Journal of Applied Linguistics and Language Research
Volume 3, Issue 1, 2016, pp. 15-35
Available online at www.jallr.com
ISSN: 2376-760X

How Different Is Arabic from Other Languages? The Relationship between Word Frequency and Lexical Coverage

Ahmed Masrai *
Assistant Professor, King Abdulaziz Academy, Saudi Arabia

James Milton
Professor, Swansea University, UK

Abstract

This study examines Zipf's law as a predictor of the relationship between word frequency and lexical coverage in Arabic. Zipf's law has been applied in a number of languages, such as English, French and Greek, and revealed useful information. However, word derivation processes are far more regular and extensive in Arabic than they are in English and it is suspected that how words are defined may significantly affect the outcome of this kind of analysis. The concept of the lemma as applied to English could be redrawn for Arabic entirely credibly. In this study, Arabic lemmatised frequency lists generated from a large Web-based corpus have been used to calculate coverage. Results show that Zipf's law does apply in Arabic, and the findings suggest that the most frequent 9,000 lemmatised words provide approximately 95% coverage, and 14,000 words give nearly 98% coverage. These results suggest that the relationship between word frequency and coverage in Arabic is comparable, to a certain degree, to English and Greek, but not to French. However, the definition of the lemma used in this study is probably more relevant to European languages than to Arabic and if this was changed it would significantly change the results.

Keywords: Arabic corpus, lexical coverage, word frequency, vocabulary, Zipf's law

INTRODUCTION

Word frequency has long been associated with language vocabulary acquisition. Palmer (1917: 123) wrote "… the more frequently used words will be the more easily learnt …" Later scholars (e.g., Mackey, 1965; McCarthy, 1990) accept this idea as self-evident and repeat Palmer's idea without reservation. It can therefore be suggested that word frequency could govern much of our understanding of which words are learned, how they are learned, how to test for word knowledge and how to establish a relationship between lexical coverage and text comprehension. Extensive work in second language vocabulary acquisition has led some researchers (e.g. Adolph & Schmitt, 2003; Laufer,

* **Correspondence:** Ahmed Masrai, Email: a.masrai@hotmail.com

1989; Milton, 2009; Nation, 2006) to explore proportions of the vocabulary a learner of English as a foreign language (EFL) needs for adequate text comprehension. Milton (2009), however, extends the work to include languages other than English to examine this relationship.

The present study contributes to this line of research by examining word frequency and potential text coverage in Arabic in the light of Zipf's law. The concept of Zipf's law is that the idea of word frequency and its usefulness in language learning is applicable in all languages. Arabic is a Semitic language and has a unique word structure; whether Zipf's law can be applied to Arabic is thus investigated in this paper. Understanding the nature of words in a language will not only help its learners, but may also raise issues for native speakers when approaching second language (L2) acquisition. This study, however, follows the methodology implemented in Milton (2009) in terms of identifying word frequency rankings and their contribution to normal text coverage. To confirm the validity of Zipf's law when applied to the Arabic lexicon we compare the data with English, French and Greek, since data is available in these languages.

The paper is organised as follows: first, the importance of word frequency in vocabulary acquisition research and the insights it provides for vocabulary testing is discussed. Then a literature background for the current study is provided, including review of the structure of morphological units in Arabic. Next, the methodology employed in the study is described, and the analysis used for the investigation is outlined. Following this, the results and discussion of the data are presented. Finally, the limitations of the study are discussed, suggestions for future research are offered and conclusions of the study are drawn.

IMPORTANCE OF WORD FREQUENCY

Word frequency lists have allowed researchers to find out what proportion of vocabulary knowledge is needed in a text before any level of comprehension would occur. However, most of the research of this kind has been executed extensively in English. Figures revealed by some of the research which investigates the relationship between coverage and normal text comprehension are found to be productive for L2 learners, particularly for EFL learners, as there is abundant research in the English language. For example, the most frequent 2,000 words, lemmas, in English generally provide coverage of around 80% of a normal text (Nation, 2001). Accordingly, Nation suggests that these 2,000 most frequent words are very important to English language learning and any effort made to make sure they are learned is worth doing. A lemma definition used in this study includes a headword and its most frequent inflections, and this process must not involve changing the part of speech from that of the headword. In English, the lemma of the verb govern, for example, would include governs, governed, and governing but not government which is a noun and not a verb and, by this method of counting, would be a different word.

There are words in languages which are used more frequently than others and the chance of encountering them in a text or a conversation is quite high. Thus, the

importance of the word frequency-based studies resides in their ability to identify which words are likely to occur more frequently than others and how often they are encountered. Certainly, to generate a reasonably reliable and useful word frequency list, one would need a large corpus of a language that covers several domains to draw on. Further, according to Milton (2009), word frequency is central to understanding the process of vocabulary learning - and how this process can be studied and assessed. He suggests that if the types of lexical items that are likely to be learned first are known and which are not, then it is feasible to construct much better measures of vocabulary knowledge.

Milton (2009) proposes that the relationship between word coverage and text comprehension in English might be reflected in other languages, but this may be in different ways. The idea of word frequency and its usefulness in language learning has been long proposed by Zipf's law. Zipf's law is known as the distribution of the probability of occurrences of words along a continuum, since some words occur very frequently, while others do not. The use of Zipf's law is not intended to be restricted to the English language alone, but should be applicable to all languages. These assumptions are discussed in more detail in subsequent sections in this paper.

It should be pointed out that Milton (2009) has made an effort to put the idea of word frequency and text coverage in some languages, such as French and Greek, in context. In English, knowledge of the most frequent 2,000 words, lemmas, which gives coverage of around 80% of normal text, can indicate whether a learner is likely to have the ability to perform at all outside the classroom setting and gains meaning form normal texts (Nation, 2001). Also, knowledge of around 5,000 word families (Laufer, 1989, 1992), giving coverage of nearly 95%, can tell whether a learner is able to comprehend a normal text in a way similar to an educated native speaker. These informative figures can help the processes of goal setting for L2 vocabulary learning. The relationship between word frequency and text coverage for French and Greek is presented in this paper. In the case of Arabic this relationship is not yet clear, as, to the best of our knowledge, no study has attempted to look at it. Nonetheless, whether figures from English hold good for Arabic will depend on whether a word is defined narrowly and includes almost exclusively the regularly formed inflections, as in English, (Bauer & Nation, 1993) or extends it to include all regularly formed inflections and derivations. A narrow lemma definition, as in English, might mean that what appears to be low coverage in Arabic would provide high comprehension since Arabic speakers can easily derive many words not covered in the definition of a lemma (Boudelaa & Marslen-Wilson, 2000; Idrisi & Kehayia, 2004).

As this kind of research can give insights into the way a learner functions in a language, then investigating more languages than those described by Milton (2009) is worthwhile. This paper, therefore, will explore if the ideas of word frequency and text coverage, and Zipf's law can translate to the Arabic language and to what degree. The study will also explore whether the same kinds of figures, disclosed in English, for minimal or more general comprehension can be produced in the Arabic language. Before going further to

present the current study, we will first review the relationship between word frequency and coverage in the light of Zipf's law.

