LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 18 No: 2, July – December 2025

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Dissecting Vocabulary Usage Across TED Talks Topics: A Corpus-based Perspective

Piyapong Laosrirattanachai[a], Chanaporn Baothong[b], Kotchakorn Laijud[c], Piyanuch Laosrirattanachai[d,*]

[a] piyapong.l@ku.th, Faculty of Hospitality Industry, Kasetsart University, Thailand
[b] chanaporn.ba@ku.th, Faculty of Hospitality Industry, Kasetsart University, Thailand
[c] kotchakorn.la@ku.th, Faculty of Hospitality Industry, Kasetsart University, Thailand
[d] piyanuch.la@ku.th, Faculty of Hospitality Industry, Kasetsart University, Thailand
[*] Corresponding author, piyanuch.la@ku.th

APA Citation:
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

Received 07/02/2025

Received in revised form 25/06/2025

Accepted 15/07/2025

ABSTRACT

Autonomous learning strategies, supported by diverse media, have become essential tools for enhancing vocabulary acquisition. Among these, TED Talks stand out due to their accessibility, topical diversity, and authentic linguistic input. Although previous research has highlighted the educational value of TED Talks, limited attention has been given to their vocabulary characteristics across different topics. This study addresses this gap by examining the lexical profile, CEFR alignment, lexical level, and lexical density of TED Talks across 20 topics using a corpus-based approach. A dataset of 1,000 Talks (2,348,137 tokens) was analysed using

computational tools and established linguistic frameworks. The findings reveal significant variation in vocabulary characteristics among topics, indicating their suitability for learners at different proficiency levels. Topics such as Relationships and Psychology cater to beginners; Education and Social Change are appropriate for intermediate learners; while advanced learners benefit from topics like Science and Government & Politics, which exhibit higher lexical density and complexity. The study highlights the importance of aligning TED Talk topics with learners' proficiency levels to optimise vocabulary acquisition. These findings provide practical guidance for educators and learners, supporting targeted language development through the tailored selection of TED Talks.

Keywords: vocabulary, TED Talks, corpus-based, lexical characteristics

Introduction

Vocabulary acquisition is central to mastering a second language (L2), underpinning communication, comprehension, and overall proficiency (Webb & Nation, 2017). For English learners, a broad vocabulary is vital for academic (Warnby, 2023), professional (Drayton & Coxhead, 2023; Roesler, 2021), and everyday success (Hanks et al., 2024). Research highlights the link between vocabulary knowledge and fluency, emphasising the need for learners to develop a diverse lexicon (Le & Miller, 2023; Otto, 2021). While traditional methods such as rote memorisation remain useful, autonomous learning strategies have gained increasing importance (Boers et al., 2023; Gay, 2022). These strategies empower learners to take control of their learning, thereby enhancing motivation and retention (Kim et al., 2024). A variety of media enables learners to develop vocabulary in context, making the learning process more dynamic and effective. Media such as movies, songs, podcasts, news articles, novels, games, and TED Talks provide authentic linguistic input and contextualised examples of vocabulary use. Among these, TED Talks have become particularly valuable instructional resources due to their accessibility, topical diversity, and potential to engage learners with meaningful content (Coxhead, 2018). Delivered by experts across disciplines, TED Talks expose learners to a range of accents, registers, and speaking styles, enriching both vocabulary and listening comprehension (Liu, 2023).

Features such as subtitles and transcripts allow learners to study at their own pace, supporting autonomous learning. Additionally, the engaging narratives often found in TED Talks help foster intellectual curiosity and support language development.

With the growing popularity of TED Talks in language learning, several studies have examined their effectiveness, particularly in academic and vocabulary development contexts. For example, TED Talks share a vocabulary profile with newspapers and novels, aligning with both academic and basic vocabulary lists (Wang, 2012), although their coverage of academic vocabulary remains limited (Coxhead & Walls, 2012). Strategies have been proposed to optimise vocabulary instruction by addressing the lexical demands required to effectively comprehend TED Talks (Nurmukhamedov, 2017). TED Talks have been used to support English for Academic Purposes instruction, including science and technology education (Liu, 2023), speaking practice for non-native postgraduate students (Zhang et al., 2024), and listening skills development (Wingrove, 2017). However, challenges remain, such as insufficient vocabulary coverage, which often necessitates supplementary materials (Madarbakus-Ring & Benson, 2024).

Although several studies affirm the usefulness of TED Talks in language learning, limited attention has been paid to the vocabulary characteristics embedded within different TED Talk topics. This gap leaves unanswered questions regarding whether and how the vocabulary used in TED Talks aligns with the needs of learners at varying proficiency levels and with different interests. The relationship between a TED Talk's topic and its vocabulary profile is particularly relevant. Different topics may naturally require distinct vocabulary types, reflecting the specific linguistic features of the subject matter. If L2 learners select TED Talks assuming that any topic will enhance their vocabulary knowledge, without considering the variation in vocabulary complexity across topics, their learning may not be optimally effective. A mismatch between the vocabulary level of the talk and the learner's proficiency may hinder learning progression, as the lexical gap may be too wide. This could ultimately make the learning experience excessively demanding. Understanding topic-specific vocabulary differences could therefore offer valuable insights for both educators and learners, facilitating the selection of TED Talks that are well-aligned with learners' proficiency levels and learning goals.

A review of the literature identifies several methods for analysing vocabulary to assess its suitability for different levels of L2 learners. These include lexical profile (LP) (Madarbakus-Ring & Benson, 2024), lexical level (LL), lexical frequency band (LFB), lexical coverage (LC) (Nurmukhamedov & Sharakhimov, 2021), CEFR level (CEFR-L) (Li et al., 2024; Treffers-Daller et al., 2018), lexical variation (LV) (Meebangsai et al., 2023), and lexical density

(LD) (Liu, 2021). To the best of our knowledge, no study has conducted an in-depth analysis of the vocabulary characteristics found within specific TED Talk topics. The most closely related work, by Coxhead and Walls (2012), examined six broad topics but focused only on vocabulary load.

To address this gap, the present study investigates the vocabulary characteristics of TED Talks across 20 topics using a corpus-based approach. It examines LP, LL, LFB, LC, CEFR-L, LV, and LD to uncover variations in vocabulary use across topics and to assess the extent to which topic-specific vocabulary features can support targeted language learning.

Theoretical Background and Related Studies

To investigate whether different TED Talk topics exhibit distinct vocabulary characteristics and to examine how each topic reflects tendencies in using vocabulary appropriate for promoting English vocabulary learning among L2 learners at varying proficiency levels, this study adopts analytical frameworks including LP, LL, LFB, LC, CEFR-L, LV, and LD, as outlined below.

Lexical Profiling

Lexical profiling is a linguistic method for categorising words in a corpus into distinct profiles, where each word is assigned to one profile only (Laufer & Nation, 1995). These profiles typically comprise high-frequency words, academic words, and outside words.

1. High-frequency words are common general English terms that occur frequently in everyday usage and are generally drawn from the General Service List (GSL) developed by West (1953).

2. Academic words are commonly used across academic texts and disciplines and are included in the Academic Word List (AWL) compiled by Coxhead (2000).

3. Outside words, listed in the Outside Word List (OWL), fall outside both the high-frequency and academic categories. These often include specialised or low-frequency vocabulary relevant for advanced learners and domain-specific contexts (Coxhead & Hirsh, 2007).

Lexical profiling has been widely applied in vocabulary research to examine the linguistic features of texts, including the analysis of LC and LP to identify the vocabulary necessary to comprehend different text types (Webb, 2021), evaluating text coverage in research articles (Wang, 2017), analysing LP within the context of beverage service (Arunvong Na Ayutthaya et al., 2022), studying the lexical characteristics of presentations in press conferences (Laosrirattanachai & Laosrirattanachai, 2023), and tracing

tourism business research trends in Scopus indexed journals (Laosrirattanachai & Laosrirattanachai, 2025). The results of lexical profiling provide valuable insights into the distribution of high-frequency, academic, and outside words in texts, enabling learners to develop targeted vocabulary strategies and helping educators select materials suited to learners' proficiency levels and learning objectives. In the present study, lexical profiling is used to identify the vocabulary composition of each TED Talk and assess its alignment with learner needs.

