From Needs Analysis to the Establishment of a University Vocabulary List for English Majors

Konul Hajiyeva
Department of Foreign Languages Teaching Methodology
Azerbaijan University of Languages
Baku, Azerbaijan
Email: hajiyeva.konul83@gmail.com

Received:12/27/2022 Accepted:04/28/2023 Published: 06/24/2023

Abstract
This paper concerns the academic vocabulary needs of undergraduate English majors in a tertiary context in Azerbaijan. Based on the results of the previous studies, it is suggested that developing an appropriate course material to meet the specific learning needs of these students will be beneficial for them. Thus, this study aims to identify the percentage of the running words and high-frequency word families beyond the 2,000 most-frequent words in the university textbook corpus to help the students to attain 98% lexical coverage. To address English majors’ scant academic vocabulary knowledge, a University Vocabulary List has been created. First, a university textbook corpus comprising the words used in 11 subject-specific course textbooks across subject areas taught at Azerbaijan University of Languages was compiled. Next, the range and frequency distribution of the words beyond the British National Corpus scale 2,000-word families was examined. Finally, the 396 most frequently occurring word families in the corpus were chosen; these formed the proposed University Vocabulary List for English majors, accounting for a 6.96% lexical coverage. The University Vocabulary List for English majors and the British National Corpus 2,000 high-frequency word families, including proper nouns, provide a text coverage of 97.28%, which is only marginally lower than the suggested lexical threshold of 98% suggested by Laufer and Ravenhorst-Kalovski (2010) and Nation (2006). It is, therefore, argued that the University Vocabulary List represents a sufficiently frequent and relevant vocabulary load for English majors whose secondary educational background and first-year general English textbooks have not equipped them for the linguistic challenges they face at the initial stages of their tertiary education. The article concludes by stating the pedagogical implications for developing these learners’ knowledge and use of the most frequent word families.

Keywords: English majors, frequency, needs analysis, range, University Vocabulary List, vocabulary development

Cite as: Hajiyeva, K. (2023). From Needs Analysis to the Establishment of a University Vocabulary List for English Majors Arab World English Journal, 14 (2) 77-89.
DOI: https://dx.doi.org/10.24093/awej/vol14no2.6
Introduction

English for Academic Purposes (EAP), the first recorded use of which dates back to 1975, is concerned with those “communication skills in English which are required for study purposes in formal education systems” (Jordan, 2012, p. 1) and the transition from “pedagogic genres to increasingly authentic genres” (Carkin, 2005, p. 85) associated with various disciplines. Despite increased professionalism in the teaching of EAP at university level in most English-speaking and non-English-speaking countries (Dudley-Evans & St. John, 1998; Jordan, 2012), it is still a “first” in a tertiary context in Azerbaijan. In other words, English language, which has been recognized as a medium of instruction in some higher education institutions (HE) in Azerbaijan, is in the process of changing from teaching a language for general purposes to teaching a language for academic and, in some instances, for Specific Academic Purposes (ESAP). Students studying within such a context have varying proficiency levels in English, different EAP needs, and a diversity of interfaces between English and their major fields of study.

The context of the present study is the Azerbaijan University of Languages (AUL) in Baku, Azerbaijan. Since the AUL joined the Bologna process in 2005, it has been a staple of its education policy that undergraduates – English majors (future English language teachers) – in the Faculty of Education follow elective and compulsory subject-specific courses through the medium of English and earn credits for these courses. This idea has led to a fundamental transformation of the curriculum and syllabus design, teaching and learning context, and in materials development. However, a hitherto unpublished investigation of course syllabi and teaching materials shows that, although AUL joined the Bologna conventions more than 15 years ago, no formal needs analysis has ever been conducted to identify the academic needs of these undergraduate students since then.

Carkin (2005) emphasizes that the main building blocks of an EAP program are academic texts, tasks, and content, which facilitate language learning. In this sense, instruction that takes the form of a highly pragmatic approach to learning, encompassing needs analyses, evaluation, academic skills, vocabulary, subject-specific content, and tasks should be included to support the students studying in a tertiary education environment.

