

Unveiling Lexical Characteristics in Diverse Movie Genres through a Corpus-based Study

KINGKAN KOR NKAMONCHANOK
 NUTTAYA LAMCHARORNPONG
 PHAWITA LAMKA
 PIYAPONG LAOSRIRATTANACHAI*

Faculty of Hospitality Industry, Kasetsart University, Thailand

*Corresponding author email: piyapong.l@ku.th

Article information	Abstract
Article history: Received: 5 Oct 2024 Last revised: 23 Jul 2026 Accepted: 25 May 2026 Available online: 5 Aug 2026	<i>This study investigates the vocabulary characteristics across eight movie genres, using a corpus of 2,318 movie scripts comprising 20,765,616 tokens. The research examines lexical profiling, lexical level, lexical variation, lexical density, and CEFR level of the vocabulary used in the movie corpus. The analysis reveals that romance movies have the highest percentage of General Service List (GSL) vocabulary at 87.62%, while action movies use the most Academic Word List (AWL) vocabulary at 1.31%. Animated movies exhibit the greatest proportion of Outside Word List (OWL) at 14.64%. Drama movies show the highest frequency of high-frequency words (95.51%), whereas animated movies utilise more medium- and low-frequency words, as well as off-list words. Horror movies demonstrate the greatest average lexical variation (16.88%), while animated movies demonstrate the highest lexical density, with content words accounting for 52.23% of the total vocabulary. In terms of CEFR levels, romance movies contain the most A1–A2 level vocabulary (94.09%), while action movies lead in B1-level words (1.68%). Animated movies display the highest percentage of B2, C1, and C2 level words. The study highlights the potential of movies as effective tools for vocabulary development, emphasising that genre selection should be aligned with specific language learning objectives to maximise vocabulary acquisition and enhance overall language proficiency.</i>
Keywords: Vocabulary Lexical characteristics Movie Corpus-based study Genre	

INTRODUCTION

Vocabulary acquisition plays a pivotal role in second language learning (Alqahtani, 2015). Wilkins (1972) asserts that while grammar may aid communication, insufficient vocabulary can result in communication failure. Therefore, mastering vocabulary not only enhances speaking and communication but is also essential for developing other language skills, such as listening, reading, and writing (Nation, 2011). Vocabulary learning extends beyond the classroom, with incidental learning through meaning-focused activities being widely recommended (Ellis, 1999; Webb & Nation, 2017). Modern learners now have access to a wide range of tools to acquire vocabulary based on their interests (Cook, 2016). For instance,

virtual games have been shown to be more effective for vocabulary learning among EFL learners than traditional methods (Alfadil, 2020), while social media also promotes vocabulary acquisition effectively for EFL and ESL learners (Nguyen, 2021). Regularly reading online news over a three-month period has been proven to increase learners' exposure to key academic terms (Dang & Long, 2024). Additionally, English teachers have found success in using songs as a tool for vocabulary instruction (Odo, 2021). Furthermore, watching discipline-specific movies or TV series supports the learning of technical vocabulary, benefiting learners in professional fields (Csomay & Petrović, 2012). According to Nation (2022), watching movies that incorporate specialised vocabulary on a weekly basis is sufficient for vocabulary acquisition.

Research supports the use of movies as an educational tool, demonstrating their ability to enhance reading, listening, speaking, and writing skills (Baratta & Jones, 2008; Ismaili, 2013; Rokni & Ataee, 2014; Yaseen & Shakir, 2015). Many learners prefer watching subtitled movies over reading texts (Gieve & Clark, 2005; Lindgren & Muñoz, 2013; Peters, 2018), as movies often provide visual context, making them easier to comprehend than written or audio materials (Chapple & Curtis, 2000; Rubin, 1994). Additionally, movies allow learners to observe characters' reactions, aiding in their understanding of conversations (Keene, 2006). This method fosters vocabulary development and motivates learners to further engage with the target language (Ismaili, 2013; Webb & Rodgers, 2009), while the engaging nature of movies stimulates learning (Chapple & Curtis, 2000).

However, the analysis of movies based on different genres remains inconclusive. Different genres present varying levels of vocabulary difficulty, making them suitable for learners at different Common European Framework of Reference for Languages (CEFR) levels (Webb & Rodgers, 2009). Selecting genres that align with the learner's knowledge is crucial for effective learning (Bailey, 2005). To address this gap, this study investigated the lexical characteristics of eight distinct movie genres through five specific objectives. First, we examined the presence of General Service List (GSL), Academic Word List (AWL), and Outside Word List (OWL) to distinguish between genres that rely on simpler vocabulary, which benefit beginners, and those containing academic or specialised language (Meebangsai et al., 2023). Second, we analysed the distribution of high-, medium-, and low-frequency words to identify which genres offer sufficient text coverage for foundational learners and which are better suited to expanding an advanced lexicon (Nation, 2022). Third, these word frequencies were mapped onto the CEFR levels, allowing teachers and students to select movies that directly align with standardised language benchmarks (Laosrirattanachai et al., 2025). Fourth, the study measured lexical variation to determine the range of diverse vocabulary across genres, helping to identify whether a genre reinforces learning through repetition or challenges learners with diverse word choices (Nonthasen et al., 2025). Finally, we evaluated content word density to pinpoint which genres provide more informative and descriptive contexts, on the basis that genres with a higher proportion of content words allow learners to explore a wider and richer range of vocabulary (Loesnopchaimongkhon et al., 2025). By integrating these five aspects, this study provides a comprehensive framework to help educators and self-directed learners select appropriate movie genres that maximise vocabulary acquisition.

THEORETICAL BACKGROUND AND RELATED STUDIES

Vocabulary acquisition is a vital element in English communication (Laufer & Nation, 1995; Lestari & Hardiyantti, 2020). Watching movies has shown to be an effective method for improving English skills (Chapple & Curtis, 2000; King, 2002), including vocabulary skills. Research indicates that movies significantly enhance listening and speaking abilities, as the integration of visuals, sound, subtitles, and narratives makes them more engaging and comprehensible than traditional learning materials (Liando et al., 2018). Furthermore, movies effectively expand vocabulary and improve vocabulary recognition (Ashcroft et al., 2018). However, non-native speakers with limited proficiency may find the language used in movies challenging (Sato, 2007). To maximise vocabulary acquisition from movies, it is crucial for learners to be aware of the vocabulary size presented in these movies, facilitating better understanding and retention (Webb & Rodgers, 2009). Consequently, the current study posits that different movie genres may offer varying levels of vocabulary difficulty, making them suitable for diverse learner groups. Methods for evaluating vocabulary levels will be comprehensively discussed in the subsequent sections.

Lexical profiling

Lexical profiling is a linguistic method that classifies words into specific target vocabulary groups using established word lists as reference lists (Laufer & Nation, 1995). This process aims to analyse and identify the characteristics of words used in diverse texts, such as television programmes, academic articles, movies, and newspapers, thereby enhancing vocabulary comprehension. Lexical profiling also supports educators and learners in achieving vocabulary learning goals and facilitates self-assessment of one's knowledge (Webb, 2021). The reference lists normally employed in lexical profiling include the General Service List (GSL), which contains the 2,000 most common words in English and constitutes a fundamental part of everyday conversation, accounting for approximately 80% of vocabulary found in written texts (Nation & Waring, 1997; West, 1953). Common examples from the GSL include words like, the, be, of, and, a, and have. The Academic Word List (AWL) comprises 570 word families frequently found in academic texts, contributing an additional 10% of written vocabulary (Coxhead, 2000). Mastering these words is beneficial for learners aiming to enhance their English vocabulary in the contexts of academic area (Coxhead & Byrd, 2007). Examples from the AWL include words such as estimate, export, analyses, and benefit. The Outside Word List (OWL) includes words outside both the GSL and AWL categories, incorporating proper nouns, technical jargon, and loanwords from other languages (Coxhead & Hirsh, 2007). Numerous studies have employed lexical profiling for vocabulary analysis. For example, it was used to examine the number of words necessary to understand academic speech (Dang & Webb, 2014). It was also employed to analyse research trends through keyword analysis (Laosrirattanachai & Laosrirattanachai, 2025). This research utilised lexical profiling to categorise vocabulary across various movie genres using the GSL, AWL, and OWL.

