Volume 6 | Issue 1 | 2026

http://langedutech.com

Language Education & Technology

A Corpus-Based Evaluation of Lexical Coverage in A2-Level EFL Coursebooks

Ömer Faruk Kaya[a] * Aylin Han [b] *

[a] Bursa Technical University, Türkiye; https://orcid.org/0000-0001-7329-5557
[b] Independent Researcher, Türkiye

Suggested citation: Kaya, Ö.F. and Han, A. (2026). A Corpus-Based Evaluation of Lexical Coverage in A2-Level EFL Coursebooks. *Language Education and Technology (LET Journal), 6*(1), 1-12.

Article Info

Date submitted: 01/02/2026

Date accepted: 03/03/2026

Date published: 04/03/2026

Research Article

Abstract

This article reports a corpus-based study of A2-level English as a Foreign Language (EFL) coursebooks, focusing on their lexical coverage, frequency-based vocabulary distribution, and cross-linguistic lexical overlap. Textbook corpora compiled from student books were analyzed using the New General Service List (NGSL), the NGSL supplementary word list, and the New Academic Word List (NAWL) to establish vocabulary load and coverage in relation to the 95% and 98% comprehension thresholds. The results show that the coursebooks rely heavily on high-frequency vocabulary. Combined coverage from the NGSL bands, supplementary list, and NAWL range from 94.47% to 95.89%, indicating that the materials support comprehension with some instructional assistance, while remaining below the level associated with largely independent reading. NGSL-only coverage remains high (92.15%–93.93%), whereas academic vocabulary contributes minimally. Analysis of Turkish–English cross-linguistic overlap reveals that cognates constitute a small proportion of the lexical input, while false cognates occur at consistently higher rates across materials. These results suggest minimal variation in lexical demands across publishers and provide empirical evidence to inform textbook selection, materials development, and vocabulary pedagogy for A2-level learners.

Keywords: lexical coverage; EFL coursebooks; vocabulary load; cognates; corpus-based analysis.

1. Introduction

Coursebooks are largely considered indispensable part of classroom instruction and are central to organizing educational plans. They often determine the progression of grammatical and lexical targets and guide the range of discourse EFL learners encounter. In many university settings, this reliance is reflected in teachers planning their lessons largely around the prescribed textbook. For example, the British Council's The State of English in Higher Education in Turkey report found that in over 70 per cent of observed lessons, instruction closely followed the textbook with little or no adaptation or supplementary material (British Council, 2015). For many learners, particularly in contexts with limited exposure to the target language beyond the classroom, textbooks are primary sources of linguistic input (Allen, 2008; Harwood,

Ömer Faruk Kaya. Bursa Technical University, Türkiye.
e-mail adress: omer.kaya@btu.edu.tr

2017). Coxhead et al. (2010) note that "the reciprocal relationship between vocabulary knowledge and textbooks is critical" (p. 4). Sun and Dang (2020) similarly emphasize that effective learning depends on learners' engagement with both textbook content and the vocabulary it contains. From this perspective, vocabulary load, defined as the proportion of known to unknown lexical items in a text (Webb & Nation, 2008), becomes a key consideration in determining whether instructional materials are appropriate for a given proficiency level.

Nation (2013) and Webb and Nation (2008) propose that learners need to know approximately 95% of a text's running words to follow its meaning with difficulty and about 98% for comfortable comprehension. These thresholds have become widely used benchmarks for evaluating the lexical demands of instructional materials using frequency-based word lists (e.g., General Service List; GSL). However, research across diverse EFL contexts, including Saudi Arabia, Vietnam, China, and Taiwan, has repeatedly shown a mismatch between the lexical coverage of textbooks and learners' actual vocabulary knowledge (Alsaif & Milton, 2012; Nguyen, 2020; Sun & Dang, 2020). For example, Nguyen (2020) analyzed the lexical characteristics of texts from 30 units of Vietnamese high school English textbooks using Nation's BNC/COCA frequency lists via Lextutor (Nation, 2016). The findings indicated that learners would need knowledge of the first 3,000- and 5,000-word families to reach the 95% and 98% coverage thresholds, respectively, levels that exceeded learners' documented vocabulary knowledge of the most frequent 2,000-word families. In addition, the textbooks provided limited opportunities for repeated encounters with newly introduced vocabulary. Similarly, a recent bibliometric analysis by Xu and Ye (2025) reports that vocabulary in textbooks for both school and university learners remains underexamined in empirical research.