Lexical Level, Frequency Band, and Coverage

Lexical level (LL) refers to the classification of vocabulary based on frequency and distribution within a corpus. Nation's (2016) framework, developed using the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), organises vocabulary into 25 base word lists. These lists are ordered by frequency, with the most frequent words in the first list and less common words assigned to subsequent lists. Schmitt and Schmitt (2014) further categorise these into three lexical frequency bands: high-frequency words (1,000–3,000), mid-frequency words (3,001–9,000), and low-frequency words (beyond 9,000), which aid in understanding vocabulary distribution across different text types.

Lexical coverage (LC), on the other hand, refers to the proportion of words in a text that a learner can understand. Research suggests that 95% LC is the minimum threshold for general comprehension, while 98% is required for optimal understanding and inferencing (Dang & Webb, 2014; MacDonald, 2019; Tegge, 2017). Nation (2006) noted that 98% coverage typically requires a vocabulary size of 8,000–9,000 words (K8–K9), while 95% comprehension requires around 5,000 words (K5) (van Zeeland & Schmitt, 2013). This highlights the relationship between LL and LC in assessing the vocabulary knowledge needed for adequate comprehension and in determining text complexity. Computational tools such as VocabProfile (Cobb, n.d.) are often used to compare texts against these base word lists. Previous studies employing this approach include Benedict and Shabdin's (2021) investigation of LL in MUET reading exams, Wingrove's (2022) study of academic LL in TED Talks, and Phung and Ha's (2022) analysis of LL in IELTS listening tests. These studies provide insights into learners' vocabulary levels and the difficulty of target texts, allowing educators to identify gaps and customise instruction accordingly. In this study, LL and LC are employed to assess how accessible or challenging TED Talks are for learners at various proficiency levels.

CEFR Level

The Common European Framework of Reference for Languages (CEFR) is an internationally recognised standard for describing language proficiency (Council of Europe, 2001). It evaluates performance in listening, reading, speaking, writing, and vocabulary, aiming to support L2 learners in achieving effective communication in English-speaking contexts (Supunya, 2022). The CEFR framework consists of six proficiency levels: A1 and A2 (Basic User), B1 and B2 (Independent User), and C1 and C2 (Proficient User), each defined by a set of descriptors outlining expected linguistic abilities. For example, learners at the C2 level can understand and use complex language with ease, while A1 learners are limited to basic expressions and structures. According to Nation and Crabbe (1991), vocabulary size correlates with CEFR levels: A1–A2 learners typically know around 1,000 words, B1 learners about 2,000–3,000 words, B2 learners approximately 4,000, C1 learners 5,000–6,000, and C2 learners 7,000–9,000 words. CEFR-level alignment has been widely adopted in research, including test development (Wudthayagorn, 2018) and participant grouping based on proficiency (Charnchairerk, 2022). The CEFR framework underscores lexical competence as a core aspect of language proficiency and provides a structured path for vocabulary development. At lower levels, learners focus on foundational vocabulary, while higher levels require mastery of more advanced and abstract terms. By mapping the vocabulary features of TED Talks to CEFR levels, educators and learners can better select content that aligns with specific proficiency levels, promoting more effective and targeted vocabulary learning.

Lexical Variation

Lexical variation (LV), or lexical diversity, refers to the range of unique words used in a corpus and serves as an indicator of linguistic competence (Schmitt & Schmitt, 2020). A common metric for evaluating LV is the type-token ratio (TTR), calculated by dividing the number of unique words (types) by the total number of words (tokens) (Laufer & Nation, 1995). Higher TTR values indicate greater lexical diversity, whereas lower values suggest more repetition. However, TTR is highly sensitive to text length, with the ratio generally decreasing as text length increases (Bayazaidi et al., 2020; Malvern et al., 2004; McCarthy & Jarvis, 2007). LV has been examined in a range of contexts, including its relationship with text length (Koizumi & In'nami, 2012), its role in EFL writing proficiency (Wang, 2014), and its distribution in scientific paper abstracts across disciplines (Viera, 2022). LV plays a key role in language acquisition, as exposure to a wide variety of lexical items promotes broader vocabulary development and enables more precise language use. In the context of TED Talks, analysing LV provides insight

into the degree of vocabulary diversity present in each topic. Higher lexical variation indicates greater exposure to a broader lexicon, which supports deeper and more robust vocabulary acquisition. By identifying TED Talks with higher LV, this study can recommend topics that offer richer linguistic input for learners seeking to expand their vocabulary knowledge.

Lexical Density

Lexical density (LD) refers to the proportion of content words relative to the total number of words in a text, serving as a measure of its complexity and informational richness (Thornbury & Slade, 2006). Content words—such as nouns, verbs, adjectives, and adverbs—carry semantic meaning, while function words, including prepositions and conjunctions, serve grammatical roles. LD is calculated by dividing the number of content words by the total number of words and is typically expressed as a percentage. For example, an LD of 50% indicates that half of the words are content words, suggesting a balanced integration of meaning and structure. Texts with higher LD are often academic or formal in nature, characterised by information-rich content (Meebangsai et al., 2023), while those with lower LD, such as conversational texts, tend to be less dense and more interactive (Johansson, 2008). LD is important for vocabulary learning because it highlights the extent of lexical content available for acquisition. Texts with high LD expose learners to a wider range of content words, promoting vocabulary development and reading comprehension. In contrast, texts with lower LD are generally more accessible and better suited for foundational learning. In this study, LD is used to evaluate the informational load and linguistic complexity of TED Talks. Talks with high LD are likely to introduce academic or specialised vocabulary, making them appropriate for advanced learners, while those with lower LD may be more suitable for learners at lower proficiency levels. Analysing LD enables the alignment of talk complexity with learner ability, thereby optimising vocabulary learning outcomes.

These frameworks are interrelated and function collectively within the present study. LP informs both LL and LC; LL and CEFR-L help to assess text difficulty and learner appropriateness; LV and LD provide measures of vocabulary richness and textual complexity. Together, they offer a comprehensive analytical lens through which to evaluate how different TED Talk topics can support vocabulary learning for L2 learners by offering appropriately challenging and educationally valuable lexical input.

Research Questions

This study investigates the vocabulary characteristics of TED Talks across various topics by addressing the following research questions:

1. What is the lexical profile in each TED Talks topic?
2. What is the lexical level in each TED Talks topic, and what lexical level is required to achieve 95% and 98% corpus coverage in each TED Talks topic?
3. What is the lexical frequency band in each TED Talks topic?
4. How is vocabulary distributed across CEFR-level in each TED Talks topic?
5. What is the lexical variation in each TED Talks topic?
6. What is the lexical density in each TED Talks topic?
7. Based on lexical characteristics, which learner proficiency level is each TED Talks topic most suitable for?

Methodology

Data Collection

TED Talk topics were further subdivided into related subtopics. The number of subtopics within each topic was balanced to ensure that the combined number of clips from each subtopic equalled 50, thereby representing the corresponding topic. The selection criteria for talks within each subtopic were based on the highest number of views. Each clip was also required to be at least 10 minutes in length. Talks that did not meet these criteria were excluded and replaced with the next most-viewed clip that satisfied the requirements. The data collection process began by compiling a list of TED Talks through the TED website. For each topic, subtopics were selected individually and sorted by "most viewed." Scripts were then retrieved from the English subtitles. As a result, a TED-corpus containing 2,348,137 tokens was created (Table 1).

