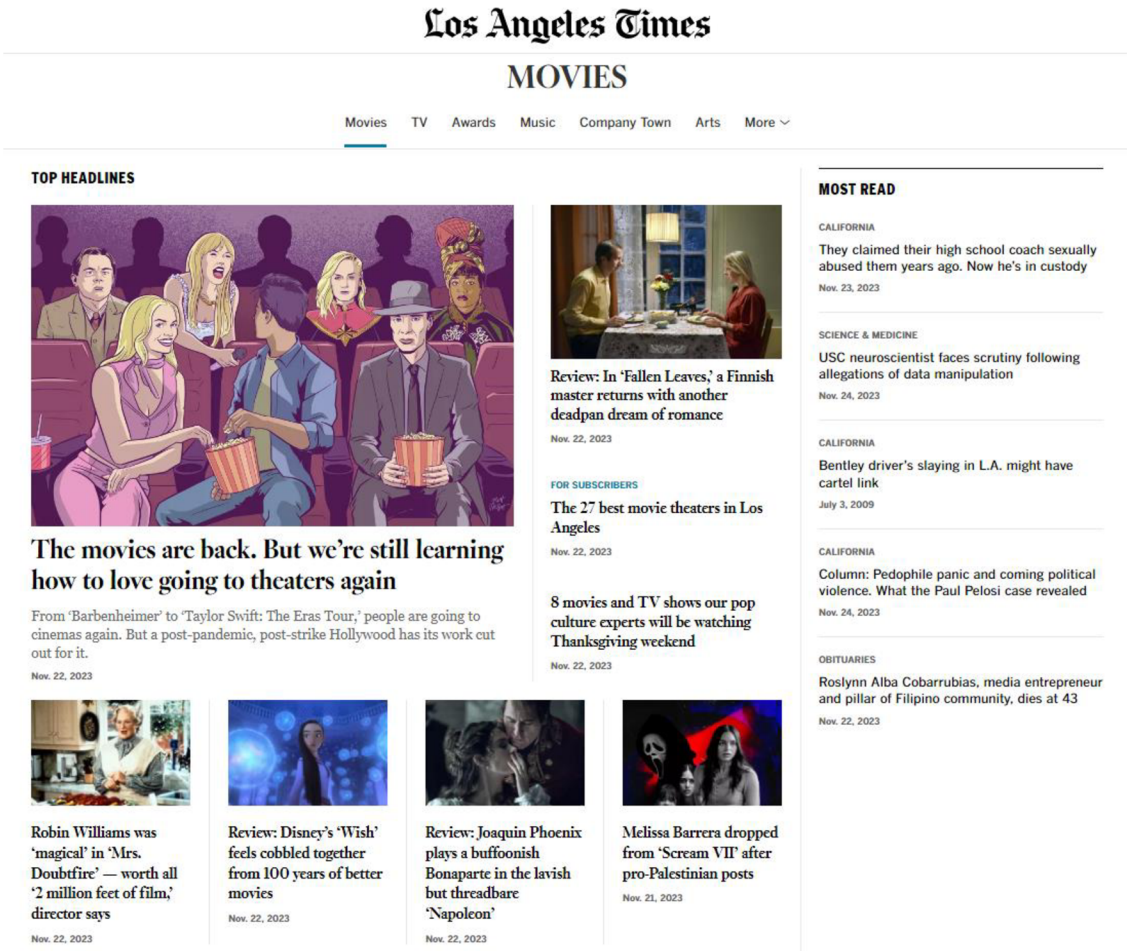

Figure 8: LaTimes.com

7.2. Chosen essays

As mentioned before, around fifty film reviews were provided by Dr Koy for analysis, seven of which were on paper and subsequently scanned while the rest were in electronic form either in .doc or .pdf formats. These reviews were mostly written by second- or third-year students of English on the University of South Bohemia who signed up for BAK1 or BAK2 classes over the last few years, however some of them were also written by at the time Erasmus students most likely from Turkey and Spain, judging by their names. The reviews also include comparisons to the movies' book version, which the reviews collected from websites may not feature. In total thirty-three different films were reviewed, with a handful of them being reviewed multiple times by different students:

- 1984
- A Christmas Carol
- A Farewell to Arms (reviewed a total of 2 times)
- A Lesson Before Dying
- A Tale of Two Cities
- All the King's Men
- American Pastoral
- Daisy Miller
- Death of a Salesman
- Elmer Gantry
- Great Expectations
- Lamb
- O Pioneers! (reviewed a total of 2 times)
- Of Mice and Men
- Pride and Prejudice
- Sense and Sensibility (reviewed a total of 2 times)
- The Age of Innocence
- The Cider House Rules (reviewed a total of 5 times)
- The Color Purple (reviewed a total of 2 times)
- The Crucible
- The Day of the Locust
- The Door in the Floor
- The Dying Animal (reviewed a total of 2 times)
- The Great Gatsby
- The House of Mirth
- The Joy Luck Club (reviewed a total of 3 times)
- The Last Tycoon (reviewed a total of 3 times)
- The Mill on the Floss
- The Quiet American
- The Red Pony
- Their Eyes Were Watching God (reviewed a total of 2 times)
- Washington Square
- Wuthering Heights (reviewed a total of 2 times)

The names of the students will not be shared, saved, or included in the analysis in any way as to not violate GDPR or any similar identity protection laws.

8. #LancsBox and used functions

#LancsBox, a freely available software package created at the Lancaster University, is custom built for the examination of language data and corpora, making it an essential tool for this study (see Figure 9). Developed by a team of talented individuals, #LancsBox boasts several key features, like the ability to handle both user-specific data and pre-existing corpora, visualise language data and corpora, compare multiple corpora, analyse data in various languages, automatically annotate data for part-of-speech, and user-friendly functionality and design. In addition to that, the #LancsBox websites is also host to numerous free tutorials explaining all the software's functionalities, available both in PDF and video form, within its comprehensive user guide (#LancsBox, 2023).

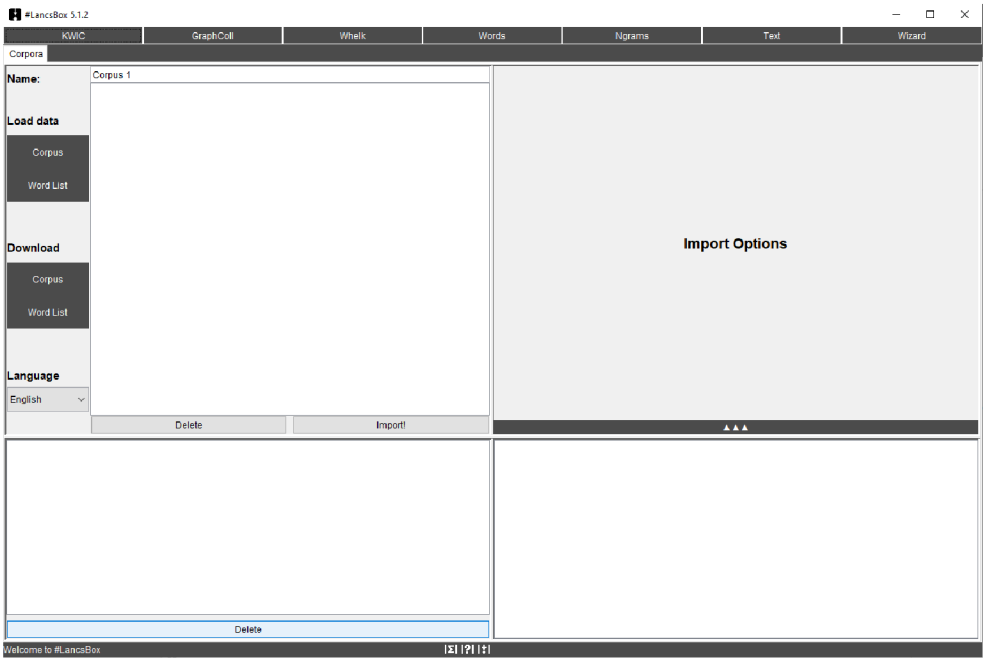

Figure 9: The default #LancsBox interface

8.1. "Words" function

One of the essential features utilized in #LancsBox is the "Words" function (see Figure 10). This function enables users to analyse the frequencies of types, lemmas,

or POS categories. Moreover, it allows the comparison of corpora through the "keywords" technique. In this thesis, the "Words" function was employed to examine the frequencies of lexemes of two corpora, one consisting of around fifty student film review essays, totalling approximately ninety-five thousand words, and the other consisting of one hundred authentic film reviews from various online websites, totalling approximately one hundred thousand words. These frequency lists were subsequently sorted from the most frequent to the least frequent for the purpose of comparison and further analysis (see Figure 11).