Comparative corpus-based evidence across major publishers and CEFR levels is particularly limited. Few studies have examined how vocabulary is distributed or recycled across textbook units, despite the importance of repetition for vocabulary development (Webb, 2021). This gap is notable given the global use of coursebook series published by Macmillan, Oxford, Cambridge, and Pearson. Although marketed as CEFR-aligned, empirical evidence confirming whether their vocabulary coverage reflects CEFR expectations remains scarce, particularly at the A2 level, where learners consolidate high-frequency lexis. Another underexplored dimension concerns cross-linguistic vocabulary. Turkish learners, in particular, may benefit from English–Turkish cognates (Uzun & Salihoğlu, 2021), although false cognates pose a risk for misinterpretation (Macizo et al., 2010). Despite their potential pedagogical relevance, little is known about whether globally used A2-level coursebooks systematically include cognates that could support early lexical development. Analyzing the distribution of cognates alongside frequency-based coverage, therefore, provides a cross-linguistic dimension for evaluating the lexical accessibility.

In response to these gaps, this corpus-based study examines the vocabulary coverage of four commercially published A2-level EFL coursebooks from Macmillan, Oxford, Cambridge, and Pearson. Using frequency-based lexical profiling, the study evaluates the distribution of high-frequency vocabulary, coverage at the 95% and 98% thresholds, lexical recurrence, and the presence of cross-linguistic cognates. By providing comparative analysis across publishers, the study provides empirical evidence for textbook evaluation, vocabulary pedagogy, and the development of CEFR-aligned materials. Guided by these, the study addresses the following research questions:

1. What is the vocabulary load of commercially published A2-level EFL coursebooks?
2. To what extent do these coursebooks reach the 95% and 98% lexical coverage thresholds based on high-frequency word lists?
3. To what extent do the A2-level commercial coursebooks include English–Turkish cognates?

2. Literature Review

Corpus-based approaches have become indispensable tools for evaluating the lexical characteristics of EFL coursebooks. With the increasing availability of large digital corpora and computational tools, researchers

can now examine vocabulary load, distribution, and accessibility with greater precision and transparency. Software such as RANGE (Heatley et al., 2002) and AntWordProfiler (Anthony, 2023), combined with widely used frequency lists, provides a replicable and systematic means of profiling the linguistic input learners encounter. Such approaches enable researchers to move beyond impressionistic judgments toward data-driven evaluations that reflect actual language use and learner exposure.

Much of the existing research on textbook vocabulary has relied on frequency lists derived from the BNC/COCA framework and organized around word families (Nation, 2012, 2016). While these lists have clear pedagogical and descriptive value—particularly due to their balance of spoken and written registers (Dang & Webb, 2016)—their reliance on morphological families may limit their applicability for lower-proficiency learners, who typically possess restricted morphological awareness. Recent research suggests that lemma-based approaches, including modified lexeme and flemma models, may offer a more transparent and developmentally appropriate representation of lexical input for beginner and lower-intermediate learners (Therova, 2020).

Concerns about the dated nature of earlier frequency lists have also motivated the development of newer lexical resources. The General Service List (GSL), originally compiled in the mid-20th century, reflects lexical priorities of its time, including items that are no longer central to contemporary English use (e.g., *telegram*, *shilling*). In response, Browne, Culligan, and Phillips (2013a, 2013b) developed the New General Service List (NGSL) and the New Academic Word List (NAWL) using larger and more contemporary corpora. The NGSL comprises 2,801 lemmas covering the most frequent words in modern English, while the NAWL contains 963 academic items not included in the NGSL and is intended to complement it. Empirical comparisons indicate that the NGSL provides more efficient coverage than the GSL across a range of educational texts, including English-medium instruction contexts (Brown et al., 2018).

A central construct in textbook vocabulary research is lexical coverage, defined as the proportion of running words in a text that learners are likely to know. A broad consensus has emerged around two threshold values: approximately 95% coverage, associated with minimally supported comprehension, and 98% coverage, associated with largely independent comprehension (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010). Importantly, recent work cautions against interpreting these thresholds as fixed cut-offs. Laufer (2020) demonstrates that coverage functions as a probabilistic condition for comprehension and inferencing rather than a deterministic guarantee, particularly at lower proficiency levels where strategic processing and instructional support play a greater role.

Corpus-based analyses of EFL coursebooks across diverse contexts consistently report that many materials cluster around the 95% coverage level but fall short of 98%. Studies from China, Vietnam, Saudi Arabia, and Indonesia show that learners would often need substantially larger vocabularies than those prescribed by curricula to achieve unassisted comprehension of textbook texts (Alsaif & Milton, 2012; Sun & Dang, 2020; Yang & Coxhead, 2020; Yu & Renandya, 2021). Sun and Dang (2020), for example, report uneven coverage and regression across textbook series, while Yang and Coxhead (2020) show that lexical demands can fluctuate markedly even within a single series. Together, these findings suggest that lexical adequacy in coursebooks is approximated rather than secured, particularly at lower levels.