Table 1

Size of TED-Corpus and its Sub-corpora

TED Talks topics	Tokens
Art & Design	117,957
Business	133,758
Climate Change	113,557
Communication	121,877
Education	115,674
Entertainment	110,433
Global Issue	135,212

TED Talks topics	Tokens
Government & Politics	111,512
Health	110,175
Leadership	116,994
Nature	108,807
Parenting	106,015
Personal Growth	118,423
Psychology	138,605
Relationships	114,792
Science	119,669
Sports & Fitness	110,747
Social Change	117,833
Technology	118,414
Women	107,683
Total	**2,348,137**

Research Instrument

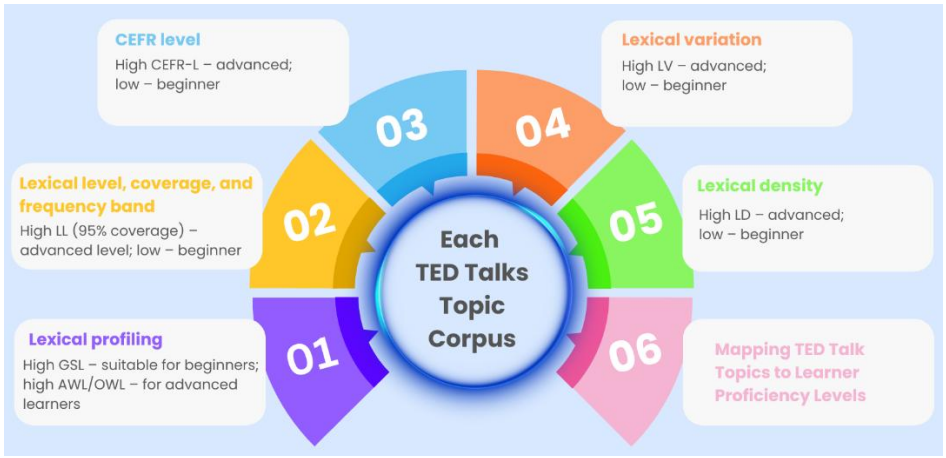
This study employed two primary tools for lexical analysis. AntWordProfiler (Anthony, 2024) was used to analyse LP, LV, and LD. This programme categorises words in a corpus into four profiles: the first two based on the GSL, divided into 1,000-word groups; the third on the AWL; and the fourth for the OWL (off-list words). Additional lists can also be added depending on research requirements. The programme provides data on both types and tokens in the corpus. VocabProfile (Cobb, n.d.) was used to analyse LL, LFB, and LC. This tool categorises vocabulary into 26 lists: the first 25 contain 1,000 words each, based on frequency rankings from the BNC and COCA, while the 26th includes infrequent, new, or misspelled words.

Data Analysis

This study examined the lexical characteristics of TED Talks across topics to assess their suitability for L2 learners at different proficiency levels. The analysis was conducted using computational tools and established linguistic frameworks to evaluate LP, LL, LFB, LC, CEFR-L, LV, and LD. The steps are summarised below.

Figure 1

Data Analysis Steps



First, AntWordProfiler was used to analyse LP using the GSL and AWL as reference lists. Each TED Talk topic was analysed individually to determine the percentage of words from each list. Words not included in the GSL or AWL were classified as OWL. A high proportion of GSL words indicated more general vocabulary and greater accessibility for beginners, whereas higher proportions of AWL and OWL words suggested suitability for intermediate or advanced learners.

Next, LL was analysed using VocabProfile. The programme identified the LL for each topic, ranging from 1K to 25K. This data was also used to determine LFB, categorising vocabulary into high-frequency (1K–3K), mid-frequency (4K-9K), and low-frequency (above 9K) bands. LC was examined by identifying the LL required to reach 95% and 98% coverage for each topic. Topics requiring higher LL were deemed more complex and better suited for advanced learners.

CEFR-L was derived from the LL data and mapped according to Nation and Crabbe's (1991) framework: A1 and A2 correspond to K1, B1 to K2-K3, B2 to K4, C1 to K5-K6, and C2 to K7-K9. This mapping facilitated

classification of TED Talk topics according to learner proficiency levels.

To analyse LV, AntWordProfiler was used to calculate the number of types and tokens. Given that TTR decreases with increasing text length (Malvern et al., 2004; McCarthy & Jarvis, 2007), the TTR was calculated for each clip and then averaged across all clips in a topic. These values were compared across topics to assess vocabulary diversity and determine suitability for vocabulary development.

LD was also analysed using AntWordProfiler, incorporating the Function Word List (Nation, 2018) to distinguish grammatical from content words. Words not included in the function list were treated as content words. LD was calculated by dividing the number of content words by the total number of words per topic. Higher LD values indicated greater lexical richness and were considered more appropriate for advanced learners, while lower LD values were associated with beginner-appropriate content.

Finally, the results from all six dimensions were used to align each topic with learner proficiency levels. Although CEFR classifies learners as A1-A2 (Basic User), B1-B2 (Independent User), and C1-C2 (Proficient User), this study adopts a more nuanced five-level categorisation: beginner, beginner-intermediate, intermediate, intermediate-advanced, and advanced. This approach acknowledges the overlapping and continuous nature of language development, allowing for a more flexible and learner-sensitive classification.

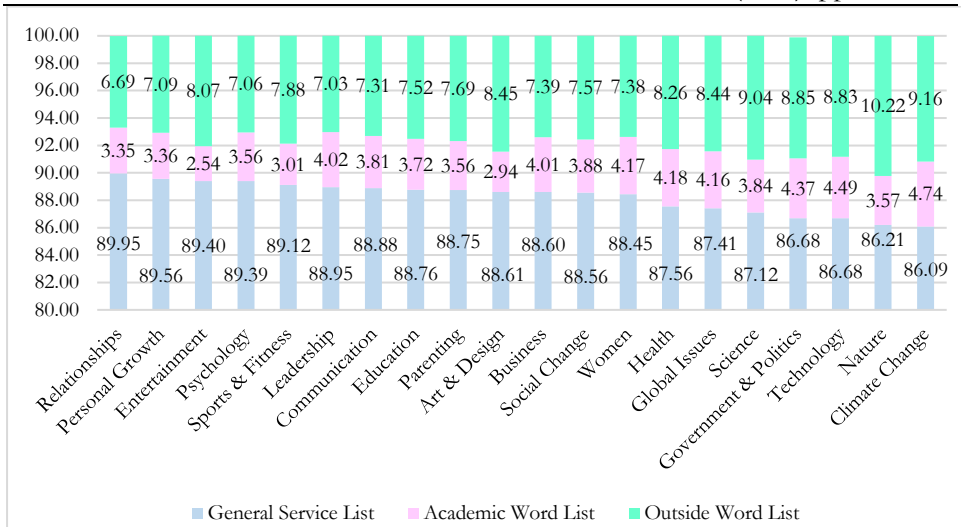
Findings

This study investigated the vocabulary characteristics of each TED Talk topic to determine their suitability as materials for enhancing vocabulary learning among L2 learners at different proficiency levels. The research addressed the following questions:

RQ1. What is the Lexical Profile in Each TED Talks Topic?

Figure 2

Lexical Profile of the 20 Topics of TED Talks



GSL words dominate across all topics, accounting for 86.09% to 89.95% of the vocabulary (Figure 2). This indicates that TED Talks rely heavily on high-frequency English words, supporting their general accessibility. AWL usage remains consistent, ranging from 2.54% to 4.74%, with Climate Change, Technology, and Health displaying relatively higher proportions. OWL words show greater variability, with Nature (10.22%) and Climate Change (9.16%) reflecting a higher presence of specialised vocabulary, while Relationships has the lowest proportion (6.69%).

Based on these findings, TED Talks can be aligned with learner proficiency levels. Topics with high GSL percentages, such as Sport & Fitness, Entertainment, and Relationships (all exceeding 88%), primarily consist of high-frequency vocabulary and are suitable for beginner learners due to their accessibility. Topics with moderate AWL usage, including Climate Change, Government & Politics, and Education (3.72% to 4.74%), contain more academic vocabulary and are appropriate for intermediate learners. Topics such as Health, Science, and Technology, with OWL percentages exceeding 9%, include more technical and specialised vocabulary and are therefore suited to advanced learners. Below is an example of the LP used in TED Talks. Words from the GSL are shown in regular text, words from the AWL are underlined, and words from the OWL are shown in bold.