Lexical level

The lexical level entails analysing vocabulary and assessing the frequency ranking of words within a text, aiding learners in comprehending the complexity and accessibility of reading materials. This analysis is intricately linked to frequency categories, specifically high, medium, and low frequencies. Vocabulary levels are categorised into three distinct tiers: high frequency, which spans K1 to K3 (1,000 to 3,000 words); medium frequency, covering K4 to K9 (4,000 to 9,000 words); and low frequency, encompassing K10 to K25 (10,000 to 25,000 words) (Schmitt & Schmitt, 2014). Importantly, four types of nouns—proper nouns, marginal words, transparent compounds, and acronyms—are excluded from the top twenty-five levels (Nation, 2020). Moreover, the lexical level is closely associated with lexical coverage, defined as the percentage of words in a text that a reader or listener comprehends. A lexical coverage of 95% and 98% is generally regarded as sufficient for comprehension (Laufer, 2013; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011; van Zeeland & Schmitt, 2013). When analysing lexical level, the BNC-COCA family, acknowledged as the most comprehensive resource for both British and American English vocabulary lists, is often employed. The BNC/COCA vocabulary list, designed for learners of English as a second language, is derived from the British National Corpus and the Corpus of Contemporary American English. It comprises twenty-five 1,000-word levels, ranging from 1,000 to 25,000 words, organised according to vocabulary frequency and range information, with the last update occurring in 2017 (Nation, 2017). This list has proven to be more reliable and effective for vocabulary acquisition than alternative resources (Dang & Webb, 2016; Dang et al., 2022). Several studies have concentrated on examining vocabulary levels. Webb and Rodgers (2009) explored word frequencies across various movies to determine the vocabulary sizes needed for 95% and 98% coverage. The results showed that learners must know 3,000 and 7,000 words from the BNC-COCA family to achieve 95% and 98% lexical coverage when watching movies and TV shows, respectively. Similarly, Ha (2022) analysed the vocabulary levels necessary for 95% and 98% comprehension in online newspapers, using data from “News on the Web,” the largest English language archive (Davies, 2021). Additionally, Xodabande et al. (2022) examined the vocabulary required for understanding 95% and 98% of MOOCs lectures based on the BNC/COCA vocabulary list. In the present study, the vocabulary levels are specifically analysed across various movie genres by examining vocabulary frequency using the BNC/COCA vocabulary list (Nation, 2017).

The CEFR levels

The CEFR is a comprehensive framework for structuring language curricula, assessments, and educational resources across Europe. It outlines the necessary knowledge and skills language learners need for effective communication, considering the cultural context in which the language is used. The CEFR also establishes proficiency levels that help assess learners’ progress in reading, writing, listening, and speaking (Council of Europe, 2001). The framework comprises six proficiency levels, from A1 to C2. Nation and Crabbe (1991) proposed a comparison between the CEFR levels and lexical frequency levels based on the BNC/COCA headword lists. Within this framework, the A1 level is aligned with Flemma + -ly, reflecting the most basic vocabulary, focused on frequently used word forms, including simple derivatives such as adverbs ending in “-ly.” For the A2 level, learners are expected to reach a partial mastery of Level 3 or K3 (the 3,000 most frequent word families), indicating the development of

a foundational vocabulary that is not yet fully functional. At the B1 level, learners should fully command a Level 3 word family size (the 3,000 most frequent words), signifying the ability to deal with familiar matters and express themselves in straightforward terms. The B2 level is associated with a Level 4 or K4 word family size (the 4,000 most frequent words), reflecting a competent grasp of language that allows for the understanding and production of clear, detailed text on a wide range of subjects. For the C1 level, learners are expected to attain Level 5 or K5 (the 5,000 most frequent words), indicating a strong command of the language and the ability to express complex ideas effectively. Finally, at the C2 level, which represents the highest degree of fluency and language mastery, learners are anticipated to have achieved Level 6 or K6 (the 6,000 most frequent words) and beyond. This suggests a highly advanced and extensive vocabulary. These word family sizes help to classify vocabulary levels across CEFR stages. Research by Hu (2021) examines features that differentiate CEFR levels in spoken performance, while Owen et al. (2021) explore how lexical variation can classify essays based on CEFR abilities. In our study, we analysed vocabulary complexity in various movie genres, aligning the results with CEFR levels. This comparison helps learners at different proficiency levels select appropriate movies to enhance their vocabulary skills based on their CEFR classification.

Lexical variation

Lexical variation, or lexical diversity, refers to the extent to which a text displays a rich and diverse vocabulary. A high level of lexical variation signifies that the speaker or writer employs a broad range of different words, minimising repetition and enhancing the depth of the discussion (Johansson, 2008). The objective of lexical variation is to assess the diversity of words in a text, thus enriching the language, and it is also referred to as lexical richness (Daller et al., 2003). The Type Token Ratio (TTR) serves as the formula for measuring lexical variation, calculated by dividing the number of unique words by the total number of words and multiplying by 100 (Bates et al., 1998; Lieven, 1978). The TTR is a widely accepted metric for evaluating word variety (Davis & Brewer, 1997; Wolf-Quintero et al., 1998), with a higher TTR indicating greater linguistic diversity. In terms of pedagogical interpretation, a high TTR indicates that a text contains a vast array of unique words with minimal repetition, thereby offering learners a greater opportunity to encounter and acquire new vocabulary. Conversely, a low TTR represents a highly repetitive text with a limited vocabulary range. These varying percentages are suited to different proficiency levels; lower-proficiency learners, who possess a smaller baseline lexicon, benefit from low-TTR texts as they need to absorb language gradually without being overwhelmed. On the other hand, advanced learners already possess a strong vocabulary foundation, making them better equipped to process and learn from high-TTR texts that challenge them with novel words.

Numerous studies have applied the concept of lexical variation. For example, Johansson (2008) explores vocabulary diversity across narrative and expository styles, while Daller et al. (2003) utilise the TTR to assess vocabulary completeness in spoken texts from bilingual Turkish-German speakers. The current study employs the TTR to evaluate lexical diversity across various movie genres. This evaluation aims to provide guidance for L2 learners of varying proficiency levels in selecting movies that are suitable for enhancing their English vocabulary according to their respective levels.

Lexical density

Lexical density refers to the ratio of content words—nouns, verbs, adjectives, and adverbs—to the total number of words within a text (Johansson, 2008). A high lexical density indicates that the text is rich in content words, reflecting a more informative and precise communication style (Johnson, 2017). In contrast, a higher proportion of function words, such as prepositions and conjunctions, leads to lower lexical density, often characteristic of conversational or informal language (Ure, 1971). In terms of pedagogical interpretation, the percentage of lexical density signifies the potential informational load of a text; a higher percentage implies that a learner will encounter a denser concentration of meaning-carrying words, whereas a lower percentage indicates a text that relies more heavily on structural or grammatical framing. Substantial variations in density across texts suggest differences in structural complexity and communicative purpose, whereas narrow gaps indicate uniform lexical configurations. The distinction between written and spoken language is notable, as spoken language typically has a lexical density of less than 40%, while written language generally exceeds 40% (Laufer & Nation, 1995). This measure is influenced by the speaker's or writer's ability to select words accurately, with less experienced writers often relying on simpler sentence structures, resulting in lower lexical density (Schmitt & Schmitt, 2020). The formula for calculating lexical density is the ratio of content words to the total number of words. Research has employed various methods to assess vocabulary density (Laufer & Nation, 1995). For instance, Johansson (2008) found that written language usually exhibits higher density than spoken language. Additionally, Nasser and Thompson (2021) investigated vocabulary density variations within English academic writing. The current study aims to evaluate the ratio of content to function words across different movie genres to identify density variations, thereby determining whether different genres introduce distinct levels of informational load to language learners.

Research questions

1. What is the proportion of GSL, AWL, and OWL in each movie genre?
2. What are the proportions of high-, medium-, and low-frequency words across different movie genres?
3. What is the distribution of vocabulary according to the CEFR levels in each movie genre?
4. What are the differences in lexical variation among movie genres?
5. What is the content word density in each movie genre?

