JSEL
Journal of Studies in the English Language

A Corpus-Based Analysis of Lexical Characteristics Across English News Categories for L2 Pedagogical Use

Rattavit Loesnopchaimongkhon
Chanapa Phrommopakorn
Pancheewa Chernchom
Piyapong Laosrirattanachai*
Kasetsart University, Thailand
*Corresponding author's email: piyapong.l@ku.th
Received April 18, 2025; revised June 23, 2025; accepted June 29, 2025

Abstract

News articles are widely regarded as valuable resources for vocabulary acquisition. However, they encompass diverse categories, each catering to specific learner needs. This study analysed the vocabulary of 3,000 news articles across 12 categories, focusing on lexical profiling, lexical level, variation, density, and CEFR level to support L2 learners. The results showed that the Health category had the highest General Service List word coverage (81.01%), while Technology featured the most Academic Word List terms (8.23%). Fashion contained the largest proportion of specialised vocabulary (18.31%) and exhibited the highest lexical variation (51.29%). High-frequency words dominated all categories (91–94.79%), while Fashion included the most mid-frequency (5.84%) and low-frequency (2.36%) words. Lexical density was highest in the Environment category (57.85%) and lowest in Sports (53.2%). The CEFR analysis indicated that A1 and A2 words comprised the majority (76.66% and 10.44%, respectively), while categories such as Fashion and Nutrition included the highest proportions of C1-C2 words (6.32% and 6.53%, respectively). These findings suggest that categories such as Health and Sports are suitable for beginner learners, while Fashion and Nutrition offer more complex vocabulary for advanced learners. This study highlights the unique lexical characteristics of news categories, providing educators and learners with guidance on selecting authentic materials to enhance vocabulary learning.

Keywords: news, corpus-based study, vocabulary, lexical characteristics, learner proficiency

Vocabulary knowledge is widely recognised as a cornerstone of language acquisition and plays a pivotal role in English learning. A robust vocabulary base is essential for learners to develop effective communication skills and achieve language proficiency (Schmitt, 2008). Without adequate lexical

knowledge, learners encounter substantial challenges in expressing themselves meaningfully, which impedes effective communication (Laosrirattanachai & Laosrirattanachai, 2025a). Furthermore, limited vocabulary not only constrains reading comprehension but also significantly hinders the acquisition of new words (August et al., 2005). Consequently, vocabulary acquisition is a critical element for L2 English learners, enabling them to learn and use the language more effectively.

This has led to various efforts to support vocabulary development through different approaches. In the modern era, English learners have access to diverse and effective methods for enhancing their vocabulary skills. For example, songs have proven to be valuable tools for beginner learners, as they support the development of listening and pronunciation skills, which in turn contribute to improved speaking abilities (Coyle & Gracia, 2014; Davis, 2017; Tegge, 2018). Similarly, watching English movies allows learners to acquire a wide range of vocabulary and syntactic structures while benefiting from visual cues such as body language and facial expressions (Baranowska, 2020; Ha, 2022a; Sayer & Ban, 2014). Among various learning tools, news articles hold particular value for language enhancement due to their diverse and content-rich nature (Dang & Long, 2023; Dang & Lu, 2024). In the digital era, access to news has become increasingly convenient, with online news offering a more affordable and flexible alternative to traditional newspapers. This accessibility makes news an appealing resource for language learners, further reinforced by its inclusion in widely recognised proficiency tests such as IELTS, which feature newspaper and magazine articles in their reading sections (Moore et al., 2015).

Two key aspects must be considered when selecting sources for teaching and learning, particularly when using news as supplementary material. The first concerns the criteria for selecting news as educational content; the second relates to the categorisation of news. Both aspects should align with learners' abilities and proficiency levels.

Regarding the first aspect, the criteria for selection should include an evaluation of factors such as lexical level (Coxhead, 2018; Nation, 2022), lexical profile (Coxhead & Hirsch, 2007; Laosrirattanachai & Laosrirattanachai, 2023), lexical variation (Meebangsai et al., 2023; Treffers-Daller et al., 2018), and lexical density (Liu, 2021; Wingrove, 2017). In addition, the present study incorporated CEFR level as a further criterion for assessing suitability. Regarding the second aspect, the literature review revealed a substantial body of research focused on the vocabulary used in news articles. For example, Ha (2022b) analysed the vocabulary knowledge required for comprehension, while Kembaren and Aswani (2022) investigated lexical density in articles from The New York Times. Despite these valuable contributions, there remains a lack of detailed analysis of news articles by category, as well as of

methods for examining vocabulary across multiple dimensions. Although previous studies confirm that news articles can facilitate English vocabulary learning, articles from different categories naturally employ vocabulary in varying ways and are therefore suitable for learners with differing proficiency levels. To the best of our knowledge, no research has yet determined which categories of news are most appropriate for L2 learners at specific proficiency levels. Such analyses could provide further insight into the development of vocabulary learning resources tailored to meet learners' needs.

The present research seeks to address this gap by analysing the vocabulary of news articles across different categories through five key aspects: lexical profiling, lexical level, lexical variation, lexical density, and CEFR level.

Literature Review

The present research explored five aspects of vocabulary in distinct types of news articles, as outlined below.

Studies on News Vocabulary

It has long been recognised that news serves as a valuable source for vocabulary teaching and learning for L2 learners. From earlier periods, when news was primarily delivered through print media (Kyongho & Nation, 1989), to the current digital era, where it is widely disseminated online (Dang & Long, 2023), its role in vocabulary acquisition has remained significant. Several studies have investigated vocabulary in news articles, particularly in relation to vocabulary coverage, lexical learning, and their potential as language learning resources.

Some studies have specifically examined the vocabulary profile and lexical coverage of news articles. For example, Ha (2022b) analysed the News on the Web (NOW) corpus and found that knowledge of the most frequent 4,000 word families from the BNC/COCA wordlist, along with proper nouns and marginal words, was required for 95% coverage of online newspaper and magazine articles. However, lexical demands varied across national contexts, especially when aiming for 98% coverage. Similarly, Hsu (2018) assessed the vocabulary level of Voice of America (VOA) news and concluded that it reached the sixth 1,000-word-family level at 98% coverage, suggesting its potential for extensive reading and mid-frequency vocabulary acquisition.

Other studies have focused on the potential of news articles for incidental vocabulary learning. For example, Dang and Long (2023) examined VOA news for its capacity to support the learning of core academic words, academic formulas, and general formulas. They found that consistent reading of VOA news could lead to frequent encounters with these lexical items, indicating that online news is a valuable resource for incidental learning. Vocabulary acquisition from news constitutes incidental learning, as the primary focus is on content

comprehension rather than deliberate word memorisation. Unlike intentional learning, it does not involve formal instruction.

Teng (2015) explored incidental vocabulary learning using BBC news texts to test the involvement load hypothesis (ILH), confirming that tasks requiring greater cognitive effort led to better vocabulary retention. This finding highlighted the importance of lexical thresholds for news comprehension. The ILH supports incidental learning by proposing that deeper processing—through need, search, and evaluation—enhances retention, aligning with meaning-focused vocabulary acquisition. Dang and Lu (2024) extended this line of research by conducting an experimental study to determine whether exposure to online news articles results in measurable vocabulary gains. Their findings confirmed learning gains in academic words and multiword sequences, with word frequency playing a substantial role in learning outcomes.

In addition to studies on vocabulary learning, some research has examined ways to optimise news articles as learning materials. For instance, Kyongho and Nation (1989) investigated how the selection of newspaper stories affected vocabulary load and word repetition. They found that selecting running stories, rather than unrelated stories, increased the repetition of low-frequency words and provided better conditions for acquiring vocabulary beyond the most frequent 2,000-word level. Astika (2015) discussed the use of the Vocabulary Profiler tool for classifying lexical items in texts into high-, low-, and academic-frequency categories, demonstrating its utility for selecting vocabulary in language teaching.

Despite these valuable insights, existing research on vocabulary in news articles has not comprehensively examined lexical characteristics across different news categories. Most studies have focused on specific aspects such as vocabulary coverage, incidental learning, or pedagogical applications, rather than providing an integrated analysis. Moreover, while some research has addressed lexical frequency and repetition, there is a lack of holistic investigation into multiple vocabulary dimensions within news articles.

Given this gap, the present study aimed to provide a more comprehensive analysis of vocabulary in news articles. Specifically, it investigated lexical profile, lexical level and coverage, CEFR level, lexical variation, and lexical density. The following sections elaborate on these aspects.

Lexical Profiling

Vocabulary studies often apply the concept of lexical profiling to categorise words according to various reference lists. Nation's (2022) framework, which classifies vocabulary into four distinct groups (high-frequency words, academic words, technical words, and low-frequency words) has significantly influenced the construction of these lists. Although the reference lists vary depending on

the specific objectives of each study, most lexical analysis software typically uses the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000) as default reference points. The GSL, comprising 2,000 high-frequency words, has played a foundational role in English vocabulary teaching, learning, and research, accounting for approximately 80% of the vocabulary in standard texts. The AWL, containing 570-word families frequently found in academic writing, generally represents about 10% of the vocabulary in academic texts.