(1) Traditional community spaces hold **Indigenous** knowledge and memories that are crucial for **ushering** in cultural **dignity**, facilitating local economy and **safeguarding** society's well-being. When we lose community spaces, we lose hope, connection with each other, and opportunities to exchange energy and build

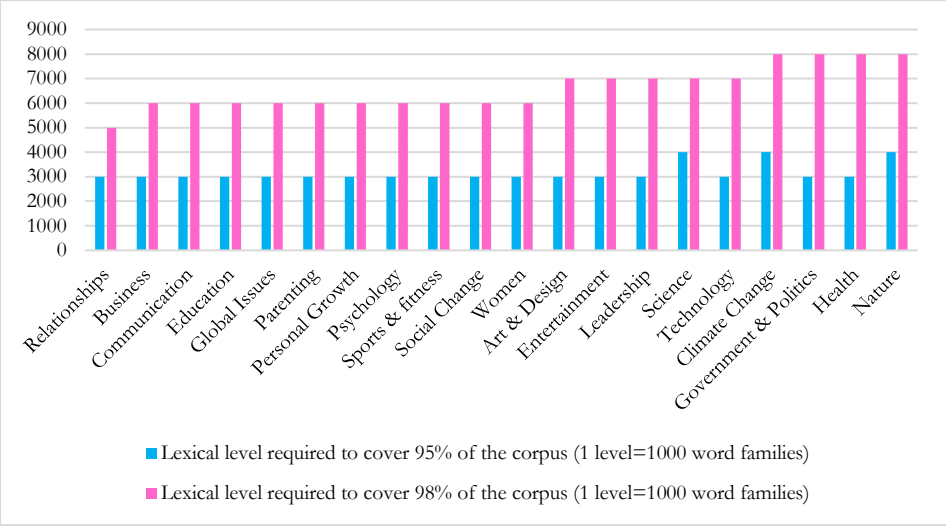
peace together. Dance is my **protest**, and **Eskista** is its contemporary expression for me.

(Sports & Fitness topic:
The ecstasy of Eskista, an ancient Ethiopian dance
by Melaku Belay, 2022)

RQ2. What is the Lexical Level in Each TED Talks Topic, and What Lexical Level is Required to Achieve 95% and 98% Corpus Coverage in Each TED Talks Topic?

Figure 3

Lexical Level Required to Achieve 95% and 98% Corpus Coverage of the 20 Topics of TED Talks



LL required to achieve 95% and 98% coverage varies across TED Talk topics (Figure 3). Topics such as Climate Change, Government & Politics, Health, and Nature demand higher LL to reach 98% coverage, reflecting more specialised or diverse vocabulary. In contrast, Business, Communication, and especially Relationships require fewer word families, indicating a more accessible vocabulary range.

In terms of learner suitability, beginner learners benefit most from topics such as Education, Global Issues, Parenting, and Relationships, which rely on smaller, more concrete, and frequently used vocabulary. These topics offer a strong foundation for gradual lexical development. Intermediate learners can benefit from topics like Art & Design, Entertainment, Leadership, and Technology, which balance accessibility with lexical

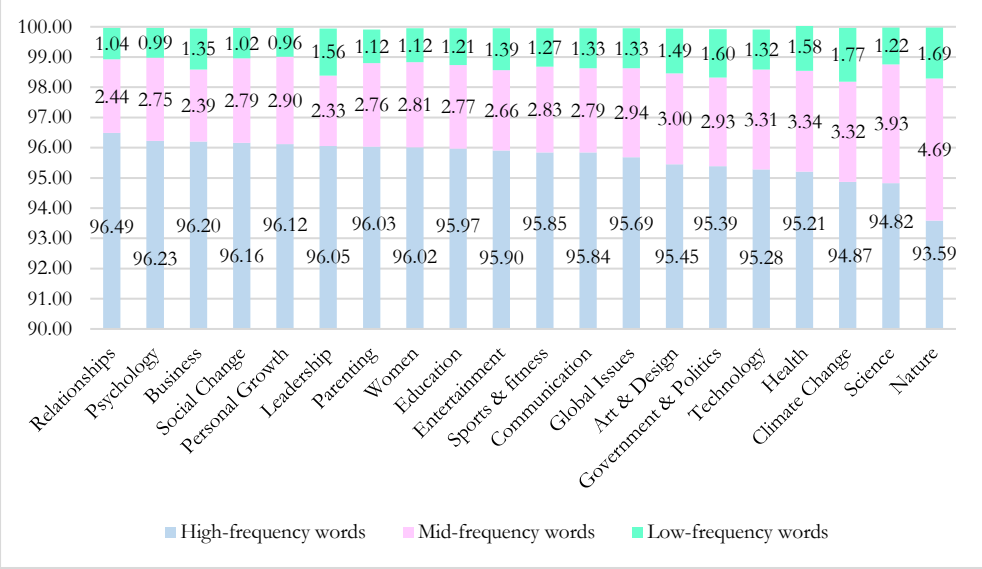
complexity and expose learners to more abstract terms. Advanced learners are best matched with topics such as Climate Change, Nature, Health, and Science. These topics require broader lexical knowledge and introduce specialised terms relevant to academic and professional domains. An example of LL and LC in TED Talks is provided: words from 1K-3K are shown in regular text, words from 4K-6K are underlined, and words beyond 6K are shown in bold.

> (2) If the soil is made up of small particles it becomes compact -- so compact, that water cannot seep in. We mix some local **biomass** available around, which can help soil become more **porous**. Water can now seep in. If the soil doesn't have the capacity to hold water, we will mix some more biomass -- some water-absorbent material like **peat** or **bagasse**, so soil can hold this water and it stays moist.
>
> (Climate Change topic:
> How to grow a forest in your backyard
> by Shubhendu Sharma, 2016)

RQ3. What is the Lexical Frequency Band in Each TED Talks Topic?

Figure 4

Lexical Frequency Band of the 20 Topics of TED Talks



This study classified vocabulary into three LFB: high-frequency (K1-K3), mid-frequency (K4-K9), and low-frequency (K10-K25 and OWL) (Schmitt & Schmitt, 2014). High-frequency words dominate all topics, ranging from 93.59% to 96.49%, indicating the overall accessibility of TED Talks (Figure 4). Topics such as Relationships, Psychology, and Personal Growth have the highest proportion of high-frequency words and are therefore particularly suitable for beginner L2 learners. These topics often deal with relatable and universal themes, further supporting comprehension.

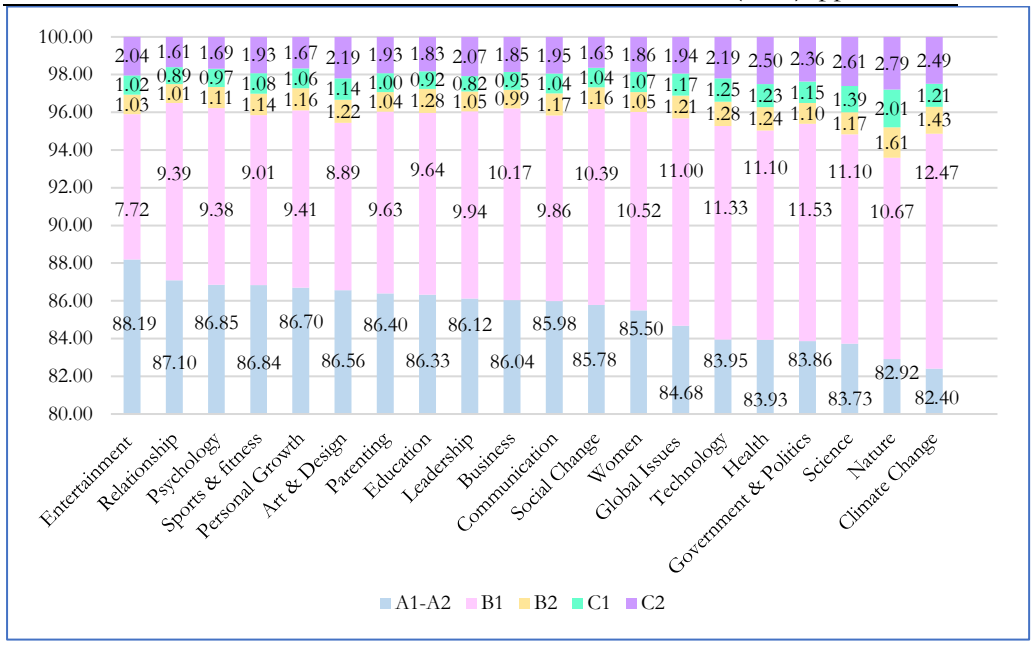
Mid-frequency words are more prevalent in topics such as Nature, Science, and Health. These topics offer a moderate level of lexical challenge and are appropriate for intermediate learners, as they promote vocabulary expansion while remaining comprehensible. Low-frequency words, which tend to be technical or specialised, are less frequent overall but are slightly more common in topics such as Climate Change, Nature, and Government & Politics. These topics are most suitable for advanced learners, offering exposure to domain-specific vocabulary and supporting the development of more complex lexical knowledge. An example of LFB use in TED Talks is provided: high-frequency words appear in regular text, mid-frequency words are underlined, and low-frequency words are shown in bold.