One factor contributing to these mismatches is the uneven distribution of lexical coverage patterns across frequency bands. The first 1000-word families are generally well covered, but the following bands receive far less attention. Considering NGSL, the 2nd and 3rd frequency band levels play a significant role in reaching the 95 and 98 percent coverage thresholds, yet they are often introduced sparsely and recycled too little to support stable learning (see O'Loughlin, 2012; Sun & Dang, 2020). The inclusion of topic-specific and low-frequency vocabulary, especially in content-oriented texts, further increases lexical burden (Yang & Coxhead, 2020). As a result, learners frequently encounter words that appear too infrequently to support acquisition, falling well below the 7–20 encounters generally associated with durable learning (Webb, 2007; Uchihara et al., 2019).

Related research highlights the importance of repetition and recycling in vocabulary development. Despite their central role in retention, analyses repeatedly show that newly introduced words receive limited and inconsistent recycling across textbook units (Al-Ahmadi & Alshumrani, 2024; Mo & Bi, 2024). This limitation extends to multiword units and formulaic sequences, which are often marginalized despite their importance for fluency and comprehension (Northbrook & Conklin, 2018; Hoang & Crosthwaite, 2024). These patterns point to systemic tensions between vocabulary breadth, repetition, and pedagogical manageability in coursebook design.

Despite the growing body of research on textbook vocabulary, several gaps remain. First, cross-publisher comparisons involving major international publishers remain limited, particularly at the A2 level, where foundational lexical development is critical. As Xu and Ye (2025) note, empirical evaluations of CEFR-aligned, globally marketed coursebooks are surprisingly scarce. Second, while frequency-based profiling is common, few studies examine internal frequency band distributions, especially using NGSL-based frameworks. Third, cross-linguistic lexical overlap remains underexamined in coursebook vocabulary design, particularly from a lexical coverage perspective. True cognates can increase effective lexical coverage and support faster form–meaning mapping, especially at lower proficiency levels (Ellis & Beaton, 1993; De Wilde et al., 2020). However, morpho-phonological similarity is not uniformly beneficial. False cognates, or false friends, on the other hand, may also lead to misinterpretation and negative transfer (e.g., Macizo et al., 2010). Experimental evidence from studies using semantic relatedness and lexical decision tasks (e.g., Brenders et al., 2011; Poort & Rodd, 2019) consistently shows that false cognates yield slower and less accurate responses than control words. Because such items create an illusion of familiarity while increasing the risk of incorrect meaning mapping, textbook discourse and vocabulary sequencing should treat them with particular care. To date, no corpus-based studies have systematically examined the distribution of English–Turkish cognates and false cognates in A2-level global coursebooks.

Existing body of literature points to challenges in aligning lexical input, learner capacity, and curricular expectations in EFL coursebooks. Addressing these challenges requires corpus-based, cross-publisher analyses that integrate frequency-based profiling with cross-linguistic considerations. By combining NGSL-based lexical analysis with a systematic examination of cognates and false cognates, the present study aims to extend current knowledge and provide empirically grounded insights relevant to textbook evaluation, CEFR alignment, and the development of EFL teaching materials.

3. Methodology

This study adopted a quantitative, corpus-based design to evaluate the lexical coverage of four A2-level EFL coursebooks published by Macmillan, Oxford, Cambridge, and Pearson. Namely, the present study used *Macmillan Education Language Hub Elementary A2 Student's Book 2nd edition, Empower Elementary/A2 Student's Book 2nd edition by Cambridge Publishing, Oxford English File Elementary Student's Book 5th edition*, and *Speakout 3rd edition A2 Student's Book by Pearson Education.* These coursebooks were selected because they are widely used in foreign language schools and university preparatory programs. For instance, Bursa Technical University and Hacettepe University use *Language Hub* in their foreign language instruction, while Sakarya University uses *Empower*, Uludağ University uses *English File*, and Yalova University uses *Speakout*. The analysis aimed to determine (a) the overall lexical coverage of the coursebooks, (b) the distribution of the first, second, and third thousand most frequent word families, and (c) the recurrence of cognates and false cognates, using established corpus-linguistic methods and frequency-based word lists.

3.1. Corpus

The corpus for this study was compiled from four commercially published A2-level English coursebooks produced by Macmillan, Oxford University Press, Cambridge University Press, and Pearson Education. These publishers were selected due to their global reach and widespread adoption in institutional and private

EFL contexts. All selected materials correspond to the A2 level of the Common European Framework of Reference for Languages, ensuring comparability across sources.