With these suggestions and the tertiary setting in mind, the researcher has undertaken a needs analysis as a “cornerstone” (Dudley-Evans & St John, 1998, p. 121) tied to a particular context and population in this research area prior to developing a course syllabus and teaching materials for the students studying at a higher education institution in Azerbaijan.

In attempting to bridge the gap, the academic needs of Azerbaijani English majors in an educational context have been investigated in several ways (Hajiyeva, 2014; 2015a, b, c). These studies include the assessment of both the receptive and the productive vocabulary knowledge of learners at the beginning of the first year of tertiary education. It also includes estimates of their vocabulary growth after one year of instruction. Other studies include a corpus-based lexical analysis of the academic texts – subject-specific university textbooks – and general English textbooks that students must be able to internalize and the measurement of lexical richness in students’ answers to examination papers. These studies focus on the frequency, range, distribution, and occurrence of high-frequency, academic and low-frequency words, and specialized vocabulary both in the subject-specific university textbooks and in student-produced texts in a tertiary context. Among the significant findings of these above-described studies are:

- the low estimates of the English majors’ vocabulary-size figures at the beginning of the first-year of tertiary education (2,104-word families receptively and 806-word families productively),
a vocabulary size growth of productive as opposed to receptive vocabulary knowledge after one year of instruction at the beginning of the second year (974-word families known productively and 1,966-word families known receptively),
• the inappropriateness (that is, an inadequate representation of high-frequency word families) of the general English textbooks,
• the lexical threshold necessary to read and comprehend the subject-specific textbooks is much higher than the students can cope with (3,500–4,000 word families for 95% text coverage and the knowledge of 6,500–7,000 word families for the coverage of 98%),
• the composition of the subject-specific university textbooks is lexically versatile (for example, low coverage of the academic words in the textbooks), and
• while there has been substantial growth in the productive vocabulary knowledge and use, the students can still not convey or use meaning, form, associations, collocations, and register accurately.

Based on the outcomes of the previous studies, it is suggested that developing tailor-made course materials to meet the specific academic needs of these students will be beneficial to them. With these ideas in mind, this study aims to develop a University Vocabulary List for English majors which will provide a basis for an EAP course syllabus.

Making a word list is not simply a mechanical task, so judgements based on well-established criteria need to be considered in the development of such word lists. Therefore, the present study chooses 98% text coverage as the primary aim, since 95% is perceived to be a relatively low threshold, considering the academic needs and demands of this particular group of students. In light of what has been said so far, this study seeks to answer the following research questions:

1. What percentage of the running words in the university textbook corpus is the UVL required to cover in order to help the students to attain 98% lexical coverage, including proper nouns?
2. Which high-frequency words beyond the 2,000 most-frequent word families make up a University Vocabulary List?

This study takes into account both the students’ and stakeholders’ (faculty members, heads of department and teachers) needs, for example, to be able to read and comprehend subject-specific textbooks, sit examinations, earn credits, adhere to the conventions of the Bologna process and bring in well-trained specialists. It is suggested that the outcome product of this study will, therefore, serve as a basis for a lexical syllabus that will be useful, in terms of text coverage and frequency distribution, to Azerbaijani English majors who have nothing like mastery of the 2,000 most frequent word families.

With these ideas in mind, current study discusses recent studies on vocabulary size, lexical coverage, types and lists in the Literature Review, describes the research methods, settings and participants in the Methods, introduces the major findings and discusses them in the Results and Discussion sections relatively, before summarising them in the Conclusion part.

