

Eihab Abu-Rabiah

Kaye Academic College of Education,
Ben-Gurion University of the Negev, Israel

 <https://orcid.org/0000-0002-8837-1089>

The Productive Vocabulary Size of Second Language Learners upon Entry into Higher Education

Abstract

Extensive vocabulary acquisition is a cornerstone of second language (L2) proficiency, directly influencing both receptive and productive language skills. However, research on the productive vocabulary size of L2 learners transitioning to higher education, particularly their mastery of high-frequency words, remains limited. This study investigated the productive Hebrew vocabulary size and frequency distribution of Arabic-speaking learners entering higher education programs where Hebrew is the primary language of instruction.

The research employed a corpus-driven approach, analyzing 156 Hebrew-language argumentative essays (18,054 orthographic words) written by native Arabic-speaking students during a college entrance examination. Automated tools were used to add contextual vocalization and disambiguate homographs, followed by manual annotation mapping each word to its corresponding lemma. The identified lemmas were then compared to established written and spoken Hebrew frequency lists. This process aimed to chart the vocabulary profile of the research population.

The study determined that learners had a productive vocabulary size of approximately 1,000 lemmas, despite completing over 1,000 hours of formal L2 instruction. A comparison with established written and spoken Hebrew frequency lists indicated that 50% of the identified lemmas fell within the 1,000 most frequent Hebrew lemmas. Additionally, the learners exhibited a typical vocabulary profile, employing more lemmas from the 1k frequency band (the 1,000 most frequent words) than from the 2k frequency band (words ranked 1,001–2,000). Similarly, their use of lemmas from the 2k band exceeded that of the 3k band (words ranked 2,001–3,000), which in turn surpassed their use of lemmas from the 4k band (words ranked 3,001–4,000). These findings highlight the learners' significant reliance on high-frequency vocabulary in L2 writing, emphasizing the need for targeted academic vocabulary instruction as they transition to higher education.

Keywords: second language acquisition, writing assessment, vocabulary assessment, frequency distribution, Zipf's law, Arabic-speaking learners, Hebrew frequency lists

Literature Review

Vocabulary Knowledge

Vocabulary knowledge serves as the fundamental underpinning of language, and, without it, both language comprehension and production become unattainable (Milton & Treffers-Daller, 2013). As aptly noted by Wilkins (Wilkins, 1972, p. 111), “without grammar, very little can be conveyed, [but] without vocabulary, nothing can be conveyed.” Vocabulary knowledge, or word knowledge, can be categorized into two primary types: receptive (or passive) knowledge and productive (or active) knowledge. Receptive knowledge pertains to words recognized when heard or read, surpassing the less extensive productive knowledge, which involves words recalled for use in speech or writing (Milton, 2009).

Words exhibit variations in multiple facets, including their sounds, letters, meanings, length, inflections, derivations, and more. However, the most pivotal aspect of these variances is the frequency with which words occur. Frequency plays a pivotal role since it determines which words learners are likely to encounter and how frequently they encounter them (Milton, 2009). Furthermore, words that are more frequent are typically acquired earlier (Ambridge et al., 2015). It is evident that certain words occur more frequently than others, requiring little knowledge about language (Schmitt & Schmitt, 2020).

Vocabulary Size

To gauge vocabulary knowledge and explore differences in word frequencies, one common approach is measuring vocabulary size or breadth. Vocabulary size signifies the number of words an individual is familiar with (Schmitt, 2014), making it a fundamental aspect of vocabulary research. Vocabulary size can be evaluated receptively and productively. Receptive vocabulary size is often evaluated through tests that assess knowledge of a sample of words (Nation & Coxhead, 2021). An example of such a tool is the Vocabulary Knowledge Scale, which uses a 5-point self-report system. The VKS allows learners to indicate their level of familiarity with specific vocabulary items, ranging from not recognizing the word at all to being able to use it confidently in a sentence (Wesche & Paribakht, 1996). Productive vocabulary size is assessable through written compositions or an individual's language output (Nation & Coxhead, 2021). It's essential to recognize that not all known words may be utilized, and longer language productions tend to incorporate a greater variety of words (Richards & Malvern, 1997; Nation & Coxhead, 2021). Studies, such as those by Pan et al. (2005) and Rowe et al. (2012), focused on written language, counting

distinct words. For example, Shakespeare's vocabulary exceeded 20,000 words, determined by scrutinizing word forms in his works (Craig, 2011). In spoken language, numerous studies transcribed and tallied words from recorded language speaker output (Nation & Coxhead, 2021).

Assessing students' receptive vocabulary size offers teachers insights into their text and listening comprehension capabilities. Conversely, understanding their productive vocabulary size provides an indication of their speaking and writing proficiency (Webb, 2008). Research consistently shows that learners' receptive vocabulary size tends to be larger than their productive vocabulary size (Fan, 2000; Webb, 2008). Studies report this ratio can vary between 50% and 80% (Milton, 2009; Nizonkiza, 2016). In general, receptive vocabulary size can be a good indicator of productive vocabulary size (Webb, 2008). It is important to note, however, that this receptive-productive relationship may be moderated by the specific language learning environment (Webb, 2008). In foreign language contexts with limited opportunities for naturalistic L2 use outside the classroom, the receptive-productive ratio could differ from patterns observed in second language immersion settings where learners have greater access to the L2 in daily life.