While technical and low-frequency words are more prevalent in specialised texts, they are rare in general usage. Within the lexical profiling framework, these two categories are collectively referred to as the Outside Word List (OWL), as they fall outside the scope of the GSL and AWL. These words, being domain-specific or infrequent, are considered indicators of higher lexical difficulty. The OWL is estimated to constitute roughly 10% of the vocabulary in a standard text (Coxhead & Hirsch, 2007). The present study applied lexical profiling to categorise vocabulary from different news categories according to the GSL and AWL. The proportion of vocabulary falling into the AWL and OWL was used as an indicator of lexical difficulty across various news categories.

Lexical Level

The classification of lexical levels is based on the frequency and distribution of vocabulary, with each level comprising 1,000-word families. High-frequency words are grouped in the initial levels, while less frequent words appear in the higher levels. A total of 25 levels, encompassing 25,000-word families, were derived from two major corpora: the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) (Nation, 2016, 2017). Lexical levels reflect the complexity of vocabulary in a text, with higher-level words indicating more advanced or specialised language. Vocabulary is commonly categorised into three frequency tiers: high-frequency words (K1-K3), mid-frequency words (K4-K9), and low-frequency words (K10-K25) (Schmitt & Schmitt, 2014).

Lexical levels are closely tied to lexical coverage, which refers to the proportion of a text that can be understood based on known vocabulary. Readers generally need to know at least 95% of a text's vocabulary for adequate comprehension, with 98% being the optimal threshold (Hu & Nation, 2000). Studies have shown that approximately 4,000 high-frequency word families, along with proper nouns, are required for 95% coverage of newspapers and novels, while 8,000 to 9,000-word families are necessary to reach 98% coverage (Nation, 2006). Ha (2022b), in analysing the News on the Web (NOW) corpus, found that knowledge of the top 4,000 BNC/COCA word

families achieved 95% coverage. However, such studies have largely focused on general news, overlooking vocabulary variation across specific categories. Selecting inappropriate news categories for vocabulary learning may hinder learner progress. Therefore, the present study employed the 25-level BNC/COCA word family lists to examine the lexical levels of news articles, aiming to guide learners in selecting articles suited to their proficiency for more effective vocabulary acquisition.

CEFR

The Common European Framework of Reference for Languages (CEFR) is a standardised system for evaluating language proficiency. Developed by the Council of Europe between 1993 and 1996 and updated in 2018, it assesses listening, speaking, reading, and writing skills. The CEFR framework consists of six levels, ranging from A1, representing basic language use, to C2, indicating near-native proficiency (Council of Europe, 2020). Each level corresponds to a specific vocabulary size and communicative ability, offering a structured approach to language assessment (Graves, 2008; Hulstijn et al., 2010).

In considering the alignment between CEFR levels and vocabulary difficulty, Nation and Crabbe (1991) proposed a framework that maps CEFR levels to lexical development based on word family size. According to this model, A1 users possess a basic repertoire (including flemma and -ly forms), while A2 to B1 users gradually expand their vocabulary to include partial and full Level 3-word families. B2 corresponds to Level 4, C1 to Level 5, and C2 to Level 6 and beyond, reflecting idiomatic proficiency and deeper lexical knowledge. A corpus-based approach can thus be used to analyse the lexical level of a text and map it to the corresponding CEFR level. The present study compared CEFR levels with the vocabulary used in various news categories to enhance the understanding of how vocabulary complexity aligns with language proficiency (Nation, 2006, 2017).

Lexical Variation

Lexical variation refers to the extent of different words used within a text, providing insight into its lexical diversity. It is traditionally measured by the Type-Token Ratio (TTR), a fundamental metric for assessing lexical richness (Davis & Brewer, 1997), which compares the number of unique words (types) to the total number of words (tokens) in a text (Bates et al., 1988). Tokens include all word occurrences, while types represent unique words—meaning that words with the same spelling are counted as one type, regardless of how often they appear.

A high TTR indicates greater lexical variation (Baker, 2006), while a lower TTR suggests limited variation (Indarti, 2017). The formula for calculating TTR is (Number of Types/Number of Tokens) × 100. For example, in the sentence: "The research article discusses the research findings from the research team in the research project," there are 15 tokens and 9 types. The repeated use of "research" reduces the TTR: (9 ÷ 15) × 100 = 60%, indicating limited lexical variety. In the present study, TTR was employed to evaluate lexical variation in online news articles. This approach enabled the analysis of vocabulary frequency and effectiveness across different news categories, offering insights into the richness and variation of language in these contexts.

Lexical Density

Lexical density measures the proportion of content words, nouns, verbs, adjectives, and adverbs, relative to the total number of words in a text (Johansson, 2008; Thornbury & Slade, 2006). It serves as an indicator of textual informativeness, with higher lexical density reflecting more precise and information-rich content. Typically, written texts exhibit lexical density above 40%, whereas spoken texts often fall below this threshold due to their greater reliance on connectors and function words (Laufer & Nation, 1995; Ure, 1971). The formula used to calculate lexical density is (Number of Content Words ÷ Number of Tokens) × 100.

In this study, lexical density was used to analyse the ratio of content to function words in news articles, offering insights into the linguistic complexity of different news categories and their suitability for learners at various proficiency levels (Nasseri & Thompson, 2021).

Based on the research gaps and literature review, the aim of the present study was to address the following research question: To what extent do the lexical characteristics, namely, lexical profile, lexical level, CEFR level, lexical variation, and lexical density, differ across 12 English news categories, and how can these differences inform the selection of news types that are pedagogically appropriate for learners at different proficiency levels?

Research Methodology

This study examined the vocabulary used in BBC and CNN online news articles (2019–2023), classifying content into specific categories and analysing five key lexical metrics: lexical profile, lexical level, CEFR level, lexical variation, and lexical density. By focusing on these dimensions, the study aimed to provide insights into vocabulary usage across different news categories to support L2 vocabulary development.

Data Collection

Previous research has classified news articles into distinct categories (e.g. Abyaad et al., 2020; Bleyer, 1916). The present study adopted the categorisation framework proposed by Kaspar and Fuchs (2021), which reflects a contemporary and reliable classification scheme commonly used across news agencies. Their framework identifies 12 categories: Economy, Entertainment, Health, Sports, Technology, Politics, Science, Environment, Nutrition, Crime, Accidents and Disasters, and Fashion.

Data were collected from two major news organisations: the British Broadcasting Corporation (BBC) and Cable News Network (CNN). As summarised in Table 1, 25 random articles were selected annually per category from each agency between 2019 and 2023, resulting in 1,500 articles per agency and a total dataset of 3,000 articles.

Table 1
Overall Information on News Articles Collected from BBC and CNN

Category	BBC (Tokens)	CNN (Tokens)	Total (Tokens)
Crime	103,006	106,010	209,016
Disaster	102,311	104,130	206,441
Economy	70,800	106,198	176,998
Entertainment	89,254	133,881	223,135
Environment	136,802	200,203	337,005
Fashion	110,319	73,547	183,866
Health	102,964	154,444	257,408
Nutrition	143,930	95,954	239,884
Politics	71,971	64,648	136,619
Science	65,769	77,178	142,947
Sports	77,607	91,410	169,017
Technology	127,017	101,346	228,363
Total	1,201,750	1,308,949	2,510,699

Data Analysis and Research Instruments

The data were analysed using a range of lexical analysis tools, each aligned with one of the five key dimensions under investigation.

For the analysis of lexical profiling, AntWordProfiler (Anthony, 2024) was used to examine the lexical composition of each news category, referencing the GSL and the AWL. Words not included in either list were categorised as belonging to the OWL. Categories with a higher proportion of GSL words were considered more accessible and therefore more suitable for beginner L2 learners, whereas those with a greater share of OWL words indicated higher lexical complexity and were deemed more appropriate for advanced learners.

In the analysis of lexical level, VocabProfile (Cobb, 2024) was employed to classify words into 25 frequency bands (K1-K25), where K1 represents the most frequently used words and K25 the least frequent. These were grouped into high-frequency (K1-K3), mid-frequency (K4-K9), and low-frequency (K10-K25) categories, with an additional classification for off-list words beyond K25. Although several frameworks have attempted to map these levels to CEFR bands, the alignment remains approximate and, at times, contested (Benigno & de Jong, 2019). Nation and Crabbe's (1991) framework, widely recognised and still in use (Li et al., 2024), provides a developmental model that correlates lexical growth with increasing language proficiency, albeit without an official reference list. One point requiring clarification is the broad range from K7 to K25 often being grouped under the C2 level. Since C2 vocabulary must be contextually appropriate and often domain-specific, this wide span is justified by the fact that specialised vocabulary in one field may be considered low-frequency in another (Nation, 2016). Thus, categories dominated by high-frequency vocabulary were considered more accessible for lower-proficiency learners, while those containing higher proportions of low-frequency items reflected increased lexical complexity.