Literature Review

Vocabulary Size and Lexical Coverage

Vocabulary is an essential component of language use, and something that anyone involved in the learning process can agree on is that learning vocabulary is an essential part of mastering a second or foreign language (Schmitt, 2010a; English, in the current study). Since we use language
to communicate, an important issue in vocabulary studies is therefore the extent of vocabulary necessary to enable communication. Vocabulary is also a good predictor of reading comprehension and other language skills (see Milton & Treffers-Daller, 2013; Qian, 2002 for further discussion). The number of words or the amount of vocabulary necessary depends on one’s learning goals. The available literature suggests two thresholds. For example, it has been estimated that, if 98% coverage of a text is needed for unassisted comprehension, then an 8,000–9,000-word family vocabulary (including proper nouns) is required. Lexical coverage here refers to the “percentage of running words in the text known by the reader” (Nation, 2006, p. 61). A vocabulary knowledge of 4,000–5,000 word families resulting in the coverage of 95% (including proper nouns) is needed for comprehension of a written text (Laufer & Ravenhorst-Kalovski, 2010; Laufer, 2020; Nation, 2006; Schmitt, 2008, 2010a; Schmitt & Schmitt, 2012). Researchers has stated that there is a lexical knowledge threshold that marks the boundary between having and not having sufficient vocabulary knowledge for “adequate reading comprehension”. Laufer and Ravenhorst-Kalovski (2010) has also stated that the term “adequate reading comprehension” (p. 16) has no clear definition since it may refer to different levels of comprehension in other contexts. A lexical threshold, therefore, depends on the predetermined lexical coverage.

The above-described two putative coverage percentages (95% and 98%) signal the possible basic and optimal lexical thresholds necessary for reading comprehension. These figures may appear daunting to both teachers and learners since each word family includes several individual forms, including the root form, its inflections and regular derivations (Nation, 2006). For example, a vocabulary size of 8,000-word families (enabling wide reading) entails knowing 34,660 words (Schmitt, 2010a). Learning and knowing this number of word families can naturally be a challenging task for those learners with limited vocabulary knowledge. It has been suggested, however, that some words deserve more attention and effort than others in different stages of learning for varying purposes (Coxhead, 2011; Nation & Waring, 1997). In other words, if learners have specific lexical gaps in and challenges with reading comprehension, then increasing the lexical coverage through targeting a more restricted vocabulary with a relatively high frequency of occurrences may be a more practical solution (Hsu, 2013).

Vocabulary Types

Over the past 20 years, much has been done in the field of vocabulary in the context of acquiring foreign or second languages (Bogaards & Laufer, 2004). Among other things, this interest has been characterized by the attention paid to the role of word frequency in vocabulary learning, for instance, the cost-benefit of learning frequent, infrequent, and specialized words (see, among others, Coxhead, 2000, 2011; Nation, 2011). For example, Wang, Liang and Ge (2008) believed that not all words are equally important and therefore they agree with Nation’s (2011) division of vocabulary into four categories:

1. high-frequency or general service vocabulary;
2. academic/sub-technical vocabulary;
3. technical vocabulary, and
4. low-frequency vocabulary.

This division points to the fact that in different phases of language learning some vocabularies need more attention than others. High-frequency word families, for example, are the words necessary to achieve basic functionality in a language and constitute the majority of all the running words in all types of text. The classic list of high-frequency words is West’s (1953) A
General Service List of English Words (GSL), which contains approximately 2,000-word families. High-frequency word families appear so frequently in the texts that sufficient time should be spent on learning these words. Academic vocabulary – also called “sub-technical vocabulary”, “specialised non-technical lexis” and “frame words” (Nation, 2011, p. 187) – is the vocabulary type that occurs with medium frequency. It is common to a wide range of academic texts, and accounts for a substantial number of words (~10%) in academic texts.

Unlike academic vocabulary, technical vocabulary contains some words that are very closely related to the topic and subject area of the text. Nation (2011) categorized the degrees of “technicalness” (p. 198) depending on how restricted a word is to a particular area. He, therefore, has classified four categories, depending on the criteria of the relative frequency of form and meaning. According to this classification, words in Category one are technical and they are unique to a particular field in both form and meaning, whereas words in Category two are technical words because the more general sense of the word when used outside the field does not provide ready access to its technical use (see Chung & Nation, 2004 for further discussion). Categories three and four are less technical because they are not unique to a particular field neither in form nor in meaning. They differ from subject area to subject area. For example, in Applied Linguistics, words such as morpheme and lemma fall in the first category, whereas words such as sense, reference, type and token (running word; see below) will be placed in the second category. Technical vocabulary comprises content-bound, frequently occurring words in a given field. In contrast, low-frequency vocabulary includes all the words that are not high-frequency words, excluding the academic words and technical words for a particular subject (Nation, 2011).

