BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Lexical Demands of Political News Reports: A Corpus-Based Lexical Profiling Inquiry

Mohammad Rashid Alfuhaid

Department of English Language and Literature, College of Languages and Humanities, Qassim University, Buraydah 52571, Saudi Arabia

ABSTRACT

Regular access and exposure to authentic written material is essential for learners of English as a foreign language (EFL) or second language (ESL) beyond the intermediate level, thus necessitating the determination of the vocabulary size required for comprehension of such material. Most research into the effect of text coverage on comprehension has used small corpora, primarily comprising simplified texts. This study demonstrates the word-family sizes necessary for the adequate and optimal comprehension of an accessible genre of authentic text-based material: online political news reports. For this purpose, a corpus of 20 million words was collected from the online New York Times newspaper over a 24-month period. Lexical profiling of the corpus was conducted via Nation's British National Corpus/Corpus of Contemporary American English (BNC/COCA) 25 word-family lists using the AntWordProfiler software. The monthly corpora reflected relatively consistent text coverage of each of the first 3,000 word-family bands and of the vocabulary sizes necessary to reach the coverage levels widely accepted as prerequisites to achieving adequate (95%) and optimal (98%) comprehension. Results of the whole-corpus analysis indicated that the 95% and 98% text coverage levels were achieved, respectively, within the 3,000–4,000 and 6,000–7,000 word-family bands. These findings render written political news reports in general, and from the New York Times in particular, a useful source of supplementary material for EFL/ESL learners to improve the breadth and depth of their English vocabulary knowledge once they master approximately 3,500 word families.

Keywords: Learning; Reading Comprehension; British National Corpus (BNC); Corpus of Contemporary American English

*CORRESPONDING AUTHOR:

Mohammad Rashid Alfuhaid, Department of English Language and Literature, College of Languages and Humanities, Qassim University, Buraydah 52571, Saudi Arabia; Email: mfhied@qu.edu.sa

ARTICLE INFO

Received: 4 March 2025 | Revised: 7 April 2025 | Accepted: 14 April 2025 | Published Online: 18 April 2025
DOI: https://doi.org/10.30564/fls.v7i4.8944

CITATION

Alfuhaid, M.R., 2025. Lexical Demands of Political News Reports: A Corpus-Based Lexical Profiling Inquiry. Forum for Linguistic Studies. 7(4): 1028–1042. DOI: https://doi.org/10.30564/fls.v7i4.8944

(COCA); Lexical Profiling

1. Introduction

Developing a sizeable vocabulary is a fundamental ongoing process for learners of English as a foreign language (EFL) or second language (ESL) to facilitate their use of the foreign language in general and to improve their reading skills in particular[1–3]. Vocabulary knowledge is a strong, if not the most significant, predictor of reading proficiency[1, 4–14]. For learners to acquire the different requirements of learning new words, they should encounter the same word frequently and in a variety of contexts[15]. The need for such exposure requires learners to engage constantly with authentic material to improve their proficiency level beyond the comprehension of simplified written or spoken texts. In this regard, newspaper reports could represent a practical source for continuous contact with authentic material.

Newspaper reports may be recommended for teachers, learners and curriculum designers since they represent an easily accessible source of authentic material that could provide learners with ample opportunities to consolidate and learn new meanings of known vocabulary and to encounter the different forms and uses of high- and midfrequency words multiple times and in multiple contexts. Research has indicated numerous advantages and benefits of newspaper reports for second-language (L2) learners[11, 16–19]. Newspapers allow learners to read about interesting topics and news on current affairs from readily accessible and possibly free resources. In addition, news reports are sometimes broadcast or published in the learners' first language (L1), allowing learners to develop reasonable background knowledge about the same topics[20]. Moreover, Nation has suggested that newspaper reports could be a good source of exposure to academic vocabulary for learners[11]. In particular, newspapers can facilitate learners' transition from reading simplified texts to coping with authentic material through narrow reading. Schmitt has recommended reading multiple authentic texts on the same topic for learners of English at the intermediate level[21], because narrow reading will allow them to encounter topic-specific words repeatedly. Narrow reading on the same topic or story has been recommended for (a) developing learners' cumulative background knowledge about the topic of focus, (b) reducing learners' lexical demand by repeatedly exposing them to recurring items of the same story and—in turn—of the same vocabulary, (c) consolidating learners' ability to cope with the incremental acquisition of differing uses of familiar words, and (d) supporting the incidental learning of new lexical items[11, 17, 20, 22–26].

Given the benefits of such reading for EFL/ESL learners, both educators and learners should understand the vocabulary size necessary for comprehending newspaper reports. To provide this knowledge, this study aims to determine the vocabulary sizes required for the adequate and optimal comprehension of written political news reports. A lexical profiling analysis of news reports can ascertain the approximate vocabulary size needed by a learner in terms of word family frequency levels to use news reports more effectively. Lexical profiling studies aim to determine the vocabulary size required to comprehend written and spoken discourse. Such studies often examine lexical coverage—the number of running words understood by a reader or listener in an item of written or spoken discourse—as an indicator of comprehension[3, 11, 26–29]. Nation, Webb, and Laufer have emphasized the significance of lexical coverage research[1, 3, 29]; they (a) highlight the impact of vocabulary knowledge on comprehension, (b) identify the vocabulary size required to comprehend different text types, and (c) allow instructors and students to plan their vocabulary teaching and learning against clear vocabulary learning targets. Therefore, the increasing research on lexical profiling has aimed to examine the vocabulary knowledge required to achieve the 95% and 98% lexical coverage thresholds linked to adequate and optimal reading and listening comprehension[11, 26–28].

1.1. Lexical Coverage and Reading Comprehension

Research has identified a strong correlation between EFL/ESL learners' lexical coverage and their comprehension of written and spoken texts[1, 2, 8, 12, 13, 26, 30–35]. In the context of comprehending spoken and written texts, the literature has proposed specific boundaries for adequate and optimal lexical coverage. However, the lexical coverage

figures commonly reported as indicators of adequate and unassisted optimal comprehension are 95% and 98%, respectively[2, 26, 31, 32, 35].

Previous research has generated varying results regarding the lexical coverage needed to avoid hindering a reader or listener's comprehension. In the context of comprehending spoken and written texts, the literature has proposed specific boundaries for adequate and optimal lexical coverage. Web noted that research on lexical coverage has reported that comprehension tends to enhance when lexical coverage increases beyond 90%[26, 30–32, 35, 36]. However, the lexical coverage figures commonly reported as indicators of adequate and unassisted optimal comprehension are 95% and 98%, respectively[2, 26, 31, 32, 35]. Numerous studies exploring the impact of the proportion of known words (i.e., lexical coverage) on learners' ability to comprehend L2 texts have proposed a lexical coverage of 95% to facilitate adequate reading comprehension and of 98% to attain optimal comprehension. Therefore, the current study has adopted these two thresholds to analyze the readability of political news reports in terms of their lexical profile.

In her pioneering study exploring how lexical coverage relates to learners' performance on a reading comprehension test of academic texts, Laufer reported that while learners with a lexical coverage of 90% had poor comprehension, the lowest grade to pass the test was achieved by those with 95% understanding of the words in the target texts[32], thus suggesting the 95% threshold as the minimal lexical coverage needed for adequate reading comprehension, requiring knowledge of approximately 5,000 lexical items. Hu and Nation explored the impact of learners' lexical coverage and their reading comprehension of a 673-word story for pleasure and found that the majority of L2 learners who knew 98% of the words demonstrated adequate comprehension in multiple-choice and cued meaning recall tests, with a few who showed adequate comprehension at 90% or 95% lexical coverage[31]. Laufer and Ravenhorst-Kalovski analyzed the relationship among test takers' vocabulary size, the lexical coverage of academic texts, and the academic English reading comprehension scores of 745 participants on a psychometric university entrance test[2]. Using Nation's word-family lists based on the BNC, they found that learners with vocabulary size of 4,000–5,000 word families achieved lexical coverage of 95% and exhibited "minimally acceptable" comprehension of the texts, whereas those with a vocabulary size of 6,000–8,000 word families achieved lexical coverage of 98% and demonstrated optimal reading comprehension[11]. Schmitt et al. tested 661 EFL/ESL learners' reading comprehension of two texts (total: 1,440 words) that required more advanced academic reading skills, finding that reading comprehension generally improved as they achieved lexical coverage over 90% and that 98% lexical coverage provided a reasonable coverage target for the comprehension of academic texts[26].