Digital PDF copies of the student books were obtained from universities and language schools that officially use these coursebooks in their English programmes. No teacher's books or supplementary materials were included. Each textbook constituted a separate sub-corpus, and the combined dataset formed a balanced instructional corpus representing A2-level input across publishers. To ensure high OCR accuracy and traceability of recognition errors, all PDF files were first converted into high-resolution page-level images using NAPS2 scanning software (https:/www.naps2.com/). Text extraction was performed using a custom Python script implementing Google Tesseract OCR.

OCR-processed texts were reviewed and corrected for spelling and segmentation errors using built-in spell-check tools and systematic Find-and-Replace procedures in Microsoft Word and Google Sheets. Typical OCR errors, including merged or split words, misrecognized characters (for example, g0 for go), inconsistent spacing, and corrupted punctuation, were corrected. Contractions were expanded to their full forms (e.g., *I'm* to *I am*), and phonetic markers and exercise headings such as Grammar Bank and Vocabulary Bank were removed to ensure compatibility with the reference word lists used in lexical profiling. All texts were converted to plain UTF-8 format, and answer keys, glossaries, visual annotations, and other non-instructional metadata were excluded prior to analysis.

Each textbook formed a sub-corpus, and the combined data created a balanced corpus representing the full instructional input offered at the A2 level by the four publishers. The corpus included only reading texts. A separate corpus was made for listening transcripts. This design allowed for a cross-publisher comparison of lexical coverage while maintaining internal consistency in proficiency level and text type. The information about corpora is detailed in Table 1.

Table 1.

Corpus size of A2-level EFL textbooks

Textbook	Publisher	Token	Type	TTR
Language Hub Elementary (A2)	*Macmillan Education*	56.964	3097	5.44
Empower Elementary (A2)	*Cambridge University Press*	60.296	2824	4.68
English File Elementary (A2)	*Oxford University Press*	56.037	3341	5.96
Speakout (A2)	*Pearson Education*	50.867	2571	5.06

3.2. Analysis

Lexical profiling was carried out using AntWordProfiler 2.2.1 (Anthony, 2024), loaded with Brezina and Gablasova's (2015) New General Service List (new-GSL), as it draws on a larger corpus that contains more contemporary vocabulary than the GSL. NGSL is designed to be used with The New Academic Word List (NAWL). The NGSL includes 2,801 words covering the most frequently used words in English, while the NAWL comprises 963 entries that are frequently used in academic texts but are not listed on the NGSL. Since its organizing principle is based on modified lexeme, it may be more useful for learners at a beginner level who possess limited morphological knowledge (Therova, 2020). The NGSL was divided into 3 frequency bands: 1st, 2nd, and 3rd 1000 thousand words based on the standard frequency index (SFI) ranks. The base form of headwords were matched with flemmas or modified lexeme list available on the website (https://www.newgeneralservicelist.com). AntWordProfiler calculated the proportion of tokens covered at the 1k, 2k, and 3k frequency bands, NAWL, supplementary list and off-list items. Lexical coverage was evaluated at 90%, 95% and 98% thresholds. In short, for each sub-corpus, the analysis identified:

(a) the proportion of tokens within the 1k, 2k, and 3k frequency bands; and
(b) the presence and recurrence of off-list items beyond the 3,000-word level.
(c) the presence and recurrence of academic words (New Academic Word List; NAWL).

Descriptive statistics (i.e., mean coverage percentages, ranges, and variability across sub-corpora) were used to summarize vocabulary load and compare lexical accessibility across publishers. Token-to-type

ratios and the recurrence patterns of off-list items were also examined to provide additional insight into lexical recycling, a key pedagogical consideration in vocabulary learning.

Lastly, to examine the presence of cross-linguistic facilitative vocabulary, the corpus was also searched for cognates. English–Turkish cognates and false cognates were identified using the validated list compiled by Uzun and Salihoğlu (2021). The list was converted into a custom wordlist and processed in AntConc to count occurrences in each sub-corpus. For each coursebook, the proportion of cognates among total tokens and types was calculated, together with the distribution of false cognates. This analysis provided a complementary perspective on the lexical accessibility and pedagogical value of the A2-level materials.

4. Results

In response to RQ1 and RQ2, Table 2 presents the lexical profiles of the four A2-level EFL coursebooks, showing cumulative coverage based on NGSL frequency bands, the NGSL supplementary list, and the New Academic Word List (NAWL). Coverage values are evaluated against the 95% and 98% thresholds commonly associated with basic and optimal text comprehension.