Examining the vocabulary size of native English speakers reveals an average vocabulary size of 10,000 to 11,000 word families among educated adult L1 English speakers (Milton & Treffers-Daller, 2013; Brysbaert et al., 2016). Generally, it is agreed that a substantial vocabulary is requisite for English language functionality, with estimates spanning from 2,000 to 10,000 word families (Milton & Treffers-Daller, 2013; Van Zeeland & Schmitt, 2013). For everyday spoken communication in English, 2,000 to 3,000 word families suffice (Van Zeeland & Schmitt, 2013), while 6,000 to 7,000 word families are needed for watching movies and TV (Webb & Rodgers, 2009a, 2009b), and 8,000 to 9,000 word families for extensive reading (Nation, 2006).

In contrast, second language learners typically possess more limited vocabularies (Laufer & Yano, 2001). Receptively, these learners typically grasp 2,000 to 4,000 word families, despite over 1,000 hours of instruction (Laufer, 2017). However, research also suggests that L2 learners at a university level can achieve larger vocabularies. For instance, Lenko-Szymanska (2002) found that Polish learners of English possessed a passive vocabulary knowledge of 8,236 to 9,218 words (first-year vs. fourth-year). Overall, vocabulary size measurements are employed as general indicators of language proficiency and are strongly correlated with proficiency in reading, writing, listening, and speaking skills (Llach & Gallego, 2009; Milton et al., 2010; Mairano & Santiago, 2020).

Importance of the Current Study

Vocabulary size is a cornerstone of fluency in any language. However, a critical gap exists in SLA research: we lack robust data on the productive vocabulary size of learners entering higher education, where the target language (L2) becomes the primary mode of instruction, particularly concerning high-frequency words. This study aims to bridge this gap by investigating the productive Hebrew vocabulary size of Arabic-speaking learners about to enter higher education. By analyzing written essays through a corpus-based approach, the research will explore vocabulary distribution and identify frequently used words. These findings, compared to existing Hebrew frequency lists, will provide a comprehensive picture of the learners' productive vocabulary landscape. By shedding light on the vocabulary size and characteristics of this specific population, this research holds significant value for educators and researchers in SLA. It can inform curriculum development, instructional practices, and ultimately bridge the gap in understanding L2 learners' preparedness for higher education programs.

Methodology

Research Questions

The present study is guided by the following main research questions:

1. Does the rank-frequency distribution of words in a corpus of L2 Hebrew writing produced by Arabic-speaking learners conform to Zipf's law, exhibiting a predominance of highly frequent function words and a greater proportion of lower frequency content words?
2. What is the productive Hebrew vocabulary size of Arabic speakers entering higher education programs, and how does it compare to established written and spoken Hebrew frequency lists?

Corpus and Population Description

For data collection, the study utilizes a corpus consisting of 156 argumentative essays, written in Hebrew by native Arabic-speaking students. These essays were obtained from an entrance examination for an Israeli academic college, designed to assess the Hebrew language proficiency of candidates. The essays revolved around the argumentative topic of whether parents should

compensate children for assisting with household chores, including cleaning and caring for younger siblings. The choice of utilizing written essays as opposed to recall tests for assessing learners' productive vocabulary knowledge was made because, in the context of essay writing, learners have a more extensive window of time to retrieve words from memory. Furthermore, they can review and make corrections or alterations to their word choices if necessary (Milton, 2009; Leijten et al., 2010).

The participants in this study attended Israeli schools where Arabic is the primary language of instruction. Hebrew is formally introduced in the second grade (ages 8–9) with 2–5 hours of weekly instruction throughout their schooling (Manor & Watad, 2024). This culminates in a final exam for Arabic-speaking students (L2) in Hebrew. Although Hebrew is officially the language used in high school Hebrew classes, teachers (often native Arabic speakers) often switch to Arabic during lessons (Abu-Rabiah et al., 2023). Their expected level of proficiency in Hebrew is B1–B2 on the CEFR (Common European Framework of Reference for Languages) scale.

Analysis Approach

In the analysis of written vocabulary, selecting the appropriate counting unit is pivotal (Nation & Coxhead, 2021; Brown et al., 2022; Maw et al., 2022). Various definitions, such as tokens, types, lemmas, flemmas, and word families, rely on how learners perceive relationships between word forms (Milton, 2009; Schmitt & Schmitt, 2020; Nation, 2021). Each word in a sample is a token, regardless of repetition, making the sample's length the count of its tokens. Types denote different (or unique) words in a sample (Jarvis & Hashimoto, 2021). Repetitions of the same tokens are counted as a single type. Tokens were not used due to susceptibility to text length and time constraints on learners, and types were avoided to prevent inflating vocabulary size counts, as noted by some researchers (Nation & Coxhead, 2021).

In this research the lemma was chosen as the counting unit. Lemma comprises diverse inflections of the same root within the same part of speech. For instance, 'מְדַרֵּךְ' (a guide.masculine) and 'מְדַרִּיכָה' (a guide.feminine) are regarded as a single lemma. This choice is based on the assumption that learners are more likely to know and use the inflections within a lemma than the derivations required to form a word family (Schmitt & Schmitt, 2020). Using lemmas as a counting unit is common in studies involving second language learners (Kremmel, 2016; Treffers-Daller et al., 2018; Jarvis & Hashimoto, 2021). Compared to larger units like flemma or word family, lemma provides greater precision and involves fewer assumptions about a learner's vocabulary knowledge (Gablasova & Brezina, 2021). Word families, which encompass both

inflections and derivations of a root, may overestimate learners' knowledge. Therefore, the lemma was selected to avoid such overestimation and ensure more accurate measurement.