Word Lists

Nation and Waring (1997) suggest that language teachers and instructors need to have clear, sensible goals for vocabulary learning. Frequency information provides a rational basis for ensuring that learners obtain the best return on their vocabulary learning effort by ensuring that any words studied will frequently be met with. Vocabulary frequency lists that take account of range – that is, the occurrence of a word across several subsections of a corpus – play an important role in curriculum and syllabus design and in setting learning goals. This does not necessarily imply that learners must be provided with large vocabulary lists as the primary source of their vocabulary learning. However, it does mean that course designers and materials developers should have lists to refer to when they consider the vocabulary component of a language program (Nation, 2004).

Word lists are essential in language teaching for various reasons; they include lists of the words or lexical phrases found in a discourse domain, often including frequency information about the use of particular words (Miller & Biber, 2015). A more recent example of a word list comprising the 2,000 most frequent words in general English is the GSL (West, 1953). The words in the GSL are primarily, but not entirely, chosen based on frequency (Nation, 2004; Schmitt, 2010a) and the list provides 75–80% text coverage. With the advances in corpus-based studies and computational techniques, developing newer and up-to-date word lists has been facilitated. For example, for general English, these include the compilation of the British National Corpus (BNC) word lists (Nation, 2004) and the recent new GSL developed by Brezina and Gablasova (2015) based on an analysis of 12 billion words taken from four general corpora. The authors state that the practical usefulness of the new GSL shows its effectiveness in covering approximately 80% of
the texts in the corpora with only 2,494 lemmas (a word and its inflected forms), with a significant reduction compared to the 4,100 lemmas that West’s GSL needs to reach that coverage.

Corpus-based word lists have also been compiled for more specialized domains. There have been several studies (for further discussion, see American University List (AUL) developed by Praninskas, 1972; University Word List (UWL) devised by Xue & Nation, 1984; Academic Word List (AWL) developed by Coxhead, 2000; Academic Vocabulary List (AVL) developed by Gardner & Davies, 2013) that have investigated the vocabulary needed for academic studies at the university level. Among these lists, Coxhead’s (2000) AWL, which provides approximately 10% coverage of tokens, has played a crucial role in setting vocabulary goals for language courses in most EAP programs. In practice, it means that the 570-word family AWL and the GSL mentioned above give the coverage of 85–90% of the tokens in the text. Nation (2004), on the contrary, states that “beginner learners at the tertiary level would be better off using materials based on the BNC lists, because of the slightly better BNC coverage” (p. 12) than the GSL + AWL text coverage. Moreover, researchers such as Hyland and Tse (2007) question the validity and utility of AWL. They argue that language instructors should instead develop lists of the most important words used in specific academic disciplines. Following this discussion, several word lists have been developed for particular disciplines, for example, for medicine by Wang, Liang, and Ge (2008), Hsu (2013), for engineering by Ward (2009), and for agriculture by Martinez, Beck, and Panza (2009).

All in all, word lists play an indispensable role in designing courses, setting learning goals, guiding the creation of simplified texts, analyzing the vocabulary in texts, analysing lexical richness, guiding the construction of vocabulary tests and, more importantly, creating specialized word lists (Nation & Webb, 2011).

Methods

Participants

The present research targeted undergraduate second-year subject-specific university textbooks for English majors in use in the year of 2015-2016 at the Faculty of Education at AUL, because university entrants (first-year students) at this university start following subject-specific courses taught through the medium of English in the second year of their tertiary education. This corpus, therefore, encompassed the subject-specific courses the second-year lecturers and students used in their classes.