To compare CEFR levels, lexical levels derived from VocabProfile were mapped to CEFR bands using Nation and Crabbe's framework. In this adapted model, K1 corresponds to A1, K2 to A2, K3 to B1, K4 to B2, K5 to C1, and K6 and above to C2. News categories with a predominance of A1-A2 level vocabulary were considered beginner-friendly, while those with a greater presence of C1-C2 level vocabulary were classified as suitable for advanced learners.

For the analysis of lexical variation, TTR was used to assess the diversity of vocabulary in each category. Because TTR is sensitive to text length, average values were calculated across all articles within each category. AntWordProfiler was used to calculate the number of types and tokens. Categories with higher average TTR values were interpreted as having greater lexical variation, thus offering more challenge and enrichment for advanced learners. Conversely, lower TTR values signalled less variation and more repetition, a feature beneficial to beginner learners for reinforcing core vocabulary.

Lexical density was measured as the proportion of content words (nouns, verbs, adjectives, and adverbs) to total words in each text. AntWordProfiler, in conjunction with Nation's (2018) Function Word List, was used for this analysis. Categories with higher lexical density were seen as more information-rich and lexically demanding, requiring learners to engage with a wider array of meaningful vocabulary. In contrast, categories with lower density placed greater reliance on function words, thereby enhancing accessibility for lower-level learners.

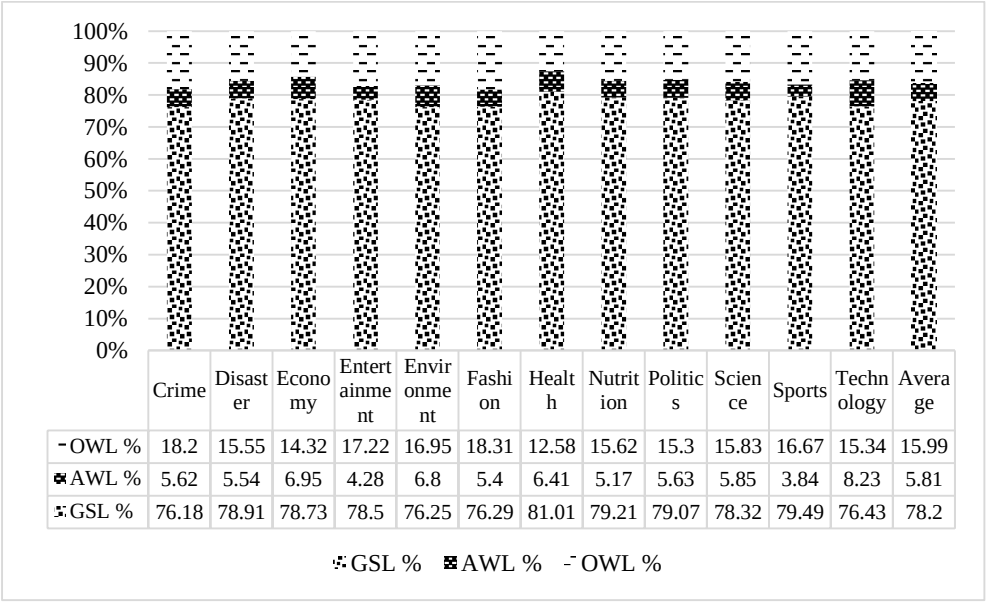
Results

The analysis of 3,000 news articles across 12 categories (Crime, Disasters, Economy, Entertainment, Environment, Fashion, Health, Nutrition, Politics, Science, Sports, and Technology) published by CNN and BBC, yielded a total of 2,510,699 running words. To address the research question, five lexical characteristics were examined, as detailed below.

Lexical Coverage in Different News Categories

Figure 1 illustrates the distribution of vocabulary across three reference lists: GSL, AWL, and OWL.

Figure 1
Lexical profiling of 12 news categories



	Crime	Disaster	Economy	Entertainment	Environment	Fashion	Health	Nutrition	Politics	Science	Sports	Technology	Average
OWL %	18.2	15.55	14.32	17.22	16.95	18.31	12.58	15.62	15.3	15.83	16.67	15.34	15.99
AWL %	5.62	5.54	6.95	4.28	6.8	5.4	6.41	5.17	5.63	5.85	3.84	8.23	5.81
GSL %	76.18	78.91	78.73	78.5	76.25	76.29	81.01	79.21	79.07	78.32	79.49	76.43	78.2

Figure 1 shows that high-frequency vocabulary (GSL) dominates across all news categories, with an average coverage of 78.20%. This suggests that most vocabulary in the analysed articles consists of commonly used words, contributing to their accessibility for a wide range of readers. In contrast, AWL usage is much lower, averaging 5.81%, with noticeable variation between categories. The Technology category contains the highest proportion of academic words, reflecting its emphasis on technical and specialised content. Conversely, the Sports category features the lowest AWL proportion, indicating a more informal, conversational style.

The OWL, which includes domain-specific and infrequent vocabulary, accounts for an average of 15.99% across all categories. Fashion articles contain the highest proportion of OWL words, highlighting the presence of specialised jargon and stylistic terminology. In contrast, the Health category contains the fewest OWL items, suggesting its use of more accessible, general-purpose vocabulary. These patterns indicate that categories such as Technology and Fashion may present higher lexical challenges for lower-proficiency L2 learners due to their reliance on specialised vocabulary. Meanwhile, categories like Health and Politics, which rely more heavily on general vocabulary, are comparatively easier to comprehend.

Examples 1–4 illustrate vocabulary patterns in categories with the highest and lowest proportions of AWL and OWL terms. In these examples, GSL words are shown in regular font, AWL words in bold, and OWL words are underlined:

Example 1: Highest AWL coverage (Technology category)

Al chatbots are also excellent at **summarizing text**. Both **versions** of ChatGPT (3.5 and 4) can **summarize** passages of **text** up to about 3,000 words. Claude 2, however, can **summarize** up to about 75,000 words, which covers the length of almost all **academic research** papers, according to Korinek. **Economists** can ask the chatbot questions on a **specific** paper such as "What are the **author**'s main **conclusions**?" or "What is the **specific evidence** supporting these points?" **Economic research** typically **involves technical tasks** such as **coding** and devising mathematical proofs. GenAI tools, such as ChatGPT Advanced **Data Analysis**, are useful at writing, explaining, translating and even debugging **code**, especially in languages such as python and R. Chatbots can set up **economic** models, **derive equations**, and explain them, though Korinek noted that genAI's **capabilities** related to math are limited at this point.

Example 2: Lowest AWL coverage (Sports category)

Now, though, there is an added complication for the game's **Generation** *Z* to consider during his absence, many admitted that they would have loved to have witnessed and gone up against—Woods in his **prime**. At the Masters, they were treated to a blast from the past. I think it is an honor to have that extra little **status** of defending champ. Ilt is what you gear your whole year around and that is something I enjoy. I enjoy the added pressure. I enjoy everything that comes along with it and I try to **focus** in on it and making the best of those weeks. It is just good for the game. I think it's good for the fans, the sponsors, the events. It's good for the tour, it's good for the players, it's good for everybody that he's winning. As players, we all knew that he was back a long time ago.

Example 3: Highest OWL coverage (Fashion category)
The garment has quickly become an essential part of the holidays, ubiquitous as Christmas lights and wrapping paper. It's obnoxious and tacky, but also fuzzy and kind of wholesome the fashion **equivalent** of a Hallmark Christmas movie (with a healthy dose of tongue-in-cheek). It took some time for the UCS to find its place in the pantheon of Christmas fundamentals, however. Christmas-themed pullovers started making an appearance in the 1950s, a nod perhaps to the holiday's growing commercialization. **Initially** referred to as "Jingle Bell Sweaters," they weren't as garish as today's iterations, and found little popularity in the market, although some TV personalities notably crooners Val Doonican and Andy Williams - really embraced the ugly side of the festive topper. The resurgence didn't last long. In the 1990s the Christmas sweater faded in popularity; it was something only your unfashionable older relatives would ever think of wearing or gifting. By the turn of the new millennium, the **item** was widely considered an eyebrow-raising sartorial mishap.