Unlike English, Hebrew allows multiple word forms (namely, a dictionary entry) to exist within a single written word (orthographic word) (Abu-Rabiah, 2020, 2022, 2023). For example, the word “בעתיד” ‘ba'atid’ includes the following word forms: ‘be’, [‘ha’], and ‘atid’ (in [the] future). To ensure accurate analysis and comparison with other languages, we used Dicta’s (Analytical tools for Hebrew texts) morphology tool to add vocalization (nikud) to each essay’s words based on context. This allowed us to distinguish homographs (words with the same spelling but different meanings) and count all individual word forms within each orthographic word. This is especially relevant in Hebrew due to its consonantal writing system (Cook, 2016), where around 23% of words are homographic when considered in isolation from their surrounding context (Shimron & Sivan, 1994). Following this automated process, the corpus was manually annotated to accurately identify and categorize each word by its corresponding lemma.

The essays in the corpus display notable length variation, ranging from 34 to 267 orthographic words, with an average of 116. The total orthographic words in the corpus amount to 18,054, representing 27,407 word forms. Length control is unnecessary in this study, as the focus is not on comparing vocabulary size among individuals; instead, all essays are treated as one population. Opting for lemmas as the counting unit.

To delineate the productive vocabulary characteristics of written Hebrew among Arabic-speaking individuals, this study employed a comparative analysis with established Hebrew frequency lists. The primary reference was the he-TenTen21 corpus (Hebrew Web Corpus 2021), which offers a frequency lemma list for written Hebrew. However, the open-access version of this list is limited to the top 1,000 most frequent lemmas. While the primary focus of the list is on words, it is worth noting that punctuation marks are also included. To assess the vocabulary size of the research population, it was necessary to supplement this with a second resource containing a broader range of lemmas.

As such, the Frequency Dictionary of Spoken Hebrew (FDOSH) (Pinto, 2018) was utilized for additional comparison. FDOSH, also based on lemmas, provides a compilation of the most frequently occurring items in conversational Modern Hebrew, derived from a corpus of film subtitles. Although FDOSH focuses on spoken rather than written Hebrew, it serves as a foundational resource for learners aiming to achieve conversational proficiency across diverse contexts. Its subtitle-based approach is well-justified, as prior studies have demonstrated that subtitle corpora closely mirror conventional spoken language (Pinto, 2018). Together, these resources enabled a more comprehensive examination of the learners’ productive vocabulary.

As an initial exploration of vocabulary size in Hebrew across both L1 and L2 contexts, this study analyzes the entire corpus of 156 learner essays as a unified dataset, rather than examining individual learners. This corpus-driven approach, aligned with prior research (e.g., Milton, 2006), aims to characterize the overall productive vocabulary knowledge of the learner sample. However, future studies could build on these findings by incorporating individual-level analyses.

Results and Discussion

1. Does the rank-frequency distribution of words in a corpus of L2 Hebrew writing produced by Arabic-speaking learners conform to Zipf's law, exhibiting a predominance of highly frequent function words and a greater proportion of lower frequency content words?

The frequency of word usage can provide valuable insights into the characteristics of a language. Table 1 displays the frequency and percentage of occurrence of lemmas in the corpus, sorted from most frequent to least frequent. Notably, the 20 most frequent lemmas on the left side of the table collectively account for nearly 59% of all word form occurrences in the corpus, while the 20 least frequent lemmas on the right side constitute less than one percent. This discrepancy signifies that the 20 most frequent lemmas are approximately 841 times more common than the 20 least frequent lemmas. Such a distribution pattern is consistent with findings in the English language, where high-frequency lemmas exhibit substantial coverage gains, while lower-frequency words exhibit diminishing gains (Schmitt & Schmitt, 2020).

For instance, the definite article 'ה' ("ha") occurs 2,977 times, covering nearly 11% of all word forms in the corpus, whereas each of the 20 least frequent lemmas appears only once, collectively covering less than half a percent of all word forms. Interestingly, there are 421 lemmas that occur only once in the corpus, and together, they account for no more than 1.54% of the total word forms. This underscores the notion that frequency serves as a reliable indicator of vocabulary utility at higher-frequency levels, but its predictiveness diminishes as one ventures into lower-frequency vocabulary in any language (Schmitt & Schmitt, 2020).

Table 1
Frequency and Percentage of the Top 20 Most and Least Frequent Lemmas in the Corpus

Order	Lemma	Translation	Occurrences	Percentage of the Full Corpus	Order	Lemma	Translation	Occurrences	Percentage of the Full Corpus
1	ה	the	2977	10.86%	1004	אסף	gathered	1	0.0007%
2	הם	they.plural.masculine	1408	5.14%	1005	בזבז	squandered	1	0.0007%
3	ו	and	1353	4.94%	1006	אתן	you.plural.feminine	1	0.0007%
4	של	of	1274	4.65%	1007	גורם	cause	1	0.0007%
5	ל	to	1255	4.58%	1008	הגן	defended	1	0.0007%
6	ב	in/at	981	3.58%	1009	הסתדר	arranged oneself	1	0.0007%
7	הורה	parent	800	2.92%	1010	הגשים	fulfilled	1	0.0007%
8	ילד	boy	779	2.84%	1011	הסתכלות	observing	1	0.0007%
9	בית	home	633	2.31%	1012	הדגמה	demonstration	1	0.0007%
10	ש	that	548	2.00%	1013	העיק	threw	1	0.0007%
11	על	on	540	1.97%	1014	אור	light	1	0.0007%
12	אני	I	515	1.88%	1015	הפחית	subtracted	1	0.0007%
13	עזר	helped	502	1.83%	1016	הובן	understood	1	0.0007%
14	הוא	he.singular.masculine	403	1.47%	1017	אקראי	random	1	0.0007%
15	את	you.singular.feminine	399	1.46%	1018	הוגדר	defined	1	0.0007%
16	זה	this.feminine	387	1.41%	1019	בטוח	safe	1	0.0007%
17	לא	no	378	1.38%	1020	בסיסי	basic	1	0.0007%
18	אנחנו	we	334	1.22%	1021	אבד	lost	1	0.0007%
19	עבודה	work	332	1.21%	1022	בעקבות	as a result of	1	0.0007%
20	כל	all	331	1.21%	1023	בנאדם	human being	1	0.0007%
Grand Total	20		16129	58.85%	Grand Total	20		20	0.015%