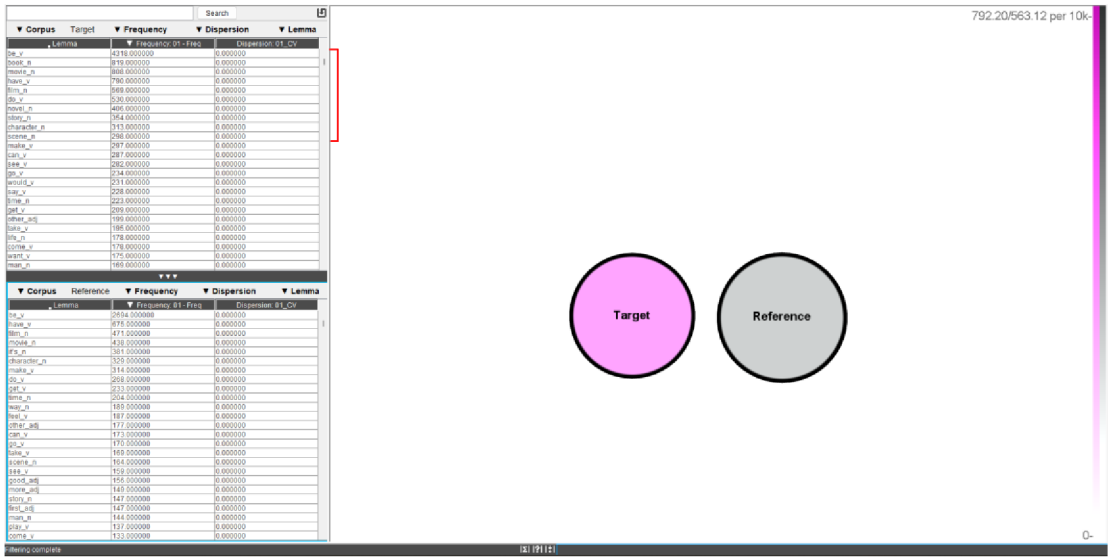

Figure 10: The "Words" function used on two corpora

▼ Corpus Target	▼ Frequency	▼ Dispersion	▼ Lemma
Lemma	▼ Frequency: 01 - Freq	Dispersion: 01_CV	
be_v	4318.000000	0.000000	
book_n	819.000000	0.000000	
movie_n	808.000000	0.000000	
have_v	790.000000	0.000000	
film_n	569.000000	0.000000	
do_v	530.000000	0.000000	
novel_n	406.000000	0.000000	
story_n	354.000000	0.000000	
character_n	313.000000	0.000000	
scene_n	298.000000	0.000000	
make_v	297.000000	0.000000	
can_v	287.000000	0.000000	
see_v	282.000000	0.000000	
go_v	234.000000	0.000000	
would_v	231.000000	0.000000	
say_v	228.000000	0.000000	
time_n	223.000000	0.000000	
get_v	209.000000	0.000000	
other_adj	199.000000	0.000000	
take_v	195.000000	0.000000	
life_n	178.000000	0.000000	
come_v	178.000000	0.000000	
want_v	175.000000	0.000000	
man_n	169.000000	0.000000	

Figure 11: Top ten most frequent words in the target corpus

8.2. “GraphColl” function

Another essential feature of #LancsBox used in this analysis is the “GraphColl” function (see Figure 12). With this function users are able to identify and display collocations of words or phrases while also providing a visual representation. In this thesis, the function was used on ten of the most frequently occurring words in both corpora, identified using the "Words" function of #LancsBox. Sequentially, these nouns were individually input into the search bar, accompanied by adjustments to the Span, Statistics, and Type parameters to fine-tune the analysis outcome. The output produced a graph illustrating the analysed word alongside its collocates, adhering to the specified parameters, including details on their frequency, position, and collocation strength (see Figure 13). Subsequently, this information was utilized in the creation of collocational profiles for each word.

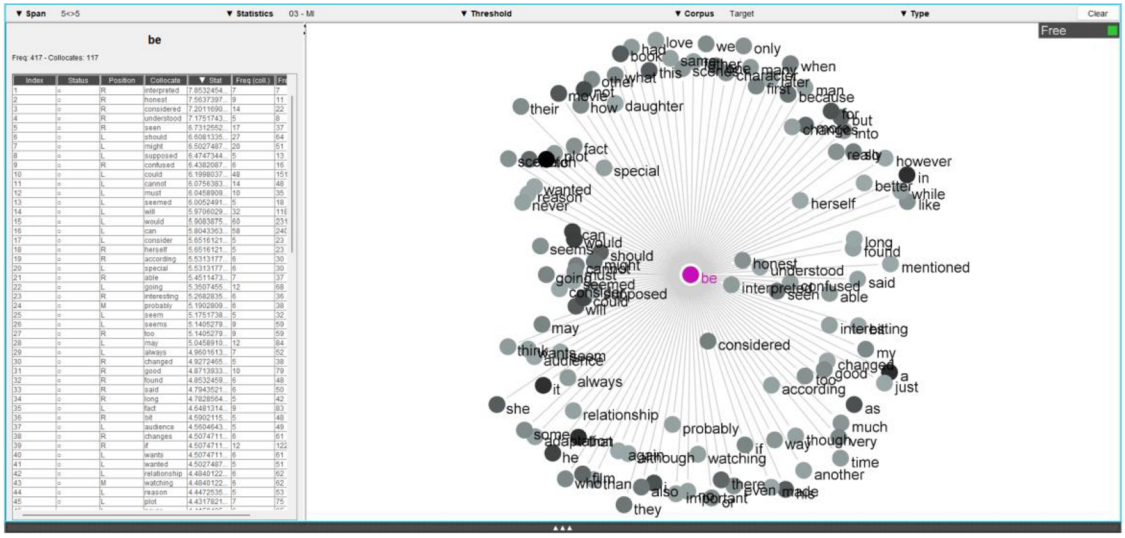

Figure 12: The "GraphColl" function

be

Freq: 417 - Collocates: 115

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corpus)
1	○	R	honest_adj	7.56373977...	9	11
2	○	R	interpret_v	7.20116969...	7	11
3	○	L	should_v	6.60813351...	27	64
4	○	L	might_v	6.56046430...	20	49
5	○	L	consider_v	6.46420359...	21	55
6	○	L	may_v	6.31719327...	10	29
7	○	L	suppose_v	6.26828338...	5	15
8	○	L	could_v	6.19980377...	48	151
9	○	R	confuse_v	6.19028130...	6	19
10	○	L	must_v	6.04589096...	10	35
11	○	L	will_v	5.97060297...	32	118
12	○	L	would_v	5.90838752...	60	231
13	○	L	can_v	5.83808616...	71	287
14	○	R	herself_pron	5.65161214...	5	23
15	○	R	accord_v	5.53131778...	6	30

Figure 13: The strongest collocates of the word "Be"

8.3. Association measures

Association measures serve as mathematical tools or formulas commonly used in identifying collocations within corpora. These measures mostly rely on statistical testing of hypotheses, however there are also measures that include both mathematically grounded and empirically motivated approaches. Notable association measures include *Dice*, *log-likelihood*, *MI-score*, *MI3*, *T-score*, etc. Due to the multifaceted nature of collocations from a linguistic and mathematical point of view, these measures may differ significantly in the way they consider important collocational patterns.

Association measures often look at the frequency the whole collocation, its individual parts, and the overall corpus size. This information is then organized in contingency tables, and the measures use a specific formula to calculate a numerical value.

The outcome value for a specific word pair in the corpus indicates the extent of association between them, and this association may be negative in certain measures, further indicating a negative association or in other words, mutual "repulsion."

Comparing numerical values between different association measures is generally not straightforward. However, for the purpose of comparison, numerical values are typically converted into ranks in a list of collocations, organized based on the numerical values of the specific measure (Český Národní Korpus, 2019).

8.3.1. MI-Score

In summary, MI-score serves as an association measure specifically applied when searching for strong collocations characterized by high relative frequency, signifying their exceptional or random nature.

There is however a drawback associated with MI-score, and that is its susceptibility to be influenced by individual word frequencies. This is not particularly uncommon as the highest values are often achieved by word pairs with lower frequencies. To address this issue, corpus management tools such as #LancsBox offer the option to establish a lower frequency limit during MI-score calculation, effectively eliminating the need to calculate the score for words falling below this limit.

MI-score values are generally positive, with negative values indicating infrequent mutual repulsion. The MI = 7 limit is commonly regarded as significant for a one hundred million corpus, suggesting a systemic collocation. In the context of this analysis, the MI = 3 limit was chosen for a one hundred-thousand-word corpus (Český Národní Korpus, 2019).

9. Analysis

9.1. Method of analysis

The first part of the analysis focused on identifying the most frequently used words in both corpora. This was accomplished by using the aforementioned "*Words*" function, selecting lemmas as the primary units, default frequency, and default dispersion. Setting the primary units as lemmas not only helps with displaying their POS, allowing for better filtering, but also displays the selected words in their base "dictionary" form. Additionally, the "*not *_other|*_con|*_pron|*_adv*" custom filter

was applied to the lemmas in order to exclude conjunctions, pronouns, adverbs and other elements such as determiners and articles from the selection process, as otherwise words like "a", "the", "or", "he/she/they" etc. would be without a doubt the most frequent in both corpora. With the filtering complete, ten of the most frequent words from both corpora were then noted and selected for further analysis in order to create their collocational profiles and their eventual comparison. It is important to note that both corpora featured the verb "be" as by far the most frequent word. This word was however excluded from the final selection, due to it being used as an auxiliary verb in the vast majority of cases, making its collocability open and being able to be distributed almost anywhere.

As for the target corpus (i.e. student film reviews), these were the ten most frequently used words by the students: *book, movie, have, film, do, novel, story, character, scene, make* (see Table 2). Of these ten words only three are verbs while the rest are nouns with the most frequent word being the noun "Book" with a total of eight hundred and nineteen occurrences.