RESEARCH INSTRUMENT AND METHODS

Corpus compilation

To effectively analyse vocabulary across varying cinematic contexts, establishing a well-defined genre classification system is an essential prerequisite. Consequently, this study utilises OpusData—a prominent, licensed commercial data service that provides highly structured and standardised entertainment-industry intelligence, particularly regarding

global box office metrics and film industry metadata. According to the foundational framework derived from this platform, movies are systematically categorised into seven primary emotional genres: action, adventure, comedy, drama, thriller, romance, and horror (Antipov & Pokryshevskaya, 2016). Nevertheless, a substantial body of educational research emphasises that animation, whilst not explicitly classified as a distinct emotional genre within the OpusData service, plays a monumental role in English language learning and teaching. Animated films heavily stimulate language acquisition (Lestari & Harahap, 2020) and serve as an exceptionally effective pedagogical tool (King, 2002) by engaging students through multimodal storytelling that combines synchronised sound, vivid expressions, gestures, and visual cues (Herron & Hanley, 1992). Recognising this significant educational value, the present study intentionally integrated animation as the eighth movie genre, thereby establishing a more holistic and pedagogically relevant corpus for vocabulary analysis. The study analyses English-language movies from the inner circle, focusing on the top 200 movies from 2014 to 2023, based on box office earnings from Box Office Mojo website. However, the number of top 200 movies listed on the Box Office Mojo website fluctuates yearly, as some high-ranking movies are in foreign languages, not English. This inconsistency complicates the data collection process. Movie scripts were sourced from Opensubtitles.org, Opensubtitles.com, and Subscene. The details of the number of entries and the size of the movie corpus are shown in Table 1.

Table 1
Number of movies and movie corpus size

Year	Movie genres								Total
	Action	Adventure	Animation	Comedy	Drama	Horror	Thriller	Romance	
2014	32	34	13	45	39	21	33	15	232
2015	35	32	13	42	51	21	35	9	238
2016	33	40	16	22	41	14	31	8	205
2017	17	22	11	28	25	13	21	8	145
2018	41	41	13	50	53	23	34	15	270
2019	46	45	10	39	48	19	41	12	260
2020	36	32	11	37	43	15	33	6	213
2021	44	37	11	46	52	13	31	17	251
2022	41	39	9	47	52	13	34	19	254
2023	40	35	12	46	48	10	32	27	250
Total	365	357	119	402	452	162	325	136	2,318
Token	3,221,552	2,826,313	1,050,975	4,325,886	4,433,037	1,164,681	2,389,473	1,353,699	20,765,616

The total number of words and text sizes varied across the eight selected movie genres. This variance in corpus sizes was primarily attributed to a combination of data collection screening criteria and inherent genre characteristics. Regarding the screening process, although the initial selection targeted the top 200 highest-grossing movies from the Box Office Mojo website, this list fluctuates yearly, and several high-ranking films are in foreign languages rather than English. Consequently, foreign-language entries were excluded, which directly contributed to the variation in final corpus sizes across the movie genres. This systematic filtering, whilst creating inconsistencies in initial numbers, was essential to ensure linguistic homogeneity within the English-only corpus. Beyond these data collection constraints, textual

variation is significantly driven by unique genre conventions. Dialogue-heavy genres like drama and comedy often utilise more words to develop plot and character, whereas genres such as horror and thriller, which rely on atmosphere and suspense, may feature fewer words, focusing more on visual storytelling. Industry trends and audience demand also affect corpus sizes; popular genres like action and adventure are produced more frequently, resulting in larger corpora, while niche genres like animation and romance have fewer productions, leading to smaller corpora. Additionally, differences in movie length impacted token counts, with longer scripts common in character-driven genres like drama, whereas action and horror movies tended to have shorter screenplays. Finally, the lexical complexity of certain genres may contribute to varying token counts, as some genres employ simpler language, reducing the overall number of words used. It is worth noting that since an individual movie may be classified into more than one category, the same film may appear in multiple sub-corpora, resulting in the total number of entries exceeding 2,000.

Data analysis

The analysis of the data collected from the eight movie genres followed a structured, multi-dimensional computational approach to ensure academic rigour.

1. Lexical profiling: To address the first objective, the corpus was categorised into specific vocabulary groups using AntWordProfiler (Anthony, 2024). The software profiled the scripts against two established baseline reference lists: the GSL and the AWL. Any vocabulary items not found in either the GSL or the AWL were categorised as the OWL. In this profiling process, a high proportion of GSL words indicates a simpler vocabulary framework suitable for beginner learners. Conversely, a prominent presence of AWL terms signifies more challenging academic language, whilst a significant representation of OWL words suggests advanced or specialised vocabulary suited to learners with higher language proficiency.

2. Lexical levels: To investigate vocabulary frequency distributions, word frequency rankings were assessed from K1 (the most frequent 1,000 words) to K25, including unlisted words. This analysis was executed using the VocabProfile software (Cobb, 2025). To ensure statistical reliability and provide explicit details regarding the source of the baseline categories, the profiling within the software utilised the comprehensive BNC/COCA headword lists developed by Nation (2017). For descriptive clarity, the analysed vocabulary was subsequently grouped into three broader frequency tiers: high-frequency (K1–K3), medium-frequency (K4–K9), and low-frequency (K10–K25) categories (Schmitt & Schmitt, 2014). Genres rich in high-frequency words are considered suitable for beginners, whereas those displaying higher frequencies of mid- and low-frequency terms are ideal for advanced learners.

3. CEFR levels: To map the vocabulary complexity onto international standards, the text coverage percentages obtained from the BNC/COCA lexical level analysis were projected onto the CEFR. This mapping was executed via an analytical framework adapted from Nation and Crabbe (1991). Rather than employing a separate software application, the vocabulary coverage statistics from the frequency levels were mathematically evaluated against standardised CEFR benchmarks. Specifically, vocabulary falling within the K1–K2 range was mapped onto the A1–A2 levels, K3 corresponded to B1, K4 to B2, K5 to C1, and K6 up to K25 was categorised as

C2. This systematic alignment served to determine the lexical suitability of each movie genre for learners at various developmental stages.

4. Lexical variation: Vocabulary variation were evaluated using the TTR. This metric was computed by processing the corpus through AntWordProfiler (Anthony, 2024), which provided the precise number of unique words (types) and total words (tokens) for each genre-specific sub-corpus. The baseline TTR formula—calculating the ratio of types to tokens multiplied by 100—was then applied to determine the percentage of lexical variation. A higher TTR percentage indicates greater linguistic diversity and less word repetition, implying a more complex linguistic environment suitable for advanced learners.

5. Lexical density: To classify lexical items in the corpus as either content words or function words, the Function Word List (FWL) developed by Nation (2022) was employed as the baseline reference archive through a deductive profiling approach implemented within AntWordProfiler (Anthony, 2024), following the framework proposed by Laosrirattanachai et al. (2026). When the movie corpus was analysed against this list, the software generated the exact percentage of text coverage accounted for by function words. Consequently, the remaining proportion of the text not covered by the FWL was identified and quantified as content words (comprising nouns, verbs, adjectives, and adverbs). A higher lexical density indicates a denser concentration of informative, content-carrying vocabulary, which is highly beneficial for advanced English learners seeking a richer lexical context.

FINDINGS

The analysis findings for the movies are presented according to each aspect of the vocabulary as follows.