Example 4: Lowest OWL coverage (Health category)
The questions were **designed** to **establish** if I was alcohol dependent, which wasn't a surprise, as I'd applied to the trial to deal with problems I had with drinking. But there were two questions that left me feeling raw and outraged. "Have you ever thought of harming yourself or trying to take your own life?" asked the **researcher**. I struggled to work out why she was asking. Everyone has such thoughts sometimes, I had **assumed**. I really couldn't understand the **significance**. But I **reluctantly** answered "yes". Then came two follow up questions: "How often did I think that?" and "Did I have plans for suicide?" I had such thoughts earlier that day and on many days, I said, but without any plans. But as strange and intrusive as this line of questioning felt, what I later discovered about how my answers would be **interpreted** was both puzzling and upsetting.

Grouping of News Categories by Lexical Level and CEFR Level

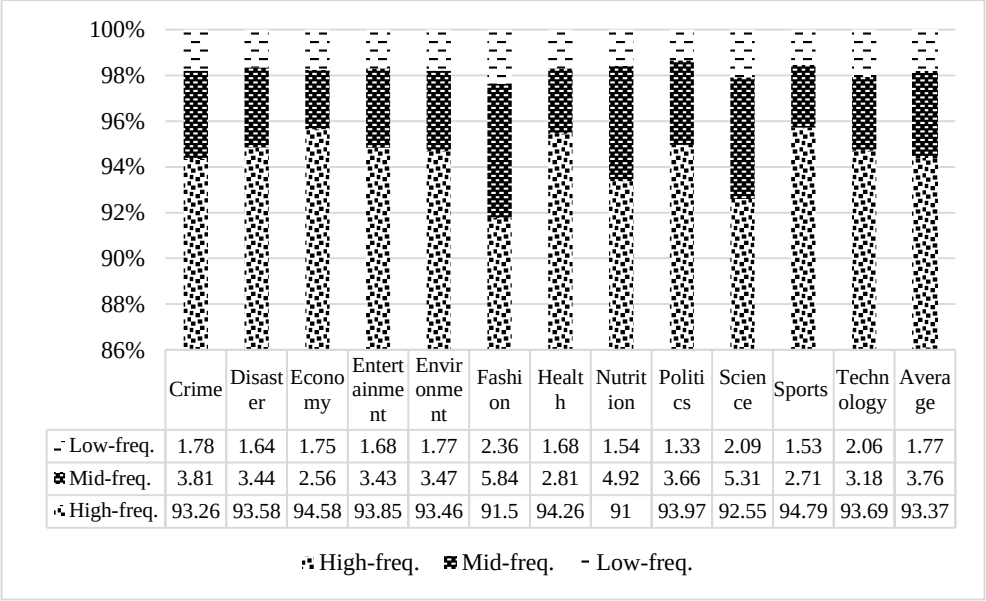
To investigate the lexical level and CEFR level of the news categories, the 12 news categories were first analysed using VocabProfile to allocate the vocabulary in each category to different levels. Then, the results of the analysis, showing the proportion of vocabulary coverage at each level, were grouped and presented in two key aspects. The first aspect involved grouping based on frequency tiers, while also determining the vocabulary levels required to achieve 95% and 98% coverage, respectively. The second aspect mapped the vocabulary levels to CEFR levels, presenting the proportions covered by each CEFR level.

Lexical Level and Lexical Coverage

The analysis of vocabulary in the 12 news categories provided a detailed breakdown of lexical levels across the K1-K25 range, including off-list words. Vocabulary was classified into three frequency tiers: high-frequency words

(K1-K3), mid-frequency words (K4-K9), and low-frequency words (K10-K25), with additional off-list items. The distribution is illustrated in Figure 2.

Figure 2
Lexical level of 12 news categories



	Crime	Disaster	Economy	Entertainment	Environment	Fashion	Health	Nutrition	Politics	Science	Sports	Technology	Average
Low-freq.	1.78	1.64	1.75	1.68	1.77	2.36	1.68	1.54	1.33	2.09	1.53	2.06	1.77
Mid-freq.	3.81	3.44	2.56	3.43	3.47	5.84	2.81	4.92	3.66	5.31	2.71	3.18	3.76
High-freq.	93.26	93.58	94.58	93.85	93.46	91.5	94.26	91	93.97	92.55	94.79	93.69	93.37

Figure 2 highlights the dominance of high-frequency words (K1–K3), which account for 91% to 94.79% of vocabulary across all categories. This indicates that the majority of news content relies on core vocabulary, enhancing its accessibility to general readers and L2 learners. However, distinctions emerge in the use of mid- and low-frequency vocabulary, which reflect more specialised or technical language. Mid-frequency words (K4–K9) comprise 2.56% to 5.84% of vocabulary across categories. Fashion and Science news exhibit the highest proportions, suggesting their reliance on moderately specialised vocabulary that may be more familiar to readers with subject-specific knowledge. Low-frequency vocabulary, including off-list words, is the least represented, ranging from 1.33% to 2.36%. Once again, the Fashion category leads in low-frequency usage, reflecting the presence of specialised and culturally specific terminology. In contrast, Nutrition and Politics contain fewer low-frequency words, relying more on general vocabulary, thereby increasing their readability.

These results suggest that while high-frequency words ensure baseline comprehensibility, the varying proportions of mid- and low-frequency words introduce different levels of lexical complexity across categories. For L2 learners, this allows for strategic selection of reading materials aligned with

their proficiency levels. Beginner learners benefit most from categories such as Politics, Nutrition, and Crime, where the dominance of high-frequency words ensures ease of understanding. Intermediate learners may be suited to categories like Environment and Economy, which incorporate more mid-frequency vocabulary and present a moderate challenge. Advanced learners can benefit from categories such as Fashion and Science, which contain a greater share of mid- and low-frequency words, offering exposure to field-specific terminology and enhancing higher-level reading skills.

In addition to frequency tier analysis, corpus coverage thresholds were calculated to determine the lexical knowledge required for effective comprehension of each category. Table 2 presents the vocabulary levels necessary to achieve 95% and 98% coverage—representing the minimum and optimal comprehension thresholds, respectively.

Table 2
Lexical coverage of 12 news categories

Category	Coverage	
	95%	98%
Crime	K4	K8
Disaster	K4	K9
Economy	K4	K8
Entertainment	K4	K9
Environment	K4	K9
Fashion	K5	K13
Health	K4	K8
Nutrition	K6	K13
Politics	K4	K7
Science	K5	K10
Sports	K4	K8
Technology	K4	K10

The table demonstrates that categories such as Crime, Economy, and Health require knowledge of vocabulary up to K4 for 95% coverage, making them relatively accessible. In contrast, Fashion and Nutrition require knowledge up to K5 or K6, reflecting greater lexical complexity. To achieve 98% comprehension, often considered sufficient for comfortable reading, most categories require vocabulary knowledge up to K8 or K9, while Fashion and Nutrition require knowledge up to K13. These results correspond with Figure 2, reinforcing that categories with greater reliance on high-frequency words demand lower lexical thresholds, whereas those with higher proportions of mid- and low-frequency words require broader lexical knowledge.

This analysis underscores the specialised nature of categories such as Fashion and Science, where technical and field-specific terms are more prevalent. Conversely, categories like Politics and Sports tend to use simpler, more general vocabulary, improving accessibility for L2 learners at lower proficiency levels. For L2 learners, aligning reading choices with the lexical demands of each category can facilitate incremental vocabulary development. Beginner learners should focus on Politics, Crime, and Sports to build confidence through exposure to high-frequency vocabulary. Intermediate learners can engage with Environment and Science for moderate lexical expansion. Advanced learners are encouraged to explore Fashion and Nutrition, where exposure to complex and less frequent vocabulary supports advanced reading fluency and lexical growth. Examples 5–7 illustrate the use of vocabulary from different frequency bands. High-frequency words are presented in regular font, mid-frequency words in bold, and low-frequency/off-list words are underlined.

Example 5: Category with high proportion of high-frequency word (Sports category)

The decision, which comes as the sporting world continues to impose sanctions on Russia and Belarus following the former's invasion of Ukraine, means athletes from the two countries will compete under the Paralympic flag and will not be included in the **medal** table. The Russian delegation must cover the Russian Paralympic Committee symbol on their uniforms in all official ceremonies and sporting competitions, the IPC said. The Belarus delegation must also cover the Belarus flags on their uniforms. The Olympic **Truce** dates back almost 3,000 years to the early days of the ancient games when the leaders of three Greek city states agreed to limit their battles. In modern times, the **Truce** has been **invoked** as a universal goal by the **UN** regularly since 1993. On January 28 **UN** Secretary General António Guterres asked nations around the world to lay down their weapons and observe an Olympic **Truce** from seven days before the Beijing Games until seven days after the end of the Paralympic Games.

Example 6: Category with high proportion of mid-frequency word (Science category)

"This study is part of ESA's comprehensive plan to make Europe a partner in global exploration in the next decade," Parker said. The mission would be a collaboration between **aerospace** scientists and **technicians** in France, Germany and Belgium. The project is now in the research phase, with scientists hoping to use an Ariane 64 **rocket** in coming years to send mining equipment to the moon. The announcement coincided with Monday's **lunar eclipse**, which treated stargazers across the **globe** to a red super blood wolf moon. The project is also part of a wider effort to **commemorate** the 50th **anniversary** of **mankind**'s first steps on the moon.