> (3) So they <u>vibrate</u> the flower, they **sonicate** it, and that releases the <u>pollen</u> in this efficient **swoosh**, and the <u>pollen</u> gathers all over the **fuzzy** bee's body, and she takes it home as food. Tomato **growers** now put **bumblebee** colonies inside the <u>greenhouse</u> to <u>pollinate</u> the tomatoes because they get much more efficient **pollination** when it's done **naturally** and they get better quality tomatoes.
>
> (Nature topic:
> Why bees are disappearing
> by Marla Spivak, 2013)

RQ4. How is Vocabulary Distributed across CEFR-level in Each TED Talks Topic?

Figure 5

CEFR Level of the 20 Topics of TED Talks



The results indicate that Entertainment contains the highest proportion of A1-A2 vocabulary (88.19%), while Climate Change features the most B1-level vocabulary (12.47%). Nature stands out for its relatively high proportions of B2 (1.61%) and C1 (2.01%) vocabulary. At the C2 level, Health (2.50%) and Science (2.61%) show the greatest lexical complexity, making them especially relevant for advanced learners seeking specialised vocabulary (Figure 5). In this study, words unclassified under any CEFR-L are treated as specialised and grouped under the C2 level.

Most TED Talk topics contain vocabulary predominantly at the A1-A2 levels, confirming their accessibility for beginner learners. However, the variation in higher-level vocabulary across topics is significant and influences their suitability for learners at different proficiency levels. Topics such as Entertainment, Relationships, and Sport & Fitness emphasise A1-A2 vocabulary and are thus ideal for beginners. These topics are also thematically familiar and engaging, helping learners retain new words and concepts with ease. For intermediate learners, topics such as Personal Growth, Psychology, and Social Change offer a more balanced distribution of A1-A2 and B1 vocabulary, with a moderate inclusion of B2 words. These topics support gradual advancement by exposing learners to more complex language while remaining accessible.

Advanced learners benefit from topics such as Nature, Science, and Health, which include noticeable proportions of C1 and C2 vocabulary. These topics present opportunities to encounter domain-specific terminology and

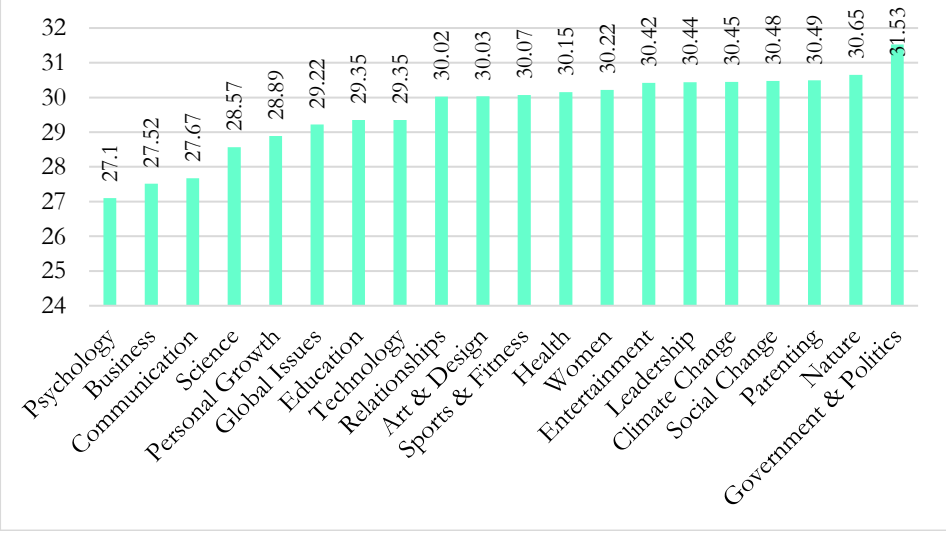
more sophisticated language structures. Their content often involves detailed explanations and abstract reasoning, encouraging learners to develop academic and professional communication skills. An example of CEFR-L use in TED Talks is provided: A1-A2 words appear in regular text, B1 words are underlined, B2 words are bold, C1 vocabulary is italicised, and C2 or unclassifiable words are shown in bold italics.

> (4) **Once** we start explaining **its** properties **in terms of** things happening inside brains and bodies, the ***apparently insoluble*** mystery **of what consciousness** is should start to **fade** away. … And then there's *conscious* self. The **specific** experience of being you or being me.
>
> (Science topic: Your brain hallucinates your conscious reality by Anil Seth, 2017)

RQ5. What is the Lexical Variation in Each TED Talks Topic?

Figure 6

Average TTR Values of the 20 Topics of TED Talks



The results provide insight into the average TTR across the 20 TED Talk topics, which indicates LV. TTR is expressed as a percentage and reflects the degree of vocabulary variety; higher values suggest greater lexical richness (Figure 6). Among the topics, Government & Politics has the highest TTR (31.53%), indicating a wide range of vocabulary. Other topics with high TTR

values include Nature (30.65%), Parenting (30.49%), and Climate Change (30.45%). In contrast, Business (27.52%) and Psychology (27.10%) show the lowest TTR percentages, reflecting a more repetitive or simplified vocabulary set.

These findings inform the alignment of TED Talk topics with learner proficiency levels. Topics with lower TTR, such as Psychology, Communication, and Business, are suitable for beginners. The reduced lexical variety supports the acquisition of high-frequency and familiar words without overwhelming learners. Topics with moderate TTR values, including Education, Technology, and Global Issues, are appropriate for intermediate learners. These topics introduce a broader range of vocabulary while maintaining a manageable level of difficulty, fostering progressive development.

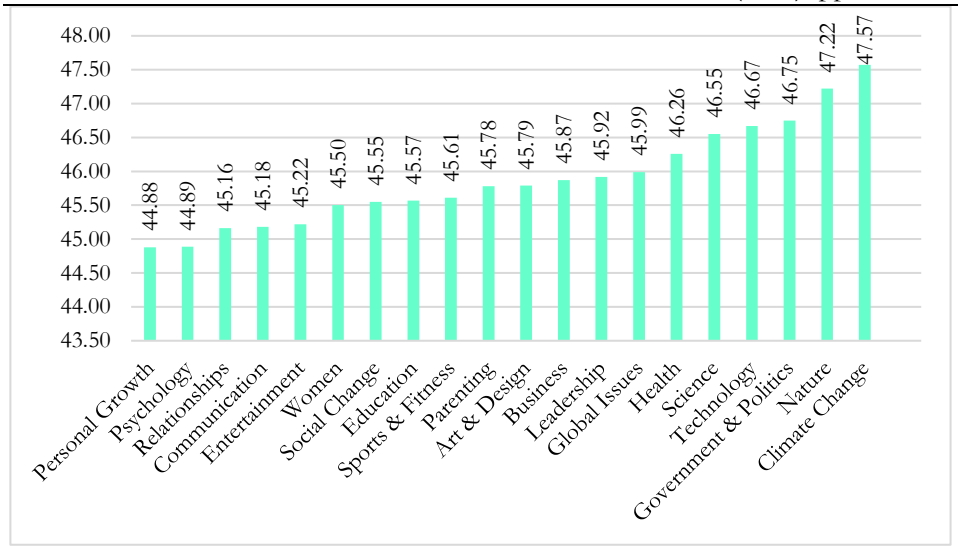
Advanced learners are best served by topics such as Government & Politics, Nature, and Social Change, where TTR values exceed 30%. These topics provide richer lexical input and more varied vocabulary, which promotes nuanced language use. The content often includes complex arguments and abstract concepts, supporting critical thinking and advanced language competence. An example of LV is provided with a set of near-synonymous words, such as produce and generate, instead of repeatedly using the same word.