1.2. Lexical Profiling of Written Texts

Webb pointed out that since the development of lexical profiling software such as RANGE, AntWordProfiler, and VocabProfile and the introduction of Nation's 14 BNC 1,000-word family frequency levels and Nation's 25 BNC/COCA 1,000 word-family levels, the lexical profiling of written and spoken discourses has attracted increasing scholarly attention[11, 36–40]. Lexical profiling studies have aimed to ascertain approximately how many word families are needed to achieve the 95% and 98% lexical coverage targets often cited as necessary for adequate and optimal comprehension, respectively. This section summarizes the main findings of a number of relevant previous studies to briefly highlight and compare their coverage results to the findings of the current study in terms of adequate and/or optimal reading comprehension. First is a discussion of the major findings of previous studies on the lexical profiles of language learning textbooks and English proficiency tests. These studies were selected on the basis of the assumption that the types of genres and texts they analyze correlate well with this study's use of news reports as part of the learning experience of EFL learners. Subsequently, a brief summary of three major studies on the lexical profiling of news reports is provided.

Previous research on the lexical profile of English language learning textbooks has determined that familiarity with the most frequent 3,000–4,000 and 5,000–6,000 word families is sufficient for achieving 95% and 98% lexical coverage, respectively. For example, Chujo used a BNC lemmatized high-frequency word list and found that the most frequent 3,200 lemmatized words provided 95% coverage of Japanese junior and senior high-school texts[41]. Webb and Macalister conducted a lexical profiling study to compare the lexical demands of written literary works for native English speakers and learners of English as an L2[42]. They reported that

the most frequent 2,000 word families in the BNC lists account for 95% coverage of L2 literary texts, while the most frequent 3,000 word families suffice for the same coverage in L1 texts. At the 98% coverage threshold, however, a significant disparity emerged between the two text types, with L1 literature necessitating a vocabulary size of 10,000 word families compared to 3,000 for L2 literature. Collins examined the lexical coverage of reading passages taken from two EFL textbooks and found that the 95% and 98% coverage levels required, respectively, vocabulary sizes of 3,000 and 6,000 word families in the BNC lists [43]. Rahmat and Coxhead investigated the vocabulary coverage and load within an EFL textbook series and found that learners required 3,000–4,000 word families in the BNC/COCA lists to attain 95% coverage, provided that support was available [44]; in contrast, 5,000–6,000 word families were necessary to achieve 98% lexical coverage, whereby the students could read the textbooks independently. Yang and Coxhead investigated a widely used English textbook series in China and found that 95% coverage required a vocabulary size of 3,000 word families in the BNC/COCA lists, whereas 98% coverage required 6,000 word families [45]. Garcia's study of the vocabulary types, progression, lexical coverage, and academic words in EFL upper secondary textbooks in Sweden showed that approximately BNC 3,000 word families were needed for 95% lexical coverage and 6,000 for 98% lexical coverage [46].

It is also valuable to consider the findings from lexical profiling research on English proficiency tests, as these may indicate the potential of news reports as preparatory materials for learners planning to undertake these tests. In their analysis of two TOEFL preparation tests (total tokens: 14,000), Chujo and Nishigaki reported that knowledge of 6,150 of the most frequent words in the BNC would be required to attain 95% lexical coverage [47]. In contrast, Kaneko found that a vocabulary size of 6,000 word families in the BNC lists (including proper nouns and defined words in context) was required to reach 95% coverage, whereas approximately 10,000 word families (including proper nouns and defined words in context) were needed for 98% coverage of reading passages in a TOEFL internet-based (iBT) exam [48]. Webb and Paribakht analyzed the lexical coverage of the reading passages from a university admission English proficiency test in Canada (CanTEST) and found similar results: achieving 95% and 98% coverage of the reading comprehension texts required, respectively, 6,000 and 14,000 word families in the BNC lists plus proper nouns and interjections [49]. Similarly, Kanzaki's lexical profiling study of 34 practice tests for the Test of English for International Communication (TOEIC) published in Japan or South Korea between 2005–2014 found that the first 3,000 word families from Nation's version of the BNC/COCA word-family lists (plus proper nouns, marginal words, transparent compounds, and abbreviations) provided up to 96.79% coverage, whereas the first 4,000 word families (plus proper nouns, marginal words, transparent compounds, and abbreviations) provided 98.24% coverage [50, 51]. Differences in the vocabulary size required for lexical coverage of seven past or official practice tests utilized as university entrance examinations in Japan—including Cambridge First, IELTS, TOEFL, and TOEIC—were identified by Kaneko, who demonstrated remarkably inconsistent results among these candidate tests [48]. According to Kaneko's results, the reading passages in these exams required between 2,000–5,000 most frequent word families from Nation's BNC/COCA word-family lists for 95% coverage, whereas 3,000–8,000 word families were required for 98% coverage [48].

In contrast to English language learning textbooks and proficiency exams, the lexical profiling of news reports has received relatively little attention. An early and influential investigation into the lexical profiling of newspapers was conducted by Nation [11]. He investigated the extent of vocabulary required to achieve 98% lexical coverage for the optimal, unassisted comprehension of five newspaper corpora (each composed of 44 news reports of approximately 2,000 words each), five fiction books, and one graded reader (approximately 10,500 words), concluding that 98% coverage required the BNC's most frequent 8,000–9,000 word families plus proper nouns. He also found that the 4,000 word-family level plus proper nouns accounted for approximately 95% of the running words. Along similar lines, Hsu's investigation of the lexical profile of the Voice of America news network and international radio broadcaster found that optimal comprehension at 98% coverage entailed a mastery of the BNC/COCA's most frequent 6,000 word families [52]. More recently, Ha's lexical profiling study of a massive corpus of online newspapers and magazines published in 20 English- and non-English-speaking countries involved

the analysis of 12 billion words from online sources within the News on the Web (NOW) corpus[53]. Ha reported that, overall, a vocabulary comprising the most frequent 4,000 word families in the BNC/COCA word family lists (plus proper nouns, marginal words, transparent compounds and acronyms) provided 95% coverage of the whole NOW corpus that he compiled, whereas the optimal lexical coverage of 98% required 7,000 word families[53]. In the analysis of individual country-based corpora, lexical coverage at the 95% point required a vocabulary size of 3,000 or 4,000 word families, with some cases requiring only the 3,000 word-family level. Ascertaining the threshold for 98% coverage, however, was more complicated. Ha reported that the word families necessary for the 98% threshold was 6,000 in three countries, 6,000–7,000 in five counties, 7,000 in seven countries, 7,000–8,000 in two countries, 8,000 in two countries, and—in one country—a remarkable 8,000–10,000 word families[53].

1.3. Rationale of the Present Study

According to Webb, lexical profiling research indicates significant differences in the lexical requirements of reading materials, which makes such research useful as it identifies the types of materials that may be most readily comprehended, along with setting vocabulary learning targets[36]. Accordingly, this study aims to contribute to lexical profiling studies on written discourse of a novel and formal newspaper genre: political news reports. The reason for the selection of political news reports is that they represent content that is usually appealing to the public, and it is hence assumed that there is high potential for EFL/ESL learners to have developed reasonable background knowledge about the topics and events discussed and reported in political news reports, especially the international political news that may also be broadcast in their L1. The present study not only builds on earlier studies but also seeks to investigate approximately how many word families in Nation's BNC/COCA word-family frequency lists are necessary for adequate and optimal comprehension of political news reports while employing more a focused, voluminous corpus (approximately 20 million words) of authentic texts[37]. The findings are expected to help educators, instructors, and learners develop vocabulary, improve reading comprehension, and expose them to a readily available resource of authentic texts of a specific genre.