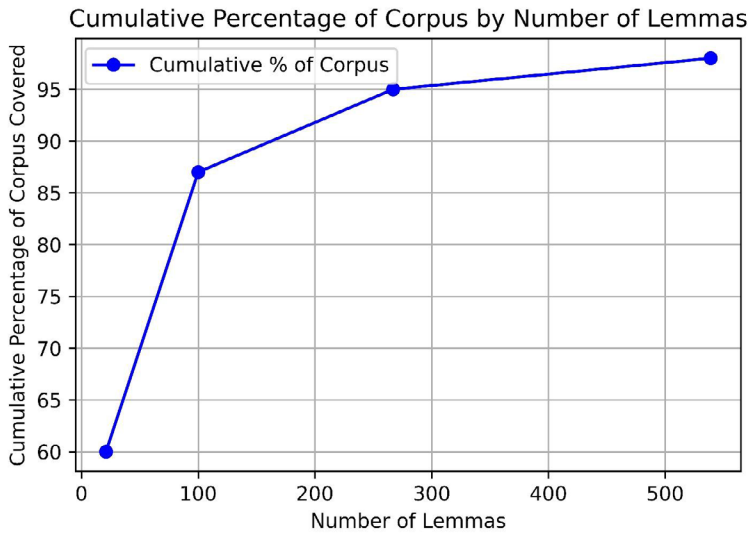
The distinction between the most frequent words and less frequent words plays a pivotal role in understanding the dynamics of vocabulary and its role in language. Vocabulary serves as the link between the tangible world and linguistic expression, connecting meanings with linguistic forms (Schmitt & Schmitt, 2020). It is primarily composed of content or lexical words, which encompass nouns, verbs, adjectives, and adverbs (e.g., ‘car’ and ‘fresh’) and are easily expandable (Cook, 2016). In contrast, function or structure words, including articles, prepositions, pronouns, and conjunctions (e.g., ‘in’ and ‘but’), appear to carry minimal inherent meaning but are indispensable for constructing grammatically and semantically coherent language (Schmitt & Schmitt, 2020). Function words are limited in number and rarely expand (Milton, 2009; Cook, 2016). Content words provide the primary semantic content of sentences and are known as lexical words or open-class words, while function words serve as connectors between content words, labeled as grammatical words or closed class words (Abdalla et al., 2010; Vejdemo & Hörber, 2016).

In actual language usage, function words tend to be the most frequent words across various corpora (Milton, 2009; Schmitt & Schmitt, 2020). In the current corpus, approximately 75% of the 20 most frequent lemmas are function words, while only 10% of the 20 least frequent lemmas belong to this category. The ubiquity of function words is observed consistently in corpora, irrespective of the general or specific nature of the language in the corpus and whether it is in written or spoken form. Content words, on the other hand, are context-specific and influenced by the thematic focus of the corpus (Schmitt & Schmitt, 2020). In the present corpus, which revolves around the topic of whether parents should provide compensation to children for assisting with household chores such as cleaning and caring for younger siblings, the frequent content words are inherently linked to this theme. Examples of these content words include ‘הורה’ (“hore”) (a parent), ‘ילד’ (“yeled”) (a son), ‘בית’ (“bayit”) (a home), and ‘עבודה’ (“avoda”) (a work).

The examination of word frequency distribution, and its adherence to established linguistic patterns such as Zipf’s law, is a crucial part of our analysis. A comprehensive survey of the entire corpus reveals that the distribution of lemma frequencies follows a pattern reminiscent of standard language distributions, wherein certain words are employed significantly more frequently than others (see Figure 1). This observed frequency distribution adheres to Zipf’s law, a well-documented linguistic phenomenon. Zipf’s law in corpus linguistics states that word frequency is inversely proportional to its rank in large language corpora (Ha et al., 2002). Under Zipf’s law, a limited number of high-frequency words are exceedingly common, with their frequency declining exponentially as one progresses through the vocabulary, resulting in the majority of words being relatively rare (Schmitt & Schmitt, 2020). While initially observed for single words, research shows it applies more accurately when combining words

and n-gram phrases (Ha et al., 2002). In addition, the law extends beyond frequency, with Zipf’s “law-of-meaning” suggesting that more frequent words tend to have more meanings (Ilgen & Karaoglan, 2007). In the current corpus, the top 21 most frequent lemmas collectively constitute 60% of the corpus, the top 100 encompass 87%, the top 267 cover 95%, and the top 539 account for 98% of the corpus.