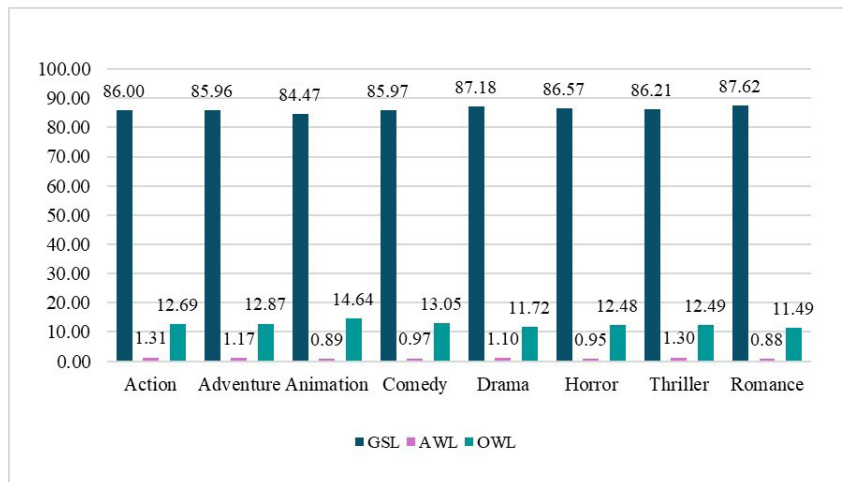


Figure 1 Lexical profiling of eight movie genres

The lexical profiling analysis reveals that vocabulary from the GSL constitutes over 80% of the language used in the examined movie genres. Romance movies lead with an average GSL

usage of 87.62%, closely followed by drama at 87.18%. Animated movies, however, have the lowest GSL average at 84.47%, while all other genres maintain averages above 85%. This indicates that romance and drama movies are particularly suitable for learners with basic English proficiency, as they can understand the content more readily. Conversely, the usage of academic vocabulary, derived from the AWL, is minimal across genres, with romance movies showing the lowest AWL usage at 0.88%, and action movies at the highest, with just 1.31%. The vocabulary from the OWL is more prevalent, encompassing proper nouns, specialised terms, and informal jargon. Animated movies exhibit the highest OWL usage at 14.64%, followed by comedy at 13.05%, while romance and drama feature lower averages at 11.49% and 11.72%, respectively. Consequently, for learners with a foundational knowledge of English, starting with romance and drama genres is advisable, as they offer a more accessible learning experience compared to animation. Example 1 illustrating GSL words in regular font, AWL words in bold, and OWL words underlined is provided for clarity.

Example 1: Excerpt from an action movie (Black Adam, 2022)

A: Why do you want Teth Adam to surrender?

B: He's the Champion of Kahndaq.

A: Who are you? The Justice Society? We have been living under **military occupation** for 27 years, and never seen you before.

B: You didn't come when intergang invaded our country, when they stole our **resources** and killed my husband.

Lexical level

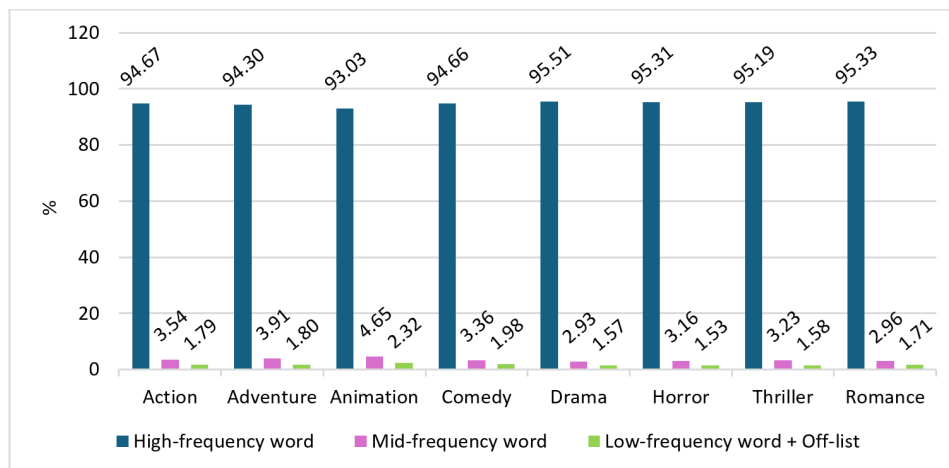


Figure 2 Lexical level of eight movie genres

The analysis presented in Figure 2 indicates that animation has the lowest average of high-frequency words among the genres, at 93.03%. In contrast, drama ranks highest at 95.51%, followed closely by romance at 95.33% and horror at 95.31%. These three genres exhibit a close frequency of high-use words, differing by less than 1%. For mid-frequency words, particularly those classified as more challenging (K4–K5), animation again leads with an average

of 4.65%, while drama and romance show lower averages at 2.93% and 2.96%, respectively. In terms of low-frequency words (K10–K25), including off-list terms, animation remains at the top with 2.32%, followed by comedy at 1.98%, and horror with the fewest at 1.53%, closely followed by drama at 1.57%.

The vocabulary coverage of 95% aligns with high-frequency words, while 98% coverage encompasses mid-frequency words. Genres such as drama, horror, thriller, and romance provide 95% coverage, making them accessible for learners with sufficient vocabulary to comprehend the movie content. In contrast, action, adventure, and comedy fall short of this coverage, suggesting they are more suitable for intermediate learners. For 98% coverage, drama, horror, thriller, and romance necessitate a vocabulary of approximately 8,000 words, while action, adventure, and comedy require around 9,000 words. Animation demands a significantly higher vocabulary of 11,000 words, making it more appropriate for advanced learners. Consequently, beginners are advised to focus on drama and romance to enhance their vocabulary skills effectively. Example 2 illustrates high-frequency words in regular font, mid-frequency words in bold, and low-frequency words underlined.

Example 2: Excerpt from a comedy movie (Abominable, 2019)

- A: You're the bad guys. Your gerbil is freaking me out.
 B: A boy that knows his gerbils. I like him.
 C: Bad? Listen, you've got this all wrong.
 A: That wild beast is unpredictable, even dangerous.
 C: We were trying to rescue your friend when the Yeti **kidnapped** her.
 A: He did **kidnap** her.
 C: Listen, you twit!

The CEFR levels

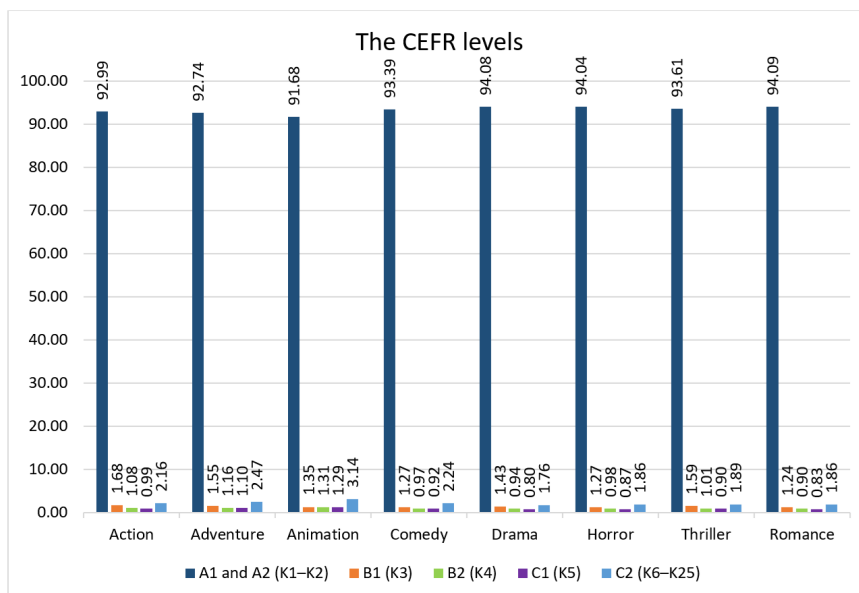


Figure 3 CEFR levels across eight movie genres

According to Figure 3, the highest values for A1 and A2 proficiency levels are observed in the genres of romance (94.09%), drama (94.08%), and horror (94.04%). These results suggest that these genres utilise simpler vocabulary, making them more accessible for beginner language learners. Conversely, the representation of B1 (intermediate) vocabulary shows a significant decline across all genres, with romance having the lowest at 1.24. This trend implies that while beginner learners may enjoy these movies, the language complexity could pose challenges as they progress to intermediate levels. For advanced learners, represented by C1 and C2 levels, animation and adventure feature the highest vocabulary complexity, particularly at C2, with 3.14% and 2.47%, respectively. This indicates a richer and more sophisticated use of language, suitable for advanced learners aiming to further expand their vocabulary. To support L2 learners in acquiring more vocabulary through movie, the selection of movies should be carefully aligned with learners' proficiency levels, ensuring that the vocabulary challenges are appropriate for their developmental stage. An example of CEFR vocabulary levels is provided, with A1 and A2 words in normal font, B1 words italicised, B2 words underlined, C1 words bolded, and C2 words both bolded and italicised.