The present study aims to contribute to corpus linguistics research by building on two significant studies conducted by Nation and Ha on newspaper corpora[1, 11]. This investigation can be considered a continuation of these two studies. It provides a renewed investigation of the lexical profile of newspapers explored in Nation[11]. In contrast to Nation's study[11], which examined a relatively small corpus of nearly 440,000 words using Nation's BNC word-family lists[11], the current study evaluates a corpus of 20 million words in newspaper reports in comparison with the updated and more comprehensive BNC/COCA word-family lists developed by Nation[37]. These updated lists are more comprehensive as they encompass both British and American English. On the other hand, the current study adopts a more focused approach in comparison to Ha's study[53], in which he utilized a massive corpora of 12 billion tokens from multiple online newspaper and magazine genres in 20 native and non-native English-speaking countries. Arguably, this massive multigenre, multisource corpus might make it unfeasible for teachers or learners to translate Ha's findings into practical decisions in terms of reading material selection[53]. Therefore, there seems to be a need for a lexical profiling study on a specific newspaper genre from a native-speaking country, from which more specific teaching and learning implications may be drawn. Furthermore, Ha stated that it may be inappropriate to evaluate the corpus's lexical demand solely based on its collective 12 billion tokens[53]. Therefore, he asserted, the variations of the corpus's lexical profile warrant a more thorough examination. In response to this call, the corpus of the current study focuses on online political newspaper reports.

To serve the study's focused purpose, political news reports from the New York Times (NYT) were selected as the corpus source. The NYT is a globally recognized newspaper from a native English-speaking country, known for its extensive coverage of various topics. However, the current study is more focused in two key ways: first, it concentrates on a specific genre—political news reports—and second, it draws exclusively from a single, reputable newspaper published in a native English-speaking context. Political news was chosen for its public relevance and the likelihood that EFL/ESL learners already possess some background knowledge of the topics, especially in the case of international political news,

which may also be broadcast in learners' first languages. Beyond its content, the NYT provides educational resources such as interactive quizzes, videos, and the NYT Learning Network, which are often tailored for classroom use and support integration into language-learning curricula. Importantly, the choice of the NYT is not due to characteristics that uniquely distinguish it from other reputable newspapers. Rather, it serves as a starting point and a call for further comparative research involving corpora from other well-established newspapers in native English-speaking contexts.

Building on the studies discussed above, the current study adopts the 95% and 98% coverage levels and aims to determine the level of word-family knowledge needed for a reader to achieve adequate (95% lexical coverage) and optimal (98% lexical coverage) comprehension of online written political news reports. It is, to the best of our knowledge, the first study to examine the lexical profile of a large sample of a specific newspaper genre (i.e., online written political news reports), with a total of more than 20 million tokens. Accordingly, this study seeks to address the following research questions:

What vocabulary size is needed to achieve 95% lexical coverage of online written political news reports against Nation's BNC/COCA word-family lists[37]?

What vocabulary size is needed to achieve 98% lexical coverage of online written political news reports against Nation's BNC/COCA word-family lists[37]?

2. Methodology

2.1. Data Collection

Developing the target corpus involved manually downloading national and international political news reports from the online version of the American New York Times (NYT) newspaper () over a two-year period between June 1, 2021, and May 31, 2023. This two-year period was selected to ensure the representativeness and diversity of the corpus and to mitigate the dominance of specific writing styles or high-frequency vocabulary related to particular events. The start date (June 1, 2021) corresponds to the commencement of the data collection process, whereas the end date (May 31, 2023) indicates the point at which the target corpus volume was achieved.

Articles were identified and selected from the Politics Section of the website, which consistently features reports related to both domestic and international political developments. Only standard news reports were included; opinion pieces, editorials, and multimedia content were excluded to maintain consistency in text type and linguistic style. The selection process was conducted manually to ensure that all texts met the criteria for genre and format. The corpus encompasses a wide range of political topics, including elections, governmental policies, legislative processes, diplomatic relations, and political responses to global issues such as the COVID-19 pandemic and climate change. It also covers major international events and conflicts, including the war in Ukraine, the civil war in Syria, political tensions in the Middle East, and other significant geopolitical developments. This thematic and geographic variety enhances the lexical diversity of the corpus and increases its pedagogical relevance for EFL/ESL learners.

In total, the data consisted of 15,285 written news reports over a 2-year period, to ensure the representativeness and diversity of the corpus and to mitigate the dominance of specific writing styles and the high frequency of specific words discussing particular topics or events. The data comprised a corpus of approximately 20 million running words in total, with an average of 636 news reports per month. The news reports differed in length but were, on average, 1,321 words. The aim is for these data to represent the largest-ever corpus developed on a specific genre of news reports.

The data collection process was conducted manually over a period of two years. Each article was individually accessed through the Politics Section of the *New York Times* website and copied into plain text format. Articles were selected specifically from the "Politics" section, using the site's built-in categorization to ensure genre consistency. Only standard written news reports were included, while opinion pieces, editorials, and multimedia-based content were excluded. Although manual collection can be time-intensive, the corpus—comprising approximately 20 million words—was compiled gradually over two years, making the process manageable for a single researcher. This approach also allowed for close monitoring of article content and structure, ensuring that all included texts met the selection criteria and enhancing the overall reliability and consistency of the corpus.

2.2. Data Analysis

The lexical coverage statistics of the 20-million-word news report corpus were analyzed via 25 1,000 word-family levels [37], drawn from the BNC/COCA word-family lists, to identify the distribution of each level within online written political news reports. For this purpose, the free AntWord-Profiler software was employed, which was recommended by Nation as the best program for lexical profiling because it is more modern, well supported, and equipped with more features than the much older RANGE program [37, 38].

Nation's BNC/COCA frequency-level word-family lists contain 25 lists of word families ranging from the first to the 25th most frequent 1,000-word families, as well as a list of proper nouns, a list of transparent compounds, a list of marginal words, and a list of abbreviations and acronyms [37]. Nation's 25 word-family lists are based on the frequency and range of occurrence of lexical items in the BNC and COCA. The word family (i.e., the basic word and its inflected forms and derivations) is used as the count unit for classifying all lexical items in the corpus according to their word-family frequency levels. This approach was followed because the word family has been suggested by Nation and Webb as the most appropriate unit through which to analyze lexical receptive knowledge [54], based on the assumption that a learner who is familiar with one or more members of a word family will most likely be able to recognize inflected forms and derivations.

The AntWordProfiler software provided statistics on the cumulative percentages of lexical coverage at Nation's BNC/COCA-based 25 word-family frequency levels (plus proper nouns, marginal words, transparent compounds, and abbreviations and acronyms) compared to the overall number of tokens in the corpus [37]. The 24 monthly corpora were first analyzed separately, following which the lexical coverage of each of the 1,000-word frequency bands within each monthly corpus was totaled to determine the coverage of all the word-family lists across the entire corpus. The lexical text coverage of each monthly corpora and of the collective corpus in the current study were calculated against Nation's BNC/COCA 25 word-family lists by counting the number and percentage of each 1,000-word frequency band until 95% and 98% lexical coverage levels were reached [37].

A number of issues were considered prior to the final corpus analysis. First, the initial analysis showed that some content words labeled by AntWordProfiler as "off list" were actually inflected forms of base words from Nation's most frequent lists (e.g., register/registrant) [37]. These were added to the appropriate frequency-level list of their base word. Second, proper nouns were retained in the input texts and included in the cumulative coverage calculations, as they were assumed to represent little to no learning load [2, 11, 38]. The totals of the proper nouns, marginal words, transparent compounds, and abbreviations and acronyms were considered as one category in the calculation of the lexical coverage percentages. Although Nation added a list of proper nouns, the initial corpus analysis showed that the vast majority of the "off list" lexical items were proper nouns [37]. This trend is unsurprising, as Hwang and Nation reported that proper nouns represented almost 10% of lexical items in news reports in newspapers [55]. Therefore, these items were added to the totals of proper nouns, marginal words, transparent compounds, and abbreviations and acronyms. Third, the news reports, in Microsoft Word format, were checked for spelling and typographic errors and corrected as necessary. Finally, the compound words that appeared off the lists were separated to allow the analysis software to reclassify them into appropriate, relevant frequency word lists.

3. Analysis and Discussion

3.1. Results of Monthly Text Coverages

Tables 1 and **2** illustrate the percentages of lexical items in each monthly corpus (from June 2021 to May 2022 and from June 2022 to May 2023). The rows in each table (from top to bottom) present the supplementary lists (SLs): proper nouns, marginal words, transparent compounds, and abbreviations and acronyms, followed by Nation's BNC/COCA 25 word-family lists, and finally the lexical items that fall outside these lists (i.e., "off lists"). The percentage of each individual list appears in the first column under each month. Then, for each month, the cumulative percentages of the coverage of each word-family level are totaled in the second column until 100% coverage is reached. The cumulative coverage calculation for each monthly corpus commences with the coverage percentage of the SLs. Since SLs were assumed to involve little or no learning load, they were included in the calculation of the 95% and 98% coverage levels.