Figure 1
Word Frequency Distribution



Zipf’s law postulates that in a natural language corpus, the frequency of the first most common word is approximately twice that of the second most common word, which, in turn, is approximately twice that of the fourth most common word (Milton, 2009). In our current corpus, this relationship holds true, where the most frequent word, the definite article ‘ה,’ (“ha”) occurs approximately twice as often as the second word ‘הם’ (“hem”) (they). It is important to note that while Zipf’s law is not without its imperfections (Milton, 2009), it underscores the idea that the most vital and practical vocabulary is concentrated among the most frequent words (Schmitt & Schmitt, 2020). This distribution pattern, encapsulated by Zipf’s law, is not exclusive to English but holds true in numerous other languages, including Greek (Hatzigeorgiu et al., 2001), French (Cobb & Horst, 2004), and many more (Jayaram & Vidya, 2008).

2. What is the productive Hebrew vocabulary size of Arabic speakers entering higher education programs, and how does it compare to established written and spoken Hebrew frequency lists?

To quantify the productive Hebrew vocabulary size of the 156 Arabic-speaking learners in the study corpus, a process of lemmatization and deduplication was implemented. Specifically, all inflected word form tokens were reduced to their corresponding lemmas. Subsequently, all instances of the same lemma type were collapsed, such that each distinct lemma was represented only once in the final vocabulary list, regardless of its frequency of occurrence in the corpus. This procedure yielded a total of 1,023 unique lemma types, constituting the productive vocabulary size estimate for this population of learners on the cusp of entering Hebrew-medium higher education programs.

The rationale for adopting this method of lemma-based vocabulary counting was to provide a conservative yet psycholinguistically motivated estimate of the learners' vocabulary knowledge. By focusing on distinct lexeme types rather than tokens, the analysis circumvented inflectional variation and generated a vocabulary size metric that more closely approximates the number of lexical stems productively commanded by these learners. This lemmatized vocabulary list therefore represents the core productive lexicon deployed by Arabic-speaking students as they prepare to engage with Hebrew as the language of academic instruction and scholarly content.

The average learner in our study had approximately 11 years of Hebrew instruction (1,188–1,980 hours). This translates to acquiring less than one productive lemma per lesson. This falls short of receptive vocabulary acquisition estimates (around 6.5–8 lemmas per lesson) reported by Vassiliu (2001). Interestingly, this size is comparable to the receptive vocabulary size reported in other studies: 1,654 word families for English majors in China (Miao, 2019) and 948–1,186 words for Spanish EFL learners (Montero-SaizAja, 2022). This suggests that our Arabic learners' productive vocabulary may be larger than the receptive vocabulary of learners in a foreign language context, potentially due to exposure to Hebrew outside classrooms. The 1,023 lemma size is also similar to the expected vocabulary size for English learners of French in the UK, where core vocabulary lists for CEFR B1 exams range from 1,000–1,500 items (Milton & Meara, 1998). However, it falls short of contexts like Hungary, where learners are expected to have 400 active lemmas by grade 8 (Krizsan, 2003, reported in Milton, 2009). In line with Laufer's (2017) suggestion that L2 learners typically grasp 2,000–4,000 word families despite extensive instruction (over 1,000 hours), a productive size of around 1,000 lemmas is not unexpected.

Research suggests receptive vocabulary size can be a good predictor of productive vocabulary, with a ratio ranging from 50% to 80% (Milton, 2009; Nizonkiza, 2016). Applying this range to our findings, the estimated receptive vocabulary size for our learners could be between 1,535 and 1,841 lemmas. However, this requires further investigation.

This study seeks to ascertain the extent of word overlap between the lemmas in the current L2 learner corpus and established written frequency lists

in both spoken and written Hebrew. It is well-acknowledged that word frequencies in oral and written language diverge (Lau et al., 2019), with conversational speech characterized by reduced formality and a less academic tone in contrast to written language (Milton, 2009).

Table 2

Comparison of the 50 Most Frequent Lemmas in the Present Corpus with FDOSH Lemmas, Indicating Overlap

Order	Corpus Lemma	Translation	FDOSH Lemma	Overlap	Order	Corpus Lemma	FDOSH Lemma	Overlap
1	ה	the	הוא	Yes	26	צריך	אז	Yes
2	הם	they.plural. masculine	ה	No	27	יש	סדר	Yes
3	ו	and	את	Yes	28	היא	צריך	No
4	של	of	ל	Yes	29	קטן	רק	No
5	ל	to	לא	Yes	30	מ	חשב	Yes
6	ב	in/at	זה	Yes	31	טוב	כאן	Yes
7	הורה	parent	ב	No	32	אח	הלך	No
8	ילד	boy	של	No	33	היה	דבר	Yes
9	בית	home	ש	No	34	או	איש	Yes
10	ש	that	היה	Yes	35	שמר	אל	No
11	על	on	מה	Yes	36	שילם	כך	No
12	אני	I	ו	No	37	אמא	יותר	No
13	עזר	helped	על	No	38	רצה	שם	Yes
14	הוא	he.singular. masculine	כול	Yes	39	הרבה	יכול	No
15	את	you.singular. feminine	ידע	Yes	40	אם	ראה	Yes
16	זה	this.feminine	כן	Yes	41	זמן	עכשיו	Yes
17	לא	no	מ	Yes	42	משפחה	אחד	No
18	אנחנו	we	יש	No	43	כסף	משהו	No
19	עבודה	work	עשה	No	44	יכול	למה	Yes
20	כל	all	אבל	No	45	ניקה	בא	No
21	עזרה	help	טוב	No	46	נתן	זאת	No
22	כי	because	רצה	No	47	דעת	או	No