The relationship between word frequency and coverage: Zipf's law

It has been noted in the previous section that the idea of word frequency is not new and dates back at least to 1917, when Palmer described the relationship between word frequency and learning. Palmer's idea suggests that the most frequent words in a language will be learned earlier than the infrequent words and that the most frequent words are the most useful to the learner to express himself/herself in an efficient way. According to Milton (2009), Palmer's (1917) concept of word frequency and word learnability is, arguably, correct in languages such as English, French and Greek. The question raised here is whether this applies to the Arabic language. In English, for example, the concept lemma is suitable because it mainly includes the base word and most common inflections, and that derivations are often infrequent and irregular, which are developed and acquired in later stages of learning (Bauer & Nation, 1993). In contrast, regularity of rules in Arabic to derive new words from roots, which are acquired in early stages of learning, may show significantly different figures of coverage. This study therefore seeks to bring to the surface some information about the nature of Arabic language vocabulary in terms of frequency and coverage.

It is indicated that words like *the* and *be*, are the most frequent two words in English (Kilgarriff, 2006) and words like *maunder* and *ecumenical* are found at the end of the frequency scale; these infrequent words are numerous. In between these two extremes of the frequency scale there is a body of words of medium frequency. This kind of word distribution is known as a Zipf's distribution and gives rise to Zipf's law. Zipf's law allows the relationship between the rank of a word in a frequency list and the number of times it occurs to be described more systematically and graphically presented (Milton, 2009: 45). The idea of Zipf's law suggests that the probability of occurrence of the word that is ranked first in a corpus is twice the occurrence of the word that is ranked second; the word ranked second is most likely to be twice as frequent as the word ranked fourth, and so on. Table 1 illustrates the rankings and frequency of the eight most frequent words in the English and Arabic corpora.

In the data presented in Table 1, Zipf's law does not seem to perfectly describe the relationship between word occurrence and frequency rankings. The table shows that the word ranked first in English is not exactly twice as frequent as the word ranked second. The frequency of both second and fourth frequent words in English are much greater than 50% of the first and second most frequent words, respectively. The eighth ranked word in English, *have*, is almost half that of the fourth word *and*. This regularity is not clear in the Arabic corpus. As can be noted from Table 1, the first ranked word in Arabic, الـ, occurred nearly five times as much as the word ranked second, وَ. Additionally, the word ranked eighth, أَنَّ, is far less than half that of the fourth word, مِن.

The relationship between coverage and frequency seems neater in the Brown corpus (Kučera & Francis, 1967) than in Kilgariff (2006), however. Here, the most frequent

word in English, *the*, occurred around 69,771 times in the whole corpus (over one million words), which accounts for about 7% of the entire corpus. The second ranked word, *of*, accounts for nearly 3.5% (36,411 occurrences of over one million words). This gives an indication that the first ranked word occurs almost twice as much as the word ranked second, which confirms Zipf's law in English.

Table 1. Ranks and frequencies of the first eight most frequent words in English and Arabic

English (Kilgariff, 2006)		Arabic (Buckwalter & Parkinson, 2011)	
the	6,187,267	الـ	5,004,793
be	4,239,632	وَ	1,110,144
of	3,093,444	فِي	92,482.3
and	2,687,863	مِن	74,519
a	2,186,369	لِـ	58,478.6
in	1,924,315	بِـ	55,323.4
to	1,620,850	عَلَى	51,869.2
have	1,375,636	أَنَّ	30,394.2

According to Milton (2009: 46), "Zipf's law is not a perfect description of language; therefore, it is an empirical law not a theoretical one". Nevertheless, Zipf's law can roughly explain how many vocabulary items are needed to reach a certain percentage of text coverage. For example, in the Brown corpus (Kučera & Francis, 1967), around 135 lexical items are needed to account for half of the entire corpus. To further clarify this, a list of frequency bands and the coverage they might provide in English is illustrated in Table 2. Figures in Table 2 show that a small proportion of vocabulary can provide EFL learners with reasonable text coverage. For example, 10 lexical items in English could provide learners with 24% of normal text coverage and around 1,000 words can increase the coverage to 74%. Furthermore, a lexical knowledge of the most frequent 2,000 words in English would increase the percentage of normal text coverage to approximately 81%.

Table 2. Typical coverage figures for different frequency bands in English (Carroll, Davies & Richman, 1971, cited in Nation, 2001)

Number of words	Text coverage (%)
10	24
100	49
1,000	74
2,000	81
3,000	85
4,000	88
5,000	89
12,000	95
44,000	99
87,000	100

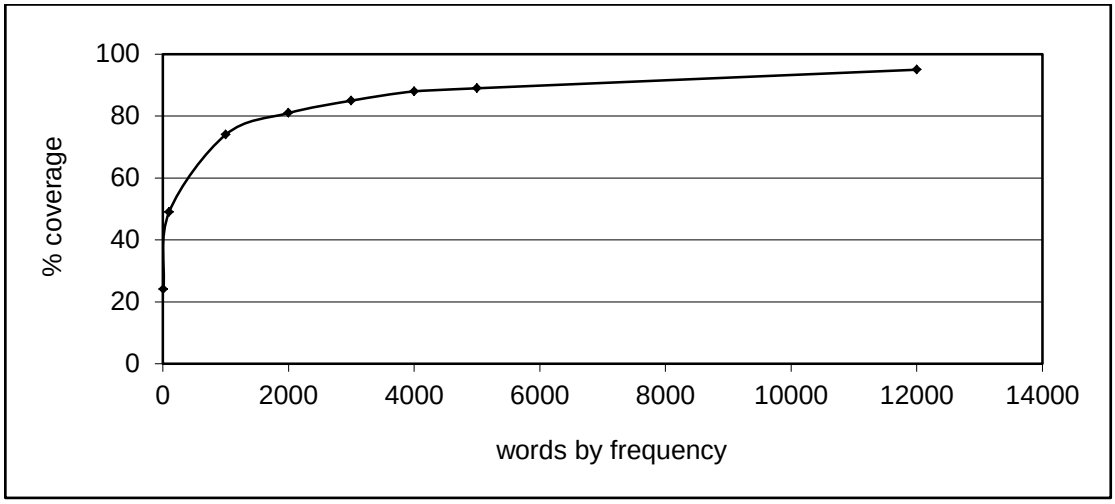

Figure 1. The most frequent bands in English and lexical coverage (Milton, 2009: 47)

The relationship in English between word frequency and coverage is clearly illustrated in Figure 1 - the curve rises sharply on the left side of the graph and starts to flatten after the 2,000 most frequent words. The steep rise of the curve on the left hand side of the graph indicates that any additional word contributes heavily to text coverage. Figures in Table 2, graphically shown in Figure 1, suggest that knowledge of the most frequent 1,000 words in English would enable EFL learners to understand around 75% of normal text and mastering the most frequent 2,000 words adds another 5-6% to their comprehension. Therefore, if the knowledge of the 2,000 most frequent words in English provides a large proportion of text coverage, then any effort in learning those words is worthwhile (Nation, 2001).

The knowledge of the 2,000 most frequent words in English appears to be very influential in normal text understanding. However, can this idea be proven to be similar in other languages, such as Arabic, and does the figure of 2,000 words apply to Arabic? There is no clear answer to this question yet. Therefore, this study explores if a similar coverage could be drawn in the Arabic language. In the following sub-section, a brief review of the relationship between word frequency and lexical coverage is provided for other available corpora, i.e., Greek and French, to compare with the outcome from Arabic.