For this study, a University Textbook Corpus (UTC) covering 11 textbooks for the courses Study skills (SS), Critical reading and effective writing (CREW), Communication and Social Interaction (CSI) and Teaching English to Young Learners (TEYL) was developed (for a complete description of the steps involved in the corpus compilation see Hajiyeva, 2015b). These are elective courses offered to students in the second year of their tertiary education, and all of them have to take any two of the courses mentioned above. Students are accredited with seven credits – covered in 105 academic hours – for each course, if they pass the final written examination. The subject areas included in the UTC are versatile; the texts are of different lengths and are intended for students with different backgrounds. All the textbooks were downloaded from internet databases or scanned using the software ABBYY Fine Reader 11 Portable. Consequently, after excluding all the bibliographies, tables, indexes, appendices, and proper nouns (which were gathered in a separate file for further analysis), a corpus of 508,802 (Table one) words comprising four sub-corpora files was created. The UTC was, therefore, not large by contemporary standards, since corpus-based research studies usually consider the representativeness of the corpus highly.
In other words, researchers have traditionally focused on two general factors while creating word lists based on corpus studies: the kinds of discourse included in the corpus (whether general or academic) and the size of the corpus, that is, its representativeness.

An analysis of the studies done by Brezina and Gablasova (2015), Coxhead (2000), Gardner and Davies (2013), West (1953), and Xue and Nation (1984) has shown that vocabulary investigations based on larger corpora do produce different results. The versatile nature of these word lists shows that by analyzing larger corpora, we discover word lists that are different from previous studies. It is assumed that these lists are more valid since they are based on more representative corpora. However, a recent study done by Miller and Biber (2015) has shown that corpus design considerations have usually focused on “issues of external representativeness” (p. 30) (representing the target domain in terms of its size), while “disregarding the issues of internal representativeness” (whether a corpus allows us to achieve reliable and stable quantitative results concerning the use of linguistic features). They, therefore, have suggested that researchers should consider the evaluation of “internal representativeness” in quantitative corpus-based studies. Of course, larger corpora will include a larger set of words in a discourse domain. However, since the aim of the study is to develop a tailor-made word list that meets specific students’ specific academic needs, it is claimed that the UTC represents all the core subject matter of second-year university textbooks needed for study at AUL. As Ward (2009) puts it, “other corpora, while considerably larger, do not address the specific needs of our students” (p. 173). Therefore, the product of this study—a word list for English majors—is likely to reflect the corpus from which it was derived, no matter how large it is.

Table 1. Tokens (running words) of the university textbooks used at AUL

<table>
<thead>
<tr>
<th>SS</th>
<th>CREW</th>
<th>CSI</th>
<th>TEYL</th>
<th>UTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens (running words)</td>
<td>94,563</td>
<td>121,428</td>
<td>141,637</td>
<td>155,992</td>
</tr>
</tbody>
</table>

**Research Instruments**

The instrument used to serve the purposes of this study is Range software developed by Heatley, Nation, and Coxhead (2002). The present corpus was run on the Range program to obtain frequency and dispersion data for all the words over four sub-corpora representing each of the university textbooks for English majors. This program breaks texts down by word frequencies into the 1st to 20th BNC and even to 25th BNC word families. The Range is installed with BASEWRD lists (each BASEWRD list contains 1,000-word families up to and including 20,000 or 25,000 BNC word family lists). These lists include proper nouns, marginal words, and word families.

The criteria used in the Range program to make word families are based on Bauer and Nation’s (1993) Level 6 word-building processes, which include inflections, the high-frequency, and regular, productive and transparent derivational affixes. These affixes can only be added to free forms. This means that the stem of each word must be able to stand as an independent word. For example, *patient* and *patience* cannot be members of the same word family since the suffix *-ence* is not added to a stem that can stand as an independent word. Researchers (Nation & Webb, 2011; Nation, 2016) suggest that word families should be regarded as an important counting unit regarding the learning load in receptive (that is for reading and listening) corpus-based studies. The rationale behind this counting unit is that the concept of a word family represents a group of words whose meanings can be inferred when the meaning of the base form in the group is known.
to the learner. For instance, the headword *appreciate* is grouped with its family members such as *appreciable, appreciably, appreciated, appreciates, appreciating, appreciation, unappreciated* and if learners know the meaning of the headword, they are likely to infer the meaning of its family members. Therefore, for this study, word family is used as a counting unit.