> (5) … this is one of the unique structures of the adult brain where new neurons can be **generated**. And if we slice through the hippocampus and zoom in, … my colleague Jonas Frisén from the Karolinska Institutet, has estimated that we **produce** 700 new neurons per day in the hippocampus.
>
> (Personal Growth topic:
> You can grow new brain cells. Here's how
> by Sandrine Thuret, 2015)

RQ6. What is the Lexical Density in Each TED Talks Topic?

Figure 7

Lexical Density of the 20 Topics of TED Talks



The results indicate that LD across the 20 TED Talk topics ranges from 44.88% to 47.57%, reflecting the proportion of content words to total word count (Figure 7). A higher LD value suggests greater lexical richness, indicating that the text contains more content words and is suitable for advanced learners. These learners typically have command of function words and foundational vocabulary, allowing them to focus on acquiring more complex and specialised terms. In contrast, lower LD values reflect a higher proportion of function words, making the text easier to process and introducing fewer unfamiliar terms. Such texts are appropriate for beginner learners, who benefit from a more gradual exposure to new vocabulary.

LD values can be grouped into three ranges for interpretive clarity: below 45.50%, between 45.50% and 45.99%, and 46.00% or above. Topics such as Climate Change (47.57%) and Nature (47.22%) exhibit the highest LD and are well suited for advanced learners. Government & Politics (46.75%) and Technology (46.67%) also provide dense lexical input, presenting a suitable challenge for learners at the higher end of the proficiency scale. Topics including Global Issues (45.99%), Social Change (45.55%), and Women (45.50%) fall into the intermediate range, balancing accessible and moderately complex vocabulary. For beginner learners, topics such as Entertainment (45.22%), Relationships (45.16%), and Personal Growth (44.88%) show the lowest LD values. These topics include a higher proportion of function words, making them easier to comprehend while still offering exposure to new vocabulary in a manageable format. An example of LD use in TED Talks is provided: function words appear in regular text, and content words are shown in bold.

(6) We can **slash emissions** from **methane** by **improving** how we **manage** the **required water**, which can be as **simple** as **maintaining** a **shallow level** of **water** in the **rice fields**. **Implementing** all of these **solutions** will **take work**. But **people** have **already started** to **act**.

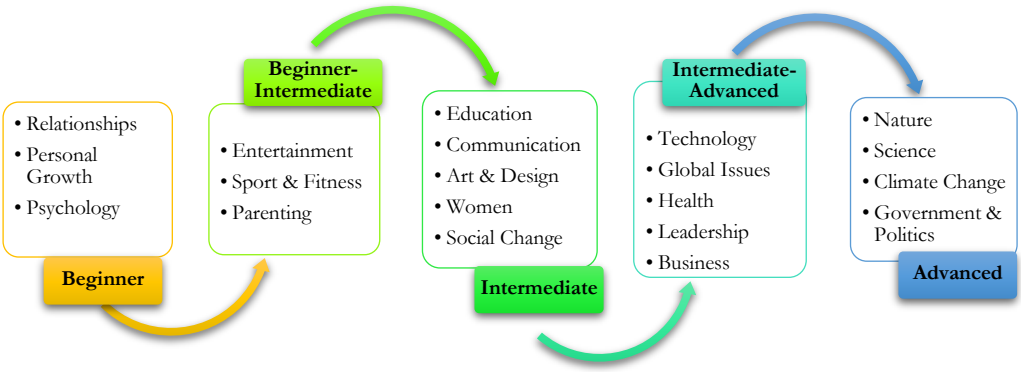
(Climate Change topic:
The fastest way to slow climate change now
by Ilissa Ocko, 2021)

RQ7. Based on Lexical Characteristics, Which Proficiency Level of Learners is Each TED Talks Topic Most Suitable for?

To address this question, the study evaluates six lexical features (LP, LL, LFB, CEFR-L, LV, and LD) across each topic to determine their overall suitability for learners at different proficiency levels (Figure 8). Learner proficiency is categorised into five levels: beginner, beginner-intermediate, intermediate, intermediate-advanced, and advanced.

Figure 8

Vocabulary in TED Talks' Topics for L2 Learner Proficiency Alignment



At the beginner level, learners are exposed to high-frequency vocabulary, low LV, basic linguistic structures, and minimal LD. CEFR-L at this level is concentrated in A1-A2. Relevant topics include Relationships, Personal Growth, and Psychology, which demonstrate low LD, high proportions of A1-A2 vocabulary, and limited lexical complexity. These topics are accessible and familiar, making them ideal for foundational vocabulary development.

The beginner-intermediate level supports learners transitioning to B1-level vocabulary. Suitable topics at this stage include Entertainment, Sport &

Fitness, and Parenting. These maintain moderate LD, include a significant amount of high-frequency vocabulary, and introduce content from the lower B1 band. They bridge the gap between foundational knowledge and more demanding lexical structures.

For intermediate learners, topics such as Education, Communication, Art & Design, Women, and Social Change offer a balance between accessibility and challenge. These topics feature a mix of high- and mid-frequency vocabulary, moderate LD, and CEFR coverage extending to B1 and B2. Learners at this level engage with moderately abstract content and gradually increase their lexical range.

The intermediate-advanced level is appropriate for learners ready to work with more complex vocabulary, higher LV, and CEFR levels in the B2-C1 range. Topics including Technology, Global Issues, Health, Leadership, and Business reflect this level of difficulty, combining elevated LD with more specialised vocabulary while remaining within reach of learners progressing toward advanced proficiency.

Finally, advanced learners benefit most from topics such as Nature, Science, Climate Change, and Government & Politics. These topics feature low-frequency vocabulary, high LD, extensive LV, and CEFR classifications reaching into C2. They demand a comprehensive lexical repertoire and are well suited for learners seeking to engage with sophisticated academic or professional discourse.

Discussion

The vocabulary used across TED Talk topics exhibits noticeable variation, shaped by the subject matter and intended audience. An analysis of LP, LL, LFB, CEFR-L, LV, and LD reveals key patterns that reflect both linguistic demands and communicative intent.

The LP results indicate that topics such as Nature and Climate Change use a higher proportion of OWL words compared to other topics, reflecting the presence of specialised or technical vocabulary. These topics often address scientific or environmental content, requiring less common lexical items to convey complex concepts. In contrast, topics like Entertainment and Personal Growth show a lower proportion of OWL words, suggesting broader accessibility. These topics rely more on GSL vocabulary to engage a wider audience, avoiding overly specialised terms. Compared with previous studies, TED Talks demonstrate relatively high GSL coverage (86.09–89.95%), likely due to their spoken format, which tends to favour more frequently used vocabulary. For instance, in Laosrirattanachai and Laosrirattanachai (2023), GSL coverage in WHO COVID-19 press conferences reached only 75.69%, while OWL usage was 18.36%. This

difference reflects the more technical vocabulary associated with medical briefings. Despite both being spoken genres, TED Talks are designed to be publicly accessible and thus use more general vocabulary.

The LL analysis further supports this finding. Topics such as Science, Climate Change, and Nature require higher LL for 95% and 98% coverage, consistent with their specialised content. Topics like Relationships and Parenting require significantly lower LL, highlighting their focus on everyday language. This study finds that 17 of the 20 topics require only 3,000 high-frequency words to achieve 95% comprehension, which contrasts with earlier estimates of 5,000 words for general comprehension (Laufer, 1989; van Zeeland & Schmitt, 2013). For 98% comprehension, TED Talks require between 5,000 and 8,000 words, slightly below Nation's (2006) suggested threshold of 8,000–9,000. This lower requirement may result from TED Talks' structured delivery and coherent topic development, which provide contextual support and reduce lexical ambiguity.