Example 7: Category with high proportion of low-frequency words (Fashion category)

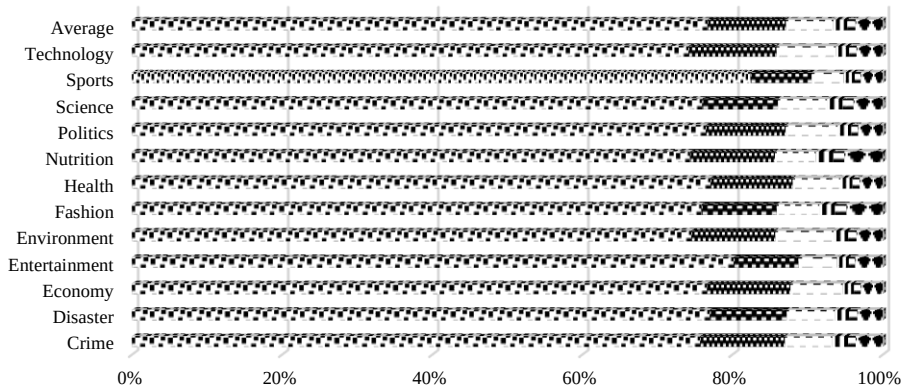
While indigo is arguably the most recognized **dye** - the plant which colored King Tutankhamun's burial **shrouds** and more recently makes your **denim** blue - there are dozens of other dyestuffs that have **incited** murder and subterfuge, made and lost fortunes and turned clothes into a status symbol for thousands of years. In a new book exploring the history of **dyes**, author and **textile** designer Lauren MacDonald weaves together the stories and science of color dating from **pre**-history to today; from the time of natural **dyeing** to modern **synthetic** production. "It's been (at least) 26,000 years since humans started to dye," the author writes. "Your great grandparents (999 removed) were stirring a **bubbling** vat of **dye**... while Woolly **mammoths** and saber-toothed cats **roamed** the earth." Indeed, in 2009, scientists found fibers of **dyed** flax up to 34,000 years old in a **cave**.

CEFR

To determine the CEFR levels represented in each news category, the lexical level results were mapped using Nation and Crabbe's (1991) framework. Figure 3 displays the CEFR classification of vocabulary across the 12 news categories.

Figure 3

CEFR of Vocabulary Used in 12 News Categories



	Crime	Disaster	Economy	Entertainment	Environment	Fashion	Health	Nutrition	Politics	Science	Sports	Technology	Average
A1	75.68	76.96	76.57	80.14	74.52	75.84	77.01	74.32	76.45	75.94	82.41	74.12	76.66
A2	11.30	10.25	11.09	8.62	11.18	10.02	10.93	11.29	10.63	10.10	8.12	11.71	10.44
B1	6.28	6.38	6.92	5.10	7.76	5.64	6.33	5.40	6.90	6.51	4.27	7.87	6.28
B2	2.12	1.91	1.52	1.60	1.94	2.10	1.61	2.37	1.84	2.09	1.47	1.78	1.86
C1	1.01	1.10	0.90	0.82	1.10	1.52	0.95	1.43	0.89	1.45	0.80	0.98	1.08
C2	3.57	3.35	2.90	3.66	3.47	4.80	3.04	5.10	3.26	3.86	2.90	3.46	3.61

A1 A2 B1 B2 C1 C2

Figure 3 categorises vocabulary in news articles by CEFR levels, revealing a dominance of basic vocabulary (A1 level) across all categories, averaging between 74% and 82%. This reflects an effort to make news articles widely accessible through commonly understood language. The Sports category has the highest proportion of A1-level vocabulary (82.14%), consistent with its informal tone. In contrast, Nutrition includes the most advanced vocabulary at the C2 level (5.10%), indicating the presence of specialised terminology.

B1-level vocabulary remains modest across all categories (4-7%), and C2-level vocabulary is generally sparse (2-5%), except in technical or specialised categories such as Nutrition, Science, and Fashion. These results have important implications for L2 learners. Beginners are likely to benefit from categories such as Sports and Entertainment, which predominantly contain A1-level vocabulary and provide an accessible entry point for reading practice. Intermediate learners, particularly those at the B1 or B2 level, may find Politics and Environment useful, as these contain slightly more advanced vocabulary and support gradual lexical development. Advanced learners aiming to refine their understanding of complex language should focus on categories such as Nutrition and Science, where the presence of C2-level vocabulary and technical terms offers opportunities to build a more sophisticated lexicon.

As the CEFR consists of six levels, and some categories show their highest coverage across multiple levels, representative examples were selected to illustrate each band. The Sports category is used to illustrate A1-level vocabulary; Technology represents A2-B1; Nutrition reflects B2; and Fashion exemplifies the C1-C2 range. In Examples 8-11, words at the A1 level are shown in regular font, A2-B1 in bold, B2 in italics, and C1-C2 are underlined.

Example 8: Category with high proportion of A1 level (Sports category)
I think we had great **hope to** start the **season** and *belief*, which we *didn't* really have the **previous** two. It **felt like** we had a **real chance** to kind of **get back** in the **fight** and I think our players felt that **way**, too. I also **felt like** We had <u>refilled</u> our cups after **missing** the <u>playoffs</u> the last two years. **Everyone** <u>came into</u> this **season** really hungry, and you **could** see the **way** we started ... we **shot out of** the **gates**, there's a great **connection** on this **team**, great <u>camaraderie</u>. In the **regular season**, we *literally* never got our **main guys** on the floor at the same time until Game 1 of the <u>Denver</u> **series**. **So** it was **sort** of a <u>rocky</u> **path** to get here, but I feel good about the *process* and our *potential* **if** we **could** get all of our key **guys** on the floor.

Example 9: Category with high proportion of A2-B1 levels (Technology category)

The *upgrades* come at a time when Apple is **gaining** *substantial* **ground** in the **traditional PC** and **laptop market** but **still** ranks **fourth** behind Lenovo, Dell and HP in the number of **products** shipped, **according to** IDC **Research**. Apple said it shipped about 26 **million** Macs in 2022, **making up** 9.1% of the *overall* **market** (up from 7.8% the year prior). At the same time, the *overall* **PC market** shipped about 292 **million** computers the same year, down 15% from the year before. The iPhone and Apple **services** such as iCloud, Apple TV+ and Apple Music—**remain** a *major* revenue driver for the **company**, but Macs and iPad **sales** have *declined* year **over** year, *largely* **due** to **weaker demand**, excess inventory and a worsening macroeconomic **climate**.

Example 10: Category with high proportion of B2 level (Nutrition category)

Keto is short for ketosis, a metabolic *state* that *occurs* when your *liver* begins to use *stored* fat to **produce** ketones for **energy**. The *liver* is **programmed** to do that when your body **loses access** to its **preferred fuel** – carbohydrates — and thinks it's *starving*. The **diet** has **actually** been **around since** the 1920s, when a doctor stumbled on it as a **way** of **controlling** seizures in children with epilepsy who *didn't respond* to other *treatment* **methods**. It was **recognized** long **ago** that *denying* the **brain access** to glucose, and *converting* to ketone-**based** metabolism, dampens **brain electrical activity**. But why on **Earth** would you want to dampen **brain electrical activity unless** you had refractory epilepsy? **Creating** ketosis is not as **simple** as *it seems*. Your *liver* is only *forced* into **producing** ketones when carb intake is drastically slashed.

Example 11: Category with high proportion of C1-C2 levels (Fashion category)

Fashion *brands* want you to be **able** to *identify* counterfeits. To **become** a "*master*" authenticator at Fashionphile- the **highest level** of **training** to weed **out designer** fakes at the luxe **online** marketplace takes more than 8,000 hours of rigorous schooling, **according to** the **company**. Trainees learn to **quickly spot** an *error* in the date **format** inside a Louis Vuitton bag, or know the **correct metal** alloy makeup of a Cartier watch. Its **competitor**, The RealReal, also *relies on* **human senses** and instinct—**recognizing** the **smell of** a $25,000 Hermès Birkin bag, or the feel of its **smooth** Barenia **leather**—but the retailer's first **round** of **checks** is undertaken **through** AI, with **software trained** on 30 **million** *images* to discern **nearly** imperceptible **differences** in the stitching or the placement of **hardware**. At the same time, an algorithm *calculates* an **item**'s *risk* **based** on **everything** from the consignor's **selling history** to the *popularity* of a **product** on the black **market**.