**Research Procedures**

As has already been mentioned above, the ultimate goal of this study is to produce the list – a list of the most essential words in the particular tertiary context of AUL. Regardless of the specific approach, vocabulary studies usually consider two “distributional properties” to identify the “important words” in a corpus: frequency and range (Miller & Biber, 2015, p. 43). Frequency is how many times readers encounter certain words, whereas range identifies words that are distributed widely across the corpus, in addition to occurring with high frequency. To explore the issues of frequency and range further in the corpus and to develop the list of the most important words, some selection principles are followed. First, the word families included have to be outside the most frequent 2,000-word families of English. Secondly, according to Miller and Biber (2015) “a word was considered “important” if it occurred with high frequency at least 20 times per million words” (p. 44). Taking this criterion for uniformity of frequency into account, then, in my 508,802-word corpus, the benchmark for uniformity of frequency for the word occurrences is a minimum of ten times in the entire corpus. Finally, members of a word family have to occur at least in three sub-corpora and at least twice in each sub-corpora.

**Results**

*RQ1 What percentage of the running words in the university textbook corpus is the UVL required to cover in order to help the students to attain 98% lexical coverage, including proper nouns?*

The university textbook corpus – the UTC – contains 508,802 running words (tokens), and 17,320-word types (individual words) and involves 6,734-word families listed in the BNC 20,000-word list. The first 2,000 word families account for 88.92% of the running words, and this shows the relatively accessible nature of undergraduate university textbooks for English majors compared to those of medical textbooks, for example, where the most frequent high-frequency word families account for 70.68% of the running words (see Hsu, 2013 for further discussion). Proper nouns give extra text coverage of 1.40%.

Taking into account the results of the previous studies (Hajiyeva, 2014; 2015b, c) which tested Azerbaijani university entrants’ receptive and productive vocabulary knowledge during a period of one year, it is evident that a vocabulary size of 2,000-word families plus the knowledge of proper nouns students can obtain a cumulative text coverage of 90.32% which is inadequate. Since the present study has chosen 98% text coverage as the primary aim for reading comprehension, then the goal for the establishment of the university vocabulary list for English majors is 7.68% lexical coverage based on what is left after the 2,000 BNC word families and proper nouns are counted (7.68% = 98% – 88.92% – 1.40%).

*RQ2 Which high-frequency words beyond the 2,000 most-frequent word families make up a UVL?*

Having identified the target cumulative text coverage of 7.68% (to help reach 98% text coverage), the next step involved selecting the most frequent word families according to the BNC 20,000-word families until the intended text coverage is attained. It is impossible to present all the data and Table two aims merely to give a sense of the results – just the most frequent words outside the BNC 2,000-word family lists.
After having sorted word frequencies in descending order and doing some mathematical calculations, 396-word families were ultimately chosen and formed the UVL for English majors (Appendix A, the list of words is given in an alphabetical order), whose cumulative coverage arrived at 6.96% and contributed to reaching 97.28% together with the rest of the word lists.

Discussion

As can be seen from the results, the cumulative text coverage of the UVL for English majors are slightly lower than the intended coverage of 7.68% to help to reach 98% text coverage. This is because word families that did not appear across more than half of the four sub-corpora (three or four out of four) were not included in the list, even though they had high frequency indicators. For example, words such as compliment (123), clench (45), assurance (45), posture (42), suspicion (32), disclosure (13), broadcast (13), gobbledygook (12), apprehensive (11): although they had a high frequency of occurrence, they were not included in the UVL since they are lexically idiosyncratic to a specific sub-corpus (CSI sub-corpus, for example). Therefore, all the word families that appeared in three sub-corpora out of four with a frequency of occurrence of at least ten times were included in the UVL (Appendix A displays a complete list of the high-frequency word families.)