Since news reports are typically loaded with the names

Table 1. Cumulative Lexical Coverage at Each Word-Family Level of the Monthly Corpora from June 2021 to May 2022.

Lists	Jun, 2021		Jul, 2021		Aug, 2021		Sep, 2021		Oct, 2021		Nov, 2021		Dec, 2021		Jan, 2022		Feb, 2022		Mar, 2022		Apr, 2022		May, 2022	
	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM
SLs	8.50	8.50	8.23	8.23	8.47	8.47	8.33	8.33	8.60	8.60	8.39	8.39	8.94	8.94	9.87	9.87	10.17	10.17	10.30	10.30	10.43	10.43	8.14	8.14
Band 1	68.80	77.24	68.98	77.21	69.20	77.67	69.09	77.42	68.61	77.21	69.09	77.48	68.10	77.04	67.10	76.97	66.62	76.79	66.78	77.08	66.41	76.84	68.84	76.98
Band 2	9.90	87.16	10.14	87.35	9.85	87.52	10.13	87.55	10.25	87.46	9.99	87.47	10.02	87.05	9.76	86.74	9.94	86.72	9.70	86.78	9.81	86.64	9.86	86.84
Band 3	7.10	94.23	6.76	94.11	6.49	94.01	6.82	94.37	7.03	94.49	6.81	94.28	6.52	93.58	6.51	93.25	6.71	93.44	6.73	93.51	6.62	93.26	7.07	93.91
Band 4	1.90	96.15	1.99	96.10	2.02	96.03	1.90	96.28	1.87	96.37	1.98	96.26	1.98	95.56	2.00	95.25	1.97	95.41	1.87	95.38	1.93	95.20	1.94	95.85
Band 5	1.10	97.27	1.13	97.22	1.10	97.13	1.01	97.28	1.03	97.40	1.02	97.29	1.14	96.70	0.94	96.20	0.98	96.39	0.99	96.37	0.97	96.17	0.97	96.82
Band 6	0.80	98.03	0.77	98.00	0.83	97.96	0.75	98.03	0.74	98.14	0.80	98.09	0.77	97.47	0.73	96.92	0.76	97.15	0.70	97.07	0.78	96.95	0.78	97.59
Band 7	0.40	98.42	0.39	98.39	0.37	98.33	0.38	98.41	0.38	98.51	0.41	98.50	0.43	97.89	0.40	97.33	0.38	97.53	0.41	97.48	0.39	97.33	0.46	98.06
Band 8	0.40	98.80	0.36	98.75	0.35	98.68	0.36	98.77	0.33	98.84	0.35	98.85	0.46	98.35	0.63	97.95	0.60	98.13	0.65	98.13	0.66	97.99	0.37	98.42
Band 9	0.20	99.00	0.24	98.98	0.20	98.88	0.20	98.98	0.21	99.05	0.23	99.07	0.21	98.56	0.23	98.18	0.22	98.35	0.20	98.33	0.23	98.22	0.21	98.63
Band 10	0.16	99.16	0.14	99.13	0.11	98.99	0.13	99.11	0.13	99.19	0.13	99.21	0.12	98.68	0.14	98.32	0.13	98.48	0.12	98.46	0.14	98.36	0.12	98.75
Band 11	0.13	99.29	0.12	99.25	0.11	99.11	0.11	99.22	0.12	99.30	0.12	99.32	0.11	98.79	0.15	98.47	0.12	98.60	0.11	98.57	0.13	98.48	0.13	98.88
Band 12	0.15	99.44	0.14	99.39	0.15	99.26	0.14	99.36	0.12	99.43	0.13	99.46	0.16	98.95	0.14	98.62	0.14	98.74	0.15	98.72	0.12	98.61	0.11	99.00
Band 13	0.06	99.50	0.09	99.49	0.21	99.46	0.14	99.50	0.09	99.51	0.08	99.54	0.07	99.02	0.08	98.70	0.06	98.80	0.07	98.79	0.06	98.67	0.07	99.07
Band 14	0.05	99.55	0.03	99.52	0.04	99.50	0.03	99.53	0.04	99.55	0.03	99.57	0.04	99.06	0.05	98.75	0.04	98.84	0.04	98.83	0.04	98.71	0.03	99.10
Band 15	0.04	99.59	0.05	99.57	0.03	99.54	0.04	99.57	0.04	99.58	0.03	99.61	0.02	99.08	0.03	98.78	0.03	98.87	0.03	98.86	0.03	98.74	0.02	99.12
Band 16	0.02	99.61	0.03	99.59	0.02	99.55	0.02	99.59	0.02	99.61	0.02	99.63	0.02	99.10	0.02	98.80	0.03	98.90	0.02	98.88	0.02	98.76	0.02	99.15
Band 17	0.01	99.63	0.01	99.60	0.02	99.57	0.02	99.61	0.01	99.62	0.02	99.65	0.01	99.12	0.02	98.83	0.02	98.92	0.01	98.89	0.02	98.79	0.02	99.17
Band 18	0.01	99.64	0.01	99.61	0.01	99.58	0.02	99.62	0.02	99.64	0.02	99.66	0.02	99.13	0.01	98.84	0.01	98.93	0.02	98.91	0.02	98.81	0.02	99.19
Band 19	0.01	99.65	0.01	99.62	0.01	99.59	0.01	99.63	0.01	99.65	0.01	99.68	0.00	99.14	0.01	98.85	0.01	98.94	0.01	98.92	0.01	98.82	0.01	99.20
Band 20	0.01	99.65	0.00	99.62	0.01	99.60	0.01	99.64	0.01	99.66	0.01	99.69	0.01	99.15	0.01	98.86	0.01	98.95	0.01	98.93	0.01	98.83	0.01	99.20
Band 21	0.00	99.66	0.01	99.63	0.01	99.61	0.00	99.65	0.02	99.68	0.01	99.70	0.01	99.16	0.01	98.87	0.00	98.95	0.01	98.94	0.00	98.84	0.01	99.21
Band 22	0.01	99.67	0.01	99.64	0.01	99.61	0.01	99.66	0.00	99.68	0.01	99.70	0.01	99.16	0.01	98.88	0.02	98.96	0.01	98.94	0.01	98.84	0.01	99.22
Band 23	0.06	99.73	0.07	99.71	0.09	99.70	0.06	99.72	0.06	99.75	0.06	99.76	0.11	99.27	0.07	98.96	0.08	99.04	0.06	99.00	0.05	98.90	0.04	99.26
Band 24	0.01	99.73	0.01	99.71	0.01	99.71	0.00	99.72	0.01	99.75	0.01	99.77	0.01	99.28	0.01	98.97	0.00	99.05	0.01	99.01	0.01	98.90	0.01	99.27
Band 25	0.03	99.76	0.02	99.74	0.02	99.73	0.02	99.74	0.02	99.77	0.02	99.80	0.02	99.29	0.03	99.00	0.03	99.08	0.02	99.03	0.08	98.98	0.02	99.29
Off lists	0.24	100.00	0.26	100.00	0.27	100.00	0.26	100.00	0.23	100.00	0.20	100.00	0.70	100.00	1.00	100.00	0.92	100.00	0.97	100.00	1.02	100.00	0.70	100.00

of people, geographical locations, events, etc., it was expected that they would include many proper nouns not included in Nation's list of proper nouns and, consequently, that these items would be classified by AntWordProfiler as "off list." The preliminary analysis of the monthly corpora confirmed this expectation, revealing that the lexical items categorized as "off list" represented 2.10%–3.14% across the monthly corpora. Notably, proper nouns comprised 5.38% of the entire corpus. In fact, a second round of investigation revealed that the vast majority of the lexical items falling outside the lists were proper nouns. These unlisted proper nouns were calculated as representing 65.94%–93.20% of the items categorized as "off list," resulting in a total coverage of listed and unlisted proper nouns of 7.42%. This result is in line with Nation's finding that proper nouns accounted for 4.55%–6.12% of the running words in newspapers[11]. Similarly, Ha found that proper nouns comprised 3.91% of the running words in a massive newspaper corpus[53]. Therefore, to avoid overestimation of the vocabulary size required for 95% and 98% lexical coverage, the unlisted proper nouns were added to the coverage percentage of the SLs for each monthly corpus.