Example 3: Excerpt from an adventure movie (Barbie, 2023)

A:to change the **Constitution** to a *government* for the Kens, of the Kens, and by the Kens.
B: You can't do this.
A: This is... Barbie.
B: If this got out that our dolls were coming to Los Angeles from Barbie Land as life-size versions of themselves... roaming the *Earth*...this *would* be very bad.
A: Barbie in the Real World? That's *impossible*.
B: We've got a definite situation on our hands.
A: **Catastrophic!** I can't stress that enough!

Lexical variation

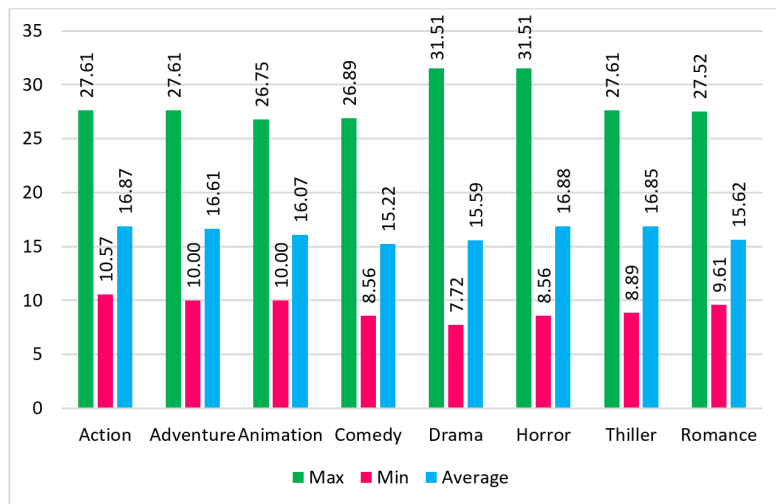


Figure 4 Lexical variation in eight movie genres

The genres with the highest lexical variation are horror (16.88%), action (16.87%), and thriller (16.85%), suggesting they expose learners to a diverse range of vocabulary and offer opportunities to practice language in various contexts. Consistently with the theoretical framework of TTR interpretation, these higher percentages mean that these genres feature a substantial number of unique words with low repetition, thereby maximising the opportunities for learners to discover and explore unfamiliar vocabulary. These genres are ideal for advanced learners who already possess a broad vocabulary base, allowing them to effectively integrate and process these new linguistic inputs. In contrast, comedy (15.22%), drama (15.59%), and romance (15.62%) have the lowest lexical variation. Comedy, in particular, shows frequent vocabulary repetition, which can benefit beginners by accommodating their limited baseline lexicons and reinforcing familiar words through repeated exposure. Thus, genres with higher vocabulary variation are more suited to advanced learners aiming for vocabulary development, while those with more repetitive language, such as comedy, drama, and romance, are better suited for building foundational language skills gradually. Example 4 illustrates the use of near-synonyms instead of repeated vocabulary, as seen in high lexical variation genres, reflects a higher TTR, indicative of diverse language use.

Example 4: Excerpted from an action movie (*Allegiant*, 2016)

- Pair 1: Do you **trust** him?
 But if you don't **believe** me, **believe** your mother.
- Pair 2: I agree. And in time, that won't be **necessary**.
 I know all of the **important** people from Chicago.
- Pair 3: And we know **absolutely** nothing about the people who sent it.
Certainly not mine.
- Pair 4: So you better start **talking** or we'll be holding your trial now.
 I'm not gonna **tell** Tris what to do.

Lexical density

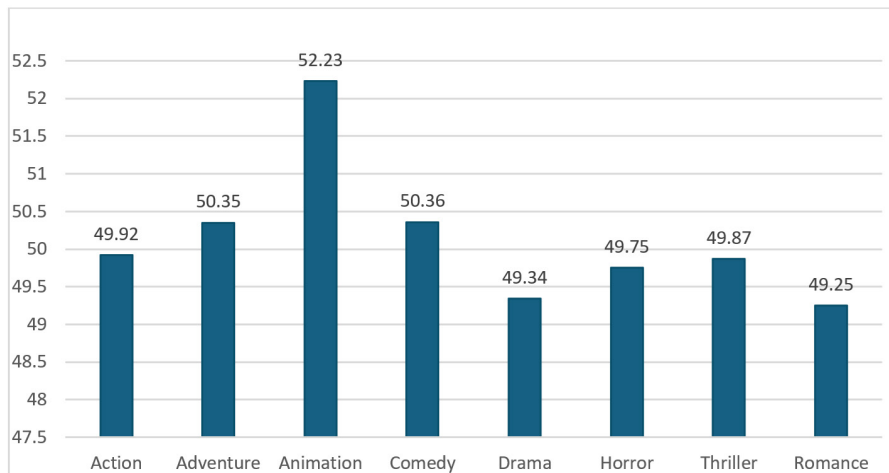


Figure 5 Lexical density in eight movie genres

According to Figure 5, the genres with the highest proportions of content words are animation (52.23%), comedy (50.36%), and adventure (50.35%), whilst romance (49.25%), drama (49.34%), and horror (49.75%) display slightly lower proportions. Crucially, the numerical values across all eight genres are remarkably consistent, remaining tightly clustered around a 50:50 ratio of content to function words, with a maximum variance of less than three per cent. This minimal variance suggests that there is no substantial or meaningful difference in lexical density across diverse movie genres. Consequently, regarding the overall concentration of information-carrying words, learners will encounter a highly stable distribution of content and function language regardless of the film genre they choose to watch. Below is an example highlighting content words in bold.

Example 5: Excerpted from an animation movie (*Despicable Me 3*, 2017)

A: Your **new leader** is **coming directly** from **head office**. **Effective immediately**. She is the **very gifted, very ambitious, Miss Valerie da Vinci**.

B: As I **look** out over all your **faces**, I am **flooded** with so many **memories**. **Oh, boy!** This is a **snooze fest!**

C: Excuse me, this is my last...

DISCUSSION

The lexical profiling analysis categorised words into the GSL, AWL, and OWL based on reference lists. Across various genres, GSL words averaged above 80%, with romance movies having the highest percentage (87.62%), followed by drama (87.18%). This suggests that GSL, which accounts for over 80% of written texts according to Nation and Waring (1997), covers an even higher proportion in movies due to their spoken nature and purpose of easy audience comprehension. The high GSL percentage in romance and drama makes these genres ideal for beginners, as the vocabulary is more accessible, facilitating vocabulary building. In contrast, AWL words appeared less frequently, with an average of just 1.31%, compared to Dang and Webb's (2014) finding of 4.1% in academic speech. This low AWL presence aligns with the entertainment-focused nature of movies (Lee, 2015), which are designed to engage rather than convey academic information. If the AWL percentage were higher, it would indicate greater difficulty and academic content. OWL vocabulary, including proper nouns, colloquial terms, and technical terms, featured more prominently, with an average of 14.64%. Animation showed the highest OWL usage, followed by comedy at 13.05%. Genres with lower OWL percentages, such as romance (11.49%) and drama (11.72%), are easier for learners to understand, while higher OWL percentages indicate more difficult content, making such movies more suitable for intermediate or advanced learners. As a result, romance, drama, and horror movies are recommended for beginners due to their high GSL and lower OWL percentages. Intermediate learners may benefit from action and adventure movies, which present more challenging vocabulary. Advanced learners can explore animation and comedy for more complex language and specialised vocabulary.