Word	***Frequency***	***Relative frequency***
Book	819	8456.9
Movie	808	8343.315
Have	790	8157.4486
Film	569	5875.4288
Do	530	5472.7192
Novel	406	4192.3096
Story	354	3655.3635
Character	313	3232.002
Scene	298	3077.1137
Make	297	3066.7877

Table 2: Top ten most frequent words (Target corpus)

As for the reference corpus (i.e. authentic "professional" film reviews), these were the ten most frequently used words by the film critics: *have, film, movie, character, make, do, get, time, way, feel* (see Table 3). Of these ten words five are verbs and five

are nouns with the most frequent word being the verb “have” with a total of six hundred and seventy-five occurrences.

Word	***Frequency***	***Relative frequency***
Have	675	6479.793
Film	471	4521.455
Movie	438	4204.6654
Character	329	3158.299
Make	314	3014.3036
Do	268	2572.7177
Get	233	2236.7283
Time	204	1958.3374
Way	189	1814.342
Feel	187	1795.1426

Table 3: Top ten most frequent words (Reference corpus)

From the analysis it is clear that there are some words that occur in both tables which is to be expected since they are either generally quite common (e.g. auxiliary verbs) or since the samples share their general topic (e.g. the word film).

The second part of the analysis was focused on creating collocational profiles for each of the previously selected words in order to determine their most frequent (and strongest) collocates. This was accomplished by using the aforementioned “GraphColl” function and searching for each in their respective corpus. To further specify the output of the function, the MI-score was utilized to identify strong collocates often associated with selected words, although the strength of collocates does not necessarily directly translate to frequency. Collocates analysed by the MI-score can be further explored using the integrated "KWIC" function. This function compiles all instances of selected collocates in the corpus, presenting them in a concise textual format. The length of the displayed text can be adjusted by modifying the Context value. It is important to note that the use of the "KWIC" function is purely optional and is not elaborated upon in this thesis.

9.2. Collocational profiles – Target corpus

These collocational profiles were created using the "*GraphColl*" function with the Span of 5<>5, MI and T Statistics, default Threshold and Lemmas as the type. The profiles show the ten most frequent collocates for each word which are ordered by their respective scores.

9.2.1. The word "Book"

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word "Book" compared to all other words they were collocated with (see Table 4).

Collocate	***MI-score***	***Freq (coll.)***	***Freq (corpus)***
Act	6.061512	11	20
Read	5.791114	57	125
Finish	5.545498	5	13
Comparison	5.454523	13	36
Correspond	5.339046	7	21
Luck	5.271932	7	22
Continue	5.207802	7	23
Compare	5.169121	16	54
Ending	5.093934	9	32
Joy	5.087507	7	25

Table 4: Collocational profile of the word "Book" (Target corpus)

The MI-score results indicate that the strongest collocate for the word "Book" was the word "Act", boasting an MI-score of approximately 6.06. This collocation occurred 11 times out of the total 20 appearances in the corpus, resulting in a relative frequency of roughly 113.584 and a probability of around 55 % of appearing as this specific collocation. The pairing of "Book" with "Act" emerged as the strongest and most prevalent collocation mostly because of the students' comparing acts of the books with the acts of the film adaptations. Additionally, students also used the word "act" as a verb, specifically when describing how someone acted in the book

compared to the film. At first glance, the MI-score results table features a relatively equal mix verbs and nouns, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.2. The word “Movie”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Movie” compared to all other words they were collocated with (see Table 5).

Collocate	***MI-score***	***Freq (coll.)***	***Freq (corpus)***
Length	5.755938	6	14
Minute	5.698223	7	17
Whereas	5.591308	13	34
Final	5.563294	6	16
Miss	5.434010	12	35
Introduction	5.393368	5	15
Cider	5.334475	8	25
Pretty	5.315366	6	19
Storyline	5.315366	6	19
Mostly	5.262124	7	23

Table 5: Collocational profile of the word "Movie" (Target corpus)

The MI-score results indicate that the strongest collocate for the word "Movie" was the word “Length”, boasting an MI-score of approximately 5.76. This collocation occurred 6 times out of the total 14 appearances in the corpus, resulting in a relative frequency of roughly 61.955 and a probability of around 42.86 % of appearing as this specific collocation. The pairing of “Movie” with “Length” emerged as the strongest and most prevalent collocation mostly because of the students’ dislike of the film’s length due to it either cutting too short and omitting crucial parts from the book or being too long and drawn out, adding unnecessary filler scenes to pad out the runtime. At first glance, the MI-score results table features a mix of verbs, nouns, adverbs, a conjunction and an adjective, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.3. The word “Have”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Have” compared to all other words they were collocated with (see Table 6).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Must	6.264903	8	35
Should	6.201541	14	64
Already	6.157147	7	33
Might	5.949401	9	49
Choice	5.934754	6	33
Imagine	5.916139	7	39
You	5.673708	22	145
They	5.625070	71	484
Problem	5.546189	5	36
Add	5.468186	5	38

Table 6: Collocational profile of the word "Have" (Target corpus)

The MI-score results indicate that the strongest collocate for the word "Have" was the word “Must”, boasting an MI-score of approximately 6.26. This collocation occurred 8 times out of the total 35 appearances in the corpus, resulting in a relative frequency of roughly 82.607 and a probability of around 22.587 % of appearing as this specific collocation. The pairing of “Have” with “Must” emerged as the strongest and most prevalent collocation mostly because of the students expressing their assumptions or opinions they gathered from the films viewing. Some of the common expressions the students used in this case were for example “they must have read the book…” or “ he/she/they must have been…”. At first glance, the MI-score results table features a mix of verbs, pronouns, and adverb and a noun, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.4. The word “Film”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Film” compared to all other words they were collocated with (see Table 7).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Maker	6.821203	11	17
Produce	6.771162	6	10
Adaptation	6.238667	70	162
Contrast	5.805378	8	25
Appreciate	5.574765	6	22
Whereas	5.531696	9	34
Throughout	5.523235	5	19
Successful	5.495038	8	31
Hard	5.449234	5	20
Shoot	5.390340	6	25

Table 7: Collocational profile of the word "Film" (Target corpus)

The MI-score results indicate that the strongest collocate for the word "Film" was the word “Maker”, boasting an MI-score of approximately 6.82. This collocation occurred 11 times out of the total 17 appearances in the corpus, resulting in a relative frequency of roughly 113.584 and a probability of around 64.706 % of appearing as this specific collocation. The pairing of "Have" with “Must” emerged as the strongest and most prevalent collocation mostly because of the students expressing what the makers of the films did or did not do to properly adapt the books into film. It is however quite interesting that even though the terms “film” and “movie” are practically interchangeable, with only negligible differences, their collocational profiles are vastly different with the word “whereas” being their only similarity. Where the collocates of “film” are mostly based around the technicalities of filmmaking (maker, produce, adaptation, shoot), the collocates of “movie” are more based around the content (storyline, length, minutes, introduction). At first glance, the MI-score results table features a relatively equal mix of nouns, verbs, conjunctions and adjectives, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.5. The word “Do”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Do” compared to all other words they were collocated with (see Table 8).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Anything	7.446295	5	18
Nothing	6.925058	6	31
Why	6.214565	11	93
Not	6.060942	89	837
Understand	6.057253	7	66
What	5.906021	17	178
Thing	5.821804	10	111
Know	5.781222	12	137
We	5.749972	21	245
Should	5.616220	5	64

Table 8: Collocational profile of the word "Do" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Do” was the word “Anything”, boasting an MI-score of approximately 7.45. This collocation occurred 5 times out of the total 18 appearances in the corpus, resulting in a relative frequency of roughly 51.629 and a probability of around 27.778 % of appearing as this specific collocation. The pairing of “Do” with “Anything” emerged as the strongest and most prevalent collocation, attributed to the students describing the plot of the films, specifically when some characters either would do anything for others or were unable to do anything in an important situation. Not far behind in terms of MI-score is also the word “nothing” which served in a similar way, either describing that nothing could be done in a given situation or that someone did nothing. At first glance, the MI-score results table features a mix of nouns, verbs, adverbs and pronouns, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.6. The word “Novel”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Novel” compared to all other words they were collocated with (see Table 9).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
American	6.888414	5	10
Comparison	6.303451	12	36
Sensibility	6.225449	6	19
Capture	6.210342	5	16
Element	5.962414	5	19
Both	5.888414	16	64
Jane	5.888414	5	20
Reflect	5.888414	5	20
Base	5.800951	8	34
Compare	5.592958	11	54

Table 9: Collocational profile of the word "Novel" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Novel” was the word “American”, boasting an MI-score of approximately 6.89. This collocation occurred 5 times out of the total 10 appearances in the corpus, resulting in a relative frequency of roughly 51.629 and a probability of 50 % of appearing as this specific collocation. The pairing of “Novel” with “American” emerged as the strongest and most prevalent collocation, attributed mostly to some of the students choosing the film American Pastoral and its book counterpart as the basis for their review. Other appearances of this collocation were in the review for the film The Quiet American. Not far behind in terms of MI-score is also the word “comparison” which served an important role when one of the students’ tasks was to compare the book/novel to its film adaptation. At first glance, the MI-score results table features a mix of nouns and verbs, and a conjunction, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.7. The word “Story”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Story” compared to all other words they were collocated with (see Table 10).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Line	6.216017	6	22
Whole	6.216017	21	77
Mainly	6.020097	5	21
Continue	5.888852	5	23
End	5.810378	68	68
Tell	5.702756	30	157
Original	5.453056	9	56
Aspect	5.427521	6	38
Begin	5.317896	6	41
Part	5.183595	16	120