4.1. Coverage of AWL and GSL Words: Comparison Between Coursebooks

As shown in Table 2, total lexical coverage across the four coursebooks ranges from 94.47% to 95.89%, indicating that all materials approach or exceed the 95% coverage threshold. Differences between publishers are relatively small. *Speakout Elementary* shows the highest overall coverage, followed by *Empower Elementary* and *Language Hub Elementary*, while *English File Elementary* yields slightly lower but comparable coverage. Despite differences in corpus size and type–token ratios, lexical coverage patterns are highly consistent across coursebooks.

When NGSL frequency bands are considered collectively, total NGSL-based coverage ranges from 92.15% to 93.93% of running words. In all materials, coverage is dominated by the first NGSL band, with progressively smaller contributions from the second and third bands. Academic vocabulary contributes only marginally to overall coverage, and the NGSL supplementary list accounts for a limited additional proportion. Despite differences in corpus size and type–token ratios, lexical coverage patterns remained highly consistent across coursebooks. None of the coursebooks reach the 98% coverage threshold based on frequency-based lists alone.

Table 2.

The Lexical Profiling of Four A2-level EFL textbooks

Coursebook	**Word Lists**					
	NGSL_1st	NGSL_2nd	NGSL_3rd	NAWL	NGSL_Sup	Total
Language Hub Elementary	84.54%	6.61%	2.07%	0.86%	1.00%	95.08%
Empower Elementary	86.96%	5.41%	1.56%	0.52%	1.05%	95.50%
English File Elementary	83.34%	6.67%	2.14%	1.06%	1.26%	94.47%
Speakout Elementary	85.35%	6.63%	1.69%	0.70%	1.52%	95.89%

The most frequent NGSL_1 items are predominantly function words, personal pronouns, and high-frequency verbs (e.g., the, a, to, you, be, and do), along with recurrent instructional items such as listen, questions, and work. High-frequency vocabulary in NGSL_2 primarily consists of content words related to classroom practices and everyday activities (e.g., sentences, correct, repeat, email, coffee, and dinner).

The third NGSL band includes lower-frequency content words and instructional labels associated with specific tasks or units. Academic vocabulary (NAWL) contributes only marginally to overall coverage (0.52%–1.06%), while the NGSL supplementary list accounts for an additional 1.00%–1.52% of running words.

4.2. Turkish–English Cognates and False Cognates Across Coursebooks

The proportion of Turkish–English cognates was relatively low across all coursebooks, ranging from 3.32% to 4.07%. English File Elementary contained the highest proportion of cognates (4.07%), followed by *Language Hub Elementary* (3.79%), *Empower Elementary* (3.59%), and *Speakout Elementary* (3.32%). In contrast, Turkish–English false cognates constituted a substantially larger proportion of the lexical items in each corpus. False cognate coverage ranged from 24.32% to 25.71%, with *Speakout Elementary* showing the highest proportion (25.71%) and *Language Hub Elementary* the lowest (24.32%). The remaining coursebooks, *Empower Elementary* and *English File Elementary*, showed false cognate coverage of 25.30% and 24.80%, respectively. Table 3. Describes the distribution of cognates across four coursebooks. None of the coursebooks reached the 98% coverage threshold based on frequency-based lists alone.

Table 3.

Turkish–English Cognates and False Cognates Across A2-Level Coursebooks

Coursebook	Cognates (%)	False Cognates (%)
Language Hub Elementary	3.90	25.00
Empower Elementary	3.66	25.84
English File Elementary	4.15	25.45
Speakout Elementary	3.35	26.18

In response to RQ3, Table 3 presents the distribution of Turkish–English cognates and false cognates across the four A2-level coursebooks. Overall, true cognates account for a small proportion of the lexical input (approximately 3%–4% of running words), with only minor variation across publishers.

In contrast, false cognates constitute a substantially larger proportion of lexical items in all four coursebooks (approximately 25%–26%), consistently outnumbering true cognates by a wide margin. This pattern is stable across publishers, indicating little variation in the cross-linguistic lexical profiles of the materials.

Across the corpora, true cognates are primarily concentrated in instructional and everyday communicative domains. Frequently occurring items include *complete*, *partner*, *exercise*, *form*, *film*, *video*, *music*, and *restaurant*, which appear repeatedly across all four coursebooks and are largely associated with classroom tasks and routine interaction. By contrast, false cognates are dominated by highly frequent grammatical function words rather than lexical content items. Across all materials, items such as *in*, *and*, and *is* occur at very high frequencies, while *do*, *not*, *of*, and *it* also appear several hundred times in each corpus.