The LFB results reinforce TED Talks' accessibility. High-frequency words dominate across all topics, especially in areas such as Relationships and Personal Growth. In contrast, topics like Science and Climate Change show slightly higher proportions of mid- and low-frequency words, reflecting their need for domain-specific vocabulary. This variation indicates that although TED Talks aim to reach a broad audience, some topics demand a more advanced vocabulary for full comprehension.

The CEFR-L analysis reveals further differences in lexical complexity. Topics such as Nature and Climate Change contain higher proportions of C1–C2 vocabulary, consistent with the abstract and technical nature of their content. Meanwhile, topics like Relationships and Entertainment rely more on A1–A2 vocabulary. This contrast suggests that TED Talk topics vary not only in lexical density and frequency but also in the proficiency level required to understand them. Scientific and environmental themes often require advanced vocabulary to explain abstract phenomena, while personal or emotional narratives can be conveyed effectively using simpler language.

The LV results indicate that topics such as Government & Politics and Nature exhibit greater lexical variation. These topics require a broader range of vocabulary to cover complex ideas and multiple perspectives. Political discussions, for example, often necessitate varied terminology to describe ideological frameworks, policy issues, and sociopolitical dynamics. Similarly, environmental discussions involve diverse terms related to scientific processes. In contrast, topics such as Psychology and Business show lower LV, likely due to the repetitive use of discipline-specific vocabulary or practical, action-oriented language. Although TTR is influenced by text length, the moderate TTR values observed may be advantageous for learners,

supporting gradual vocabulary acquisition without overwhelming cognitive load.

The analysis of LD supports these observations. Topics such as Climate Change, Nature, and Science show the highest LD, suggesting a higher concentration of content words. These topics are generally information-rich, often involving technical descriptions and data-driven explanations. Conversely, topics like Personal Growth and Psychology have lower LD, reflecting more narrative or conversational styles. Such language tends to include more function words and informal constructions, aligning with the purpose of connecting personally with the audience. These results align with previous research. Meebangsai et al. (2023) found that written academic texts had an average LD of 57.52%, while Johansson (2008) noted that conversational language typically has lower LD. Ure (1971) reported an average LD of less than 40% for spoken texts. In contrast, TED Talks in this study had an average LD of 45.90%, higher than Ure's spoken language benchmark but lower than that of written academic discourse. This positions TED Talks as a valuable middle ground—spoken yet content-rich—making them suitable for vocabulary development in L2 contexts. Given that L2 learners often produce language with lower LD (To et al., 2013; Zhang et al., 2021), TED Talks with LD values above 40% provide an effective resource for enriching learners' lexical output.

Overall, the findings of this study provide clear guidance on how TED Talks can be used to support vocabulary development for L2 learners at varying proficiency levels. By analysing lexical characteristics such as LP, LL, LFB, CEFR-L, LV, and LD, the study identifies which topics are best aligned with beginner, intermediate, and advanced learners. For example, beginners (CEFR A1–A2) benefit most from topics such as Relationships, Personal Growth, and Psychology, which contain high GSL coverage, low LD, and minimal LV. Intermediate learners (CEFR B1–B2) can engage with topics like Entertainment, Parenting, and Education, which introduce more mid-frequency words and moderate LD. Advanced learners (CEFR C1–C2) are best suited to topics like Science, Nature, and Climate Change, which include high LD, more lexical variation, and greater reliance on OWL vocabulary. Importantly, the lower-than-expected LL required for 95%–98% comprehension suggests that L2 learners may be able to engage with authentic spoken content earlier than previously assumed. TED Talks' structured delivery, contextual clarity, and thematic coherence likely contribute to this accessibility, making them an effective resource for vocabulary instruction across all proficiency levels.

Pedagogical Implication

The findings on vocabulary use across 20 TED Talk topics offer valuable guidance for autonomous L2 learners aiming to expand their vocabulary. The results highlight a clear progression from beginner to advanced levels, enabling learners to select topics that align with their current proficiency while promoting gradual development of vocabulary knowledge. From these findings, practical and implementable pedagogical strategies can be derived.

For beginner learners, topics such as Relationships, Personal Growth, and Psychology are most appropriate for building foundational vocabulary. These topics are characterised by high-frequency vocabulary and lower lexical complexity. Effective learning strategies at this level include recognising common word forms, practising vocabulary in context, and completing listening-based summarisation activities (Liu, 2023; Wingrove, 2017). As learners move into the beginner-intermediate stage, topics such as Entertainment, Sport & Fitness, and Parenting offer slightly more lexical challenge while remaining accessible. Activities such as creating personal glossaries, participating in guided discussions, and completing targeted reading tasks can support vocabulary retention and expansion. At the intermediate level, topics including Education, Communication, Art & Design, Women, and Social Change are appropriate for learners developing a deeper understanding of vocabulary in context. Strategies such as analysing collocations (Phoocharoensil, 2013), constructing phrases, and engaging in reflective writing (Szenes & Tilakaratna, 2021) help learners understand how words are used in real-world discourse. Intermediate-advanced learners can explore topics such as Technology, Global Issues, Health, Leadership, and Business. These topics support further lexical development through structured tasks such as critical listening, thematic writing, and organised debates. These activities allow learners to apply vocabulary in cognitively demanding and communicative tasks. Advanced learners benefit most from topics such as Nature, Science, Climate Change, and Government & Politics. These topics present low-frequency vocabulary and more complex structures. Learners at this level can engage in synthesising information across texts, composing academic-style essays, and delivering oral presentations to refine vocabulary control in authentic, content-rich contexts (Laosrirattanachai & Laosrirattanachai, 2024).

In classroom contexts, TED Talks can be effectively integrated into instruction by aligning topic difficulty with learners' CEFR levels. For beginner learners, tasks such as vocabulary pre-teaching, comprehension checks, and working with simplified transcripts are recommended. Beginner-intermediate and intermediate learners can complete word recognition or gap-fill tasks focused on mid-frequency vocabulary. At the intermediate-advanced and advanced levels, learners can engage in summarising, critical discussion,

and vocabulary expansion exercises, particularly targeting OWL items and abstract expressions. To promote more natural language use, teachers can introduce near-synonyms based on vocabulary identified through LV analysis. Learners should be trained to identify recurring linguistic patterns within each topic. For example, topics such as Science and Climate Change frequently contain nominalisations and technical collocations, while topics like Relationships and Entertainment often use narrative sequences and idiomatic expressions. Recognising these patterns supports learners in predicting meaning and acquiring topic-specific vocabulary.

Furthermore, analysing lexical variation and density in TED Talks encourages learners to observe how language functions in context. Exposure to topics with high LV allows for broader vocabulary acquisition, while topics with higher LD provide sustained engagement with content-rich texts. This type of exposure is essential for learners preparing for academic and professional communication, where both vocabulary range and density are critical to success.

Limitations and Recommendations for Future Studies

The first limitation of this study concerns the duration of the selected talks. Only TED Talks with a minimum length of 10 minutes were included in the corpus. However, shorter talks may exhibit different lexical characteristics. Given the increasing popularity of brief digital content, future research could explore vocabulary usage in TED Talks under 10 minutes. This is particularly relevant considering current trends favouring short-form video formats, which prioritise conciseness and rapid consumption. Such formats have gained substantial traction on platforms such as TikTok and Instagram Reels (Cervi, 2021; Yang et al., 2019). It is possible that TED Talks and similar educational media may also adapt to these formats, thereby warranting investigation into their lexical features.

The second limitation is the study's lack of consideration for the speaker's linguistic background. The analysis did not distinguish between native and non-native English speakers. Vocabulary usage and stylistic choices may vary based on the speaker's language background, potentially influencing the lexical complexity and authenticity of the content. Future studies could compare vocabulary features in talks delivered by native and non-native speakers to assess these differences systematically. Such analyses may provide insight into how speaker identity influences lexical choices and could inform the design of more effective vocabulary learning materials tailored to L2 learners' needs.