Table 10: Collocational profile of the word "Story" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Story” was the word “Line”, boasting an MI-score of approximately 6.22. This collocation occurred 6 times out of the total 22 appearances in the corpus, resulting in a relative frequency of roughly 61.955 and a probability of around 27.272 % of appearing as this specific collocation. The pairing of “Story” with “Line” emerged as the strongest and most prevalent collocation, attributed most likely to a misspelling of the word “storyline” by one or more students, separating it into two words which incidentally boosted its MI-score. The next best collocate, which is not a misspell, is the word “whole” which students mostly used in expressions like “the whole story…” or “the story as a whole…”. At first glance, the MI-score results table features a mix of nouns, verbs, adjectives and an adverb, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.8. The word “Character”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Character” compared to all other words they were collocated with (see Table 11).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Development	8.614366	8	14
Wells	7.220088	5	23
Main	6.580419	18	129
Play	5.573724	5	72
Another	5.334258	8	136
Homer	5.292438	8	140
Important	5.075946	6	122
Only	4.563740	6	174
This	4.438728	16	506
What	4.267916	5	178

Table 11: Collocational profile of the word "Character" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Character” was the word “Development”, boasting an MI-score of approximately 8.61. This collocation occurred 8 times out of the total 14 appearances in the corpus, resulting in a relative frequency of roughly 82.607 and a probability of around 57.143 % of appearing as this specific collocation. The pairing of “Character” with “Development” emerged as the strongest and most prevalent collocation, attributed to the students expressing their thoughts about the personal development of the films or books characters or lack thereof. The table also features two quite interesting words, which being “Homer” and “Wells”, which are the first and last names of the titular character from the book Cider House Rules and its film adaptation. This can be attributed to the popularity of the title, as it was chosen by the students a total of five times, making the most film to review by the students. At first glance, the MI-score results table features a mix of nouns, adjectives and a verb, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.9. The word “Scene”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Scene” compared to all other words they were collocated with (see Table 12).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Final	7.727935	6	16
Extra	7.204373	6	23
Where	6.087119	19	158
Whole	5.683541	7	77
Appear	5.582258	5	59
Next	5.534163	5	61
Another	5.377438	10	136
This	5.289262	35	506
There	5.129510	20	323
Which	4.674203	14	310

Table 12: Collocational profile of the word "Scene" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Scene” was the word “Final”, boasting an MI-score of approximately 7.73. This collocation occurred 6 times out of the total 16 appearances in the corpus, resulting in a relative frequency of roughly 61.955 and a probability of around 37.5 % of appearing as this specific collocation. The pairing of “Scene” with “Final” emerged as the strongest and most prevalent collocation, attributed mostly to the students describing the endings of the films, often comparing the finals scenes of the films to those of the books. Another relatively frequent collocate was the word “extra”, which was used mainly to illustrate the differences or additions that the films had compared to the books. At first glance, the MI-score results table features a mix of adjectives, adverbs, determiners and a verb, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.2.10. The word “Make”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Make” compared to all other words they were collocated with (see Table 13).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Easy	7.761419	6	25
Sense	7.168236	7	44
Audience	6.732850	6	51
Own	6.522632	6	59
Us	6.387353	5	54
Any	6.119873	5	65
Feel	5.347825	5	111
More (adj.)	5.309351	5	114
Good	5.166870	6	151
More (adv.)	5.092392	6	159

Table 13: Collocational profile of the word "Make" (Target corpus)

The MI-score results indicate that the strongest collocate for the word “Make” was the word “Easy”, boasting an MI-score of approximately 7.76. This collocation occurred 6 times out of the total 25 appearances in the corpus, resulting in a relative frequency of roughly 61.955 and a probability of 24 % of appearing as this specific collocation. The pairing of “Make” with “Easy” emerged as the strongest and most prevalent collocation, attributed mostly to two phrases, being “easy to make” and “make it easy”, which the students used to describe some choices the film makers made either to attract the audience or to better convey the film’s plot, that may have been a bit too convoluted in the book. At first glance, the MI-score results table features a mix of adjectives, nouns, a pronoun, an adverb and a verb, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3. Collocational profiles – Reference corpus

These collocational profiles were, similarly to the ones created for the target corpus, created using the "*GraphColl*" function with the Span of 5<>5, MI and T Statistics, default Threshold and default type. The profiles show the ten most frequent collocates for each word which are ordered by their respective scores.

9.3.1. The word "Have"

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word "Have" compared to all other words they were collocated with (see Table 14).

Collocate	***MI-score***	***Freq (coll.)***	***Freq (corpus)***
Lie	7.027756	6	17
Don't	6.829817	8	26
Superhero	6.337611	7	32
Could	6.267222	20	96
Would	6.193973	20	101
May	6.070825	14	77
Doesn't	5.700182	9	64
Might	5.530257	9	72
Little	5.430721	7	60
Really	5.208329	9	90

Table 14: Collocational profile of the word "Have" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word "Have" was the word "Lie", boasting an MI-score of approximately 7.03. This collocation occurred 6 times out of the total 17 appearances in the corpus, resulting in a relative frequency of roughly 57.598 and a probability of around 35.294 % of appearing as this specific collocation. The pairing of "Have" with "Lie" emerged as the strongest and most prevalent collocation, attributed mostly to phrases regarding the target demographic of films and that the film makers or producers sometimes "have to lie" to their audience to sell them the film for example through making the trailers much more bombastic than the film actually is. The next best collocate, the word "don't" is

another example of the “GraphColl” function considering it a new lexeme other than a lemma of “do”. At first glance, the MI-score results table features a mix of nouns, verbs and adjectives, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.2. The word “Film”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Film” compared to all other words they were collocated with (see Table 15).

Collocate	***MI-score***	***Freq (coll.)***	***Freq (corpus)***
Entire	6.104928	9	36
Throughout	5.935004	6	27
Open	5.671969	5	27
Marvel	5.519966	5	30
Begin	5.486019	7	43
Less	5.382462	5	33
Problem	5.339394	5	34
Final	5.178929	5	38
Course	5.104929	5	40
First	5.075181	18	147

Table 15: Collocational profile of the word "Film" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Film” was the word “Entire”, boasting an MI-score of approximately 6.1. This collocation occurred 9 times out of the total 36 appearances in the corpus, resulting in a relative frequency of roughly 86.397 and a probability of 25 % of appearing as this specific collocation. The pairing of “Film” with “Entire” emerged as the strongest and most prevalent collocation, attributed the reviewers wanting to describe either the entirety of the film (either in positive or negative light) or something recurring throughout the film’s runtime. The same goes for the next strongest collocate, the word “Throughout”, which the word “entire” also often appeared next to. At first glance, the MI-score results table features a mix of nouns, verbs, adjectives and a

conjunction, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.3. The word “Movie”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Movie” compared to all other words they were collocated with (see Table 16).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Disney	7.044128	7	17
Star	6.236773	12	51
So	5.587270	9	60
Original	5.288612	5	41
Since	5.253846	5	42
Much (adj.)	5.154311	8	72
Watch	5.044128	7	68
This	4.993846	51	513
Action	4.943414	12	125
Much (adv.)	4.897971	8	86

Table 16: Collocational profile of the word "Movie" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Movie” was the word “Disney”, boasting an MI-score of approximately 7.04. This collocation occurred 7 times out of the total 17 appearances in the corpus, resulting in a relative frequency of roughly 67.197 and a probability of around 41.176 % of appearing as this specific collocation. The pairing of “Movie” with “Disney” emerged as the strongest and most prevalent collocation, attributed most likely to the popularity of Disney films in general. In addition, Disney being the film and entertainment giant it is, there is hardly a month or two where a new Disney film does not come out. At first glance, the MI-score results table features a mix of nouns, adjectives, conjunctions, an adverb and a determiner, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.4. The word “Character”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Character” compared to all other words they were collocated with (see Table 17).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Main	8.032309	9	24
Development	8.032309	6	16
Every	6.107496	8	81
Black	5.936384	5	57
Really	5.277421	5	90
Play	5.156669	7	137
Even	4.646446	8	223
His	4.311893	20	703
As	4.277421	20	720
Than	4.161944	6	234

Table 17: Collocational profile of the word "Character" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Character” was the word “Main”, boasting an MI-score of approximately 8.03. This collocation occurred 9 times out of the total 24 appearances in the corpus, resulting in a relative frequency of roughly 86.397 and a probability of around 37.5 % of appearing as this specific collocation. The pairing of “Character” with “Main” emerged as the strongest and most prevalent collocation, attributed solely to reviewers describing the main characters of the films, either describing them or expressing their liking or disliking of their behaviour. This pair of words being the strongest collocation is not very surprising as nearly every film has a main character. Another very strong collocate of the word “Character” was the word “Development”, having the same MI-score as “Main” but a bit lower frequency. This also is not very surprising as character development tends to be a common plot point in films, making the characters in films appear more realistic. At first glance, the MI-score results table features a mix of nouns, conjunctions, adverbs, and adjective, a pronoun and a determiner, though

semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.5. The word “Make”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Make” compared to all other words they were collocated with (see Table 18).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Enough	6.882490	5	42
Help	6.679861	6	58
Them	5.831864	10	174
Want	5.751246	5	92
Would	5.616596	5	101
How	5.084983	5	146
They	5.065354	10	296
Feel	4.990947	6	187
Some	4.831864	5	174
To	4.767579	70	2547

Table 18: Collocational profile of the word "Make" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Make” was the word “Enough”, boasting an MI-score of approximately 6.88. This collocation occurred 5 times out of the total 42 appearances in the corpus, resulting in a relative frequency of roughly 47.998 and a probability of around 11.905 % of appearing as this specific collocation. The pairing of “Make” with “Enough” emerged as the strongest and most prevalent collocation, attributed surprisingly not to the phrase “… make enough (of something)” but to the phrase “… enough to make (something)”. Some of the phrases used by the reviewers include: “… was enough to make her into a supervillain …” or “… enough to make an in-the-know horror fan stop …”. The next strongest collocate was the word “Help” which reviewers mostly used when describing elements of the films that either helped it make sense or something that happened in the plot. At first glance, the MI-score results table features a mix of

verbs, pronouns, determiners, an adverb and an adjective, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.6. The word “Do”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Do” compared to all other words they were collocated with (see Table 19).