The analysis of lexical levels, divided into high-frequency (K1–K3), medium-frequency (K4–K5), and low-frequency (K10–K25 and off-list) words, reveals critical insights into the vocabulary demands of different movie genres. High-frequency words cover an average of 95% of the vocabulary in genres such as drama (95.51%), romance (95.33%), horror (95.31%), and thriller (95.19%). These genres are ideal for beginner learners due to their accessible language. However, animation, adventure, comedy, and action genres fall short of this 95% coverage, making them more challenging for learners relying on high-frequency vocabulary. Research suggests that 98% coverage is the ideal goal for understanding almost all content, with studies by Laufer and Ravenhorst-Kalovski (2010) and Schmitt et al. (2011) confirming this benchmark. Webb and Rodgers (2009) found that a vocabulary size of 3,000 words covers 95% of vocabulary in movies, while 6,000–7,000 words cover 98%. The current study builds on these findings, showing that genres like drama, romance, horror, and thriller require 3,000 words to reach 95% coverage, making them accessible for beginners. However, genres like adventure, comedy, and action demand 4,000 words, and animation requires 5,000 words, demonstrating higher complexity, particularly in modern animated movies. Medium-frequency words, associated with 98% coverage, are more prominent in animation (4.65%) and adventure (3.91%), requiring larger vocabulary sizes to fully understand these movies. Notably, animation requires 11,000 words for 98% coverage, 1,000 more than previous studies indicated, showing it is particularly demanding for learners, especially beginners. Low-frequency words, while lower in percentage, are also prevalent in animation (2.32%) and comedy (1.98%). These words not only include proper nouns and colloquial terms but also technical terms, which can make comprehension difficult for learners with limited vocabulary. Selecting appropriate genres based on learners' vocabulary levels is crucial for effective language learning, as too advanced content may hinder progress. Thus, drama and similar genres provide better opportunities for building foundational vocabulary, while animation and comedy are more suitable for learners seeking to expand their advanced lexicon.

The analysis of vocabulary across different genres in relation to CEFR levels highlights that romance, drama, and horror primarily cater to A1 and A2 learners, exhibiting vocabulary percentages exceeding 94%. This high percentage suggests that these genres provide accessible language suitable for beginners, allowing them to engage with content that is both entertaining and educational. Conversely, the vocabulary distribution for higher CEFR levels, particularly B2 and above, shows a significant decrease across all genres. Animation stands out as the most lexically rich genre at the C2 level, with a vocabulary percentage of 3.14%, indicating its potential benefits for advanced learners. In contrast, genres such as comedy, horror, and romance display relatively lower combined percentages for B1 and B2 levels, suggesting they may not adequately support intermediate learners. These findings have important implications for language teaching and learning. While romance and drama serve as excellent entry points for A1 and A2 learners, instructors should consider students' diverse interests when selecting content. Integrating genres like animation and adventure, which feature a higher vocabulary percentage for advanced learners, could enhance the learning experience for C1 and C2 students. Moreover, teachers must be aware of the vocabulary complexity inherent in various genres. While action and adventure movies may engage young learners, they might not sufficiently challenge advanced students. Consequently, it is advisable for teachers to curate a balanced curriculum that incorporates a variety of genres,

ensuring that learners across all CEFR levels encounter appropriate vocabulary challenges, thus facilitating a more comprehensive and effective language learning experience.

Regarding lexical variation, the categorisation of the eight movie genres into high, moderate, and low categories was established through an internal comparative profiling approach based on the empirical hierarchy of the collected data, rather than an arbitrary universal benchmark. Given that absolute TTR values are naturally dependent on corpus size, the relative distribution within this corpus offers a statistically valid baseline for internal stratification. The analysis reveals that horror, action, and thriller movies occupy the upper stratum of this hierarchy, exhibiting the highest average lexical variation (16.88%, 16.87%, and 16.85%, respectively). This statistical placement suggests a more varied and complex vocabulary framework within these genres. From a narrative perspective, this higher variation can be attributed to the diverse situational context shifts inherent in these genres, which frequently move across complex plotlines, investigative dialogues, and descriptive suspense elements, thereby requiring a broader lexical repertoire. In contrast, genres such as comedy, drama, and romance demonstrate the lowest average lexical variation within the corpus (15.22%, 15.59%, and 15.62%, respectively), indicating a more repetitive vocabulary configuration. This relative compression occurs because these genres heavily rely on domestic register, slice-of-life interactions, and interpersonal conversational routines. These communicative contexts naturally draw upon a more confined, recurrent set of everyday vocabulary, leading to higher word repetition and a lower TTR. Positioned between these two clear dimensions, adventure and animated movies maintain a moderate lexical variation threshold, averaging at 16.61% and 16.07%, respectively. This mid-tier statistical distribution reflects a balanced narrative structure that bridges action-driven descriptive terms with accessible, character-centric dialogues. Consequently, this internal grouping provides a systematic, evidence-based continuum indicating that genres with relatively higher vocabulary variation are structurally suited to advanced learners, while more repetitive configurations are optimised for foundational skills.

The evaluation of lexical density revealed that all eight movie genres share a highly uniform lexical structure, with content words accounting for approximately half of the total vocabulary (49.25% to 52.23%). Contrary to potential expectations that action or adventure might be less dense due to their fast-paced narratives, the lack of substantial variance indicates that genre characteristics do not significantly influence the overall distribution of meaning-carrying versus structural words. This striking homogeneity is likely because all these materials belong to the same overarching register of spoken, conversational English embedded within screenplays. Although written and carefully edited, movie scripts fundamentally mirror authentic interactive discourse, which naturally maintains a balanced equilibrium between information-delivery (content words) and grammatical cohesion (function words) to ensure immediate auditory comprehension for a general audience. This finding builds upon Johansson's (2008) and Laufer and Nation's (1995) assertions that spoken language typically displays lower density than academic texts, yet it qualifies their benchmarks by demonstrating that movie dialogue sits at a stable mid-point threshold of around 50%. For language learners, this lack of variance carries a significant pedagogical implication: rather than relying on lexical density to determine text difficulty, educators and self-directed students should focus on

other multi-dimensional metrics, such as lexical variation and CEFR levels, as the cognitive load regarding content-to-function word ratios remains constant across all film genres.

A notable pattern across the analyses is the consistent lexical complexity of animation, a genre often assumed to use simple language because it is linked to child audiences. This assumption does not match how mainstream animated films are actually made. Unlike vocabulary-controlled materials such as graded readers, animated films are usually written for both children and the adults who bring them, since adult ticket buyers are the main paying audience. This pattern is known as dual-audience appeal in family films, where content is designed to work on two levels at once, one for children and one for adults. While visual humour appeals to children, the dialogue often contains wordplay, pop-culture references, and adult-level humour aimed at the accompanying adults (Meeusen, 2018; Schober, 2015). This may explain why animation shows a higher OWL percentage and more medium- to low-frequency words in the present corpus. Animated stories also tend to build fictional worlds, which require invented terms, proper nouns, and specialised descriptive words to make imaginary characters, places, and technology understandable to viewers. This need for world-building vocabulary likely adds further to the genre's high OWL percentage. It may also explain the genre's high lexical density, since descriptive content words are needed to establish a fictional setting, unlike genres such as romance or drama that can rely on everyday, real-world conversation. These factors suggest that the linguistic complexity of animation does not come from a lack of simplification, but from a different communicative purpose compared to other genres.

Pedagogical implication

The findings of this research provide a valuable foundation for further investigation into using movies as a tool for language learning, particularly for vocabulary expansion and communication skills enhancement. Movies have long been utilised by educators to boost learner engagement and make learning more enjoyable (Chenchen, 2011). By analysing vocabulary across different movie genres, this study offers insights that can improve language teaching strategies in classrooms or self-directed learning environments. When selecting movies for educational purposes, it is crucial to align the content with the learner's proficiency level (Bailey, 2005). Appropriate movie selection enhances learning effectiveness, while overly challenging vocabulary can discourage students and reduce motivation. Moreover, this vocabulary analysis benefits self-learners, allowing them to choose movies that match their language abilities, thereby maximising the effectiveness of their practice and making language learning more accessible and enjoyable.