Ratios of Lexical Variation and Lexical Density in News Categories

Lexical variation and lexical density provide deeper insight into the characteristics of vocabulary usage in news articles. Unlike lexical profile, lexical level, and CEFR alignment, these metrics do not directly indicate vocabulary difficulty. Instead, they reflect the diversity and repetitiveness of vocabulary in a given text, which are factors that influence the potential for vocabulary acquisition and learner engagement. These features can help determine which categories are most suitable for L2 learners at different proficiency levels.

Figure 4

Lexical Variation of 12 News Categories

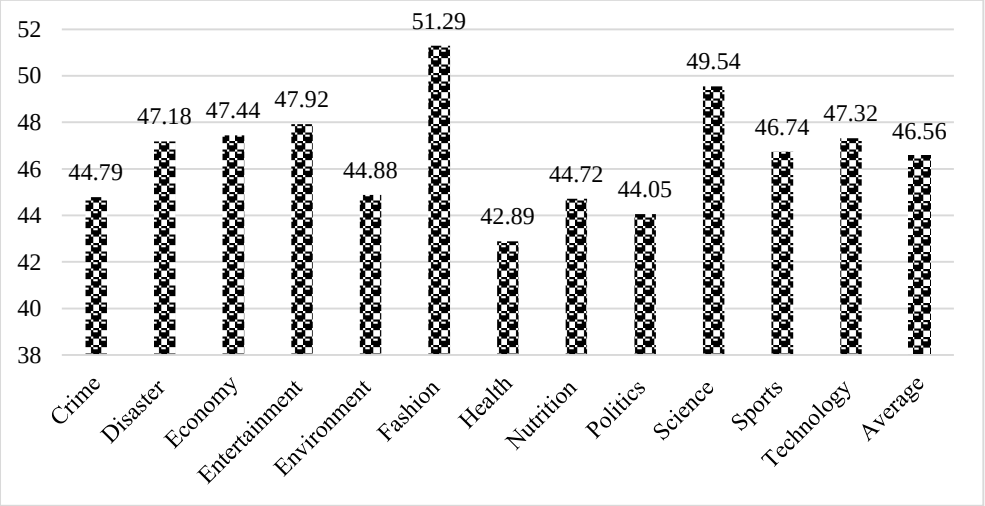


Figure 4 illustrates lexical variation, measured using the Type-Token Ratio (TTR), across the 12 news categories. The average TTR was 46.56%, indicating that nearly half of the words in news texts are unique, reflecting a moderate level of lexical variation overall. However, variation between categories was substantial.

Fashion demonstrated the highest lexical variation (51.29%), suggesting a rich and diverse vocabulary often driven by creative and descriptive language. This diversity may be attributed to the dynamic and expressive nature of fashion journalism. By contrast, Health showed the lowest TTR (42.89%), possibly due to the repeated use of technical or topic-specific terms, which can aid accessibility while reducing lexical variety.

It is worth noting that this study draws from BBC and CNN, which represent different English dialects, British and American English, respectively. These dialectal differences may influence lexical variation, providing learners

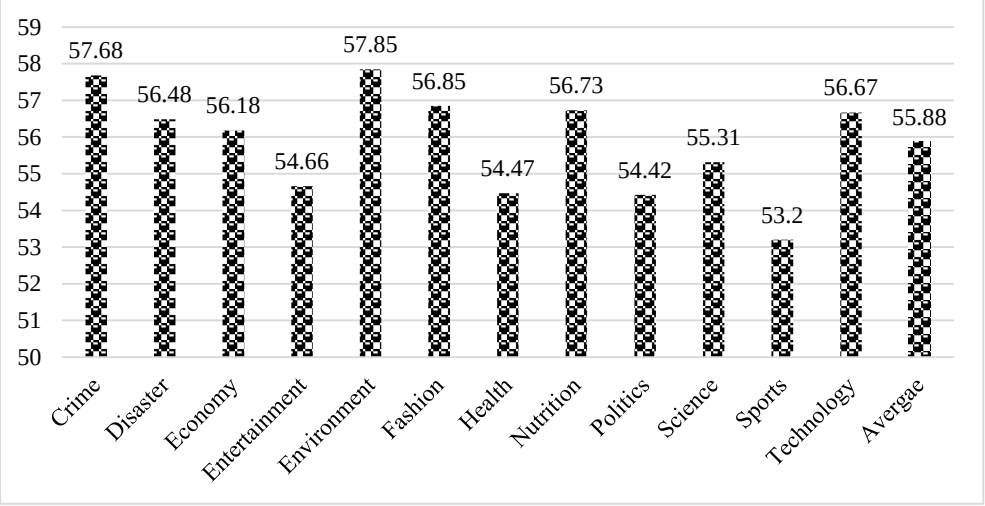
with exposure to a broader range of English usage. Such variation can support vocabulary development by familiarising learners with lexical patterns from multiple varieties of English.

In terms of pedagogical application, learners at the beginner level may benefit from reading categories such as Health and Technology, where lower lexical variation and repetitive terminology support vocabulary reinforcement. Intermediate learners could explore categories such as Sports and Environment, which offer a balanced level of lexical variation. Advanced learners may be challenged by categories like Fashion and Entertainment, where creative and varied vocabulary supports lexical expansion and more nuanced comprehension. Example 7 illustrates the use of the near synonyms "support" and "encourage" in a news article.

> **Example 7: The highest maximum TTR (Entertainment category)**
> "Covid Bristol music venues and nightclubs "need more **support**" Nightclubs and music venues say more financial **encourage** is needed to help protect their sector."

Lexical Density

Figure 5
Lexical density of 12 news categories



The analysis revealed that the lexical density varied only slightly across the 12 news categories, with the relatively narrow range suggesting limited lexical contrast (Figure 5). The highest lexical densities were in Environment (57.85%) and Crime (57.68%), while the lowest was in Sports (53.2%). However, the percentage differences are small and do not reflect any major contrasts. This limited variation may be attributed to the distinctive nature of language use in

news discourse, which typically maintains a consistent balance between content and function words across topics. Since lexical density measures only the proportion of these two-word types, it may not fully reflect the deeper lexical variation that exists between categories.

Discussion

When selecting a source for primary or supplementary use in vocabulary instruction, it is essential to assess that source across multiple lexical dimensions. Prior research has typically examined vocabulary through individual aspects such as lexical profile, lexical level, lexical variation, or lexical density (Coxhead, 2000, 2018; Coxhead & Hirsch, 2007; Liu, 2021; Meebangsai et al., 2023; Nation, 2016, 2022; Treffers-Daller et al., 2018). However, most studies have addressed only one or two of these dimensions in isolation (Coxhead & Walls, 2012; Ha, 2022b; Li et al., 2024; Madarbakus-Ring & Benson, 2024; Vuković-Stamatović & Čarapić, 2024). By incorporating a broader set of criteria, a more comprehensive understanding of lexical characteristics can be achieved, which facilitates more targeted recommendations for learners with varying proficiency levels or domain-specific interests. The present study expands this scope by analysing five lexical aspects: lexical profile, lexical level, CEFR alignment, lexical variation, and lexical density. This approach provides pedagogically relevant insights for both classroom instruction and autonomous learning.

The GSL is estimated to account for approximately 80% of vocabulary in general texts (West, 1953; Nation & Waring, 1997). In this study, the average GSL coverage across all news categories was 78.2%, closely aligning with this benchmark, though no single category exceeded the 80% threshold. Health came closest (81.01%), while Fashion (76.29%) and Crime (76.18%) featured more vocabulary beyond the GSL, reflecting a broader lexical range. The AWL, typically expected to account for about 10% of academic text vocabulary (Coxhead, 2000), averaged only 5.81% across the news categories. Even in more technical categories, such as Technology (8.23%) and Environment (6.8%), AWL coverage fell short of expectations. This underscores the AWL's limited applicability to journalistic writing, which tends to prioritise general accessibility over academic rigor. In contrast, the OWL accounted for an average of 15.99% of the vocabulary, exceeding the anticipated 10% (Coxhead & Hirsch, 2007). OWL usage was particularly high in categories such as Fashion (18.31%) and Crime (18.20%), suggesting that specialised or context-specific vocabulary is frequently used in journalistic writing and must be accounted for in instructional design.

Understanding 95% of vocabulary in a text is generally regarded as the minimum threshold for comprehension, with 98% coverage considered optimal

for fluent reading (Coxhead & Demecheleer, 2018; Hu & Nation, 2000; Schmitt et al., 2011; van Zeeland & Schmitt, 2013). Based on the present analysis, achieving 95% lexical coverage across most news categories requires knowledge of approximately 4,000 to 6,000 word families, averaging 4,300. This aligns with previous estimates that 4,000 high-frequency word families, supplemented by proper nouns, are sufficient to achieve 95% coverage in newspapers and novels. For 98% coverage, the average lexical requirement increases to 9,300 word families, supporting Nation's (2006) estimation of 8,000 to 9,000 word families. However, certain categories deviate notably from this pattern. For instance, Fashion and Nutrition require substantially more lexical knowledge, with 5,000-6,000 word families needed for 95% coverage and up to 13,000 for 98%. These figures point to the presence of specialised vocabulary not adequately captured by general high-frequency lists. Such variability underscores the limitations of relying solely on generalised vocabulary thresholds when analysing diverse or domain-specific content. While previous research, such as the analysis of the NOW corpus by Ha (2022b), supports the 4,000-word family benchmark for general news, the present study shows that this does not apply uniformly across all categories. For example, Science and Technology require a minimum of 5,000 word families for 95% comprehension, highlighting the need for more nuanced, category-specific lexical assessments.