Collocate	***MI-score***	***Freq (coll.)***	***Freq (corpus)***
Made	9.942763	8	9
Devil	9.653257	8	11
Me	8.282614	9	32
What	6.337901	16	219
Want	6.174089	6	92
We	5.965847	7	124
Thing	5.864761	6	114
Find	5.762191	5	102
I	5.762191	5	102
They	5.710590	14	296

Table 19: Collocational profile of the word "Do" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Do” was the word “Made”, boasting an MI-score of approximately 9.94. This collocation occurred 6 times out of the total 8 appearances in the corpus, resulting in a relative frequency of roughly 76.797 and a probability of around 27.272 % of appearing as this specific collocation. The pairing of "Do" with “Made” emerged as the strongest and most prevalent collocation, attributed solely to the film *The Conjuring: The Devil Made Me Do It,* which was being reviewed and the title often repeatedly referred to. Being part of the film’s title may be the reason, why the “*GraphColl*” function considered it an entirely new lexeme and not falling under the lexeme “make”. The next two strongest collocates share the same fate as “Made”, however with a bit lower frequency and more overall corpus appearances. At first glance, the MI-score results

table features a mix of nouns, verbs, and pronouns, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.7. The word “Get”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Get” compared to all other words they were collocated with (see Table 20).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
We	5.862127	7	124
You	5.761041	10	190
Thing	5.761041	6	114
How	5.626499	7	146
Out	5.437427	9	214
Can	5.381696	7	173
What	4.819144	6	219
Even	4.793031	6	223
Do	4.750235	7	268
Up	4.669119	6	243

Table 20: Collocational profile of the word "Get" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Get” was the word “We”, boasting an MI-score of approximately 5.86. This collocation occurred 7 times out of the total 124 appearances in the corpus, resulting in a relative frequency of roughly 67.197 and a probability of around 5.645 % of appearing as this specific collocation. The pairing of “Story” with “Line” emerged as the strongest and most prevalent collocation, attributed mostly to phrases such as “we get introduced”, “we get to know” etc. The MI-scores and frequencies of the other collocates are also generally low, which indicates that even though “Get” was used a total of 233 times, its use was quite varied in terms of what It was collocated with and it was not “stuck” with a handful of very strong collocates. At first glance, the MI-score results table features a mix of pronouns, adverbs, Verbs and a noun, though semantic nuances

may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.8. The word “Time”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Time” compared to all other words they were collocated with (see Table 21).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Travel	8.949823	7	8
Spend	7.505037	9	28
Die	7.142467	9	36
Screen	6.557505	7	42
Run	6.335112	5	35
Much	6.294470	10	72
Long	6.216468	5	38
Same	6.142467	11	88
During	5.909807	5	47
No	5.894540	12	114

Table 21: Collocational profile of the word "Time" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Time” was the word “Travel”, boasting an MI-score of approximately 8.95. This collocation occurred 7 times out of the total 8 appearances in the corpus, resulting in a relative frequency of roughly 67.197 and a probability of around 87.5 % of appearing as this specific collocation. The pairing of “Time” with “travel” emerged as the strongest and most prevalent collocation, attributed most likely to a select few reviews regarding a sci-fi film *The Adam Project,* which features time travel elements, as one of the main plot points of the film is the main character traveling to the past and meeting his younger self. At first glance, the MI-score results table features a mix of nouns, verbs, adjectives and a conjunction and a determiner, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.9. The word “Way”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Way” compared to all other words they were collocated with (see Table 22).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Along	7.615387	9	31
Long	6.473660	5	38
Find	5.897159	9	102
Give	5.688164	10	131
Them	5.278643	10	174
Out	5.243154	12	214
Try	5.198025	5	92
Would	5.063375	5	101
Go	4.990268	8	170
Through	4.978195	7	150

Table 22: Collocational profile of the word "Way" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Way” was the word “Along”, boasting an MI-score of approximately 6.22. This collocation occurred 9 times out of the total 31 appearances in the corpus, resulting in a relative frequency of roughly 86.397 and a probability of around 29.032 % of appearing as this specific collocation. The pairing of “Way” with “Along” emerged as the strongest and most prevalent collocation, attributed solely to the reviewers’ usage of the phrase “along the way” when describing either the progression of the film’s plot or details about the film’s development. The next strongest collocate, the word “Long”, was also used in similar situations although with a bigger focus on the film making process with phrases such as “it would go a long way if…”, mostly pointing out the shortcomings of the films. At first glance, the MI-score results table features a mix of conjunctions, verbs, a pronoun, an adjective and an adverb, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

9.3.10. The word “Feel”

These represent the top ten collocates, determined by the highest MI-score and relative frequency, making them the strongest and most frequent collocates of the word “Feel” compared to all other words they were collocated with (see Table 23).

Collocate	*MI-score*	*Freq (coll.)*	*Freq (corpus)*
Less	7.945227	5	33
Like	6.534668	22	386
Can	6.402990	9	173
You	5.905193	7	190
They	5.458240	8	296
Even	5.188721	5	223
Make	5.180427	7	314
Not	4.917265	6	323
Movie	4.477869	6	438
That	3.990922	13	1330

Table 23: Collocational profile of the word "Feel" (Reference corpus)

The MI-score results indicate that the strongest collocate for the word “Feel” was the word “Less”, boasting an MI-score of approximately 7.95. This collocation occurred 5 times out of the total 33 appearances in the corpus, resulting in a relative frequency of roughly 47.998 and a probability of around 15.151 % of appearing as this specific collocation. The pairing of “Feel” with “Less” emerged as the strongest and most prevalent collocation, attributed to both positive and negative reactions to some of the film making choices present in the films. These reactions included phrases such as “it made action feel even less consequential”, “finale won’t fee l any less satisfactory” and more. At first glance, the MI-score results table features a mix of adverbs, pronouns, verbs, a noun and a determiner, though semantic nuances may be challenging to discern in some instances due to the absence of contextual information in the table.

10. Discussion of Results

When it comes to the results, the first difference can be already seen from the ten most frequent words both groups used. While this is to be expected, as it is quite unlikely from two completely distinct groups of people with diverse backgrounds to use identical vocabulary describing a certain topic, it is however important to note that in addition to simply reviewing the film, the students were also to compare it to its book version. This fact explains the presence of the words "book" and "novel" in the target corpus with such high frequencies and the lack of these words in the reference corpus. Contrary to this, there are also some similarities, specifically the words "Have", "Film", "Movie", "Do", "Make" and "Character". These words however often appear with significantly lower frequencies (for absolute frequency see Figure 14, for relative frequency see Figure 15) in the reference corpus then in the target corpus, which is quite interesting as the reference corpus was in fact a bit larger than the target one, by approximately five thousand words. This could be the result of a few possibilities. One possibility could be that the professional reviewers simply just use more varied sentences that do not often repeat words and try to convey their thoughts in different ways, while the students tend to use repetitive sentence structures that they familiar with and are easier to use. Another possibility might be that even though the reference corpus is larger and comprises one hundred reviews, they are individually not as long as the ones in the target corpus. This means that since the reviews are shorter (and from a wider range of authors), there may not be as much space for repetition as in the ones written by the students, which are often multiple pages long. There may be other possibilities and factors at play, this thesis however does not explore these possibilities further.