The lexical coverage analyses of the corpora for the 24 months across Nation's BNC/COCA 25 1,000-word frequency bands were compared to identify any similarities or differences in their coverage[37]. These monthly corpora showed a considerable degree of coverage agreement across the 25 lists plus the SLs. For example, the SLs provided a coverage of 8.14%–10.43%. Similarly, the coverage of the first 1,000-word frequency band, which constituted the highest text coverage of all the lists, was 66.41%–69.43%. However, the lexical coverage per subsequent word frequency band decreased. For example, the second and third 1,000-word frequency bands represented coverage ranges of 9.70%–10.52% and 6.49%–7.10%, respectively. Another notable result was the 93.25%–94.5% coverage range provided collectively by the first 3,000 bands.

The 95% coverage level required a vocabulary size of between 3,000–4,000 word families, whereas the vocabulary size required for 98% coverage was 6,000 in 75% (i.e., 18 of 24) of the monthly corpora. As shown in **Table 1**, in some cases, requirements for 98% coverage varied by month, necessitating between 7,000–9,000 word frequency bands: 1 month required 7,000, 4 months 8,000, and 1 month 9,000 word families. The data presented in bold in **Tables 1** and **2** highlight where the 95% and 98% coverage points were reached. The coverage percentage of the fourth 1,000-word frequency band diminished to 1.83%–2.04%, while from the fifth band onward, the coverage dropped to ≤ 1%.

3.2. Analysis and Discussion of Overall Corpus Coverage

Table 3 displays the overall cumulative lexical coverage percentages of the entire corpus achieved at each 1,000-word frequency band until the vocabulary sizes necessary

Table 2. Cumulative Lexical Coverage at Each Word-Family Level of the Monthly Corpora from June 2022 to May 2023.

Lists	Jun, 2022		Jul, 2022		Aug, 2022		Sep, 2022		Oct, 2022		Nov, 2022		Dec, 2022		Jan, 2023		Feb, 2023		Mar, 2023		Apr, 2023		May, 2023	
	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM	%	CUM
SLs	8.74	8.74	8.5	8.5	8.76	8.76	8.22	8.22	8.44	8.44	8.85	8.85	8.78	8.78	8.97	8.97	9.1	9.1	9.08	9.08	8.91	8.91	9.02	9.02
Band 1	68.82	77.56	69.01	77.51	68.28	77.04	68.91	77.13	68.98	77.4	68.47	77.32	68.67	77.45	68.7	77.68	68.7	77.8	67.96	77.04	68.06	76.97	68.27	77.29
Band 2	10.02	87.58	10.14	87.65	10.35	87.39	10.28	87.41	10.17	87.6	10.33	87.65	10.44	87.88	10.34	88.02	10.17	87.97	10.52	87.56	10.4	87.36	10.23	87.52
Band 3	6.87	94.45	6.84	94.49	7.03	94.43	6.83	94.25	6.76	94.4	6.97	94.62	6.68	94.56	6.55	94.57	6.64	94.61	6.87	94.42	6.98	94.34	6.83	94.35
Band 4	1.96	96.41	1.91	96.4	1.97	96.39	2.04	96.28	2.04	96.4	1.93	96.55	1.87	96.43	1.86	96.43	1.87	96.48	1.83	96.26	1.85	96.19	1.84	96.19
Band 5	1.01	97.42	1.00	97.4	0.99	97.38	1.00	97.28	0.98	97.4	0.93	97.48	0.97	97.4	0.98	97.41	0.98	97.46	1.03	97.28	0.99	97.18	1.11	97.29
Band 6	0.74	98.16	0.72	98.12	0.83	98.21	0.8	98.07	0.76	98.1	0.76	98.25	0.73	98.13	0.7	98.1	0.72	98.18	0.86	98.14	0.84	98.03	0.72	98.01
Band 7	0.42	98.58	0.42	98.54	0.41	98.62	0.42	98.49	0.41	98.5	0.38	98.62	0.41	98.54	0.39	98.49	0.4	98.58	0.41	98.56	0.4	98.42	0.4	98.41
Band 8	0.35	98.93	0.35	98.89	0.35	98.97	0.34	98.84	0.34	98.9	0.36	98.98	0.36	98.9	0.39	98.88	0.38	98.96	0.38	98.93	0.39	98.82	0.38	98.8
Band 9	0.22	99.15	0.23	99.12	0.23	99.2	0.26	99.09	0.27	99.2	0.22	99.2	0.2	99.09	0.22	99.1	0.24	99.2	0.21	99.15	0.20	99.02	0.22	99.02
Band 10	0.12	99.27	0.14	99.27	0.11	99.32	0.14	99.23	0.15	99.3	0.13	99.33	0.13	99.22	0.13	99.23	0.12	99.32	0.12	99.27	0.15	99.17	0.13	99.15
Band 11	0.11	99.38	0.11	99.38	0.12	99.43	0.12	99.35	0.11	99.4	0.10	99.43	0.12	99.34	0.11	99.34	0.11	99.43	0.10	99.37	0.10	99.27	0.11	99.26
Band 12	0.11	99.49	0.11	99.49	0.09	99.52	0.11	99.46	0.11	99.5	0.10	99.53	0.12	99.46	0.10	99.44	0.11	99.54	0.10	99.47	0.10	99.37	0.11	99.36
Band 13	0.06	99.55	0.06	99.54	0.05	99.57	0.07	99.52	0.07	99.6	0.06	99.59	0.06	99.52	0.07	99.51	0.07	99.61	0.05	99.52	0.06	99.43	0.06	99.42
Band 14	0.04	99.58	0.03	99.57	0.04	99.61	0.04	99.56	0.04	99.6	0.05	99.64	0.04	99.55	0.04	99.55	0.04	99.64	0.04	99.56	0.05	99.48	0.04	99.46
Band 15	0.02	99.61	0.03	99.6	0.02	99.63	0.02	99.58	0.02	99.7	0.03	99.67	0.04	99.59	0.04	99.6	0.04	99.68	0.04	99.6	0.05	99.52	0.05	99.51
Band 16	0.02	99.63	0.02	99.62	0.02	99.65	0.02	99.6	0.02	99.7	0.02	99.69	0.03	99.62	0.02	99.62	0.03	99.71	0.01	99.61	0.02	99.55	0.02	99.53
Band 17	0.02	99.65	0.02	99.64	0.02	99.67	0.02	99.63	0.03	99.7	0.03	99.72	0.03	99.65	0.02	99.64	0.02	99.73	0.02	99.64	0.02	99.57	0.03	99.56
Band 18	0.01	99.67	0.02	99.66	0.01	99.68	0.02	99.64	0.02	99.7	0.01	99.73	0.02	99.67	0.02	99.65	0.02	99.75	0.03	99.66	0.03	99.59	0.02	99.58
Band 19	0.01	99.67	0.01	99.67	0.01	99.68	0.01	99.65	0.01	99.7	0.01	99.74	0.01	99.68	0.01	99.66	0.01	99.76	0.01	99.67	0.01	99.6	0.01	99.59
Band 20	0.01	99.68	0.01	99.68	0.01	99.69	0.01	99.66	0.01	99.7	0.01	99.75	0.01	99.69	0.00	99.66	0.02	99.77	0.02	99.69	0.01	99.61	0.01	99.6
Band 21	0.01	99.69	0.01	99.68	0.01	99.71	0.00	99.66	0.00	99.7	0.01	99.76	0.01	99.7	0.01	99.67	0.01	99.78	0.01	99.7	0.01	99.62	0.00	99.6
Band 22	0.01	99.7	0.01	99.69	0.01	99.72	0.01	99.68	0.01	99.8	0.01	99.77	0.01	99.7	0.01	99.67	0.01	99.79	0.01	99.7	0.01	99.63	0.00	99.61
Band 23	0.03	99.73	0.04	99.73	0.03	99.75	0.03	99.7	0.03	99.8	0.02	99.8	0.04	99.74	0.03	99.7	0.03	99.82	0.03	99.73	0.03	99.65	0.03	99.64
Band 24	0.01	99.73	0.01	99.74	0.00	99.76	0.00	99.71	0.00	99.8	0.01	99.8	0.00	99.74	0.00	99.7	0.00	99.82	0.01	99.74	0.00	99.66	0.01	99.65
Band 25	0.03	99.77	0.02	99.75	0.01	99.77	0.01	99.72	0.02	99.8	0.02	99.82	0.02	99.77	0.02	99.73	0.02	99.84	0.05	99.79	0.03	99.69	0.02	99.66
Off lists	0.23	100	0.25	100	0.23	100	0.28	100	0.2	100	0.18	100	0.23	100	0.27	100	0.16	100	0.21	100	0.31	100	0.34	100

for 95%, 98%, and 100% lexical coverage were reached. The number of tokens at each frequency level within all the monthly corpora was totaled to attain the overall results for the entire corpus. The first column in **Table 3** presents the total number of tokens for each list; the second column shows the individual coverage percentage of each list compared to all tokens of the entire corpus; and the third column displays the cumulative lexical coverage of the bands, indicating at what points the 95%, 98%, and 100% levels were reached.