Coverage in Greek and French corpora

The discussion of coverage and comprehension has been, so far, almost exclusively on the English language corpora and research into EFL learners. As digitised texts become increasingly available, some well-developed corpora other than English also become available. This has allowed word count and frequency lists to be generated and compared. Milton (2009) has compared the French and Greek languages to English and found, although some variation has emerged, that Zipf's law appears to be remarkably robust. He generally proposes that the feature of all languages appears to be a small

number of words that are very frequent and similarly offer a large amount of text coverage.

Milton (2009: 67-68) presents a comparison of coverage between three corpora: English corpus (Carroll, Davies & Richman, 1971), Hellenic National Corpus, Greek (Hatzigeorgiu, Mikros & Carayannis, 2001) and Baudot's (1992) French corpus. The comparison is illustrated in Figure 2. The figure shows the line of coverage between these three corpora when lemmatised present approximately an equivalent list. At the beginning, the first few words appear to be comparatively more frequent in Greek and French than in English. For instance, the definite article in Greek is relatively more frequent than in English (Milton, 2009). Therefore, Greek vocabulary provides proportionately less coverage and, as the graph shows, the lines cross over. In English for example, the most frequent 5,000 words provides 89% text coverage (Carroll et al., 1971), whereas the same proportion of words only provides 82.6% coverage in the Greek corpus. One distinct aspect of the Greek language is the existence of the large number of *hapax legomena* (word or phrases which are logged as having been used only once), which make up about 49.4% of the entire corpus. Additionally, the corpus from which lists were generated was constructed from a massive number of small items on many different topics from online journalism without inclusion of spoken elements (Hatzigeorgiu et al., 2001). In English and similarly in French, *hapax legomena* comprise about 30%, which is substantially less than the Greek. Nevertheless, the data shown in Figure 2 strongly suggests that the details of coverage and the lines in the graph are similar in both English and Greek, which confirm that Zipf's law functions in Greek as in English (Hatzigeorgiu et al., 2001).

French, on the other hand, indicates that the levels of coverage, shown in Figure 2, are very similar to English. However, comparable to Greek, the most frequent few words in French are more frequent than the most frequent words in English but, unlike Greek, the distribution trend is not lost at the less frequent levels, and it can be seen the two lines in the graph are almost identical. Yet the same proportion of words constantly provides a slightly higher coverage in French than in English. In spite of the fact that there are some differences in detail between languages, Zipf's law seems to work well with languages other than English, here, Greek and French. Will Zipf's law work well in a language like Arabic as well? This study examines whether Zipf's law can give a similar distribution in Arabic and whether the same number of words provides similar coverage to that in English.

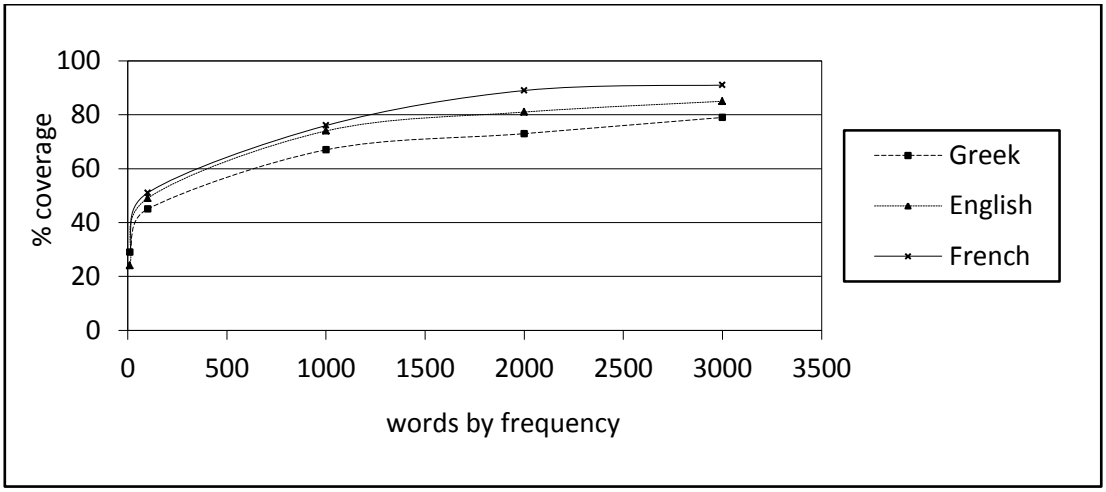

Figure 2. Comparing coverage between Carroll et al.'s (1971) English corpus; Hatzigeorgiu et al.'s (2001) Greek corpus and Baudot's (1992) French Corpus

MORPHOLOGICAL UNITS IN THE ARABIC MENTAL LEXICON

There is little evidence regarding how words are likely to be stored in the native Arabic speakers' mental lexicon. One of the few studies found in the literature is Idrissi and Kehayia (2004), which investigated the role of morphology in the organisation and representation of lexemes in Arabic speakers' mental lexicon with patients who suffer from deep dyslexia. The motive for the study was the assumption proposed by Bohas (1997) that the core meaning of words in Arabic is encoded within the biliteral component of the root: '*etymon*' (Idrissi & Kehayia, 2004: 185). It has long been known that the root, which is generally comprised of three ordered consonants, is core in forming words in Arabic. Findings from Idrissi and Kehayia's study suggest, based on priming experiments and speech error, that the morphemic/lexical status of the three consonantal root is dominant.

Priming experiments conducted in Semitic languages (Arabic and Hebrew) have discovered that the root morphemes are particularly responsible for a resilient morphological priming effect (Boudelaa & Marslen-Wilson, 2000; Deutsch, Frost & Forster, 1998; cited in Idrissi & Kehayia, 2004). Participants in these studies were found to react more quickly when target and prime words shared the same root than when they did not. This kind of effect thus suggests that the role of the root in lexical access and similarly lexical and morphological representation are crucial. Further evidence to support that the root is responsible for word formation in Arabic is data from speech errors from both impaired and unimpaired speech. In a study by Prunet, Beland and Idrissi (2000), which investigated the speech of an Arabic-French patient with deep dyslexia, evidence was found to support the status of the abstract consonantal root as a morphological/lexical unit in Arabic.

Pertinent to the current study is how Arabic words might be stored in Arabic speakers' mental lexicon; are they stored as base words, or may they be stored as *lemmas*? In English, for example, there is broad evidence from slips of the tongue and aphasic

patients that base forms and the rules for inflection are responsible for forming the basis of words and that derived forms tend to be stored and accessed independently (Aitchison, 2003). Derivations in English tend to be less frequent and less regular and words derived with these affixes are thought to be stored separately from the base word (Aitchison, 1987; Gardner, 2007). From the little evidence we have about the role of roots to form the basis for words in Arabic, one can loosely assume that words in Arabic might also be stored as base words, with rules for inflecting or deriving these words being stored separately. Nonetheless, the large number of highly regular derivations that can be applied to the base words in Arabic, results in a large lexicon size. If a lemma is taken to include highly regular inflections and derivations, then in Arabic the lemma might be much more extensive than in English and would result in a bigger vocabulary size and provide more coverage than in English.