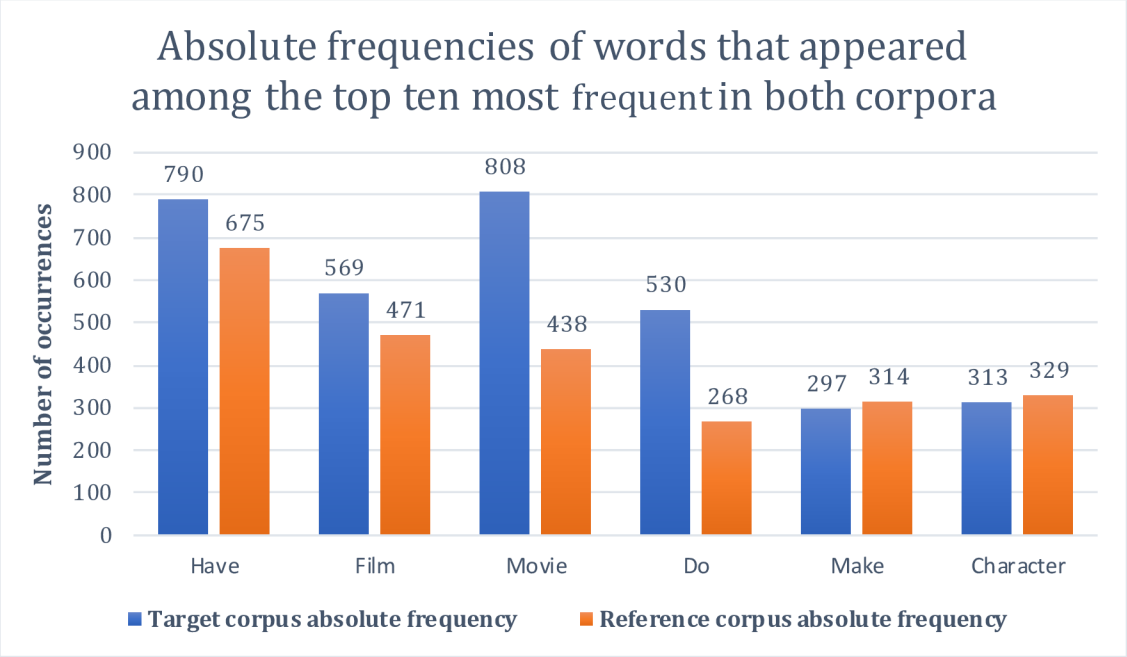

Figure 14: Absolute frequencies of words that appeared among the top ten most frequent in both corpora

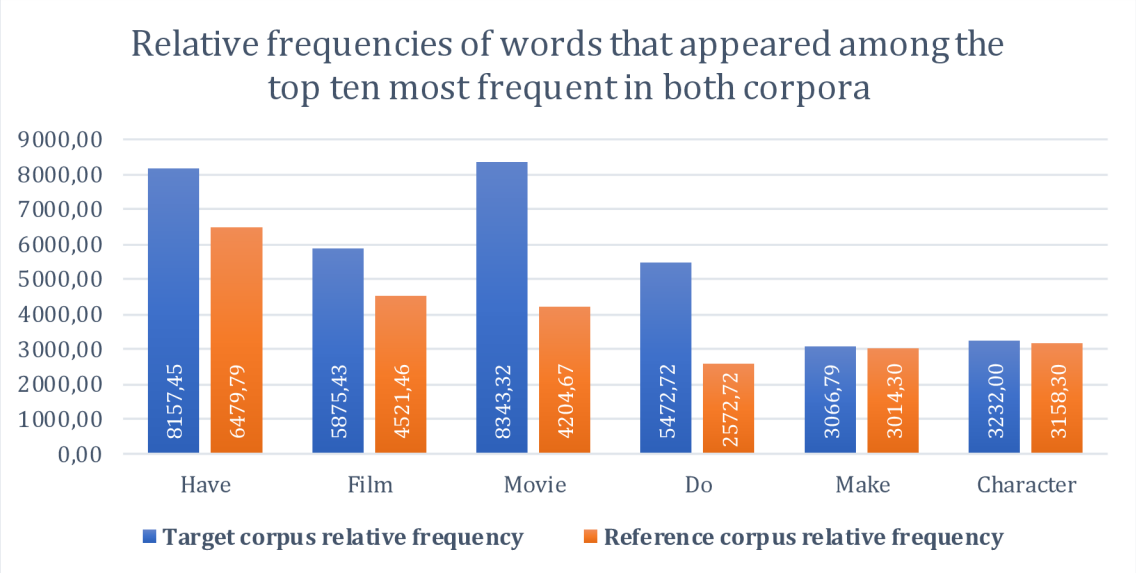

Figure 15: Relative frequencies of words that appeared among the top ten most frequent in both corpora

Sticking to these words, we can also examine their collocates to see if there are any major differences between both corpora. Starting with the reference corpus and the word "Have", the first notable feature that can be seen is the usage of "don't" and "doesn't" which, although incorrectly assessed as separate lexemes by #LancsBox, shows the tendencies of the reviewers to use these forms instead of the more formal "do not" or "does not" the students are taught to use in their papers for them to appear more "academic". While film reviews are hardly a perfect example of

academic writing, it is still interesting that the words "don't" and "doesn't" do not appear among the frequent collocates, nor even the lexeme "do" itself. Another interesting difference is in the use of modal verbs as collocates. While the professional reviewers preferred the use of "could", "would" and "may", the students preferred the use of "must" and "should, with both groups sharing only the use of "might" (see Figure 16). Not only was there a difference in the verbs themselves, but there was also a difference in their frequencies, as even though both groups' usage of modal verbs resulted in similar MI-scores, the frequencies of these verbs (both as specific collocates and their total count in the corpus) in the reference corpus was in general much higher, attributed most likely to the use of epistemic modality. This can be interpreted as the students, while perfectly able to use modal verbs, preferring to use them only when necessary to avoid longer and more complicated verb phrases and in turn avoiding longer and more complicated sentences (either due to not being confident enough to use them or perhaps to avoid unnecessary mistakes). This is however not a problem for the native speakers.

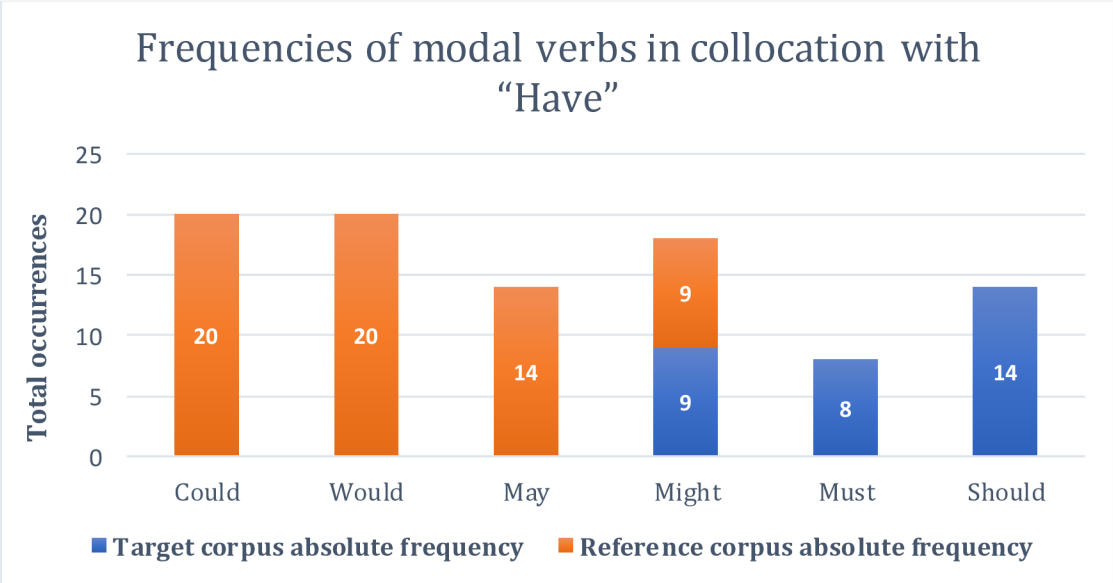

Figure 16: Frequencies of modal verbs in collocation with "Have"

When it comes to the word "Film", the only collocate that was shared by both groups was the word "throughout" with nearly the same frequency (5 for the target corpus, 6 for the reference corpus). As for the rest of the collocates there is quite a striking difference. While the students used collocates like "Maker", "Produce" and "Shoot" (with generally noticeably higher MI-scores, thus having stronger collocations),

focusing more on the film making process and its specific details, the reviewers focused more on the general descriptions of the film with collocates like "Entire", "Begin" or "Final". The focus on the specificities of film making is something that would be more often than not expected from professional in-depth film reviews or film school students rather than from future English teachers. This difference in descriptions could have had many reasons. It could have either been a mandatory part of the students' assignment to also include a more in-depth description of the films, or perhaps it could have been due to the length of the student reviews, as most of the reviews were approximately five pages longs, the students used these in-depth description to "fill in" the space. One more reason that also comes to mind when talking about the length of reviews is the fact that the professional reviews were almost rather on the shorter side, being a maximum of one or one and a half pages long. So perhaps if the professional reviewers were made to write longer reviews, they would be more likely to write more in-depth descriptions.

The word "Movie" was one of the two words that greatly differed in frequencies among both corpora. Along with the word "Do", these words had almost double the frequencies in the target corpus compared to the reference one. This may be yet another indication of the students' proneness to repetition of "safe" or important words, rather than referring to them through other means. Another interesting feature is that similarly to the word "Film", it seems that the students yet again focused more on the technicalities of the films, discussing its runtime as well as some broader topics like the films' finales, introductions and storylines. Contrary to that, the professional reviewers focused more on the film industry in general, having "Disney" and "Star" as the strongest collocates. Interestingly though, the word "Much" appears in the reference corpus' list of top ten most frequent collocates of the word "Movie" twice, once as an adjective (with MI-score of approx. 5.15) and once as an adverb (with MI-score of approx. 4.9). This may indicate that the professional reviewers are not afraid of using some words as various parts of speech based on context, as it is something they are used to from everyday life, while the students may stick to just one "version" of a word and use other words instead where a part of speech shift would be necessary. This may be a result of them just simply wanting to avoid unnecessary mistakes or them generally not knowing the word could be used in this way.