Apart from the frequency of occurrence and range, Table two also displays the word levels along the BNC scale. An interesting observation should be emphasized at this point: some of the most frequent words included in the UVL for English majors were listed among the words that belong to the 1,000 or 2,000 BNC word family lists, although they have not been included in the BASEWRD 1 or BASEWRD 2 lists in the Range program. For example, achieve belongs to the BNC 1,000 word band, but it is not included in the BASEWRD 1 list. The same applies to absence, which belongs to the BNC 3,000-word level, but, surprisingly, is placed in the BASEWRD 2 list.

All in all, there were 40 such word families from the BNC 1,000 and 49-word families from the BNC 2,000-word families that were categorized by the Range program as the word families falling beyond the 2,000-word families and these word families were, therefore, included in the UVL vocabulary list. This outcome shows that there is inconsistency in the classification of the word families in terms of the BNC scale and Range BASEWRD lists. Researchers should take into account this inconsistency in future studies.
Referring back to the vocabulary list of word families listed in the UVL for English majors, which includes 396-word families dispersed among the BNC word lists, it covers 6.96%. It contributes to reaching 97.28% text coverage, including proper nouns. It should be noted that there were some words such as *to brainstorm*, *wiki*, and the *Internet* outside the BNC lists that occur frequently across the UTC texts included in the list. Analysis of the most frequent word families in the list showed some examples of technical vocabulary based on Nation’s (2011a) classification of words in terms of their technicality. For example, words such as *discourse*, *semantic*, *synonym*, *linguistic*, *pedagogy*, *coherence*, *subordinate*, and *syllabus* were identified as Category one technical vocabulary, according to Nation’s (2011a) classification.

It should also be noted that, compared to the results obtained in the previous study (Hajiyeva, 2015c), which observed the frequency distribution of academic words in the university textbook corpus and identified 127 most frequent academic word families, in this study, there was an overlap of 168- word families with Coxhead’s (2000) 570 AWL family members. Since in this study, the frequency of occurrence of the words to be included in the list was set at those occurring at least ten times compared to the previous study, which set this criterion at 16 times, this outcome seems rather logical. Decreasing the number of occurrences from 16 to 10 has resulted in the inclusion of an additional 41 academic word families from Coxhead’s AWL (2000). Ward (2009) states that it is the frequency criterion that determines the length of the resulting word list, but he also emphasizes that “there is no reliable way to determine exactly how long it should be” (p. 177). Schmitt (2010b) states that if a student could learn 50 words per week, then in 40 weeks of school, basic vocabulary (2,000 words) could be introduced. Bearing in mind that the target students in this study are “familiar” with most of the 2,000-word families (based on the vocabulary size test results) and they have 120 face-to-face hours of instruction at their disposal, then teaching and learning 396-word families is a relatively, modest target vocabulary for a semester’s work (25-26 word families per week). Of course, it should be noted that 396-word families include approximately 2,522 individual words and demand extra work in that sense.

Referring back to the points raised in the literature section of this paper in terms of the ‘internal representativeness’ of the corpus as suggested by Miller and Biber (2015, p.34), the Researcher decided to evaluate the reliability of the results as a prerequisite to considerations of their validity. The UTC already comprises four sub-corpora based on various course textbooks, although the topics, subject areas, and texts in the sub-corpora do not share common ground. The main objective of this study is to cater for the needs of both the students and the stakeholders and the underlying concern now is this: Does the word list that is developed allow teachers to achieve stable quantitative findings concerning the text coverage across the four sub-corpora?