This study will explore the relationship between word frequency and lexical coverage, and postulates that if words are stored as base words in Arabic learners might achieve comprehension with less coverage when words are calculated as lemmas; Zipf's law will be implemented in this study to show the distribution of words by frequency and their relationship to coverage. Furthermore, to examine the distribution of word frequency against its rank in the Arabic corpus (Sharoff, 2006), log-log Zipf's distribution for the most frequent 100,000 word, types, was performed. Further, to allow the comparison of text coverage provided by types and lemmas to be made, analysis of the types corpus was run. *Al-Morid Al-Qarib*, one of the largest Arabic dictionaries indicates that Arabic language has only about 10,000 base words. This suggests that the Arabic language is a highly inflected and derivative language and rules for generating new words from base words are applied extensively (Habash, 2010). Therefore, it is predicted to see far less text coverage provided by types than in the case of lemmas.

METHODOLOGY

This study aims to achieve its objectives by examining:

1. The application of Zipf's law in the most frequent 12,000 (lemmas) in Arabic to find out if it is applicable in Arabic language.

2. The frequency distribution against the probability of text coverage a certain number of words might provide.

3. Whether the lemma definition of a word count is suitable to predict text coverage in Arabic.

4. Log-log Zipf's distribution of word rank and the total number of occurrences.

5. The distribution of the 20 most frequent words (types) in Arabic around the Zipf's curve.

Concerning the third research question, in Modern Standard Arabic (MSA) almost all words are broadly based on a root morpheme, which is typically composed of three or four consonant letters (e.g., *d-r-s* is the root morpheme for the general concept to *study*)

(Abu-Rabia, 2002; Shimron, 1999), when creating hundreds of words. Thus, it is assumed that fewer words in Arabic would provide greater coverage because Arabic speakers are likely to draw heavily on roots and store words in base forms rather than lemmas.

As research investigating Arabic lexicon is generally scarce, areas such as the relationship between word frequency and text coverage and how words might be stored in Arabic lexicon are under-explored. Therefore, this study hopes to form a grounding for carrying out future research in this area, such as investigating coverage and levels of comprehension in Arabic, which is beyond the scope of the current research.

Procedures and method of analysis

The lemmatised frequency lists analysed in the current study were generated from a large Arabic web-based corpus (around 180 million words, tokens). This Arabic corpus is one of a series of nine language web-based corpora (Sharoff, 2006, 2007). As the Arabic language has one of the most sophisticated writing systems, generating frequency lists and hence producing lemmatised lists is quite a challenging task. However, in a project called 'Kelly', Kilgarriff et al. (2011) have generated unlemmatised frequency lists for nine different languages; Arabic is one of these. After the Arabic frequency list has been generated from the large web-based corpus, in subsequent work, Sawalha and Atwell (2011) produced a 100,000 lemmatised words list from Kilgarriff et al.'s lists.

However, the most frequent 20,000 words in Arabic, lemmas were placed into a Microsoft Excel spreadsheet to calculate the occurrences of these words in the entire corpus. In order to establish a clear comparison with the other languages, which have been already investigated, in terms of frequency and coverage, percentages of occurrences for the 20,000 words were calculated. The probability of text coverage was calculated separately for 10, 100, 1,000, 2,000, 3,000, 4,000, 5,000, 9,000, 12,000 and 14,000 words. It is believed these different proportions of words should make a comparable distribution, using Zipf's law, of the other languages intended for comparison. Zipf's law was mainly used in the current study because it is one of the most known tools in quantitative linguistics that links the rank of the word to its frequency. Also, the validity of Zipf as an empirical law has been observed in a large spectrum of phenomena, including natural languages, economics, biological systems, and even in statistics of Web usage (Mikros, Hatzigeorgiu & Carayannis, 2005: 171). Moreover, the validity of Zipf's law has been verified for a considerable number of languages (Miller, Newman & Friedman, 1958; Rousseau & Zhang, 1992).

The same methodology used for analysing lemmatised words was also implemented for words based on the *type* definition of a word count. Nonetheless, Zipf's distribution was only conducted for a sample of the corpus of types.

RESULTS AND DISCUSSION

Coverage of different frequency bands in Arabic

Table 3 presents the likely coverage by the most frequent 12,000 words, lemmas, in Arabic. These words were broken down to nine different word frequency levels. It can be seen from Table 3 that knowledge of the most frequent 1,000 words in Arabic can provide 66% coverage of a normal text. In other words, knowing the most frequent 1,000 words in Arabic means understanding about 66% of the words appearing in normal text or even everyday conversation. Knowledge of the most frequent 2,000 words in English is suggested to be of paramount importance (Nation, 2001). Knowing these words in English could enable the learner to understand about 80% of the words presented in a normal text; therefore, Nation (2001: 16) suggests that anything that can be done to make sure that they are learned is worth doing.

In Arabic on the other hand, the most frequent 2,000 words appear to give less text coverage than the most frequent 2,000 words in English- 4% less. However, 76% text coverage, which is the knowledge of the most frequent 2,000 words in Arabic, still appears to be considerable. Unlike English, the second thousand frequency band contributes largely to text coverage in Arabic, by an additional 10%. In English the second thousand adds about 7% coverage (Carroll et al., 1971). Nonetheless, in both English and Arabic, the most frequent 2,000 words appear to give broad text coverage.

Table 3. Coverage figures for different frequency bands in Arabic

Number of words	Text coverage (%)
10	12
100	34
1,000	66
2,000	76
3,000	82
4,000	86
5,000	89
9,000	95
14,000	98

Furthermore, the most frequent 3,000 words in Arabic provide coverage of about 82% and the 4,000 frequency level adds another 4% to increase the coverage to 86%. It therefore becomes clear that as the words become less frequent their contribution to coverage reduces. For example, the fifth thousand only adds 3% to the overall coverage. Interestingly, the most frequent 5,000 words in Arabic provide exactly the same coverage as the most frequent 5,000 words in English (Carroll et al., 1971). Both provide coverage of about 89%. In Arabic, however, learners seem to hit the 95% coverage of the text when they know around 9,000 words, lemmas. Moreover, if 98% of text coverage is necessary for comprehension, then learners should know about 14,000 words, lemmas.

Application of Zipf's law for 14,000 words, lemmas, in Arabic

One of the aims of the current study was to investigate whether Zipf's law can be applied in a language like Arabic and would work equivalently well as in other languages. For this purpose, proportions of words reported in Table 3 are shown in Figure 3 for clarification. In the graph, the curve rises very steeply on the left and in this area each additional word seems to contribute considerably to text coverage. Knowing the first 1,000 words in Arabic means understanding almost two thirds of the words presented in a normal text and knowing the second 1,000 words will increase the likelihood of understanding to nearly 80%. Therefore, it could be argued that if a learner masters these words, then he/she would know a large quantity of texts he/she reads or hears and might even largely understand them.