As these results emphasize, proper nouns represented a high percentage of the tokens classified as "off list," with an average 83.12% coverage of the unlisted tokens. These were added to the coverage of the SLs, which represented 8.93% of the entire corpus, for a total of 1,847,211 tokens. This was the third-greatest list coverage after the first and second 1,000-word frequency bands. More than 14 million tokens were classified under the first 1,000-word frequency band, providing a coverage of 68.32%. The second 1,000-word frequency band included 2,092,650 tokens and constituted 10.11% of the corpus. However, the third 1,000-word frequency band highlighted a considerable decrease in coverage, accounting for only 6.79% and totaling 1,405,895 tokens. The subsequent coverage percentages declined consistently thereafter. In fact, the decline was less than 1% from the sixth frequency band onward, indicating only minute cumulative increases in lexical coverage. Another interesting finding in the current study is the broad lexical coverage provided by the 3,000 most frequent word families, which comprised 94.15% of the corpus.

The first question that the present study sought to answer concerned the necessary vocabulary size to attain 95% lexical coverage of online written political news reports. As evidenced in **Table 3**, by including the coverage of the SLs within the cumulative coverage of the first 1,000-word frequency band, the present study shows that 95% coverage could be achieved between the BNC/COCA's third and fourth 1,000-word frequency bands. Therefore, it could be speculated that a vocabulary size of approximately 3,500 word families would enable minimal reading comprehension of political news reports. The second question concerned the vocabulary size required for 98% lexical coverage of online written political news reports. Cumulative 98% lexical coverage was found to be attained between the BNC/COCA's sixth and seventh 1,000-word frequency bands, indicating that a vocabulary size of approximately 6,500 word families could be sufficient for optimal comprehension of political news reports.

With regard to previous lexical profiling studies on news reports, the 95% coverage figure in the present study echoes Nation [11], who found that the BNC's most frequent 4,000 words plus proper nouns accounted for approximately 95% of the running words of newspapers and novels. Similarly, the current study reinforces the finding of Ha that a vocabulary comprising the most frequent 3,000–4,000 word

Table 3. Cumulative Lexical Coverage at Each Word-Family Level of the Entire Corpus.

Lists	Total Token at Each Band	Coverage % at Each Band	Cumulative Coverage % of the Lists	Lists	Total Token at Each Band	Coverage % at Each Band	Cumulative Coverage % of the Lists
SLs	1,847,211	8.93	8.93	Band 14	8,017	0.04	99.38
Band 1	14,138,626	68.32	77.25	Band 15	6,987	0.03	99.41
Band 2	2,092,650	10.11	87.36	Band 16	4,581	0.02	99.44
Band 3	1,405,895	6.79	94.15	Band 17	4,049	0.02	99.46
Band 4	399,486	1.93	96.08	Band 18	3,573	0.02	99.47
Band 5	209,722	1.01	97.1	Band 19	1,878	0.01	99.48
Band 6	158,295	0.76	97.86	Band 20	1,955	0.01	99.49
Band 7	83,180	0.4	98.26	Band 21	1,511	0.01	99.5
Band 8	85,627	0.41	98.68	Band 22	1,777	0.01	99.51
Band 9	45,759	0.22	98.9	Band 23	10,039	0.05	99.56
Band 10	27,064	0.13	99.03	Band 24	1,161	0.01	99.56
Band 11	24,008	0.12	99.15	Band 25	5,360	0.03	99.59
Band 12	25,033	0.12	99.27	Off lists	85,306	0.41	100
Band 13	15,672	0.08	99.34	Total	20,694,422	100	

families in the BNC/COCA list provided 95% coverage of a massive corpus of online newspapers and magazines published in different English- and non-English-speaking countries [53].

Regarding the vocabulary size required for optimal comprehension coverage at the 98% coverage level, the results of the present study (i.e., 6000–7000 word families) were less demanding than the figures (8,000–9,000 words plus proper nouns) reported by Nation [11]. In general, previous research has reported higher figures than those derived from the current study. Kaneko, for example, found that 98% lexical coverage of reading passages from an authentic TOEFL iBT examination necessitated approximately 10,000 word families plus proper nouns and defined words [56]. Collins reported an even higher figure of 12,000 word families needed to achieve 98% coverage [43]. Webb and Paribakht reported comparable findings, indicating that the reading comprehension texts of CanTEST, an English proficiency examination for Canadian university entry, necessitated 14,000 word families in the BNC (in addition to proper nouns and interjections) to achieve 98% coverage [49]. Nevertheless, the 98% coverage levels identified in the present study replicate those of Hsu [52], who found that the Voice of America news network and international radio broadcaster required 6,000 word families for optimal comprehension in the BNC/COCA (i.e., 98% lexical coverage). Moreover, they align generally with the results of Ha's lexical profiling study on web-based newspapers and magazines [53], which reported that the optimal lexical coverage of 98% in most cases required between 6,000–7,000 and 7,000–8,000 word families.

The two coverage levels of 95% at the 3,000–4000 and 98% at the 6,000–7,000 word-family frequency bands reported in the current study are also comparable to those from research on the lexical profiling of English learning textbooks. They reflect similar coverage levels to those reported by Chujo on Japanese junior and senior high-school texts [41], by Rahmat and Coxhead on an Indonesian EFL textbook series [44], by Yang and Coxhead on an English textbook series in China [45], and by Garcia on EFL upper secondary textbooks in Sweden [46]. These comparable findings may render news reports a practical, often inexpensive source of accessible and authentic supplementary material for EFL teaching and learning settings at the intermediate and advanced levels.

Regarding whether news reports can be recommended for EFL/ESL students preparing for English proficiency examinations, in general, studies on the lexical coverage of the reading passages of such exams have tended to yield somewhat more demanding results concerning the requisite vocabulary size for comprehension than the present study, with a few exceptions. For example, the lexical profiling of political news reports in the current study suggests a lower 95% coverage level of the BNC/COCA lists (between the third and fourth word-family frequency bands) than the 5,900–6,300 word families identified by Chujo [41], the 6,000 word families suggested by Webb and Paribakht [49], the 6,000 word families determined by Collins [43], the 6,000 word families found by Kaneko [56], and the 6,150 word families ascertained by Chujo and Nishigaki [47]. However, some previous studies have reported comparable figures to the results of the present study. Kaneko, for example, found that among seven English proficiency tests utilized as university entrance examinations

in Japan, the reading passages required 2,000–5,000 most frequent word families from Nation's BNC/COCA lists for 95% coverage and 3,000–8,000 word families for 98% coverage[48]. Similarly, Kanzaki found that the first 3,000 word family bands from Nation's 2012 version of the BNC/COCA word-family lists (plus SLs) provided up to 96.79% coverage for 34 TOEIC practice tests published in Japan or South Korea between 2005–2014, whereas the first 4,000 word families (plus SLs) provided 98.24% coverage[50, 51]. The variation in lexical demand between the reading passages in language proficiency tests and written political news reports can be attributed to the fact that political news reports are designed to be more accessible and engaging for a broad public audience, often prioritizing clarity and reader engagement over academic complexity.

In comparison to the findings of previous studies on genres other than news reports, there are two noteworthy broad interpretations of the findings of the current study. First, the current study identified relatively moderate requirements for both the 95% and 98% coverage levels. In fact, the vocabulary size needed for 95% lexical coverage of news reports in the current study is either similar to or smaller than most previous studies. Moreover, the vocabulary size required for 98% lexical coverage of news reports is remarkably lower. Consequently, as a second observation, unlike the vast majority of previous lexical profiling research, the current study revealed a comparatively moderate gap between the vocabulary sizes required for the 95% and 98% coverage levels. This differentiation could be due to the general nature of the topics covered by political news reports, as they are supposed to cover and discuss general issues and matters that are more appealing to a broad public.