Lexical density across the analysed categories averaged 55.88%, aligning with the 40–65% range typical of nonfiction texts as reported by Stubbs (1986). A critical comparison with this benchmark indicates both consistency and areas for further inquiry. The relatively narrow range of lexical densities suggests that news articles tend to balance information content with readability, which is an expected characteristic of journalistic writing intended for a general audience. Among the categories, Environment (57.85%) and Fashion (56.85%) exhibited the highest lexical densities, suggesting a greater concentration of content words and potentially more cognitively demanding texts. This may reflect the technical vocabulary often present in environmental journalism or the descriptive, trend-specific language typical of fashion writing. Conversely, Sports (53.2%) and Entertainment (54.66%) had lower lexical densities, indicating a more conversational tone that prioritises accessibility over technical precision. It is important to note, however, that lexical density alone does not fully capture text complexity. Texts with high density may still be accessible due to supportive context, familiar topics, or simplified syntax. Conversely, texts with moderate density may be more challenging if they include rare or domain-specific vocabulary. For example, Nutrition (56.73%) and Technology (56.67%) had similar lexical densities, yet their specialised terminology likely demands higher levels of reader expertise.

Pedagogical Implication

The vocabulary analysis of 12 news categories provided insights into five key aspects of lexical assessment: lexical profile, lexical level, CEFR alignment, lexical variation, and lexical density. These findings offer a basis for aligning news content with L2 English learners at different proficiency levels (beginner, intermediate, and advanced) and for deriving practical pedagogical recommendations.

For beginner learners, the most suitable categories are Sports, Health, and Entertainment. These categories featured high proportions of high-frequency vocabulary (over 94%), strong alignment with CEFR A1 and A2 levels (above 75%), and relatively low lexical density (below 55%), enhancing accessibility. Pedagogically, beginner learners can benefit from using graded news reading applications or websites offering simplified versions of these categories. Supplementary tools such as flashcards and wordlists may reinforce foundational vocabulary (Chung, 2009).

Intermediate learners would benefit from categories such as Environment, Disaster, and Technology, which balance mid-frequency vocabulary with moderate lexical variation and CEFR B1 coverage (approximately 6-8%). These categories support vocabulary expansion while remaining accessible to learners transitioning from lower levels. Effective strategies include extensive reading of recommended categories alongside the use of digital tools for collocation and corpus-based exploration. Self-directed tasks such as glossary development and summary writing can further enhance vocabulary retention (Na Ayutthaya et al., 2022).

Advanced learners are best suited to categories such as Science, Fashion, and Nutrition. These categories include higher proportions of low-frequency words, substantial lexical variation (above 49%), and significant representation of CEFR B2, C1, and C2 vocabulary. Their higher lexical density (above 56%) also signals increased cognitive demands. Advanced learners can benefit from reading authentic, ungraded news texts, supported by tools such as concordances and advanced online dictionaries to facilitate the acquisition of nuanced and domain-specific vocabulary (Laosrirattanachai & Laosrirattanachai, 2025b). Tasks such as writing critical responses or producing news-style articles can further deepen their lexical mastery.

The findings of this study also hold value for CLIL (Content and Language Integrated Learning) and EMI (English-Medium Instruction) contexts, particularly in Thai higher education across disciplines such as Law, Medicine, Sports Science, Fashion, and Fine Arts. Learners in these subject areas benefit from exposure to authentic OWL vocabulary relevant to their academic domains. Instructors may begin by evaluating the lexical characteristics of relevant news content, selecting materials that span a range of difficulty levels. Vocabulary

instruction can be scaffolded by introducing high-frequency GSL items at the CEFR A1 level and gradually incorporating OWL vocabulary as learners progress. This stepwise approach enables the systematic integration of lexical profiling, frequency, CEFR alignment, and variation into content-specific vocabulary instruction.

It is important to note that the recommendation of news categories by proficiency level in this study is intended primarily to support vocabulary development. However, this structured approach may not always align with learners' individual interests. Teachers, acting as facilitators, and learners themselves, as active participants, should consider the overarching goal of vocabulary growth when selecting reading materials. If learner motivation is driven by personal interest, prioritising relevance over lexical simplicity may be more effective. Conversely, if the goal is to improve vocabulary knowledge and enhance comprehension of global issues, a gradual progression from simpler to more complex categories remains a sound, research-informed strategy.

Conclusion, Limitations, and Recommendations for Future Studies

This study examined the extent to which lexical characteristics, specifically lexical profile, lexical level, CEFR level, lexical variation, and lexical density, differ across 12 English news categories, and how these differences can inform the selection of pedagogically suitable texts for L2 learners at varying proficiency levels. Through a large-scale lexical analysis of 3,000 news articles, the study offers a comprehensive empirical account of lexical variation across news categories.

The findings revealed significant differences in lexical profile, lexical level, CEFR alignment, and lexical variation, while lexical density remained relatively stable. Lexical profiling highlighted the predominance of high-frequency vocabulary (GSL and K1-K3), which supports general accessibility. However, categories such as Fashion and Science featured higher proportions of mid- and low-frequency words, along with academic or domain-specific vocabulary (AWL and OWL), indicating increased lexical complexity. The CEFR-based analysis reinforced this distinction: beginner-friendly categories such as Sports and Health were dominated by A1-level vocabulary, while categories such as Nutrition and Fashion contained a larger share of C2-level vocabulary.

Lexical variation was also greater in categories with creative or technical content (Fashion, Entertainment), while categories with more formulaic content (Health, Technology) showed lower variation. Although lexical density exhibited only minor differences across categories, it contributed to a broader understanding of vocabulary complexity in news texts. Overall, the study provides a useful

reference for educators and learners in selecting authentic news materials tailored to vocabulary learning goals at different proficiency levels.

Despite the contributions of this research, several limitations should be noted. First, the analysis focused solely on two prominent news organisations (BBC and CNN). While both are highly credible, they may not fully represent the global diversity of news media, particularly from regional or non-mainstream sources. This could limit the generalisability of the findings. Second, the study employed a fixed framework of 12 news categories based on Kaspar and Fuchs (2021). Although reliable, this framework may not capture the full range of evolving news genres, such as opinion pieces or multimedia content, potentially omitting relevant lexical patterns. Third, the dataset was limited to news articles published between 2019 and 2023. This timeframe may not reflect longer-term lexical trends or the impact of major historical events prior to 2019.

Future research should consider expanding the scope of analysis to include regional and non-mainstream news outlets. This would support a more comprehensive understanding of vocabulary usage across diverse contexts and writing styles. It is also recommended that future studies explore additional or alternative news categories to better reflect the dynamic nature of news discourse. Extending the timeframe of data collection would allow for longitudinal analyses that examine diachronic changes in vocabulary, including the influence of significant historical, technological, or societal events. Further studies might also investigate the role of dialectal variation by comparing lexical characteristics across news outlets representing different English varieties. Such work could offer deeper insights into regional differences in lexical choice and usage. Finally, experimental research should be conducted to evaluate the pedagogical impact of using different news categories at various proficiency levels. This would help bridge the gap between corpus-based lexical analysis and practical vocabulary instruction in L2 settings.

References

Abyaad, R., Kabir, M. R., & Hasan, S. (2020). A novel approach to categorize news articles from headlines and short text. *Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Bangladesh,* 2020, 162–165. https://doi.org/10.1109/TENSYMP50017.2020.9230675

Anthony, L. (2024). *AntWordProfiler* (Version 2.2.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntWordProfiler

Astika, G. (2015). Profiling the vocabulary of news texts as capacity building for language teachers. *Indonesian Journal of Applied Linguistics, 4*(2), 123–134. https://doi.org/10.17509/ijal.v4i2.689

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20*(1), 50–57. https://doi.org/10.1111/j.1540-5826.2005.00120.x

Baker, P. (2006). *Using corpora in discourse analysis.* Bloomsbury Publishing.

Baranowska, K. (2020). Learning most with least effort: subtitles and cognitive load. *ELT Journal, 74*(2), 105–115. https://doi.org/10.1093/elt/ccz060

Bates, E., Bretherton, I., & Snyder, L. S. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge University Press.