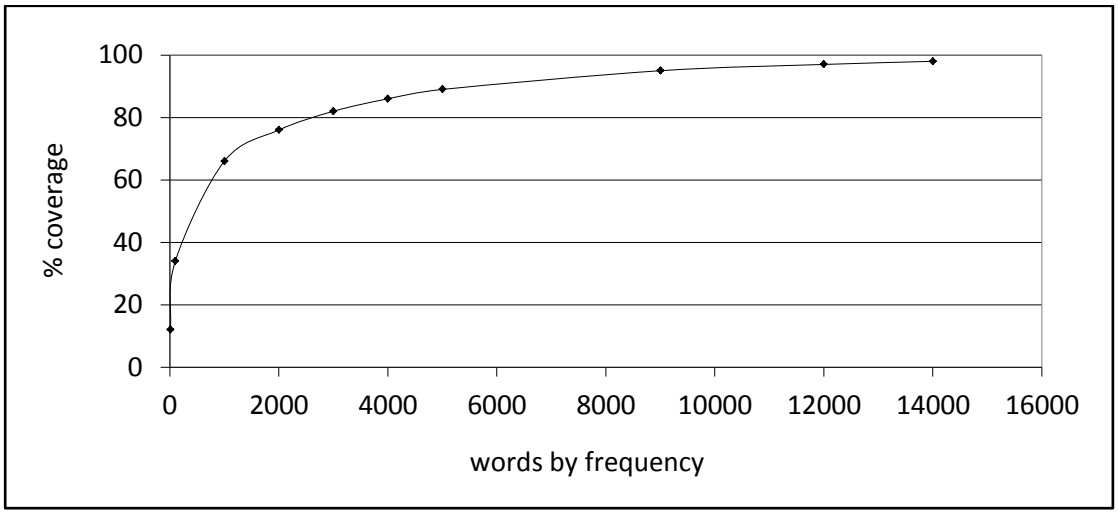

Figure 3. Coverage of the most frequent bands in the Arabic corpus

In accordance with the first research question in this study, whether Zipf's law is able to allow the relationship between the ranking of a word in a frequency list and the number of times it occurs to be systematically described and graphed up, Zipf's law seems to be remarkably robust when applied to the Arabic corpus, in spite of the fact that there are variations between languages. The Arabic language appears to share the same features as other languages in the fact that a small number of words are very highly frequent and can provide a large proportion of text coverage. However, to further illustrate this result, the researchers have compared Arabic with English, French and Greek to see clearly the relationship between the number of words when ranked by frequency and the potential coverage of the text they provide.

Figure 4 overlays the line for coverage from Carroll et al.'s (1971) corpus of English with the Arabic Web-based corpus (Sharoff, 2006) after it has been lemmatised to give a roughly equivalent list. At the outset, the first few words are comparatively more frequent in English than in Arabic. Thereafter, Arabic vocabulary provides proportionately similar coverage until it overlapped at the point of the 5,000 most frequent words, where the difference becomes indistinguishable. The lines continue in an identical trend to near the 12,000 most frequent words where the Arabic vocabulary begins to provide slightly higher coverage. The 5,000 most frequent words provide

exactly the same coverage in both English and Arabic (89%), whereas the most frequent 12,000 words provide coverage of 95% in English and nearly 97% in Arabic.

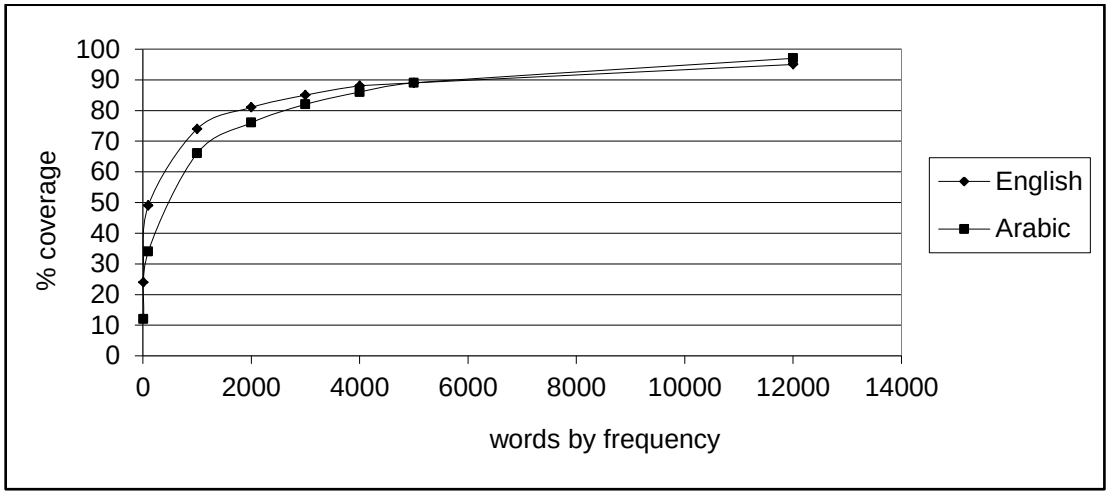

Figure 4. Comparing coverage between Carroll et al.'s (1971) English corpus and Sharoff's (2006) Web-based corpus

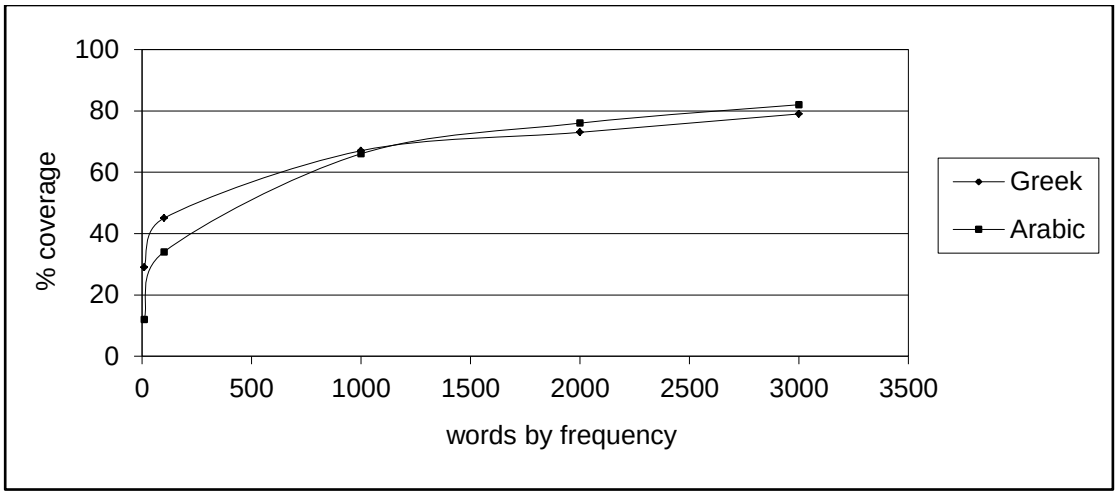

Figure 5. Comparing coverage between Sharoff's (2006) Web-based corpus and the Hellenic National Corpus (Hatzigeorgiu et al., 2001)

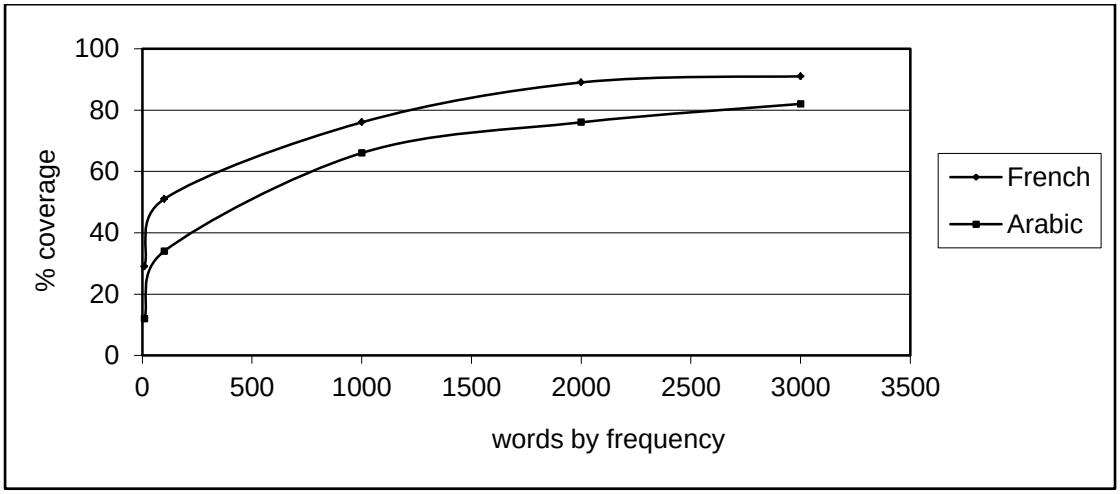

Figure 6. Comparing coverage between Baudot's (1992) French corpus and Sharoff's (2006) Arabic web-based corpus