4. Teaching Implications, Limitations and Suggestions for Future Research

The findings of the current study suggest a number of pedagogical implications. First, the notable cumulative coverage levels of the first three most frequent word-family bands in Nation's BNC/COCA lists support the value of these word families as gradual prioritized targets for EFL learners[37]. For this purpose, as a second implication, the current study underscores that news reports represent a useful source of accessible, authentic supplementary material for EFL/ESL explicit and the implicit teaching and learning of vocabulary at the intermediate and advanced levels. Learners can benefit from the different linguistic and intralinguistic advantages of accessible and possibly free authentic material that provides substantial repetition and multiple encounters with different forms of the target frequent word families. Third, the coverage rarity of words from the sixth frequency band onward suggests that students can incidentally learn words from this band onward by practicing extensive reading to maximize their exposure to mid- and low-frequency vocabulary, although this approach should be accompanied by training in comprehension compensation strategies[2]. Fortunately, lexical coverage at the 95% level has been reported to have improved learners' chances of successfully guessing new words. For example, Liu and Nation reported that learners' inference of new words improved at 96% text coverage compared to 90%[57]. Similarly, Laufer found that successful lexical inferencing of unknown words improved at the 95% and 98% coverage levels compared to 90%[29]. A closely related strategy to the improvement of inferring the meaning of unknown words is the study of the word formation system in English. This is important because the findings of this study are based on the assumption that recognizing one or more family members of a new word will most likely allow learners to recognize its inflected forms and derivations. Fourth, an essential teaching implication from the current study is the need for teachers' careful selection of the target texts for their students. Laufer, in this regard, has suggested two steps[1]. First, teachers should measure their learners' vocabulary size at the beginning of their study using easily accessible computerized vocabulary-level tests. Then, teachers should evaluate the difficulty of the target texts for their students using freely available lexical profiling programs. This approach will assist teachers in initially identifying their learners' lexical knowledge gaps. Subsequently, they can devise a practical plan to enhance their students' vocabulary size through word-focused instruction, aiming to achieve one of the lexical thresholds suggested by previous research, specifically, 95% or 98% lexical coverage.

Although it is assumed that 95% text coverage allows learners to adequately understand the text and successfully guess new words, they need to recognize 98% of the vocabulary within a text to read it independently and without

assistance. Consequently, a crucial teaching and learning implication concerns the gap reported in the current study between the vocabulary size of 3,000–4000 word families in Nation's BNC/COCA lists required for the 95% lexical coverage level and the 6000–7000 required for the 98% level [37]; learners would need to double their efforts to bridge this gap through extensive reading of political news reports.

The findings of the present study should, however, be considered tentative due to three main methodological limitations. First, the corpus was limited to one source of authentic material (i.e., the online American version of the *New York Times*). Future research could expand the geographical representation of the corpus by considering the lexical profiles of written political news reports from other American and non-American news sources, thus enabling better generalization of the study findings. Second, the AntWordProfiler software is unable to calculate the coverage figures of multiword units or differentiate between homographs [11]. Therefore, despite the feasibility of corpus-informed reading material selection, language instructors should consider these limitations and plan the necessary amendments and explanations prior to using such materials in the classroom. A third limitation of this research is that some previous studies used a different vocabulary list (predominantly the BNC word-family frequency lists), potentially compromising the comparison and alignment of this study's conclusions with those of earlier studies.

A final note on future research: It would be interesting to illuminate the lexical profiling of other sections of online newspapers, especially those that could prove beneficial for EFL/ESL academic settings and engage with learners' interests or specialty (e.g., business, the arts, technology, science, and sports).

5. Conclusions

Lexical profiling research has recently witnessed significant interest because it underscores the importance of vocabulary development for independent comprehension of written and spoken discourse and, in turn, provides insightful vocabulary learning objectives for teachers and learners. The current study sought to determine the vocabulary size, in terms of word families, required for learners of English to attain adequate (95% lexical coverage) and optimal (98% lexical coverage) comprehension of online written political news reports in English. For this purpose, a corpus of 20 million tokens was collected from the online American *New York Times* newspaper over a two-year period. The lexical coverage levels of each monthly corpus and that of the entire corpus were analyzed against Nation's BNC/COCA 25 word-family lists using the AntWordProfiler software [37, 38].

The analysis of the monthly corpora revealed relatively consistent lexical coverage by the 25 word-family frequency bands. A vocabulary size of 3,000–4,000 word families is necessary to reach 95% coverage throughout the monthly corpora. In contrast, the 98% coverage is less stable, with 18 months requiring a vocabulary size of 6,000 word families, whereas 1 month required 7,000, 4 months 8,000, and 1 month 9,000 word families.

The overall analysis of the entire corpus revealed that the vocabulary size required for 95% lexical coverage of online written political news reports in the entire corpus may be achieved between the BNC/COCA's third and fourth 1,000-word frequency bands, including the supplementary lists (i.e., a vocabulary size of approximately 3,500 words). The incremental 98% lexical coverage, on the other hand, was achieved between the BNC/COCA's sixth and seventh 1000-word frequency bands, suggesting that a vocabulary size of approximately 6,500 words is necessary for optimal comprehension of online written political news reports.

The 3,000–4,000 word families required for the 95% coverage in the current study are comparable to the findings of previous research on newspaper corpora [11, 53]. Similarly, the 6,000–7000 word families reported in the current study for 98% coverage replicates those of Hs and, largely, those of Ha, but are less demanding than the 8,000–9,000 word families reported by Nation [11, 52, 53]. Likewise, when compared to the findings from research on the lexical profiling of English learning textbooks, the two coverage levels reported in the current study are similar to the coverage levels reported by a number of previous studies. The similarity in text coverage levels between the current study and previous research on English textbooks suggests that the texts analyzed in this study can serve as valuable resources for vocabulary instruction. However, when compared with the results of previous research on the lexical profiling of reading passages from English proficiency tests, this study indicates that political news reports are significantly less challenging.

Funding

This work was supported by the Deanship of Graduate Studies and Scientific Research at Qassim University grant number [QU-APC-2025].

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study was manually compiled by the author through extensive personal effort over an extended period. To protect the integrity of this work and due to related project considerations, the data is not publicly available. Nonetheless, reasonable requests for access may be considered by the corresponding author on a case-by-case basis.

Conflicts of Interest

The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] 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. DOI: https://doi.org/10.1002/tesq.140

[2] 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(2), 15–30.

[3] Nation, I.S.P., 2013. Learning vocabulary in another language, 2nd ed. Cambridge University Press: Cambridge, UK.

[4] Choi, Y., Zhang, D., 2021. The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. International Review of Applied Linguistics in Language Teaching. 59(1), 1–30.

[5] Dagnaw, A.T., 2023. Revisiting the role of breadth and depth of vocabulary knowledge in reading comprehension. Cogent Education. 10(1), 2217345.

[6] Dong, Y., Tang, Y., Chow, B.W.Y., et al., 2020. Contribution of vocabulary knowledge to reading comprehension among Chinese students: A meta-analysis. Frontiers in Psychology. 11, 525369.

[7] Ha, H.T., 2021. Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. Language Testing in Asia. 11(1), 20.

[8] Laufer, B., 1992. How much lexis is necessary for reading comprehension? In: Arnaud, P.J.L., Béjoint, H. (eds.). Vocabulary and applied linguistics. Palgrave Macmillan: London, UK. pp. 126–132. DOI: https://doi.org/10.1007/978-1-349-12396-4_12

[9] Li, H., Gan, Z., 2022. Reading motivation, self-regulated reading strategies and English vocabulary knowledge: Which most predicted students' English reading comprehension? Frontiers in Psychology. 13, 1041870.

[10] Nation, I.S.P., 2001. Learning vocabulary in another language. Cambridge University Press: Cambridge, UK. DOI: https://doi.org/10.1017/CBO9781139524759

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

[12] Qian, D.D., 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. The Canadian Modern Language Review. 56(2), 282–308. DOI: https://doi.org/10.3138/cmlr.56.2.282

[13] Qian, D.D., 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. Language Learning. 52(3), 513–536. DOI: https://doi.org/10.1111/1467-9922.00193

[14] Spencer, M., Quinn, J.M., Wagner, R.K., 2017. Vocabulary, morphology, and reading comprehension. In: Cain, K., Compton, D., Parrila, R. (eds.). Theories of reading development. John Benjamins Publishing Company: Amsterdam, Netherlands. pp. 239–256.