Limitation of the study and recommendations for future studies

This research has some limitations. First, it focuses on English-language movies produced by native speakers, specifically from the inner circle, where English is the primary language. Additionally, movies are often categorised into multiple genres, which leads to repeated data analysis. For instance, the 2023 Barbie movie is classified as both an adventure and a comedy, complicating efforts to isolate genre-specific vocabulary. Future research could explore how

the emotional impact of movies, such as excitement or sadness, influences vocabulary retention and learning. Investigating the role of cultural context in movies may also provide insights into how cultural settings enhance language learning by presenting vocabulary in diverse and varied contexts.

CONCLUSION

This study provides a comprehensive, multi-dimensional corpus analysis of vocabulary characteristics across eight distinct movie genres, offering crucial empirical insights for L2 vocabulary acquisition. By synthesising the findings across lexical profiling, frequency levels, CEFR alignments, lexical variation, and lexical density, distinct linguistic patterns emerge regarding how specific genres behave across different metrics. Notably, genres such as romance, drama, and horror demonstrate highly consistent behaviour across all structural analyses; they exhibit the highest concentrations of the GSL vocabulary (exceeding 87%), the greatest proportion of high-frequency words (above 95.3%), and a dominant representation at the CEFR A1–A2 levels (surpassing 94%). This absolute alignment across all domains definitively classifies romance, drama, and horror as the most lexically accessible genres, making them ideal entry points for foundational and beginner language learners. Conversely, animation stands out as an exceptionally complex and anomalous genre, occupying the opposite end of the linguistic spectrum across multiple analyses. Despite traditional assumptions of simplicity, animated movies within this contemporary corpus yield the lowest percentage of GSL words, the highest proportion of the OWL vocabulary (14.64%), and an extensive reliance on medium-to-low frequency vocabulary—requiring an advanced baseline vocabulary size of 11,000 words to achieve 98% text coverage. Furthermore, animation leads in CEFR C2-level vocabulary complexity (3.14%) and exhibits the highest lexical density (52.23%). This multidimensional density and complexity position animation as a highly challenging yet rewarding resource tailored specifically for advanced learners aiming to expand their sophisticated lexicon. Meanwhile, genres like action and thriller display a unique lexical configuration; while they do not possess high lexical density, they exhibit the highest lexical variation (approximately 16.8%), indicating that they expose learners to a highly diverse and non-repetitive vocabulary across shifting narrative contexts. In contrast, comedy features low lexical variation but high lexical density, meaning it delivers dense information-carrying contexts through a more recurrent, repetitive vocabulary framework that can aid retention through exposure. Across all eight genres, lexical density remains remarkably uniform at around a 50:50 ratio of content to function words, indicating a stable baseline register of spoken cinematic discourse. Ultimately, this study demonstrates that movie genres do not possess uniform difficulty; rather, their lexical demands shift systematically. Aligning these specific genre profiles with standardised language learning benchmarks allows educators and self-directed learners to move beyond arbitrary choices, optimising genre selection to target precise vocabulary acquisition goals and maximise communicative proficiency.

THE AUTHORS

Kingkan Kornkamomchanok is an undergraduate student in the English for Service Industry programme at the Department of Service Innovation and Intercultural Communication, Faculty of Hospitality Industry, Kasetsart University, Kamphaeng Saen Campus, Thailand. Her academic interests include English for specific purposes, tourism and hospitality communication, service industry communication, and applied linguistics.
kingkan.ko@ku.th

Nuttaya Lamcharornpong is an undergraduate student in the English for Service Industry programme at the Department of Service Innovation and Intercultural Communication, Faculty of Hospitality Industry, Kasetsart University, Kamphaeng Saen Campus, Thailand. Her academic interests include English for specific purposes, tourism and hospitality communication, service industry communication, and applied linguistics.
nuttaya.lam@ku.th

Phawita Lamka is an undergraduate student in the English for Service Industry programme at the Department of Service Innovation and Intercultural Communication, Faculty of Hospitality Industry, Kasetsart University, Kamphaeng Saen Campus, Thailand. Her academic interests include English for specific purposes, tourism and hospitality communication, service industry communication, and applied linguistics.
phawita.l@ku.th

Piyapong Laosrirattanachai, Ph.D., is an Associate Professor in the English for Service Industry Programme at the Department of Hospitality Innovation and Intercultural Communication, Faculty of Hospitality Industry, Kasetsart University, Kamphaeng Saen Campus, Thailand. His research interests include corpus linguistics, vocabulary studies, word list development, English for specific purposes, and move analysis.
piyapong.l@ku.th

REFERENCES

- Alfadil, M. (2020). Effectiveness of virtual reality game in foreign language vocabulary acquisition. *Computers & Education*, 153, Article 103893. <https://doi.org/10.1016/j.compedu.2020.103893>
- Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education*, 3(3), 21–34. <https://doi.org/10.20472/TE.2015.3.3.002>
- Anthony, L. (2024). *AntWordProfiler* (Version 2.2.1) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software/AntWordProfiler>
- Antipov, E. A., & Pokryshevskaya, E. B. (2016). How to measure the power of actors and film directors? *Empirical Studies of the Arts*, 34(2), 1–13. <https://doi.org/10.1177/0276237416628904>
- Ashcroft, J. R., Garner, J., & Haddingham, O. (2018). Incidental vocabulary learning through watching movies. *Australian Journal of Applied Linguistics*, 1(3), 135–147. <https://doi.org/10.29140/ajal.v1n3.89>
- Bailey, K. M. (2005). *Practical English language teaching: Speaking*. McGraw-Hill.
- Baratta, A., & Jones, S. (2008). Using films to introduce and develop academic writing skills among UK undergraduate students. *Journal of Educational Enquiry*, 8(2), 15–37.
- Bates, E., Bretherton, I., & Snyder, L. (1998). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge University Press.
- Box Office Mojo. (2024, September). *Box Office Mojo*. <https://www.boxofficemojo.com/>

- Chapple, L., & Curtis, A. (2000). Content-based instruction in Hong Kong: Student responses to film. *System*, 28(3), 419–433. [https://doi.org/10.1016/S0346-251X\(00\)00021-X](https://doi.org/10.1016/S0346-251X(00)00021-X)
- Chenchen, X. (2011). How to use English movies to improve a student's listening and speaking ability in Chinese ESL culture learning classrooms [Master's thesis, University of Wisconsin-Platteville]. MINDS@UW.
- Cobb, T. (2025). *Vocabprofile*. [Computer program]. <http://www.lexutor.ca/vp/>
- Cook, V. (2016). *Second language learning and language teaching* (5th ed.). Routledge. <https://doi.org/10.4324/9781315883113>
- Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing*, 16(3), 129–147. <https://doi.org/10.1016/j.jslw.2007.07.002>
- Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, 12(2), 65–78.
- Csomay, E., & Petrović, M. (2012). “Yes, your honor!”: A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, 40(2), 305–315. <https://doi.org/10.1016/j.system.2012.05.004>
- Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2), 197–222. <https://doi.org/10.1093/applin/24.2.197>
- Dang, T. N. Y., & Long, X. (2024). Online news as a resource for incidental learning of core academic words, academic formulas, and general formulas. *TESOL Quarterly*, 58(1), 32–62. <https://doi.org/10.1002/tesq.3208>
- Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, 33(1), 66–76. <https://doi.org/10.1016/j.esp.2013.08.001>
- Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL - International Journal of Applied Linguistics*, 167, 132–158. <https://doi.org/10.1075/itl.167.2.02dan>
- Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, 26(4), 617–641. <https://doi.org/10.1177/1362168820911189>
- Davies, M. (2021). *Corpus of News on the Web (NOW)*. Brigham Young University. <https://www.english-corpora.org/now/>
- Davis, B. H., & Brewer, J. P. (1997). *Electronic discourse: Linguistic individuals in virtual space*. State University of New York Press.
- Ellis, R. (1999). *SLA research and language teaching*. Oxford University Press.
- Gieve, S., & Clark, R. (2005). The Chinese approach to learning: Cultural trait or situated response? The case of a self-directed learning programme. *System*, 33(2), 261–276. <https://doi.org/10.1016/j.system.2004.09.015>
- Ha, H. T. (2022). Lexical profile of newspapers revisited: A corpus-based analysis. *Frontiers in Psychology*, 13, Article 800983. <https://doi.org/10.3389/fpsyg.2022.800983>
- Herron, C. A., & Hanley, J. (1992). Using video to introduce children to a foreign culture. *Foreign Language Annals*, 25(5), 419–426. <http://dx.doi.org/10.1111/j.1944-9720.1992.tb01122.x>
- Hu, X. (2021). Predicting CEFR levels in L2 oral speech, based on lexical and syntactic complexity. *Asia Pacific Journal of Corpus Research*, 2(1), 35–45.
- Ismaili, M. (2013). The effectiveness of using movies in EFL classrooms: A study conducted at South East European University. *Academic Journal of Interdisciplinary Studies*, 2(4), 121–132. <https://doi.org/10.5901/ajis.2012.v2n4p121>

- Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing*, 37, 13–38. <http://dx.doi.org/10.1016/j.jslw.2017.06.001>
- Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers*, 53, 61–79.
- Keene, M. D. (2006). Viewing video and DVD in the EFL classroom. *Bunkyo Gakuin University Journal*, 8(1), 217–234.
- King, J. (2002). Using DVD feature films in the EFL classroom. *Computer Assisted Language Learning*, 15(5), 509–523. <http://dx.doi.org/10.1076/call.15.5.509.13468>
- Laosrirattanachai, P., Baothong, C., Laijud, K., & Laosrirattanachai, P. (2025). Dissecting vocabulary usage across TED Talks topics: A Corpus-based perspective. *LEARN Journal: Language Education and Acquisition Research Network*, 18(2), 943–973. <https://doi.org/10.70730/XUFI2893>
- Laosrirattanachai, P., & Laosrirattanachai, P. (2025). Tracing tourism business research trends in Scopus-indexed journals using corpus-based and judgement-based approaches. *Humanities, Arts and Social Sciences Studies*, 25(1), 32–53. <https://doi.org/10.69598/hasss.25.1.268122>
- Laosrirattanachai, P., Rittikote, N., & Laosrirattanachai, P. (2026). Developing a nature tourism word list for L2 learners in the tourism sector. *Australian Review of Applied Linguistics*, Advance online publication. <https://doi.org/10.1075/aral.25109.lao>
- Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly*, 47(4), 867–872. <https://doi.org/10.1002/tesq.140>
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size, and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.
- Lee, S. (2015). Lexical demand of animation movies and vocabulary learning through watching them: A corpus-driven approach. *STEM Journal*, 16(1), 23–47.
- Lestari, D., & Harahap, D. I. (2020). The use of animation movies to improve students' motivation in learning English. *MELT Journal*, 5(1), 40–53.
- Lestari, I. W., & Hardiyanti, N. (2020). Vocabulary learning autonomy through incorporation of English songs: Indonesian EFL students' perspective. *3L: The Southeast Asian Journal of English Language Studies*, 26(2), 91–104. <https://doi.org/10.17576/3L-2020-2602-07>
- Liando, N. V. F., Sahetapy, R. J. V., & Maru, M. G. (2018). English major students' perceptions towards watching English movies in listening and speaking skills development. *Advances in Social Sciences Research Journal*, 5(6), 1–16.
- Lieven, E. V. M. (1978). Conversations between mothers and young children: Individual differences and their possible implication for the study of child language learning. In N. Waterson & C. Snow (Eds.), *The development of communication: Social and pragmatic factors in language acquisition* (pp. 173–187). Wiley.
- Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, 10(1), 105–129. <https://doi.org/10.1080/14790718.2012.679275>
- Loesnopchaimongkhon, R., Phrommopakorn, C., Chernchom, P., & Laosrirattanachai, P. (2025). A corpus-based analysis of lexical characteristics across English news categories for L2 pedagogical use. *Journal of Studies in the English Language*, 20(2), 36–65. <https://so04.tci-thaijo.org/index.php/jsel/article/view/279647>

- Meebangsai, D., Pongtin, P., Kitipoontanakorn, P., & Laosrirattanachai, P. (2023). Investigating proficiency of academic English in student writing: A comparative case study on vocabulary utilization in student research article writing vis-à-vis national and international research. *PASAA*, 67, 66–100.
- Meeusen, M. (2018). “Unless someone like you” buys a ticket to this movie: Dual audience and aetonymativity in picturebook to film adaptations. *Children’s Literature in Education*, 49(4), 485–498. <https://doi.org/10.1007/s10583-017-9334-6>
- Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 47, Article 100511. <https://doi.org/10.1016/j.asw.2020.100511>
- Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, 44(4), 529–539. <https://doi.org/10.1017/S0261444811000267>
- Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists*. Victoria University of Wellington. <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- Nation, I. S. P. (2020). *About the BNC/COCA headword lists*. Victoria University of Wellington. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists>
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.
- Nation, I. S. P., & Crabbe, D. (1991). A survival language learning syllabus for foreign travel. *System*, 19(3), 191–201.
- Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge University Press.
- Nguyen, N. T. T. (2021). A review of the effects of media on foreign language vocabulary acquisition. *International Journal of TESOL & Education*, 1(1), 30–37.
- Nonthasen, N., Samakketwit, S., Srisompak, P., & Laosrirattanachai, P. (2025). Exploring vocabulary characteristics across nine music genres: A corpus-based study for vocabulary learning. *Studies in English Language and Education*, 12(3), 1507–1524. <http://dx.doi.org/10.24815/siele.v12i3.41375>
- Odo, D. M. (2021). Examining the influence of English songs on English L2 lexical learning: A quantitative meta-analytic review. *Language, Culture and Curriculum*, 35(4), 386–403. <https://doi.org/10.1080/07908318.2021.2022684>
- OpusData. (2020, May). *Movie genre*. https://www.opusdata.com/documentation/index.php/Movie_genre
- Owen, N., Shrestha, P., & Bax, S. (2021). *Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Languages (CEFR) levels assessed in the APTIS test*. British Council ARAGs Research Reports Online. <https://oro.open.ac.uk/75228/>
- Peters, E. (2018). The effect of out-of-class exposure to English language media on learners’ vocabulary knowledge. *ITL-International Journal of Applied Linguistics*, 169(1), 142–168. <https://doi.org/10.1075/itl.00010.pet>
- Rokni, S. J. A., & Ataee, A. J. (2014). Movies in EFL classrooms: With or without subtitles. *The Dawn Journal*, 3(1), 715–726.
- Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal*, 78(2), 199–221. <https://doi.org/10.2307/329010>
- Sato, M. (2007). *Exploring effective English teaching methods using film scripts among poor learners of English* (Research Report No. 12, pp. 218–231). Waseda University.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, 95(1), 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484–503. <https://doi.org/10.1017/S0261444812000018>
- Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching* (2nd ed.). Cambridge University Press.
- Schober, A. (2015). ‘Why can’t they make kids’ flicks anymore?’: Willy Wonka and the Chocolate Factory and the dual-addressed family film. In N. Brown & B. Babington (Eds.), *Family films in global cinema: The world beyond Disney* (pp. 53–68). Bloomsbury Publishing.

- Ure, J. (1971). Lexical density and register differentiation. *Contemporary Educational Psychology*, 5, 96–104.
- van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457–479. <https://doi.org/10.1093/applin/ams074>
- Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, 33(2), 278–293.
- Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.
- Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, 30(3), 407–427. <https://doi.org/10.1093/applin/amp010>
- West, M. (1953). *A general service list of English words*. Longman.
- Wilkins, D. (1972). *Linguistics in language teaching*. Edward Arnold.
- Wolf-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i, Second Language Teaching & Curriculum Center.
- Xodabande, I., Ebrahimi, H., & Karimpour, S. (2022). How much vocabulary is needed for comprehension of video lectures in MOOCs: A corpus-based study. *Frontiers in Psychology*, 13, Article 992638. <https://doi.org/10.3389/fpsyg.2022.992638>
- Yaseen, B. H., & Shakir, H. (2015). Movie effects on EFL learners at Iraqi schools in Kuala Lumpur. *International Journal of Education and Literacy Studies*, 3(3), 31–36. <https://doi.org/10.7575/aiac.ijels.v.3n.3p.31>