The answer is straightforward. If the same quantitative findings across those four sub-corpora of CSI, CREW, SS and TEYL are obtained, then it can be stated that confident that the corpus and the list reliably represent the linguistic patterns of variation. This issue is further explored by running the *Range* program against the UVL for English majors with the four sub-corpora. The results show that the UVL gives a cumulative text coverage of

- 8.63% in the sub-corpus of Critical Reading and Effective Writing (CREW);
- 7.68% in the sub-corpus of Study Skills (SS);
- 6.89% in the sub-corpus of Teaching English to Young Learners (TEYL), and
- 4.65% in the sub-corpus of Communication and Social Interaction (CSI).

Given these findings, it can be stated that the UVL provides some reliable estimates of the number of different word families in this limited-size but a versatile corpus of university textbooks.
for English majors. In terms of the sub-corpus of CSI, though, the results show a deviation from the suggested pattern of 6.96%. The main reason for this pattern is that, as Miller and Biber (2015) state, some words might be very frequent in a corpus because they are extremely frequent in a single text or chapter, reflecting the specific topic and content of that text. These usually tend to be technical terms related to a particular topic and are not so essential or frequent in the corpus in general. Including these lexically idiosyncratic words in the list might make the list overwhelmingly long for students.

The UVL is intended for a specific context, though, where both secondary and tertiary education (early stages) have failed to bring students near to the kind of lexical threshold needed to succeed in an academic context (Hajiyeva, 2014; 2015 b, c). In their study, Miller and Biber (2015) could not “achieve a reliable word list” from their corpus of three million words derived from psychology textbooks within a restricted domain and they concluded that there is not a single, reliable list of the most important words in a discourse domain.

With these constraints in mind, it can be stated that, although the University Vocabulary List still leaves students some way short of the lexical knowledge required of them to read textbooks in the sub-corpus of *Communication and Social Interaction* with understanding, it is a modest, but reasonable lexical target for a semester’s work.

**Conclusion**

This paper aimed to develop a University Vocabulary List for students to meet their academic needs. In this paper, observations of the vocabulary content in a series of second-year university textbooks for English majors at the Faculty of Education at AUL are presented. A University Vocabulary List of the most-frequent 396-word families, which gives an additional 6.96% coverage of a variety of English-medium university textbooks beyond the BNC most frequent 2,000-word families was developed. The UVL represents a relatively frequent and relevant vocabulary load for English majors whose secondary educational background and first-year general English textbooks could not equip them for the linguistic challenges they face at the initial stages of their tertiary education. It is, therefore, concluded that, should they master the most frequent 2,000 high-frequency word families and also the UVL with a text coverage of 6.96%, students can attain a text coverage of 97.28%. As described in the literature review part of the study, 98% coverage provides adequate comprehension of a text, with coverage being dependent on the level of difficulty of a text. In other words, 98% coverage is sufficient for unassisted comprehension of a relatively easy text. The cumulative text coverage of 97.28% obtained in this study falls only marginally short of the suggested coverage of 98%. The results show that there are some words in the corpus which occur very frequently; however, these words were not included in the final list since they are extremely frequent in a single text or chapter and merely reflect the specific topic and content of that text or chapter. Miller and Biber (2015) argue that these words usually tend to be technical terms related to a specific topic and not as important or frequent in the corpus as a whole. The inclusion of these lexically eccentric words in the list might therefore make the list overwhelmingly long for these particular students. However, it is believed that this list will fulfil Azerbaijani students’ need to have a command of the vocabulary required of undergraduate students for whom English is a foreign language to be able to read for and internalize their subject-specific courses.
About the Author:

Konul Hajiyeva is a post-doc researcher in Applied Linguistics and senior lecturer at Azerbaijan University of Languages. Her research interests include hybrid teaching, vocabulary development, and continuing professional development. She is also involved as a head of the unit State language and foreign languages at State Examination Center of the Azerbaijan Republic.

https://orcid.org/0000-0001-5748-0189

References


Hajiyeva, K. (2015b). Exploring the relationship between receptive and productive vocabulary sizes and their increased use by Azerbaijani English majors. English Language Teaching, 8(8), 31-45. doi:10.5539/elt.v8n8p31