The word “Do”, as previously mentioned, was the second of two words that experienced a significant difference in frequencies between both corpora. However, the differences do not end there, as the word’s collocates also greatly differ. As for the reference corpus, the top three strongest collocates, being “Made”, “Devil” and “Me”, all had staggeringly high MI-scores of approximately 9.9, 9.7 and 8.3 respectively. This most certainly caused by the film *The Conjuring: The Devil Made Me Do It* being reviewed, as mentioned in the “Do” word’s specific collocation profile chapter. As for the remaining collocates, there are also slight differences between the corpora. At first glance one can spot the difference in pronoun usage, where the professional reviewers can be seen using “Me” (although highly contextual in this case), “We”, “I” and “They” more frequently, while the students mostly stuck to “We” as their most frequent choice of a pronoun (see Figure 17). Yet again, we can see a difference in variety, although it might be explained by the students wanting their reviews to appear more academic.

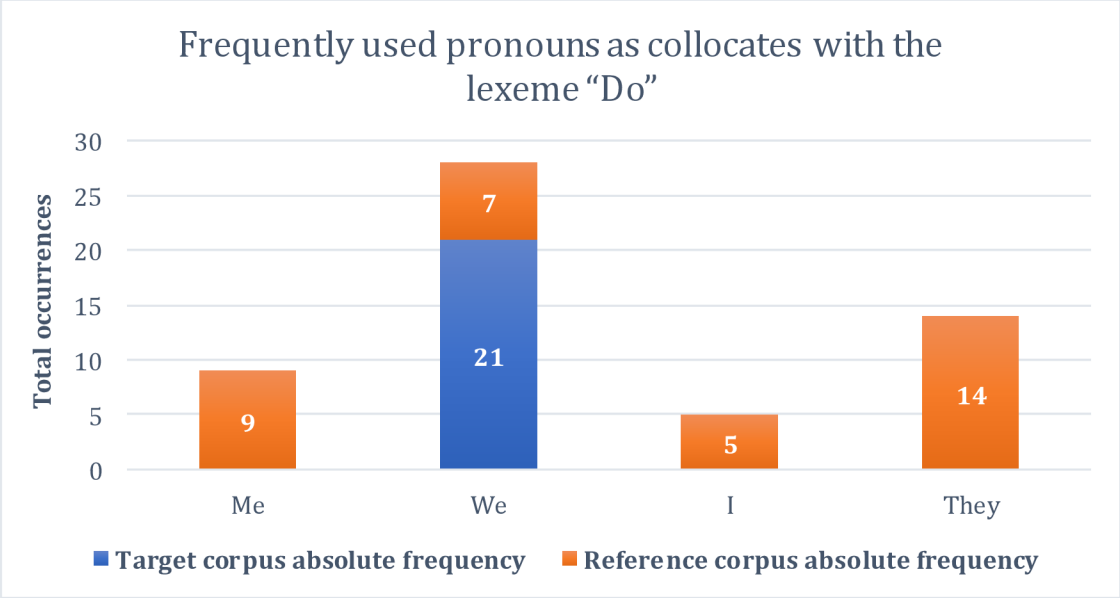

Figure 17: Frequently used pronouns as collocates with the lexeme "Do"

When it comes to the words “Make”, the first striking feature that can be seen in its collocational profiles is the use “More” as both an adjective (with MI-score of 5.3) and an adverb (with MI-score of 5.3), however the difference being that it now appeared in the target corpus. This goes against the previous theory that the students perhaps don’t feel as confident to use certain words as different parts of speech based on context. Another difference between the corpora is the most

frequent parts of speech among the collocates. Where the students collocated "Make" most frequently with adjectives, such as "Easy" or "Good", the professional reviewers mostly used other verbs, such as "Help" or "Want", as collocates. It is a possibility that using other verbs as collocates to the word "Make" feels more natural to native speakers, which is something non-native speakers simply do not feel, although an exact answer to this would probably require a more in-depth analysis of a larger data sample. Lastly, the word "Make" is one of the words that appear more frequently in the reference corpus than in the target one, although not by much.

Finally, the word "Character", the second of the two words that appeared more frequently in the reference corpus than in the target one. Among its frequent collocates, three were shared between both corpora, which being "Development", "Main" and "Play". While "Development" and "Play" were used roughly with the same frequency, "Main" was used twice as often by the students than the professional reviewers, specifically eighteen times in this specific collocation (and one hundred and twenty-nine times in total) in the target corpus compared to nine times in this specific collocation (and sixteen times in total) in the reference corpus. This may be yet another example of the students' proneness to repetition, since it appeared that much frequently in the target corpus, although perhaps a larger sample and a more in-depth analysis would be required to say for certain.

IV. CONCLUSION

The aim of this thesis was to show and describe the potential differences and similarities in written English between that of English students on PF JČU and native speakers. This was done through corpus analysis of texts written by both groups on a similar topic, which being film reviews. To gather the required data sample to analyse and compare, a total of one hundred professional film reviews and forty-five student film reviews were obtained through various means and imported into #LancsBox. Two corpora were then created, a target and a reference corpus, and used in the creation of collocational profiles for the most frequently used lexemes in both corpora. These collocation profiles were subsequently used to compare these corpora and illustrate any potential differences and similarities.

Initially, prior to the creation of collocational profiles, the “Words” function of #LancsBox was used with specific filters applied to create frequency lists of the most frequent lexemes in both corpora. Subsequently, ten lexemes with the highest frequencies in both corpora were selected, and had collocational profiles created for each, utilizing the "GraphColl" function of #LancsBox. The results of the “GraphColl” function were generated using MI-score as the chosen statistic.

The collocational profiles showed not only differences in overall lexeme usage but also difference in frequencies among the lexemes shared by both corpora. As for the shared lexemes, which included the lexemes “Have”, “Film”, “Movie”, “Do”, “Make” and “Character”, their collocational profiles were selected and compared with their counterparts in the other corpus a provide a more in-depth look at the differences and similarities of their usage.

The results of the comparison showed that the students, although quite adept at writing English texts, still struggle with repetition and overuse of certain words or phrases, perhaps in attempt to avoid possible mistakes. Another difference could be seen in the use of modal verbs as collocates, which the professional reviewers used almost twice as often compared to the students. Lastly, there were also differences in the usage of pronouns and verbs as collocates, as in some cases the professional

reviewers preferred the use of pronouns as collocates significantly more frequently than the students and in other cases preferred the use of verbs as collocates compared to the students' preference for adjectives.

Overall, the research showed that the students' written texts, while in some select cases similar to the ones of the native speakers, possibly still suffer from several factors keeping them from appearing "native-like", with the most prominent factor being repetition. There however may be even more factors not touched upon here, which could possibly be explored in a larger, more in-depth research.

V. RESUMÉ

Tato diplomová práce pojednává o problematice přirozenosti jazyka studujících angličtiny v psaných textech, konkrétně se tedy jedná o studující angličtiny na katedře anglistiky PF JČU, a do jaké míry se jejich písemné projevy podobají těm od rodilých mluvčí. V práci bylo toto docíleno korpusovou analýzou textů obou skupiny a následným porovnáním. Práce se mimo vytyčení podobností a odlišností mezi analyzovanými texty věnuje také možným důvodům vzniku odlišností či nastínění častých problémových oblastí pro studenty.

Teoretická část se zprvu zabývá nastíněním klíčových pojmů, jako jsou „nativelike selection" a „idiom principle". V rámci „nativelike selection" je zde popsaný i tzv. koncept „puzzle of nativelike selection", který pojednává o problémech, se kterými se studující angličtiny potkávají při volbě a rozpoznávání přirozeně znějícího jazyka od toho nepřirozeného. Dále mimo popisu idiomů a srovnání „idiom principle" a „open-choice principle" se tato část také věnuje oblasti korpusové a textové lingvistiky se zaměřením na koncepty „keyness", neboli vlastnost slova či fráze být klíčovým slovem v daném kontextu, a „aboutness", neboli vlastnost věty či textu sdělit jeho hlavní pointu. Část také zahrnuje obecný pohled na kolokace, jejich typy, kolokability a frazémy. V závěru pak teoretická část identifikuje časté problémy, s nimiž se nerodilí mluvčí často potýkají při psaní anglických textů. Ty problémy zahrnují například chybný slovosled, doslovné překlady vět a slov, či gramatické chyby nebo vynechávání členů.

Praktická část se věnuje korpusové analýze esejí studentů anglistiky na PF JČU na téma filmové recenze a porovnává je s autentických filmovými recenze z internetu. Studentské eseje pro analýz poskytnul PhDr. Christopherem Koyem, M.A., Ph.D., jak ve fyzické, tak v elektronické podobě a jsou použity pro sestavení cílového korpusu. K porovnání je použit referenční korpus utvořený z autentických filmových recenzí z internetu, které jsou získány z celkem osmi webových stránek různé popularity a pocházejí od široké škály profesionálních recenzentů a autorů. Nachází se zde také krátký popis jednotlivých webových stránek i stručný obecný popis studentských esejí včetně pro recenze zvolených filmů. Dále se je stručně popsán i program

#LancsBox včetně funkcí a parametrů použitých pro prvotní analýzu a pro porovnávání, tj. funkce „Words“ a „GraphColl“ a meřítko „MI-score“.