5. Discussion

The The present study examined whether four commercially published A2-level EFL coursebooks provide lexical coverage consistent with commonly cited comprehension thresholds, and whether English–Turkish cross-linguistic overlap (cognates and false cognates) meaningfully shapes lexical accessibility. Overall, the lexical profiling results show a strong dependence on high-frequency vocabulary, with total coverage from NGSL bands, NAWL, and the NGSL supplementary list ranging from 94.47% to 95.89%. Without the other lists, the total NGSL-based coverage ranges from 92.15% to 93.93% of running words, which is still higher than the 90% threshold. Importantly, lexical coverage represents a necessary but not sufficient condition for comprehension. As previous research has shown, vocabulary knowledge interacts with other factors such as grammatical competence, discourse familiarity, background knowledge, and strategic processing, meaning that coverage thresholds should be interpreted as probabilistic indicators rather than guarantees of comprehension (Laufer, 2020; Schmitt et al., 2011). In this sense, the observed coverage levels suggest that A2-level coursebooks are designed to support comprehension under instructional

conditions, where lexical knowledge is supplemented by contextual cues and pedagogical scaffolding, rather than enable fully autonomous reading. This finding aligns with Laufer's (2020) argument that lexical coverage functions as a probabilistic condition for comprehension, particularly at lower proficiency levels where pedagogical scaffolding and contextual support play a central role.

A closer examination of frequency bands reveals that coverage is overwhelmingly driven by the first NGSL band, which alone accounts for over 83% of running words across all coursebooks. The progressively smaller contributions of the second and third NGSL bands, together with the marginal role of academic vocabulary, indicate that lexical accessibility is achieved primarily through dense concentration and recycling of core high-frequency items rather than through systematic expansion into mid-frequency ranges. Evidence from secondary-level and intermediate contexts further reinforces this interpretation. Sun and Dang (2020) found that Chinese high school EFL textbooks provided extensive exposure to the most frequent 1000-word families but underrepresented the second- and third-frequency bands and recycled them inconsistently. Similarly, Bergström, Norberg, and Nordlund (2022) showed that Swedish intermediate EFL textbooks reached the 95% coverage threshold largely through high-frequency vocabulary, while offering insufficient exposure to mid-frequency items necessary for sustained lexical development. Comparable patterns have been reported in Indonesian and Japanese contexts, where textbooks achieve high NGSL-based coverage yet cover only a limited proportion of the NGSL itself, repeatedly recycling familiar items rather than expanding lexical breadth (Nakayama, 2021; Rahmat & Coxhead, 2021). Taken together, these studies suggest that high lexical coverage in textbooks often reflects lexical control rather than lexical richness.

The NGSL-based coverage values observed in this study are relatively high compared to studies using the original GSL or BNC/COCA word-family lists. This difference can be attributed to both proficiency level and list design. A2-level materials are expected to prioritize general vocabulary more strongly than content-oriented texts, resulting in denser coverage of frequent items. In addition, the NGSL's lemma-based organization and expanded coverage of contemporary English likely inflate coverage estimates relative to older lists (Browne, 2014). However, as Aziz and Roslim (2021) caution, although most NGSL items are mapped to CEFR levels, level-specific expectations for NGSL coverage remain empirically underdefined. High NGSL-based coverage at the A2 level should therefore not be interpreted as direct evidence of full CEFR lexical alignment, but rather as an artefact of list breadth and frequency concentration.

The cognate analysis further refines this picture of lexical accessibility. True Turkish–English cognates account for only a small proportion of lexical input, suggesting that cross-linguistic facilitation plays a limited role in supporting comprehension at this level. In contrast, false cognates constitute a substantial proportion of running words, largely due to the dominance of highly frequent grammatical function words. While these items are essential for grammatical development, their prevalence also highlights the potential for cross-linguistic ambiguity, particularly for learners relying heavily on form-based processing at early stages. Importantly, the stability of both cognate and false cognate distributions across publishers indicates that cross-linguistic lexical profiles, like frequency-based profiles, show slight variation in global A2 coursebook design.

This pattern of frequency-driven coverage must also be interpreted in relation to text type, proficiency level, and analytical framework. Hu, Gao, and Qiu (2021) demonstrate that secondary-level English-medium science textbooks contain substantially higher proportions of academic vocabulary and correspondingly lower proportions of general vocabulary than EFL textbooks. In contrast, A2-level EFL materials are expected to prioritize general high-frequency vocabulary more strongly, resulting in denser coverage of core lexical items. From this perspective, the comparatively high coverage values observed in the present study reflect deliberate lexical control appropriate to beginner-level instruction rather than greater lexical richness. In addition, the use of the New General Service List likely contributes to higher coverage estimates relative to studies employing the original GSL or BNC/COCA word-family lists. NGSL

provides broader coverage through its inclusion of approximately 400 additional word families and its lemma-based organization, which captures inflectional variation more transparently (Browne, 2014).