Figures 5 and 6 illustrate the comparison of word coverage in Arabic with both Greek and French. In Greek, the first few words give higher coverage than Arabic, but Arabic appeared to provide slightly higher coverage in the later frequency bands. In French, on the other hand, the difference in coverage is noticeable from the outset. The most frequent 3,000 words in French appeared to provide relatively higher text coverage when compared to Arabic. This might be a product of the fact that Baudot's (1992) corpus is relatively small, as it includes only 1.2 million words and this may not be sufficient to give stable figures beyond a few thousand words in the frequency lists. Therefore, is does not necessarily mean that Arabic speakers need more words than French speakers to achieve comparable text comprehension. Arabic speakers would most probably utilise the high regularity of derivation to know words beyond the restriction of lemma definition.

Rank-frequency distribution in the Arabic corpus

The last part in the current study describes the Zipf's distribution of the words frequency against their ranks in the Arabic corpus. Figure 7 shows the log-log distribution of the most frequent 100,000 types. As can be seen, the descent of word frequencies is almost a perfect 45-degree slope. However, there is a slight deviation of low frequency words at the end of Zipf's distribution. This kind of deviation is probably attributable to the fact that words at the end of the frequency scale have the same number of occurrences in the corpus being analysed. This phenomenon, according to Baroni (2009), is common in natural languages. Nonetheless, it can be observed here that the sample words from the Arabic corpus are convincingly distributed across the Zipf's law fit line until it reaches the plateau of word occurrences at the end of the frequency scale.

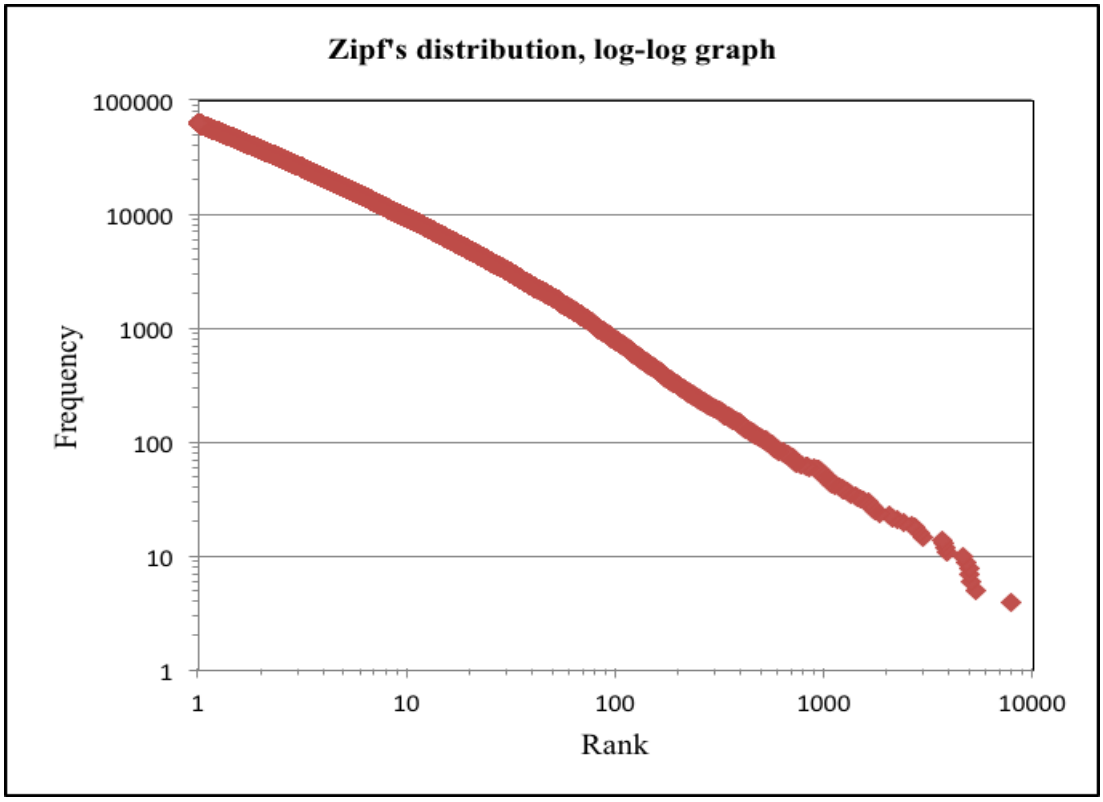

Figure 7. Zipf distribution log-log graph for types across the web-based corpus (Sharoff, 2006)

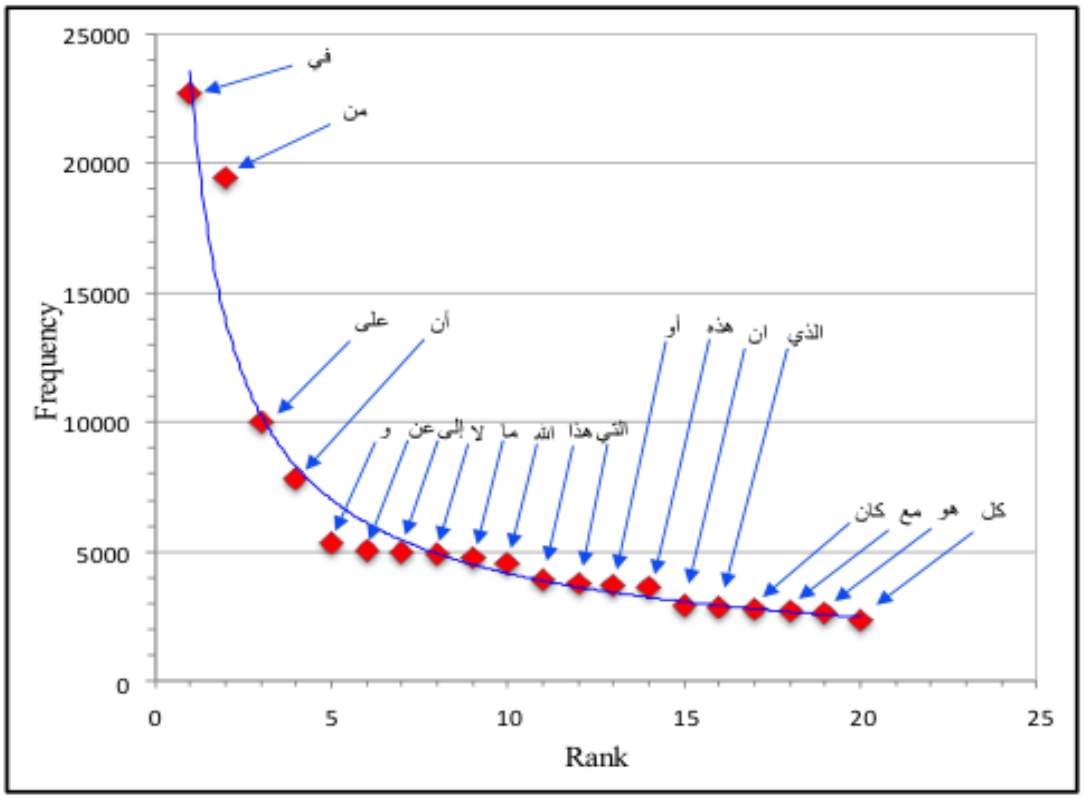

Figure 8. Distribution of the 20 most frequent Arabic words around the Zipf curve