V rámci analýzy je zvoleno deset nejčastěji používaných lexémů z obou korpusů a každému lexému je právě pomocí #LancsBox vygenerován kolokační profil, ze kterého je zřejmě, jaké jsou nejsilnější kolokace každého z lexémů. Ke každému lexému je tedy vytvořena vlastní tabulka, která tato zjištěná data zobrazuje, tedy MI-score, udávající sílu kolokace, frekvenci výskytu právě v této kolokaci a celkovou frekvenci v korpusu. Ke každé tabulce je přítomen také krátký popis, ve kterém jsou výskyty jednotlivých slov popsány včetně možných odůvodněný.

V závěru praktické části jsou získané výsledky znázorněny, porovnávány a diskutovány na nejčastějších lexémech, které jsou sdíleny oběma korpusy. Z analýzy vzešlo hned několik oblastí, ve kterých se recenze studentů liší od recenzí profesionálních recenzentů (a rodilých mluvčí), ze kterých se asi nejčastěji projevovalo opakování zaběhlých slovních spojení a frází a obecné nadměrné používání určitých slov. Mezi další zjištěné odlišnosti patří například rozdíly v preferovaných modálních slovesech či v používání zájmen. Tato zjištění jsou zde znázorněna v několika grafech, aby byla přehlednější a bylo možné snadněji interpretovat získané informace. V neposlední řadě jsou ke každému zjištění také nabídnuty možné příčiny vzniku a odůvodnění. Závěrem práce je poté poukázáno na fakt, že ač studenti jazyk v mnohých případech velmi dobře ovládají, pořád se najdou některé specifické oblasti, na kterých je potřeba, ač samostudiem či univerzitní výukou, zapracovat, aby se písemný projev studentů blíže přibližoval písemnému projevu rodilých mluvčí.

List of tables and figures

VI. REFERENCES

Primary Sources

BARFIELD, A., & GYLLSTAD, H. Researching Collocations in Another Language: Multiple Interpretations. Palgrave Macmillan, 2009. ISBN 9780230245327.

BECK, David & MEL'ČUK, Igor. Morphological phrasemes and Totonacan verbal morphology. Linguistics, 2011, roč. 49. DOI: 10.1515/ling.2011.005.

BENSON, M., BENSON, E., ILSON, R. (Eds.). The BBI Combinatory Dictionary of English: A Guide to Word Combinations. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1993. ISBN 9780915027804.

CHOMSKY, N. Syntactic structures. Mouton, 1957. ISBN 978-1614278047.

CRYSTAL, David. A Dictionary of Linguistics and Phonetics. 3. B. Blackwell, 1991. ISBN 9780631178712.

ČERMÁK, František a Michal ŠULC. Kolokace. 2. Praha: Nakladatelství Lidové noviny, 2006. ISBN 80-7106-863-2.

ČERMÁK, František. Collocations, Collocability and Dictionary. Turín: Edizioni dell'Orso, 2006. ISBN 88-7694-918-6.

ČERMÁK, František. Frazeologie a idiomatika - česká a obecná. 1. Praha: Karolinum, 2007. ISBN 978-80-246-1371-0.

ERMAN, B., & WARREN, B. The idiom principle and the open choice principle, 2000. ISSN 0165-4888.

GABRIELATOS, Costas. Keyness analysis: Nature, metrics, and techniques. 2018. ISBN 9781315179346.

GEERAERTS, Dirk. Theories of Lexical Semantics. New York: Oxford University Press, 2010. ISBN 978-0-19-870030-2.

GRANT, L. and NATION, I.S.P. How many idioms are there in English? In ITL

International Journal of Applied Linguistics. 151. 1-14. DOI: 10.2143/ITL.151.0.2015219, 2006.

LEVORATO, M. C., ROCH, M., & NESI, B. A longitudinal study of idiom and text comprehension. In Journal of Child Language. ISSN 0305-0009, 2007.

LIU, Haitao. Dependency Distance as a Metric of Language Comprehension Difficulty. Journal of Cognitive Science. 9. 159-191. DOI: 10.17791/jcs.2008.9.2.159, 2008.

McENERY, T. & HARDIE, A. What is corpus linguistics? In Corpus Linguistics: Method, Theory and Practice (Cambridge Textbooks in Linguistics, pp. 1-24). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511981395.002, 2011.

MEL'ČUK, Igor. Phraseology in the language, in the dictionary, and in the computer. In Yearbook of Phraseology. 3. DOI: 10.1515/phras-2012-0003, 2012.

NESSELHAUF, N. Collocations in a Learner Corpus. Amsterdam, Philadelphia: John Benjamins Publishing Company, 2005. ISBN 9789027222855.

PAWLEY, Andrew & SYDER, F. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Language and Communication. ISBN 9781315836027, 1983.

PHILLIPS M. Lexical structure of text. English Language Research. ISBN 9780704410435, 1989.

POSLUŠNÁ, Lucie. Nejčastější chyby v angličtině a jak se jich zbavit. Brno: Computer Press, 2009. ISBN 978-80-251-2427-7.

SCOTT, M. & THOMPSON, G. Patterns of Text: in honour of Michael Hoey, Amsterdam: Benjamins. ISBN 90-272-2572-9, 2001.

SCOTT, Mike & TRIBBLE, Christopher. Textual Patterns: Key Words and Corpus Analysis in Language Education. ISBN 9789027293633, 2006.

SINCLAIR, John a Les SINCLAIR. Corpus, Concordance, Collocation. 2. Kalifornská univerzita: Oxford University Press, 1991. ISBN 9780820473451.

Internet Sources

What are collocations? [online]. Future Learn, Macquarie University, 2021 [cit. 2023-10-18].
Dostupné z: https://www.futurelearn.com/info/courses/improve-ielts-speaking/0/steps/98854

McARTHUR, T. (1992). *The Oxford Companion to the English Language*. [online] Oxford University Press. [cit. 2023-10-18].
Dostupné z: https://www.cambridge.org/core/journals/language-in-society/article/abs/tom-mcarthur-ed-the-oxford-companion-to-the-english-language-oxford-new-york-oxford-university-press-1992-pp-xxix-1184/4D4806A81E2909AD4CCF2BBC36561966

What is a corpus, what is corpus linguistics? [online]. University of Helsinki, 2016 [cit. 2023-10-18].
Dostupné z: https://www.futurelearn.com/info/courses/improve-ielts-speaking/0/steps/98854

Introduction to Text Linguistics [online]. King Saud University, 2014 [cit. 2023-10-18] Dostupné z: https://fac.ksu.edu.sa/sites/default/files/booklet_part_one-an_introduction.doc

Text linguistics [online]. Wikipedia, The Free Encyclopedia, 2023 [cit. 2023-10-18].
Dostupné z: https://en.wikipedia.org/wiki/Text_linguistics

Keyword (linguistics) [online]. Wikipedia, The Free Encyclopedia, 2023 [cit. 2023-10-18].
Dostupné z: https://en.wikipedia.org/wiki/Keyword_(linguistics)

GOMEZ, Cristobal. *Collecting collocations*, Kaplan International Languages [online]. Kaplan International Languages, 2021 [cit. 2023-10-18]. Dostupné z: https://www.kaplaninternational.com/blog/learning-languages/eng/collecting-collocations-speak-like-a-native

WEI, Yong. *Teaching Collocations for Productive Vocabulary Development*. [online] ERIC - Institute of Education Sciences, 1999. [cit. 2023-10-18]. Dostupné z: https://eric.ed.gov/?id=ED457690

Phraseology [online]. Wikipedia, The Free Encyclopedia, 2023 [cit. 2023-10-18].
Dostupné z: https://en.wikipedia.org/wiki/Phraseology

Phraseme [online]. Wikipedia, The Free Encyclopedia, 2023 [cit. 2023-10-18].
Dostupné z: https://en.wikipedia.org/wiki/Phraseme

The Mayfield Handbook of Technical & Scientific Writing [online]. Massachusetts Institute of Technology, 1997 [cit. 2023-10-18]. Dostupné z: https://web.mit.edu/course/21/21.guide/esl-link.htm

MEHROTRA, A. The 10 Most Common Errors Non-English Speakers Make in Their

Academic Writing. Academic Language Experts, 2023 [online]. Dostupné z: https://www.aclang.com/blog/the-10-most-common-errors/

Cvrček, Václav - Richterová, Olga (eds). *Český národní korpus* [online]. Příručka ČNK; 2019 Apr 9, 13:10 GMT [Citováno 2023 Nov 10]. Dostupné na: http://wiki.korpus.cz/doku.php?id=pojmy:asociacni_miry&rev=1554815423.

Other Sources

SVOBODA, Martin (2021). Collocability of the most frequent nouns in COVID-19 forum threads. Bakalářská práce. Katedra anglistiky PF JČU, České Budějovice.

MALÁ, Markéta & BRŮHOVÁ, Gabriela & VAŠKŮ, Kateřina (2022). Reporting Verbs in L1 and L2 English Novice Academic Writing. ELOPE: English Language Overseas Perspectives and Enquiries. 19. 127-147. 10.4312/elope.19.2.127-147.