Recent work linking the NGSL to the CEFR further contextualizes these results. Aziz and Roslim (2021) report that 97.2% of NGSL headwords are mapped to CEFR levels ranging from A1 to C2. However, they also emphasize that level-specific coverage expectations for the NGSL have not yet been empirically established. Consequently, high NGSL-based coverage at the A2 level should not be interpreted as direct evidence of full CEFR lexical alignment, but rather as a reflection of the NGSL's broad functional scope across proficiency levels. The present findings, therefore, contribute to ongoing discussions about how CEFR descriptors are operationalized in vocabulary design and evaluation.

Taken together, the results suggest that commercially published A2-level EFL coursebooks converge on a shared lexical profile characterized by heavy reliance on high-frequency vocabulary, limited mid-frequency expansion, and minimal systematic exploitation of cross-linguistic overlap. These findings reinforce a well-established pattern in textbook research and extend it by demonstrating that convergence persists when lexical coverage is examined through NGSL frequency bands and cognate distributions.

6. Conclusion

This study examined the lexical coverage profiles of A2-level EFL coursebooks using frequency-based analysis and a cross-linguistic perspective. The findings show strong convergence across materials, with lexical input overwhelmingly dominated by high-frequency vocabulary. Combined coverage approaches the 95% threshold associated with minimally supported comprehension, while remaining below the 98% level typically required for largely unassisted reading. In line with lexical threshold theory, these results suggest that A2 coursebooks are calibrated for instructional contexts in which comprehension is mediated by pedagogical support rather than by lexical sufficiency alone. The limited contribution of mid-frequency and academic vocabulary further indicates constrained lexical expansion at this level.

Several limitations should be noted. The analysis focused solely on lexical coverage and did not account for other factors that contribute to comprehension, such as grammatical complexity, discourse features, or learner background knowledge. In addition, the corpus was restricted to a small set of student books, and coverage estimates were not directly linked to learner performance. Future research should integrate learner corpora validated at the A2 level, examine vocabulary recycling and progression across proficiency levels, and explore how coverage interacts with comprehension and production in instructional settings.

References

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal*, *40*(1), 21–33. https://doi.org/10.1080/09571736.2012.658221

Alshumrani, H. A., & Al-Ahmadi, N. M. (2022). The Representation of Vocabulary Knowledge Aspects in Saudi EFL Textbooks. *Arab World English Journal*, *13*(4), 325–340. https://doi.org/10.24093/awej/vol13no4.21

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

Aziz, A., & Roslim, N. (2021). Relevance of the New General Service List in Selecting Reading Passages for ESL Students. *European Journal of English Language Teaching*, *6*(6), 223-241.

Bergström, D., Norberg, C., & Nordlund, M. (2025). Do textbooks support incidental vocabulary learning?–a corpus-based study of Swedish intermediate EFL materials. *Education Inquiry*, *16*(1), 69-87. https://doi.org/10.1080/20004508.2022.2163050

Brenders, P., Van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child second language learners: Evidence from cognates and false friends. *Journal of experimental child psychology*, *109*(4), 383-396. https://doi.org/10.1016/j.jecp.2011.03.012

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for?. *Vocabulary learning and Instruction*, 3(2), 1-10.

Browne, C., Culligan, B., & Phillips, J. (2013a). The New Academic Word List. Retrieved from http://www.newgeneralservicelist.org

Browne, C., Culligan, B., & Phillips, J. (2013b). The New General Service List. Retrieved from http://www.newgeneralservicelist.org

Doff, A., Thaine, C., Puchta, H., Stranks, J., & Lewis-Jones, P. (2022). *Empower Elementary/A2 student's book* (2nd ed.). Cambridge University Press.

Eales, F., & Oakes, S. (2023). *Speakout A2 student's book* (3rd ed.). Pearson Education Limited.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign vocabulary learning. *Language Learning*, *43*(4), 559–617. https://doi.org/10.1111/j.1467-1770.1993.tb00627.x

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE* [Computer software].

Hoang, H., & Crosthwaite, P. (2024). A comparative analysis of multiword units in the reading and listening input of English textbooks. *System, 121*, 103224. https://doi.org/10.1016/j.system.2024.103224

Hu, J., Gao, X., & Qiu, X. (2021). Lexical coverage and readability of science textbooks for English-medium instruction secondary schools in Hong Kong. *SAGE Open, 11*(1), 1–9. https://doi.org/10.1177/21582440211001867

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. https://doi.org/10.64152/10125/66973

Latham Koenig, C., Oxenden, C., Lambert, J., & Seligson, P. (2018). *English file elementary student's book*. Oxford University Press.