Table 2 continued

Order	Corpus Lemma	Translation	FDOSH Lemma	Overlap	Order	Corpus Lemma	FDOSH Lemma	Overlap
23	דבר	thing	אם	Yes	48	בן	זמן	No
24	עשה	did	עם	Yes	49	מה	נכון	Yes
25	גם	and	אמר	No	50	כמו	כמו	Yes
Overlap								50%

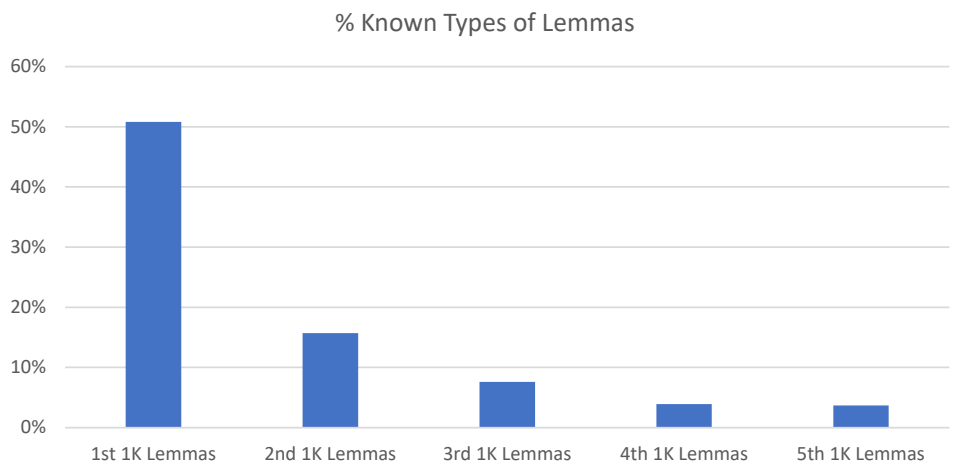
When comparing the 50 most frequent lemmas in the present corpus with the 50 most frequent lemmas in the FDOSH (Frequency Dictionary of Spoken Hebrew), 50 percent overlap is observed (see Table 2). This implies that half of the 50 most frequent lemmas in the written Hebrew utilized by Arab learners coincide with the 50 most frequent lemmas in spoken Hebrew. A similar pattern of overlap emerges when considering all 1,023 lemmas in the current corpus against the 5,000 lemmas in FDOSH (see Table 3). Specifically, half of the lemmas in the corpus are among the top 1,000 most frequent lemmas in spoken Hebrew, while 30 percent of the lemmas fall within the remaining 4,000 lemmas, distributed as 15 percent in the 2,000 lemmas, 7 percent in the 3,000 lemmas, and 4 percent in both the 4,000 and 5,000 lemmas. In essence, 80% of the 1,023 lemmas in our current corpus are included in the 5,000 lemmas of FDOSH. The additional 20% of the lemmas do not appear in the FDOSH’s 5,000 most frequent lemmas. This set of words encompasses not only content words like ‘ניקיון’ (“nikayon”) (cleaning), ‘הרהר’ (“hirher”) (ponders), and ‘חירות’ (“herut”) (freedom) but also function words like ‘אנחנו’ (“anahnu”) (we).

Table 3
Comparison of Lemmas in the Current Corpus and FDOSH

Total Lemmas in the Current Corpus	FDOSH Lemma Frequency	Overlapped Lemmas	Percentage of Overlap: (Overlapped Lemmas \ Total Lemmas in the Current Corpus)	Additional Overlap Percentage from Previous Frequency Group
1023	1000	508	50%	–
	2000	665	65%	15%
	3000	741	72%	7%
	4000	780	76%	4%
	5000	817	80%	4%
	Out of the list	206	–	–
	Total	1023	80%	–

When we segment the FDOSH dataset into five distinct frequency bands, each consisting of 1,000 (1k) lemmas, and examine the occurrence of the 1,023 lemmas from the current corpus within these bands, a discernible pattern emerges. This pattern aligns with the well-documented concept of a typical learner’s vocabulary knowledge, in which individuals tend to possess a more extensive command of high-frequency words compared to their familiarity with lower-frequency ones (Meara, 1992; Kweon & Kim, 2008 but cf. Crossley et al., 2019) and less sophisticated words (Abu-Rabiah, 2024). This distribution of lemmas within the frequency bands exhibits a characteristic profile, as depicted in the following graphical representation (see Figure 2). Notably, the learner’s knowledge is notably robust in the higher-frequency columns on the left side, while it diminishes as one progresses towards the less frequent columns on the right side. This distinctive downward slope in vocabulary knowledge, from left to right, underscores the common observation in language acquisition research (Milton, 2009). This phenomenon highlights that learners tend to excel in acquiring and employing words that are encountered more frequently in spoken and written language, whereas their proficiency gradually declines as they confront less common and less frequently encountered lemmas.

Figure 2
Distribution of Lemmas in the Current Corpus across 5 Frequency Bands, Each Consisting of 1000 (1k) Lemmas, According to FDOSH



A comparison between the 1,023 lemmas in the current corpus and a shorter list of the 1,000 most frequent lemmas in written Hebrew, as documented in the heTenTen21 corpus, has yielded significant findings (see Table 4). The analysis identified 511 lemmas common to both datasets, representing a 50% overlap.