[15] Nation, I.S.P., 2008. Teaching vocabulary: Strategies and techniques. Heinle Cengage Learning: Boston, MA, USA.

[16] Klinmanee, N., Sopprasong, L., 1997. Bridging the EFL vocabulary gap between secondary school and university: A Thai case study. Guidelines. 19(1), 1–10.

[17] Kyongho, H., Nation, P., 1989. Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. Reading in a Foreign Language. 6, 323–335.

[18] Schmitt, N., Carter, R., 2000. The lexical advantages of narrow reading for second language learners. TESOL Journal. 9(1), 4–9. DOI: https://doi.org/10.1002/j.1949-3533.2000.tb00220.x

[19] Sternfeld, S., 1989. The University of Utah's immersion/multiliteracy program: An example of an area studies approach to the design of first-year college foreign language instruction. Foreign Language Annals. 22(4), 341–352. DOI: https://doi.org/10.1111/j.1944-9720.1989.tb02757.x

[20] Day, R., Bamford, J., 1998. Extensive reading in the second classroom. Cambridge University Press: Cambridge, UK.

[21] Schmitt, N., 2000. Vocabulary in language teaching. Cambridge University Press: Cambridge, UK.

[22] Cho, K.S., Ahn, K.O., Krashen, S., 2005. The effects of narrow reading of authentic texts on interest and reading ability in English as a foreign language. Reading Improvement. 42(1), 58–65.

[23] Kang, E.Y., 2015. Promoting L2 vocabulary learning through narrow reading. RELC Journal. 46(2), 165–179. DOI: https://doi.org/10.1177/0033688215586236

[24] Krashen, S.D., 2004. The power of reading: Insights from the research, 2nd ed. Libraries Unlimited: Westport, CT, USA.

[25] Rodgers, M.P., Webb, S., 2011. Narrow viewing: The vocabulary in related television programs. TESOL Quarterly. 45(4), 689–717. DOI: https://doi.org/10.5054/tq.2011.268062

[26] 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. DOI: https://doi.org/10.1111/j.1540-4781.2011.01146.x

[27] Laufer, B., 2020. Lexical coverages, inferencing unknown words and reading comprehension: How are they related? TESOL Quarterly. 54(4), 1076–1085. DOI: https://doi.org/10.1002/tesq.3004

[28] Nurmukhamedov, U., Webb, S., 2019. Lexical coverage and profiling. Language Teaching. 52(2), 188–200. DOI: https://doi.org/10.1017/S0261444819000028

[29] Webb, S., Nation, I.S.P., 2013. Computer-assisted vocabulary load analysis. In: Chappelle, C. (ed.). Encyclopedia of applied linguistics. Wiley-Blackwell: Oxford, UK. pp. 844–853.

[30] Bonk, W.J., 2000. Second language lexical knowledge and listening comprehension. International Journal of Listening. 14(1), 14–31. DOI: https://doi.org/10.1080/10904018.2000.10499033

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

[32] Laufer, B., 1989. What percentage of text-lexis is essential for comprehension? In: Laurén, C., Nordman, M. (eds.). Special language: From humans thinking to thinking machines. Multilingual Matters: Clevedon, UK. pp. 316–323.

[33] Laufer, B., Sim, D.D., 1985. Measuring and explaining the reading threshold needed for English for academic purposes texts. Foreign Language Annals. 18(5), 405–411. DOI: https://doi.org/10.1111/j.1944-9720.1985.tb00973.x

[34] Stæhr, L.S., 2009. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. Studies in Second Language Acquisition. 31(4), 577–607. DOI: https://doi.org/10.1017/S0272263109990039

[35] 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. DOI: https://doi.org/10.1093/applin/ams074

[36] Webb, S., 2021. Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. Reading in a Foreign Language. 33(2), 278–293. DOI: https://doi.org/10125/67407

[37] Nation, I.S.P., 2017. The BNC/COCA Level 6 word family lists (Version 1.0.0). Available from: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx (cited 15 May 2021).

[38] Anthony, L., 2014. AntWordProfiler (Version 1.4.1) [software]. Available from: https://www.laurenceanthony.net/software/antwordprofiler/ (cited 15 May 2021).

[39] Heatley, A., Nation, P., Coxhead, A., 2002. RANGE (Version 1.0) [software]. Victoria University of Wellington: Wellington, New Zealand. Available from: https://www.victoria.ac.nz/lals/research/projects/Range (cited 15 May 2021).

[40] Cobb, T., 2021. VocabProfile [software]. Available from: https://www.lextutor.ca/vp/ (cited 15 May 2021).

[41] Chujo, K., 2004. Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In: Nakamura, J., Inoue, N., Tabata, T. (eds.). English corpora under Japanese eyes. Brill: Leiden, The Netherlands. pp. 231–249.

[42] Webb, S., Macalister, J., 2013. Is text written for children appropriate for L2 extensive reading? TESOL Quarterly. 47(2), 300–322. DOI: https://doi.org/10.1002/tesq.70

[43] Collins, J.B., 2017. Applying the lexical coverage hypothesis to establish the suitability of EFL reading materials: A case study of the TOEFL (ITP). APU Journal of Language Research. 3, 29–39. DOI: https://doi.org/10.34409/apujlr.3.0_29

[44] Rahmat, Y.N., Coxhead, A., 2021. Investigating vocabulary coverage and load in an Indonesian EFL textbook series. Indonesian Journal of Applied Linguistics. 10(3), 804–814. DOI: https://doi.org/10.17509/ijal.v10i3.31768

[45] Yang, L., Coxhead, A., 2022. A corpus-based study of vocabulary in the New Concept English textbook series. RELC Journal. 53(3), 597–611. DOI: https://doi.org/10.1177/0033688220964162

[46] Garcia, D.V., 2023. Exploring the vocabulary content of upper secondary EFL textbooks in Sweden: A corpus-based analysis of types, lexical coverage, progression, and academic words [Master's thesis]. Dalarna University: Falun, Sweden. Available from: https://du.diva-portal.org/smash/get/diva2:1783297/FULLTEXT01.pdf (cited 11 April 2024).

[47] Chujo, K., Nishigaki, C., 2003. Bridging the vocabulary gap: From EGP to EAP. JACET Bulletin. 37, 73–84.

[48] Kaneko, M., 2020. Lexical frequency profiling of high-stakes English tests: Text coverage of Cambridge First, EIKEN, GTEC, IELTS, TEAP, TOEFL, and TOEIC. JACET Journal. 64, 79–93. DOI: https://doi.org/10.32234/jacetjournal.64.0_79

[49] Webb, S., Paribakht, T.S., 2015. What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? English for Specific Purposes. 38, 34–43. DOI: https://doi.org/10.1016/j.esp.2014.11.001

[50] Kanzaki, M., 2017. Lexical coverage of the TOEIC. In: Brooks, G. (ed.). The 2016 PanSIG Journal. JALT: Tokyo, Japan. pp. 126–133.

[51] Nation, I.S.P., 2012. The BNC/COCA word family lists. Available from: http://www.victoria.ac.nz/lals/about/staff/paul-nation (cited 15 May 2021).

[52] Hsu, W., 2019. Voice of America news as voluminous reading material for mid-frequency vocabulary learning. RELC Journal. 50(3), 408–421. DOI: https://doi.org/10.1177/0033688218764460

[53] Ha, H.T., 2022. Lexical profile of newspapers revisited: A corpus-based analysis. Frontiers in Psychology. 13, 800983. DOI: https://doi.org/10.3389/fpsyg.2022.800983

[54] Nation, I.S.P., Webb, S.A., 2011. Researching and analyzing vocabulary. Heinle, Cengage Learning: Boston, MA, USA.

[55] Hwang, K., Nation, P., 1989. Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. Reading in a Foreign Language. 6(1), 323–335.

[56] Kaneko, M., 2014. Is the vocabulary level of the reading section of the TOEFL Internet-Based Test beyond the lexical level of Japanese senior high school students? Vocabulary Learning and Instruction. 3(1), 44–50. DOI: http://doi.org/10.7820/vli.v03.1.kaneko

[57] Liu, N., Nation, I.P., 1985. Factors affecting guessing vocabulary in context. RELC Journal. 16(1), 33–42.