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

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly, 47*(4), 867–872. https://doi.org/10.1002/tesq.140

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. https://doi.org/10.64152/10125/66648

Le, N. T. M., & Dinh, H. T. (2022). Vocabulary coverage in a high school Vietnamese EFL textbook: A corpus-based preliminary investigation. *Vietnam Journal of Education, 6*(2), 102–113. https://doi.org/10.52296/vje.2022.187

Macizo, P., Bajo, T., & Martín, M. C. (2010). Inhibitory processes in bilingual language comprehension: Evidence from Spanish–English interlexical homographs. *Journal of Memory and Language*, *63*(2), 232-244.

Maggs, P., Smith, C., & Tennant, A. (2024). *Language hub elementary A2 student's book* (2nd ed.). Macmillan Education.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. Routledge.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

Mo, J., & Bi, P. (2025). Evaluation of vocabulary use in EFL textbooks: Evidence from curriculum words. *English Teaching & Learning*, *49*(1), 1-16. https://doi.org/10.1007/s42321-024-00170-3

Nakayama, S. (2021). A quantitative analysis of vocabulary taught in Japanese EFL textbooks. *Research Square*, 1-21 https://doi.org/10.21203/rs.3.rs-688772/v1

Nation, I. S. P. (2006). How large a vocabulary is needed for 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. (2012). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nation, I. S. P. (2013). *Teaching and learning vocabulary*. Heinle Cengage Learning. https://doi.org/10.1017/CBO9781139858656

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

Nguyen, C.-D. (2020). Lexical features of reading passages in English-language textbooks for Vietnamese high-school students: Do they foster both content and vocabulary gain? *RELC Journal*, 1-10. https://doi.org/10.1177/0033688219895045

Northbrook, J., & Conklin, K. (2018). "What are you talking about?": An analysis of lexical bundles in Japanese junior high school textbooks. *International Journal of Corpus Linguistics*, *23*(3), 311–334. https://doi.org/10.1075/ijcl.16024.nor

O'Loughlin, R. (2012). Tuning in to vocabulary frequency in coursebooks. *RELC Journal*, *43*(2), 255–269. https://doi.org/10.1177/0033688212450640

Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, *22*(8), 974–991. https://doi.org/10.1080/13670050.2017.1325834

Poort, E. D., & Rodd, J. M. (2019). Towards a distributed connectionist account of cognates and interlingual homographs: Evidence from semantic relatedness tasks. *PeerJ*, *7*, e6725. https://doi.org/10.7717/peerj.6725

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 102279. https://doi.org/10.1016/j.system.2020.102279

Tesseract OCR contributors. Tesseract OCR. https://github.com/tesseract-ocr/tesseract. Released May 25, 2025. Accessed September 15, 2025.

Therova, D. (2020). General word lists: Overview and evaluation. *Vocabulary Learning and Instruction*, *9*(1), 51-61. https://doi.org/10.7820/vli.v09.1.therova

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning, 69*(3), 559–599. https://doi.org/10.1111/lang.12343

Uzun, L., & Salihoğlu, U. M. (2021). A list of English–Turkish cognates and false-cognates. *Poznan Studies in Contemporary Linguistics*, *57*(2), 325-327. https://doi.org/10.1515/psicl-2021-0014

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2021). The lemma dilemma: How should words be operationalized in research and pedagogy?. *Studies in Second Language Acquisition*, *43*(5), 941-949. https://doi.org/10.1017/S0272263121000784

Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, *16*, 1–10.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press. https://doi.org/10.25170/ijelt.v12i1.1458

Xu, J., & Ye, F. (2025). Two decades of research on ELT textbook content: A bibliometric and content analysis. *Language Teaching*, *0*, 1-26 https://doi.org/10.14746/ssllt.39278

Xu, J., & Ye, F. (2025). Two decades of research on ELT textbook content: A bibliometric and content analysis. *Studies in Second Language Learning and Teaching*, *1*, 1-15. https://doi.org/10.14746/ssllt.39278

Yang, L., & Coxhead, A. (2020). A corpus-based study of vocabulary in the New Concept English textbook series. *RELC Journal, 53*(3), 597–611. https://doi.org/10.1177/0033688220964162

Yu, M., & Renandya, W. A. (2021). A corpus-based study of the vocabulary profile of high school English textbooks in China. *LEARN Journal: Language Education and Acquisition Research Network, 14*(1), 28–49.

Zhang, P. (2022). How does repetition affect vocabulary learning through listening to the teacher's explicit instruction? The moderating role of listening proficiency and preexisting vocabulary knowledge. *Language Teaching Research*, *30*(2), 716-739. https://doi.org/10.1177/13621688221140521 (Original work published 2026)