About the Authors

Piyapong Laosrirattanachai: An Assistant Professor at Kasetsart University in Thailand. His primary research interests include corpus linguistics, vocabulary development, word lists, multi-word expressions, move analysis, and English for the hospitality sector.
Chanaporn Baothong: A member of the English for Service Industry programme within the Faculty of Hospitality Industry at Kasetsart University. Her academic focus lies in English for the hospitality industry and vocabulary studies.
Kotchakorn Laijud: A member of the English for Service Industry programme under the Faculty of Hospitality Industry at Kasetsart University. Her research interests include media studies and vocabulary development.
Piyanuch Laosrirattanachai: A lecturer in the English for Service Industry programme at Kasetsart University, Thailand. Her current research areas include vocabulary studies, multi-word expressions, move analysis, and sociolinguistics.

References

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

Arunvong Na Ayutthaya, J., 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

Bayazaidi, A., Ansarin, A.-A., & Mohammadnia, Z. (2020). Lexical complexity as a function of task type and proficiency level in the speech monologs of Iranian EFL learners. *Journal of Modern Research in English Language Studies, 7*(1), 29–44. https://doi.org/10.30479/jmrels.2019.11406.1418

Benedict, M. C., & Shabdin, A. A. (2021). A comparison of the vocabulary learning strategies employed by high and low proficient pre-university students in Malaysia. *LEARN Journal: Language Education and Acquisition Research Network, 14*(2), 222–246.

Boers, F., Bui, T., Deconinck, J., Stengers, H., & Coxhead, A. (2023). Helping learners develop autonomy in acquiring multiword expressions. *The Modern Language Journal, 107*(1), 222–241. https://doi.org/10.1111/modl.12829

Charnchairerk, C. (2022). Key defining linguistic features in the writing performance of first-year university students across different language proficiency levels. *LEARN Journal: Language Education and Acquisition Research Network, 15*(2), 858–891.

Cervi, L. (2021). Tik Tok and generation Z. *Theatre, Dance and Performance Training, 12*(2), 198–204. https://doi.org/10.1080/19443927.2021.1915617

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

Council of Europe. (2001). *The Common European of Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Coxhead, A. J. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Coxhead, A. J. (2018). *Using TED Talks in the classrooms*. RELC. SEAMEO.

Coxhead, A. J., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé*, *12*(2), 65–78.

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

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

Drayton, J., & Coxhead, A. (2023). The development, evaluation and application of an aviation radiotelephony specialised technical vocabulary list. *English for Specific Purposes, 69,* 51–66. https://doi.org/10.1016/j.esp.2023.01.006

Gay, F. (2022). Investigating the effects of technology-enhanced vocabulary learning strategy instruction on supporting mixed-ability EMI learners studying a journalism and communication major: An action research project at a university in China. *Journal of English for Academic Purposes, 55*, 101068. https://doi.org/10.1016/j.jeap.2021.101068

Hanks, E., Hashimoto, B., & Egbert, J. (2024). The contracts word list: Integral vocabulary for reading and writing English contracts. *English for Specific Purposes, 75*, 37–48. https://doi.org/10.1016/j.esp.2024.03.002

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

Kim, M., Black, S., & Clifton, J. (2024). Evidence-based holistic approach to supporting international students' English language proficiency needs and wellbeing: The impact of personalized autonomous

learning. *Applied Linguistics,* amae038, https://doi.org/10.1093/applin/amae038

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System, 40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

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. (2024). The triangle of language use: A corpus-based analysis of hotel responses to reviews. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 28*(1). https://doi.org/10.55593/ej.28109a3

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

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Laurén & M. Nordmann (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–332.

Le, C. N, N., & Miller, J. (2023). A core meaning-based analysis of English semi-technical vocabulary in the medical field. *English for Specific Purposes, 70*, 252–266. https://doi.org/10.1016/j.esp.2023.01.006

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,* 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

Liu, C. Y. (2023). Specialized vocabulary in TED talks and TED-Ed animations: Implications for learning English for science and

technology. *Journal of English for Academic Purposes, 65*, 101293. https://doi.org/10.1016/j.jeap.2023.101293

MacDonald, E. (2019). An analysis of vocabulary level in reading passages of the National Center Test. *Shiken, 23*(2), 19–27.

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

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Sage Journals, 24*(4), 459–488. https://doi.org/10.1177/0265532207080767

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. DOI:10.58837/CHULA.PASAA.67.1.3

Nation, I. S. P. (2006). How large a vocabulary is needed to reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

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

Nation, I. S. P., & Crabbe, D. (1991). A survival language learning syllabus for foreign travel. *System*, *19*(3), 191–201.

Nurmukhamedov, U. (2017). Lexical coverage of TED Talks: Implications for vocabulary instruction. *TESOL Journal, 8*(4). 768–790.

Nurmukhamedov, U., & Sharakhimov, S. (2021). Corpus-based vocabulary analysis of English podcasts. *RELC Journal, 54*(1), 1–15. https://doi.org/10.1177/0033688220979315

Otto, P. (2021). Choosing specialized vocabulary to teach with data-driven learning: An example from civil engineering. *English for Specific Purposes, 61*, 32–46. https://doi.org/10.1016/j.esp.2020.08.003

Phoocharoensil, S. (2014). Exploring learners' developing L2 collocational competence. *Theory and Practice in Language Studies, 4*(12), 2533–2540.

Phung, D. H., & Ha, H. T. (2022). Vocabulary demands of IELTS listening test: An in-depth analysis. *SAGE Open, 12*(1), 1–13. https://doi.org/10.1177/21582440221079934

Roesler, D. (2021). When a bug is not a bug: An introduction to the computer science academic vocabulary list. *Journal of English for Academic Purposes, 54*, 101044. https://doi.org/10.1016/j.jeap.2021.101044

Schmitt, N., & Schmitt, D. (2014). A Reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484–503.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching* (2nd ed.). Cambridge University Press.

Supanya, N. (2022). Towards the CEFR action-oriented approach: Factors influencing its achievement in Thai EFL classrooms. *3L: The Southeast Asian Journal of English Language Studies, 28*(2), 33–48. http://doi.org/10.17576/3L-2022-2802-03

Szenes, E., & Tilakaratna, N. (2021). Deconstructing critical reflection in social work and business: Negotiating emotions and opinions in reflective writing. *Journal of English for Academic Purposes, 49*, 100931. https://doi.org/10.1016/j.jeap.2020.100931

Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System, 67*, 87–98.

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 register 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

Viera, R. T. (2022). Lexical richness of abstracts in scientific papers in anglophone and non-anglophone journals. *3L: The Southeast Asian Journal of English Language Studies, 28*(2), 224–239.

Wang, P. (2017). A corpus-based study of English vocabulary in art research articles. *Journal of Arts and Humanities, 6*, 47–53.

Wang, X., (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*, 65–88.

Wang, Y. (2012). An exploration of vocabulary knowledge in English short talks: A corpus-driven approach. *International Journal of English Linguistics, 2*(4), 33–43.

Warnby, M. (2023). Academic vocabulary knowledge among adolescents in university preparatory programmes. *Journal of English for Academic Purposes, 61,* 101203. https://doi.org/10.1016/j.jeap.2022.101203

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), 287–302.

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

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

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

Wingrove, P. (2022). Academic lexical coverage in TED talks and academic lectures. *English for Specific Purposes, 65*, 79–94. https://doi.org/10.1016/j.esp.2021.09.004

Wudthayagorn, J. (2018). Mapping the CU-TEP to the Common European Framework of Reference (CEFR). *LEARN Journal: Language Education and Acquisition Research Network, 11*(2), 163–180.

Yang, S., Zhao, Y., & Ma, Y. (2019). Analysis of the reasons and development of short video application——taking Tik Tok as an example. *9th International Conference on Information and Social Science (ICISS 2019).* DOI: 10.25236/iciss.2019.062

Zhang, Y., Li, T., & Li, H. (2024). Using TED Talks in 'English Corners' to boost academic discussion skills of postgraduate students in a Chinese EMI university. In B. Zou & T. Mahy (Eds.), *English for Academic Purposes in the EMI Context in Asia* (pp. 231–257). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-63638-7_10