The final task in this study was to examine Zipf's law regarding predictability of the 20 most frequent words in Arabic around the Zipf curve, and to compare text coverage provided by lemmas and types. Figure 8 shows the scatter of the most frequent 20 Arabic types around the Zipf curve. The words are distributed in almost a descending order around the curve. Although the words are not perfectly scattered on the curved Zipf line, it is visually clear that the accuracy of distribution is considerable.

With reference to text coverage, the definition of a word seems influential in specifying the percentage of coverage for a proportion of a text. The results show a marked difference between coverage provided by lemmas and types in the Arabic corpus. Figure 9 illustrates that when words are counted as types far less coverage is reached than when words are counted as lemmas. The difference is expected, as lemmas include headword and other commonly inflected forms of a word. However, in Arabic this difference appears to be very large. The most frequent 14,000 words provide coverage of about 98% in the lemma count and around 60% when words are counted as types. This indicates that reliance on the regularity of rules in Arabic to derive a massive number of new words from roots is very substantial. Therefore, if a broader definition of lemma is taken into account, a highly predictable coverage would emerge by the first 1,000 frequent words. Nonetheless, it should be pointed out here that the Arabic morphological system is quite complex and producing word family lists is somehow difficult. Writing an Arabic morphological analysing tool is a challenging and sophisticated task (Jaafar & Bouzoubaa, 2014), but creating word family lists in Arabic would offer useful information for researchers and educators.

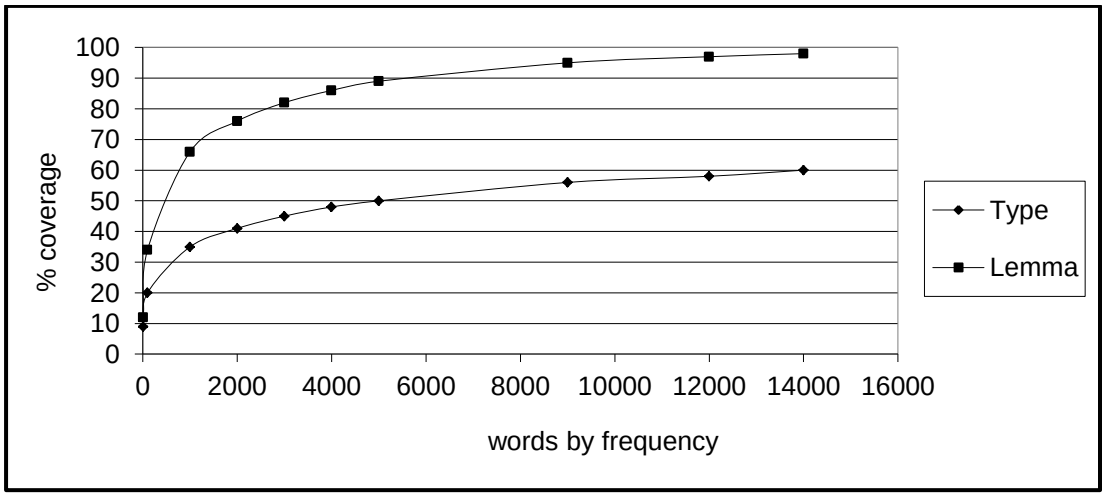

Figure 9. Coverage of the most frequent words (type and lemma) in the Arabic corpus

Findings from this study, however, suggest that Arabic native speakers or/and learners would achieve text comprehension with less coverage, as calculated in lemmas, because they depend heavily on roots in word formation, as suggested by some studies (e.g. Boudelaa & Marslen-Wilson, 2000; Idrissi & Kehayia, 2004; Prunet et al., 2000). Therefore, fewer words give greater coverage when counted as lemmas. Results from the current study tentatively imply that native Arabic speakers tend to learn the morphological rule at the outset of learning the language and apply these rules

extensively to generate new words from the roots. It could be assumed here that the structure of the Arabic speakers' mental lexicon might differ in a way from the structure of the mental lexicon of monolingual English speakers. Arabic speakers appear to store words as base words and utilise the morphological rules, which are acquired at the very early stages of language learning, to make up new vocabularies. Contrastingly, evidence suggests that the base words and the rules for inflection forming the basis of words in English and derived forms tend to be stored and accessed separately by monolingual English speakers (Aitchison, 1987; Gardner, 2007).

As the number of roots in Arabic is very low (nearly 10,000) and that Arabic speakers or/and learners draw heavily on the implementation of morphological rules in new word formations, calculating words using lemmas might, to an extent, underestimate Arabic speakers' vocabulary size and in turn underestimate coverage of the text in Arabic. In this study the relationship between word frequency and the coverage of the text in Arabic was compared with English, Greek and French. Close comparison was undertaken with English because it is one of the most researched languages and there is a body of research related to text coverage and comprehension (e.g. Laufer, 1989, 1992; Nation, 2001, 2006; Schmitt, Jiang & Grabe, 2011).

The vocabulary size estimates as well as text coverage figures that emerged in Arabic indicate the way in which lemmas are accessed in Arabic might be significantly different from the way lemmas are accessed in English. In English, for example, children and L2 learners first learn base words and the most inflected forms of these words and then develop derivations during later stages of learning. It seems from the types of error that speakers produce, derivations are stored and accessed as separate words (Aitchison, 1987). From this perspective, morphological awareness development seems to link with cognitive development and children will not utilise this feature until they reach a certain age (Casalis & Louis-Alexandre, 2000; Tyler & Nagy, 1989, cited in McBride-Chang et al., 2008). Similar obstacles are probably experienced by L2 learners, as they need first to grow a larger vocabulary and learn morphological rules before they can develop the ability to work out the derived forms of words. This suggests that the concept of the lemma might serve as a practical definition of a word to be used in vocabulary size estimates in English.

In Arabic, on the other hand, children's utilisation of the morphological feature is accessed at the initial stage of learning. The root [k t b,كتب] is not learned as a word but as a root comprised of three consonantal letters. Children or/and learners of Arabic, for example, need to link the article [ال] to the root [k t b] to form the word [الكتاب, book] and link the three consonantal letters [k t b] to form the word [كتب, wrote]. This pattern is applied to most words in Arabic, as it is heavily dependent on the roots to form new words (Boudelaa & Marslen-Wilson, 2010). Hence, the derivational morphology is acquired in a very early phase of learning the language. Therefore, vocabulary size tests based on lemma definitions of a word tend, but only tend, to underestimate Arabic speakers and learners' vocabulary.