Benigno, V., & de Jong, J. (2019). Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model. In A. Huhta, G. Erickson, & N. Figueras (Eds.), *Development in language education: A memorial volume in honour of Sauli Takala* (pp. 8–29). Jyväskylä University Printing House.

Bleyer, W. G. (1916). *Types of news writing*. Houghton Mifflin.

Chung, M. (2009). The newspaper word list: A specialised vocabulary for reading newspapers. *JALT journal, 31*(2), 159–182. https://doi.org/10.37546/JALTJJ31.2-2

Cobb, T. (2022). *Vocabprofile.* [Computer program]. http://www.lextutor.ca/vp/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2018). *Vocabulary and English for Specific Purposes research: Quantitative and qualitative perspectives*. Routledge. https://doi.org/10.4324/9781315146478

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., & Demecheleer, M. (2018). Investigating the technical vocabulary of Plumbing. *English for Specific Purposes, 51*, 84–97. https://doi.org/10.1016/j.esp.2018.03.006

Coxhead, A. & Hirsch, D. (2007). A pilot science word list for EAP. *Revue Française de linguistique appliqueé, 12*(2), 65–78. https://doi.org/10.3917/rfla.122.0065

Coxhead, A., & Walls, R. (2012). Ted talks, vocabulary, and listening for EAP. *TESOLANZ Journal, 20*(1), 55–67.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment–Companion volume.* Strasbourg.

Coyle, Y., & Gracia, R. G. (2014). Using songs to enhance L2 vocabulary acquisition in preschool children. *ELT Journal*, *68*(3), 276–285. https://doi.org/10.1093/elt/ccu015

Dang, T.N.Y., & Long, X. (2023). 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., & Lu, C. (2024). Learning academic vocabulary through reading online news. *International Review of Applied Linguistics in Language Teaching*, 1–21. https://doi.org/10.1515/iral-2023-0206

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76. https://doi.org/10.1016/j.esp.2013.08.001

Davis, G. M. (2017). Songs in the young learner classroom: a critical review of evidence. *ELT Journal*, *71*(4), 445–455. https://doi.org/10.1093/elt/ccw097

Davis, B. H., & Brewer, J., & Brewer, J. P. (1997). *Electronic discourse: Linguistic individuals in virtual space.* Suny Press.

Graves, K. (2008). The language curriculum: A social contextual perspective. *Language Teaching*, *41*(2), 147–181. https://doi.org/10.1017/S0261444807004867

Ha, H. T. (2022a). Vocabulary demands of informal spoken English revisited: What does it take to understand movies, TV programs, and soap operas? *Frontiers in Psychology*, *13*, Article 831684. https://doi.org/10.3389/fpsyg.2022.831684

Ha, H. T. (2022b). Lexical profile of newspapers revisited: A corpus-based analysis. *Frontiers in Psychology*, *13*, Article 800983. https://doi.org/10.3389/fpsyg.2022.800983

Hsu, W. (2018). Voice of America News as voluminous reading material for mid-frequency vocabulary learning. *RELC Journal*, *50*(3), 408–421. https://doi.org/10.1177/0033688218764460

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Hulstijn, J. H., Charles, A. J., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In I. Bartning, M. Martin, & I. Veddar (Eds.), *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research* (pp. 11–20). Eurosla.

Indarti, D. (2017). Lexical richness of the Jakarta Post opinion articles: Comparison between native and non-native writers. *Wanastra*, *4*(2), 138–142. https://doi.org/10.31294/w.v9i2.2550

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers*, *53*, 61–79.

Kaspar, K., & Fuchs, L. A. M. (2021). Who likes what kind of news? The relationship between characteristics of media consumers and news interest. *SAGE Open, 11*(1), 1–12. https://doi.org/10.1177/21582440211003089

Kembaren, F. R., & Aswani, A. N. (2022). Exploring lexical density in the New York Times. *Journal of English Language, Literature, and Teaching*, *7*(2), 110–119. https://doi.org/10.32528/ellite.v7i2.8795

Kyongho, H., & Nation, I. S. P. (1989). Reducing the Vocabulary Load and Encouraging Vocabulary Learning through Reading Newspapers. *Reading in a Foreign Language*, *6*, 323–335.

Laosrirattanachai, P., & Laosrirattanachai, P. (2023). Analysis of vocabulary use and move structures of the World Health Organization Emergencies press conferences on Coronavirus Disease: A corpus–based investigation. *LEARN Journal: Language Education and Acquisition Research Network, 16*(1), 121–146. https://so04.tci-thaijo.org/index.php/LEARN/article/view/263436

Laosrirattanachai, P., & Laosrirattanachai, P. (2025a). Unveiling the distinction of near synonymy: A corpus-based analysis on attempt, endeavor, strive, and try. *PASAA*, *70*, 132–163.

Laosrirattanachai, P., & Laosrirattanachai, P. (2025b). 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

Laufer, B., & Nation, I. S. 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

Li, Z., Li, J. Z., Zhang, X., & Reynolds, B. L. (2024). Mastery of listening and reading vocabulary levels in relation to CEFR: Insights into student admissions and English as a medium of instruction. *Languages*, *9*(7), 239. https://doi.org/10.3390/languages9070239

Liu, C. Y. (2021). Examining the implementation of academic vocabulary, lexical density, and speech rate features on OpenCourseWare and MOOC lectures. *Interactive Learning Environments*, *31*(8), 4924–4939. https://doi.org/10.1080/10494820.2021.1987274

Madarbakus-Ring, N., & Benson, S. (2024). TED Talks and the textbook: An in-depth lexical analysis. *Languages*, *9*(10), 309. https://doi.org/10.3390/languages9100309

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. https://doi.org/10.58837/CHULA.PASAA.67.1.3

Moore, T., Morton, J., Hall, D., & Wallis, C. (2015). Literacy practices in the professional workplace: implications for the IELTS reading and writing tests. *IELTS Research Reports Online Series*, *46.* https://ielts.org/researchers/our-research/research-reports/literacy-practices-in-the-professional-workplace-implications-for-the-ielts-reading-and-writing-tests

Na Ayutthaya, J. A., Kunthonjinda, K., Somwang, K., & Laosrirattanachai, P. (2022). Making beverage service word list for English for Specific Purposes classroom. *rEFLections*, *29*(2), 325–343. https://doi.org/10.61508/refl.v29i2.259524

Nation, I. S. P. (2006). How large a vocabulary is needed to reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* John Benjamins. https://doi.org/10.1075/z.208

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists* (Version 1.0.0) [Data file]. http://www.victoria.ac.nz/lalsstaff/paul-nation.aspxl

Nation, I. S. P. (2018, April, 10). *Resources*. https://www.wgtn.ac.nz/lals/resources.

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009093873

Nation, I. S. P., & Crabbe, D. (1991). A survival language learning syllabus for foreign travel. *System*, *19*(3), 191–201. https://doi.org/10.1016/0346-251X(91)90044-P

Nation, 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.

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

Reynolds, B. L., Xie, X., & Pham, Q. H. P. (2022). Incidental vocabulary acquisition from listening to English teacher education lectures: A case study from Macau higher education. *Frontiers in Psychology*, *13*, 1–18. https://doi.org/10.3389/fpsyg.2022.993445

Sayer, P., & Ban, R. (2014). Young EFL students' engagements with English outside the classroom. *ELT Journal*, *68*(3), 321–329. https://doi.org/10.1093/elt/ccu013

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

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

Stubbs, M. (1986). Lexical density: A computational technique and some findings. In M. Coulthard (Ed.), *Talking about text* (pp. 27–48). University of Birmingham.

Tegge, F. (2018). Pop songs in the classroom: time-filler or teaching tool? *ELT Journal*, *72*(3), 274–284. https://doi.org/10.1093/elt/ccx071

Teng, F. (2015). EFL vocabulary learning through reading BBC news: An analysis based on the Involvement Load Hypothesis. *English as a Global Language Education (EaGLE) Journal*, *1*(2), 63–90. https://doi.org/10.6294/EaGLE.2015.0102.03

Thornbury, S., & Slade, D. (2006). *Conversation: from Description to Pedagogy*. Cambridge University Press.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327. https://doi.org/10.1093/applin/amw009

Ure, J. (1971). Lexical density and category differentiation. ln G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443–452). Cambridge University Press.

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

Vuković-Stamatović, M., & Čarapić, D. (2024). Vocabulary profile, lexical density and speech rate in science podcasts: How appropriate are science podcasts for EAP and EST listening? *Ibérica*, *47*, 201–226. https://doi.org/10.17398/2340-2784.47.201

Wingrove, P. (2017). How suitable are TED talks for academic listening? *Journal of English for Academic Purposes*, *30*, 79–95. https://doi.org/10.1016/j.jeap.2017.10.010

West, M. (1953). *A general service list of English words*. Longman.