Table 4

Comparison of Lemmas in the Current Corpus and heTenTen21, Indicating Overlap and Additional Overlap by Frequency Band

Total Lemmas in the Current Corpus	heTenTen21 Lemma Frequency	Overlapped Lemmas	Percentage of Overlap (Overlapped Lemmas / Total Lemmas in the Current Corpus)	Percentage of Additional Overlap from the Previous Frequency Band
1023	50	45	4%	–
	100	93	9%	5%
	500	343	34%	25%
	1000	511	50%	16%
	Out of the list	31	–	–
	Total	1023	50%	–

A more in-depth examination focusing on the 50 most frequent words indicated that 45 lemmas were common to both datasets, constituting a notable 90% concordance. This level of agreement surpasses the level observed when contrasting these lemmas with the FDOSH dataset, which exhibited a 50% overlap. Similarly, the scrutiny of the 100 most frequent words unveiled 93 shared lemmas, signifying a significant 93% concurrence, compared to the 89% alignment with the FDOSH dataset.

When juxtaposing the initial 1,000 most frequent lemmas of heTenTen21 with the FDOSH dataset, a 58% overlap is observed. Furthermore, when evaluating the 1,000 most frequent lemmas of heTenTen21 against the 5,000 most frequent lemmas of FDOSH, a significant 91% level of overlap is recorded. Importantly, it is worth noting that this degree of overlap may potentially be even higher in practice, given the possible imperfections in the cleanliness of the heTenTen21 corpus, as previously acknowledged.

Collectively, these findings emphasize the significant overlap among the 1,000 most common lemmas in both written and spoken Hebrew, extending to the written Hebrew of Arabic speakers. It is evident that a considerable portion of the written vocabulary utilized by Arabic-speaking learners of Hebrew relies on these high-frequency words. However, their command of less common vocabulary appears to be more restricted. Half of their productive vocabulary is derived from the 1,000 most frequent written and spoken words in Hebrew. The majority of these words do not include academic vocabulary, suggesting that they may not be adequately prepared for higher education. This limitation could impede their ability to articulate nuanced concepts and navigate the intricacies of academic writing.

Conclusions

This study investigated the vocabulary size and characteristics of L2 learners entering higher education, with a specific focus on Arabic speakers entering Hebrew-medium programs. By analyzing 156 argumentative essays written in Hebrew, the research revealed a productive vocabulary size of approximately 1,023 lemmas, with a distribution following Zipf's law. Notably, despite exceeding 1,000 hours of instruction, learners in contexts where the target language is used outside classrooms may still acquire a limited productive vocabulary (around 1,000 lemmas).

The analysis also revealed a significant overlap between the learners' vocabulary and established Hebrew frequency lists, such as heTenTen21 and FDOSH, which encompass both written and spoken language. In addition, the learners exhibited a typical vocabulary profile, demonstrating greater use of lemmas from the 1k frequency band compared to the 2k band, with usage progressively declining in subsequent frequency bands. The findings further indicated that learners demonstrated a strong command of high-frequency words, which were crucial for everyday communication in Hebrew. However, their mastery of less frequent vocabulary, essential for expressing nuanced ideas in academic contexts, was notably limited.

Based on these findings, it is recommended to deliberately teach academic vocabulary at the beginning of the academic journey to support learners in producing discipline-specific language and engaging effectively with academic content. Additionally, frequency lists can be utilized to assess and guide vocabulary instruction, helping to set realistic expectations for language learners. Future research should explore the gap between L2 learners' receptive and productive vocabulary sizes, particularly as they transition to higher education. Furthermore, examining how productive vocabulary size varies across different proficiency levels may yield valuable insights into the developmental trajectory of L2 lexical knowledge.

References

- Abdalla, F., Robb, M. P., & Al-Shatti, T. (2010). Stuttering and lexical category in adult Arabic speakers. *Clinical Linguistics & Phonetics*, 24(1), 70–81.
- Abu-Rabiah, E. (2020). Lexical measures for testing progress in Hebrew as Arab students' L2. *Journal of Language and Linguistic Studies*, 16(3), 1096–1114. <https://doi.org/10.17263/jlls.803551>

- Abu-Rabiah, E. (2023). Evaluating L2 vocabulary development features using lexical density and lexical diversity measures. *LLT Journal: A Journal on Language and Language Teaching*, 26(1), 168–182. <https://doi.org/10.24071/llt.v26i1.5841>
- Abu-Rabiah, E. (2024). Exploring lexical sophistication in second language: An analysis of vocabulary using a word-rating method. *Theory and Practice of Second Language Acquisition*, 10(2), 1–24. <https://doi.org/10.31261/TAPSLA.15623>
- Abu-Rabiah, E., Gafter, R., & Henkin, R. (2023). Where syntactic interference persists: The case of Hebrew written by native Arabic speakers. *L1-Educational Studies in Language and Literature*, 23, 1–19. <https://doi.org/10.21248/llesll.2023.23.1.399>
- Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2), 239–273.
- Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, 43(3), 596–602.
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7, 1116, 1–11.
- Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In T. Cobb & M. Horst (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). John Benjamins B.V.
- Cook, V. (2016). *Second language learning and language teaching*. Routledge.
- Craig, H. (2011). Shakespeare's vocabulary: Myth and reality. *Shakespeare Quarterly*, 62(1), 53–74.
- Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, 41(4), 721–744.
- DICTA. Analytical tools for Hebrew texts: <https://morph-analysis.dicta.org.il/>
- Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELJ Journal*, 31(2), 105–119.
- Gablasova, D., & Brezina, V. (2021). Words that matter in L2 research and pedagogy: A corpus-linguistics perspective. *Studies in Second Language Acquisition*, 43(5), 958–961.
- Ha, L., Sicilia-Garcia, E., Ming, J., Smith, F. (2002). Extension of Zipf's law to words and phrases. In *COLING 2002: Proceedings of the 19th International Conference on Computational Linguistics*.
- Hatzigeorgiou, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics*, 8(3), 175–185.
- heTenTen21 corpus: Hebrew Web corpus 2021: <https://www.sketchengine.eu/hetenten-hebrew-corpus/>
- Ilgen, B., & Karaoglan, B. (2007). Investigation of Zipf's "law-of-meaning" on Turkish corpora. In *22nd International Symposium on Computer and Information Sciences (ISCIS 2007)* (pp. 1–6).
- Jayaram, B. D., & Vidya, M. N. (2008). Zipf's law for Indian languages. *Journal of Quantitative Linguistics*, 15(4), 293–317.
- Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, 7(1), 163–194.
- Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, 50(40), 976–987.
- Kweon, S. O., & Kim, H.R. (2008). Beyond raw frequency: Incidental vocabulary acquisition in extensive reading. *Reading in a Foreign Language*, 20(2), 191–215.