LIMITATIONS OF THE STUDY AND SUGGESTIONS FOR FUTURE RESEARCH

It was suggested in this study that Zipf's law is practical in confirming aspects of quantifying the use of Arabic. Nonetheless, it seems that the concept of lemma might underestimate text coverage in Arabic because of the regularity of derivation processes that apply extensively in Arabic. Moreover, the study did not directly examine the levels of reading and listening comprehension in connection with the text coverage figures that emerged in Arabic. Thus, a further research goal that could be pursued in future studies is to identify the relationship between text coverage and comprehension, operationalising the word family definitions of words. Such research should scrutinise more accurately the number of words known in a text, the percentage of coverage and the level of comprehension.

CONCLUSIONS

This paper has considered some basic quantitative characteristics of the relationship between the rank of word frequency in Arabic and its contribution to the text coverage, applying Zipf's law. The study has examined whether Zipf's law is applicable to the Arabic corpus and whether it could produce comparable figures to languages such as English, French and Greek. Zipf's law was found to be valid to empirically work with Arabic frequency lists. It provided believable figures when compared with other languages. Investigating the validity of Zipf's law in Arabic in this paper is only an initial attempt to find out something about the nature of vocabulary in the Arabic language. Using Zipf's law in thorough corpus analysis in Arabic might reveal some more interesting ideas about Arabic vocabulary.

This study also explored the relationship between the number of words and the percentage of text coverage in Arabic. The results suggest that the first 2,000 most frequent words in Arabic can contribute enormously to text coverage (about 76%). Additionally, the most frequent 9,000 words appear to be very important figure in order to get a threshold of text comprehension - providing around 95% coverage of texts. Nonetheless, in Arabic, a learner might need to know around 14,000 words, which yields a coverage of 98% of normal texts, to reach a good level of comprehension.

Nevertheless, it is worth noting that Arabic lemmas are accessed differently from English lemmas. It was suggested throughout this study that the inflection and derivative rules in Arabic are learned at the very early stages of learning; therefore, using the lemma count of words might, to an extent, underestimate the coverage of normal text in Arabic.

To sum up, the findings suggest that the relationship between word rank and coverage of a text can be established in Arabic when implementing Zipf's law. Additionally, findings suggest that most frequent words in Arabic are the most important words and that Arabic learners might achieve a good level of comprehension with less Arabic words when calculated in lemmas. Yet these text coverage figures in Arabic need to be further validated in further empirical studies (e.g., reading comprehension and listening

comprehension studies) to know more accurately how many words are needed to get a full comprehension when reading or listening. This study only forms some grounding for research in this particular area, and more research needs to be carried out.

REFERENCES

Abu–Rabia, S. (2002). Reading in a root–based–morphology language: the case of Arabic. *Journal of Research in Reading, 25***,** 299-309.

Adolphs, S. & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24***,** 425-438.

Aitchison, J. (1987). *Words in the mind: An introduction to the mental lexicon.* Oxford: Blackwell.

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon* (3rd ed). Oxford: Blackwell.

Baroni, M. (2009). Distributions in text. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 803-821). Berlin: Mouton de Gruyter.

Baudot, J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain.* Montréal: Les Presses de l'université de Montréal.

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography, 6*(4), 253-279.

Bohas, G. (1997). *Matrices, etymons, racines.* Leuven: Peeters.

Boudelaa, S. & Marslen-Wilson, W. D. (2000). Non-concatenative morphemes in language processing: Evidence from Modern Standard Arabic. *Proceedings of SWAP, Nijmengen: Max-Plank-Institute for Psycholinguistics, 1***,** 23-26.

Buckwalter, T. & Parkinson, D. (2011). A frequencu ditionary of Arabic core vocabulry for learners. London and New York: Routledge.

Carroll, J. B., Davies, P. & Richman, B. (1971). *The american heritage word frequency book.* Boston: Houghton Mifflin.

Casalis, S. V., & Louis-Alexandre, M. F. (2000). Morphological analysis, phonological analysis and learning to read French: A longitudinal study. *Reading and Writing*, *12*, 303-335.

Deutsch, A., Frost, R. & Forster, K. (1998). Verbs and nouns are organized and accessed differently in the mental lexicon: evidence from Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24***,** 1238-1255.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28***,** 241-265.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies, 3***,** 1-187.

Hatzigeorgiu, N., Mikros, G. & Carayannis, G. (2001). Word length, word frequencies and Zipf's Law in the Greek language. *Journal of Quantitative Linguistics, 8***,** 175-185.

Idrissi, A. & Kehayia, E. (2004). Morphological units in the Arabic mental lexicon: Evidence from an idividual with deep dyslexia. *Brain and Language, 90***,** 183-197.

Jaafar, Y., & Bouzoubaa, K. (2014). Benchmark of Arabic morphological analyzers challenges and solutions. *In Proceedings of the Intelligent Systems: Theories and Applications (SITA-14), 9th International Conference*. 1-6.

Kilgarriff, A. (2006). *BNC database and word frequency lists* [Online]. http://www.kilgarriff.co.uk/bnc-readme.html#lemmatised. [Accessed 09/10 2014].

Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Bondi, J., Khalil, S., Johansson, S., Lew R., Sharoff, S., Vadlapudi, R. & Volodina, E. (2011). Corpus-based vocabulary lists for language learners for nine languages. *LREJ special issue*.

Kučera, H. & Francis W. N. (1967). *A computational analysis of present-day american English.* Providence, RI: Brown University Press.

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

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In. H. Bejoint & P. Arnaud (eds.) *Vocabulary and applied linguistics.* (pp. 126-132). London: Macmillan.

Mackey, W. (1965). *Language teaching analysis.* London: Longman.

McBride-Chang, C., Tardif, T., Cho, J., Shu, H., Fletcher, P., Stokes, S., Leung, K. (2008). What's in a word? Morphological awareness and vocabulary knowledge in three languages. *Applied Psycholinguistics, 29*(3), 437-462.

McCarthy, M. (1990). *Vocabulary.* Oxford: Oxford University Press.

Mikros, G., Hatzigeorgiu, N. & Carayannis (2005). Basic quantitative characteristics of the modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics, 12***,** 167-184.

Miller, G., Newman, E. & Friedman, E. (1958). Length-frequency statistics for written English. *Information and Control, 1***,** 370-389.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol: Multilingual Matters.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, P. (2006). How large vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63***,** 59-82.

Palmer, H. E. (1917). *The scientific study and teaching of languages.* London: Harrap.

Prunet, J., Beland, R. & Idrissi, A. (2000). The mental representation of Semitic words. *Linguistic Inquiry, 31***,** 609-648.

Rousseau, R. & Zhang, Q. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics, 24***,** 201 - 220.

Sawalha, M. & Atwell, E. (2011). *Accelerating the processing of large corpora: Using grid computing technologies for lemmatizing 176 million words Arabic Internet corpus*. *Advanced Research Computing Open Event.* University of Leeds, Leeds, UK.

Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95***,** 26-43.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In. M. Baroni & S. Bernardini (eds.) *WaCky! Working papers on the Web as corpus.* (pp. 63-98). Gedit, Bologna.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In. C. Fairon, H. Naets, A. Kilgarriff & G.-M. De Schryver (eds.) *Building and exploring Web corpora.* (pp. 83-95). Louvain-la-Neuve: Cahiers du Cental.

Shimron, J. (1999). The role of vowel signs in Hebrew: Beyond word recognition. *Reading and Writing, 11***,** 301-319.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28*, 649-667.