- Lau, S. H., Huang, Y., Ferreira, V. S., & Vul, E. (2019). Perceptual features predict word frequency asymmetry across modalities. *Attention, Perception, & Psychophysics*, 81, 1076–1087.
- Laufer, B. (2017). From word parts to full texts: Searching for effective methods of vocabulary learning. *Language Teaching Research*, 21(1), 5–11.
- Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language*, 13(2), 549–566.
- Leijten, M., Van Waes, L., & Ransdell, S. (2010). Correcting text production errors: Isolating the effects of writing mode from error span, input mode, and lexicality. *Written Communication*, 27(2), 189–227.
- Lenko-Szymanska, A. (2002). How to trace the growth in learners' active vocabulary? A corpus-based study. In B. Kettermann & M. Georg (Eds.), *Teaching and learning by doing corpus analysis* (pp. 217–230). Rodopi.
- Llach, M. P. A., & Gallego, M. T. (2009). Examining the relationship between receptive vocabulary size and written skills of primary school learners. *Atlantis*, 31(1), 129–147.
- Mairano, P., & Santiago, F. (2020). What vocabulary size tells us about pronunciation skills: Issues in assessing L2 learners. *Journal of French Language Studies*, 30(2), 141–160.
- Manor, R., & Watad, A. (2024). The complexity of teaching Hebrew in Israel's Arab school system. *Language Teaching*, 58(1), 14–26.
- Maw, T. M. M., Clenton, J., & Higginbotham, G. (2022). Investigating whether a flemma count is a more distinctive measurement of LD. *Assessing Writing*, 53, 100640.
- Meara, P. M. (1992). *EFL vocabulary tests*. University College Swansea: Center for Applied Language Studies.
- Miao, J. (2019). A study on English majors' vocabulary size and its correlation with scores on test for English majors grade-four. In *2nd International Conference on Contemporary Education, Social Sciences and Ecological Studies* (pp. 202–208). Atlantis Press.
- Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, 16(2), 187–205.
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.
- Milton, J., & Meara, P. (1998). Are the British really bad at learning foreign languages?. *Language Learning Journal*, 18(1), 68–76.
- Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review*, 4(1), 151–172.
- Milton, J., Wade, J. & Hopkins, N. (2010). Aural word recognition and oral competence in English as a Foreign Language. In R. Chacon-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (83–98). Multilingual Matters.
- Montero-SaizAja, A. (2022). The productive vocabulary of multimodal and unimodal English as a Foreign Language learners. *Porta Linguarum: Revista Internacional de Didáctica de Las Lenguas Extranjeras. Monográfico*, 4, 139–153.
- Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82.
- Nation, I. S. P., & Coxhead, A. (2021). *Measuring native-speaker vocabulary size*. John Benjamins Publishing Company.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge.
- Nation, P. (2021). Thoughts on word families. *Studies in Second Language Acquisition*, 43(5), 969–972.
- Nizonkiza, D. (2016). First-year university students' receptive and productive use of academic vocabulary. *Stellenbosch Papers in Linguistics*, 45, 169–187.

- Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, 76(4), 763–782.
- Pinto, J. D. (2018). *Creating a frequency dictionary of spoken Hebrew: A reproducible use of technology to overcome scarcity of data*. Doctoral dissertation, University of Texas at Austin.
- Richards, B. J., & Malvern, D. D. (1997). *Quantifying lexical diversity in the study of language development*. University of Reading, Faculty of Education and Community Studies.
- Rowe, M. L., Raudenbush, S. W., & Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, 83(2), 508–525.
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64(4), 913–951.
- Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press.
- Shimron, J., & Sivan, T. (1994). Reading proficiency and orthography evidence from Hebrew and English. *Language Learning*, 44(1), 5–27.
- Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 39(3), 302–327.
- 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.
- Vassiliu, P. (2001). *Lexical input and uptake in the low level EFL classroom*. Unpublished PhD dissertation, University of Wales Swansea.
- Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PloS ONE*, 11(1), e0147924.
- Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(1), 79–95.
- Webb, S., & Rodgers, M. P. (2009). The lexical coverage of movies. *Applied Linguistics*, 30(3), 407–427.
- Webb, S., & Rodgers, M. P. (2009). Vocabulary demands of television programs. *Language Learning*, 59(2), 335–366.
- Wesche, M., & Paribakht, T. S. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *The Canadian Modern Language Review*, 52, 154–178.
- Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold.