Journal of Applied Language and Culture Studies
Volume 6, Special Issue 3, 2023, pp. 1-19
https://revues.imist.ma/index.php?journal=JALCS

pISSN 2605-7506
e2605-7697

Word Frequency and Lexical Coverage in English and Arabic

Ahmed Ech-Charfi
Faculty of Education, Rabat, Morocco

To cite this article: Ech-Charfi, A. (2023). Word frequency and lexical coverage in English and Arabic. *Journal of Applied Language and Culture Studies, 6*(3), 1-19.

Abstract

This paper addresses the issue of language corpora and words occurrence frequency. The relation between word frequency and word order in a frequency list is reported to be proportional. However, corpora compiled in English and some other languages all attest to the universality of Zipf's law. In Arabic, however, the tools to extract word lists with frequencies are rather less developed and the definition of what a word is in Arabic is still in need of consensus. The paper makes an interesting comparison between coverage in English, as reported in some studies on the basis of some large corpora, and coverage in Arabic on the basis of a small corpus of modern Arabic prose compiled by Landau (1959). The main finding of the paper is that coverage in both languages is relatively similar in that the most frequent 2000 words tend to constitute around 80% of a corpus, leaving only 20% for the remaining words in the frequency list. Focus in this paper is on Arabic and this is due to the fact that there is little research on Arabic.

Keywords: Language corpora, vocabulary, words occurrence frequency.

0. Introduction

The importance of vocabulary in language learning cannot be overestimated, whether the target language is a first, a second or a foreign language. Research over the last three decades in this area has shown that there is a strong correlation between vocabulary knowledge and the four language skills. More specifically, the correlation with listening reaches .61, accounting for around 37% of the variance in the data, and .64 with reading, accounting for 40% of the variance, but hits .70 with writing, accounting for almost 50% of the variance (cf. Schmidt 2010). This fact indicates that knowledge of vocabulary as a construct is closely related to the other components of language proficiency, if not an aspect of those components, and an essential one for that matter. After all, it is hardly imaginable that reading, listening, writing or speaking can ever happen at all without some knowledge of words. What learners really know when they are believed to have learned a word is a complex issue that cannot be discussed here for space constraints (cf. Nation 2013, for taxonomy)? However, it can be safely asserted that the more learners know about words, the more likely they will develop high order strategies like inferring word meaning from context or "reading between the lines".

But not all words are of equal importance in the process of language learning. Researchers as well as learners are well aware that some vocabulary items are more likely to be encountered in reading or listening than others. In speaking and writing as well, acquaintance with some words can be more urgent because of their key function in the articulation of ideas. It is this key role that makes such words more frequently used than others and, consequently, they have higher text coverage; that is, they constitute a high percentage of written or spoken texts. But word frequency is not easily amenable to operationalization. It is precisely this issue that the present paper will deal with. The paper will be articulated as follows. Section 2 will discuss the notion of word in English and Arabic. Section 3 will provide a brief state of the art review of English and Arabic corpus linguistics. Section 4 will compare lexical coverage in the two languages and, finally, Section 5 will summarize the main points tackled in the paper and propose some pedagogical implications and applications.

1. **What is a Word?**

What do we mean when we refer to some form as being a word in a given language? The answer to this question is not straightforward; it all depends on what we want to measure. In this respect, researchers distinguish four units of vocabulary. These are **token**, **type, lemma** and **word family**. These four units of measurement, however, may not be easily adaptable to languages like Arabic, as will be discussed below.

In writing, a token refers to any set of letters separated by spacing, irrespective of its length. Thus, in English the indefinite article "*a*" is as much a word as the noun "*procrastination*". In speaking, the spacing is probably reflected in the relative silence between words, though that silence can only be measured in milliseconds. In addition to being distinct in writing and speaking, any occurrence of a token is counted as a distinct word despite being exactly similar to other occurrences. For example, there are 11 word tokens in the following sentence: "*the man in the car is the dean of the university*", though the definite article "*the*" occurs 4 times. Obviously, as a unit of measurement, the token is not of great use to an estimation of learners' vocabulary knowledge, as it would produce different results each time a text changes. Besides, there is clearly no point in claiming that a learner knows four times the article "*the*", for example! But tokens are useful in measuring the length of a text or the size of a corpus, for instance. Situations in which such measurement is needed do arise, as when students are required to write essays of some specific length or when research submissions are restricted to a particular number of words.

In Arabic, the same definition of a word token can be maintained, though there are clear differences in comparison with English. As is the practice in other languages, Arabic also separates different words by spacing in written texts and by relative silence in spoken discourse. But unlike English words, an Arabic word could stand for a whole sentence, as in "وصلنا" (we arrived). In this example, the word is constituted of a perfect form of the verb to which is attached a pronominal suffix functioning as a subject. In English, the subject can only be realized as a different word. On the other hand, there are cases which would be counted as different words in English but which have affixal equivalents in Arabic and, consequently, are not counted as word

tokens. Such is the case of the definite article "ال" and the preposition "ب" (with), to cite only a couple of instances. Because they are affixed to the following noun, these and similar cases are not treated by a word processor, for example, as distinct words. Such differences would make text length in the two languages incommensurable. Regarding knowledge of vocabulary, a leaner who knows "بالكرة" (with the ball) cannot be said to know exactly the same number words as someone who knows only "كرة" (ball), though the two forms constitute one token. These differences in word structure between the two languages will certainly have bearings on choosing the optimal unit of vocabulary measurement in each of them, as will be argued later on.

Unlike tokens, types treat as different words only forms that are different in some respect, in writing or in pronunciation. So, in a sentence like the one cited above and which included 11 tokens, there are only 7 word types; the four occurrences of the definite article are counted as one type only. Different conjugations of the same verb are also considered as different types when they are marked differently for person, gender, number or tense. For example, the various forms of the verb "*to be*" are all considered different words; viz. *am, is, are, was, were, be, being*. But if two forms do not differ, though used for different morphological categories, they are counted as one word type; e.g. "*speak*" in "*I, you, we, they speak*".

This definition of word types, however, cannot be carried over to Arabic words in a straightforward manner. This is mainly because the Arabic script does not include short vowels. A word form such as "كتب", for example, could be pronounced at least in four different ways, namely "*kataba*" (he wrote), "*kutiba*" (it was written), "*kattaba*" (he wrote intensively) or "*kutub*" (books). To be sure, vowels can be added as diacritics to letters in order to disambiguate written forms, but they are rarely used in normal texts; only the Quran and readers intended for beginners usually include them. Therefore, counting word types in Arabic written texts is likely to be problematic as words need to be pronounced first before any decision could be made regarding their similarity or difference with other forms.

As can be concluded from the above discussion, the word type cannot be used adequately to estimate learners' vocabulary knowledge. It is

certainly more adequate than the word token, especially with beginners who may not be able to associate different but related forms, but it will obviously result in overestimation of the number of words familiar to other categories of learners. Generally, learners with some knowledge of the grammar of a language will be able to infer the meaning of a word form when first encountered, or derive it when needed in speaking or writing, even if they have never heard of it. For instance, someone who knows that the plural form in English can be derived by the suffixation of "*s*" to the singular noun will be able to guess that "*apostate*" and "*apostates*" are the respective singular and plural forms of the same word, though these are very infrequent forms. Therefore, to say that this person knows two words: "*apostate*" and "*apostates*" would be to confuse lexical knowledge and grammar knowledge. Generally, the lexicon is conceived of as consisting only of idiosyncratic features while grammar includes features that can be predicted by some rule and, therefore, can be applied redundantly to all forms satisfying the rule description. With respect to acquisition, grammar rules are thought to reduce the learning load for exactly that reason. It is for that reason as well that grammar and the lexicon are usually considered to form different constructs.

On the basis of what has just been said, researchers propose to group related forms under one word category known as *lemma* (cf. Milton 2009, Schmidt 2010, Nation 2013 among others). A lemma is a head word and its inflected forms. In English, verb inflection includes tense, aspect, person and number while noun inflection includes mainly number and case. Thus, the verb lemma "*work*" includes the forms: *work, works*, *worked* and *working*; the noun lemma "*book*" subsumes the singular form "*work*", the plural "*works*" and the genitive forms "*work's*" and "*works* '". Irregular forms, however, are treated as distinct lemmas precisely because they are idiosyncratic, as explained above. The verb "*to be*", for example, includes the following lemmas: *am, is, are, was, were and been*; similarly, both "*child*" and "*children*" are considered two distinct lemmas.

A similar definition cannot apply easily to Arabic, however. Inflection in Arabic is not as regular as it is in English. Thus, although the so-called strong verbs are fairly regular in their conjugation, the weak verbs are known for their irregularity. The equivalent of the verb "*to*

be", for instance, can have as base forms "كان" or "كن" in the perfect aspect, depending on person, number and gender. In the imperfect aspect as well, its base form can be "كون" or "كن", also depending on the same factors. In this example, the weak consonant is medial; verbs with initial and final weak consonants also exhibit irregularity, but of a different nature. Grammar books provide lengthy discussions of such phenomena, which students often spend considerable time wrestling with. Similarly, plurals in Arabic are categorized into two types: sound and broken plurals. Sound plurals are called so because they bear a plural suffix, as opposed to broken plurals in which number is expressed by vowel change (compare "*ʔamrīkī*" vs. "*ʔamrīkiyūn*" (American(s)) and "*mayribī*" vs. "*mayāribah*" (Moroccan(s)). It would be irrelevant to dwell too long on these and similar issues; suffice it to note that the number of these irregularities is much larger than in English, a fact which should be born in mind when comparing lexical coverage in the two languages.

Relations between words can also be expressed through derivation. Unlike inflection, derivation often results in the change of the grammatical category; and when it does, the result is usually treated as a different lemma. But derived word forms, though characterized by a certain degree of idiosyncrasy, are nonetheless regular to a noticeable extent. Native speakers of English, for example, do realize that "govern", "government", "governmental", "governmentally" and "ungoverned" are related in meaning. Non-native learners also start developing this knowledge at a certain point in time, and when they do, they do not need any longer to be taught what each form with similar affixes mean: if the stem is familiar to them, it takes no more than a simple stroke of imagination to guess what the derived forms mean. Therefore, researchers think that it would be more convenient to include both inflectional and derivational forms of a head word under a single unit of measurement. This unit is the *word family*. Obviously, once this unit is adopted for the estimation of vocabulary knowledge, a learner's vocabulary size would shrink significantly than would be the case if the lemma is adopted.

Derivation in Arabic also shows significant differences in comparison with English derivation. But while English derivational morphology tends to be idiosyncratic, Arabic derivational morphology is fairly

predictable. From the root "*k-t-b*", for example, can be derived the perfect verb form "*katab*", the active participle "*kātib*", the passive participle "*maktūb*", and the noun of place/time "*maktab*"/ "*maktabah*". The nouns "*kitāb*" and "*kuttāb*" seem to be less regular than the other forms. Grammar books provide lengthy chapters to the discussion and illustration of these derivational rules, which may obviously have exceptions. So, if the learning load is the major criterion for distinguishing the lemma from the word family, it would be more consistent to include in the Arabic lemma the regular derivational forms as well. But if grammatical category and meaning are taken into consideration as well, they should be included under the word family. In one respect, this would be the more convenient decision to make, given that some derived forms may have a different meaning from that of the base, and therefore, constitute an additional vocabulary resource for the language user. Take the verb form *"ʔaðhab*" (to make something disappear), for instance; it is derived from "*ðahab*" (to go) by the affixation of *"ʔa-*", a fairly regular process. But the two verbs have different valences: the base is intransitive while the augmented form is transitive. Besides, while the base is frequent and is likely to be familiar even to the beginner, the augmented form is rarely used even by native speakers. Therefore, it would make more sense to treat the two as distinct words.

Of the four units of vocabulary measurement mentioned in this section, each can be used for a different purpose. Tokens, for example, can be used to measure text length, but they are of no significant use for the estimation of learners' vocabulary size. In comparison, types can be useful in measuring the vocabulary size of beginners, but are certainly not appropriate for advanced learners who are already familiar with the grammar of the target language. As to the word family, it is probably more suitable for use with native speakers or with very advanced second language learners, while the lemma is more appropriate with foreign language learners. Before ending this section, it should be pointed out that, while there is a general consensus among researchers on what constitutes a lemma or a word family in English, no such a consensus has been reached yet regarding these categories in Arabic. Consequently, many instruments for researching vocabulary in English are simply lacking for Arabic.

2. Corpora of English and Arabic

Since the advent of the Generative theory of grammar, linguists have become less enthusiastic about using corpora, a method that was widely in practice among the previous generations of structural linguists (cf. Harris 1951). The claim behind this shift away from corpus analysis is that knowledge of language (i.e., competence) is not always reflected in language use (i.e. performance), and that native speakers' intuitions are much more reliable in accessing the underlying grammar (cf. Chomsky 1959). This paradigm shift was probably the main reason behind the relative delay in compiling large corpora of different languages.

As a methodology, corpus linguistics is essentially empirical in nature. Crystal (1980/2008, 117) defines a language corpus as "A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a LANGUAGE". As such, any collection of written or oral texts that can serve as a reference point for linguistic analysis can be qualified as a corpus, irrespective of its size or storage. The collection of poems used by Arab grammarians to describe the grammar of Classical Arabic, for example, can be considered as a corpus of this language. But what characterizes modern corpora is that they are machine readable. Obviously, the advent of computer technology has made it possible to search language corpora in a much faster and easier way, compared to non-computer readable corpora. Besides, tagging words for their part of speech and other syntactic and semantic features can enable researchers to identify much regularity in language that would not be noticeable by intuitions alone.

Today, corpus linguistics has grown into a discipline of its own, as more and more corpora in different languages are being compiled and more software is developed. In English, the first electronic corpora, such as the Brown Corpus, were collected in the early sixties. But these corpora were mostly of small size, e.g. Brown Corpus contained only one million words. Besides, the tools to search them were not much developed, a fact which explains why the first research based on the Brown Corpus was published years after the compilation, viz. Kučera and Francis (1967). Large and representative corpora started to be collected only in the early 1990s. The British National Corpus (BNC),

for example, is composed of 100 million words of written and spoken texts. The texts represent various genres of 20th century British English. The Corpus of Contemporary American English (COCA) is even larger, as its compilers continue to add texts to it on a yearly basis so much so that it exceeds one billion words now representing different genres of modern American English.

Along with the compilation of large electronic corpora, computational linguists are also developing tools to search those corpora. Among the most useful of these tools is the concordancer, which extracts key words in context. A concordancer enables researchers to study the phraseology of a language by investigating the co-occurrence restrictions between various lexical items (i.e. collocations), and between lexical items and grammatical constructions (i.e. collostructions). Calculations can be made on the basis of the data provided by the concordancer to determine the strength of association between different lexical, morphological, syntactic and other linguistic features.

But the tools most relevant to the topic of this paper concern word frequency. In this respect, the lemmatizer is one of the most useful of these tools, as it classifies all the tokens of the inflected forms of a head word into a single lemma and computes its frequency in a corpus. With the help of such software, a word list can be generated from a corpus in which lemmas (or word families) are classified according to their frequency of occurrence. Word lists are extremely useful for pedagogical purposes, as will be explained later; they are also useful for our immediate concerns because lexical coverage can only be computed if word frequency is known. Regarding English, many word lists are available, depending on the corpus from which they have been extracted. This fact is a good indication that frequency is not absolute, but relative to a certain source, i.e. corpus.

In comparison, "Arabic corpus linguistics as a research endeavour is still in its infancy", as stated by McEnry, Hardie and Younis (2019, 1). This is by no means an indication that there are very few corpora of Arabic, because many of them have been compiled by individual researchers or by institutions. Some of these corpora are even of considerable size, as is the case of KACST Arabic Corpus, for instance, which is composed of a billion words of Classical and Modern Arabic

texts compiled by researchers at King Abdualaziz City for Science and Technology in Saudi Arabia. The issue with such corpora is that they tend to be unbalanced. For example, some of them are composed only of newspaper articles (e.g. Arabic Corpus; Watan-2004 Corpus) while others rely exclusively on material extracted from the internet, with the major requirement that it be written in the Arabic script (e.g. Sketch Engine; Leeds Arabic Internet Corpus) (For a review, see Zaghouani 2017). The result is that more weight may have been given to some genres at the expense of others and, consequently, the corpus does not reflect the general use of the language. But the most weakness that Arabic corpus linguistics suffers from is undoubtedly the lack of tools to annotate and search the existing corpora. This lack may be partly due to the unsatisfactory state of research on theoretical issues. For example, a difficulty was pointed out earlier to apply the definition of English lemma and word family to Arabic. Without widely accepted definitions of these key notions, computational linguists cannot embark on designing software such as an Arabic lemmatizer, and without such a lemmatizer, word frequency in the language cannot be computed.

The above comparison between English and Arabic corpus linguistics, though being very brief, indicates that there is a wide gap between the two. Consequently, any comparison of lexical coverage in the two languages can only be rough at this stage.

3. Lexical Coverage in English and Arabic

The objective of comparing coverage in English and Arabic is to see the number of high frequency words required to cover a high percentage of texts in the two languages. It is true that a dictionary of a standard language like English and Arabic include a large number of entries, but many of those entries tend to occur once only; and this happens only in large corpora. These low frequency words have come to be known by the name of *hapax legomena*, which have an extremely low coverage, compared to high frequency words.

Although high frequency words are generally limited in number, they tend to constitute a high percentage of texts. These are usually closed-class items that carry grammatical meaning and are, therefore, indispensable to any well-formed sentence in the language. This fact

seems to be true of all languages, as exemplified by the eight most frequent words in English, French and Arabic in the following table:

Table 1: The most frequent eight words in English, French and Arabic

	Kilgariff (2006)		Baudot (1992)		Backwalter and Parkinson (2011	
1	the	6,187,267	de	68,373	ال	5,004,793
2	be	4,239,632	le	42,419	و	1,110,144
3	of	3,039,444	être	26,897	في	924,823
4	and	2,687,862	un	26,613	من	745,190
5	a	2,186,369	avoir	23,570	ل	584,786
6	in	1,942,315	à	23,475	ب	553,234
7	to	1,620,850	et	23,325	على	518,692
8	have	1,375,636	les	19,230	أن	303,942

In the three corpora representing the three languages, the most frequent words are all grammar words: articles, conjunctions, auxiliaries and prepositions. In the three languages, the list of high frequency words tends to be constituted of dozens of closed-class items, though there may be differences between them in how long the list of such items is. Their high frequency is obviously due to the fact that sentences in language in general cannot be well-formed without the use of grammar items.

Another crucial point about word frequencies in the table relates to their distribution. As can be noticed, the difference between the different ranks is not a matter of a few occurrences, but is rather greater than would be the case if chance was the only factor. For example, while the definite article occurs more than six million times in the English list, the copula occurs only a little more than four million times, with a difference of almost two million occurrences. The same remark can be made in relation to the two most frequent words in the other two lists. In fact, the first item in the Arabic list is about four times more frequent than the second. These remarks indicate clearly that the lower a word's rank is, the higher its frequency will be.

This is exactly what Zipf's Law attempts to capture. This law is not a model based on mathematical assumptions, but rather an empirical one. Zipf (1936, 1949) claims that, in natural language a corpus, the

frequency of a word tends to be inversely proportional to its rank. For example, the first word in the list will be twice as frequent as the second word and four times as the fourth, etc. To a certain extent, the data in the table above supports this prediction, though not as precisely as we would expect from a mathematical equation. In fact, the Arabic list contains data that do not accord fully with the law. In particular, the fifth, the sixth and the seventh ranks only differ minimally in frequency. But despite these and similar recalcitrant data, there seems to be no doubt that the distribution of frequencies in language corpora tend to follow some pattern. This is the reason why researchers have been trying to capture that pattern in more precise ways ever since the publication of Zipf's seminal work
(cf. Piantadosi 2014 for a review).

A major consequence of Zipf's Law is that a small number of high frequency words will constitute a large percentage of natural language texts. The comparison of word frequency and coverage in the Brown Corpus, shown in the following table, illustrates this idea clearly:

Table 2: Frequency and coverage in Brown Corpus (Nation 2013)

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

The figures indicate that the first 100 words in the frequency list account for 49% of the corpus. Most of these words are likely to be closed-class items in English. But as we move downward, more and more words are needed to account for less and less coverage. For example, while the first 1000 words form 74% of the corpus, the second 1000 add only 7% to that figure. Similarly, the same number of words accounts only for 1% of the difference between 4000 and 5000 words,

and words' contribution to coverage gets less and less so much so that the difference between 99% and 100% coverage requires the addition of 43000 words. Most of these words are likely to be *hapax legomena*. The calculation of coverage in other English corpora may be slightly different from the figures in Table 2, but they are unlikely to diverge radically from them.

Unlike the case of English, lexical coverage in Arabic has not been studied at any significant depth. Perhaps the only study carried so far is Masrai and Milton (2016). Even if no adequate definition of the units of vocabulary measurement has been widely accepted and used in the literature and, consequently, no frequency list is currently available, the researchers used a preliminary list extracted from the Leeds Corpus of Internet Arabic and lemmatized by Sawalha and Atwell (2011), which is available at: http://corpus.leeds.ac.uk/query-ar.html. The list is composed of 100,000 items, but Masrai and Milton (2016) limited their calculations to the first 20,000 items. Their findings are summarized in the following table:

Table 3: Lexical coverage in Arabic according to Masrai and Milton (2011)

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

As can be noticed, the first band including 1000 words accounts for 66% of the corpus with 8% less than the coverage of the first band in English, as shown in Table 2 above. The second band adds another 10%, thus resulting in 76% coverage, while 5000 words account for a total of 89% of the corpus. To reach 98% coverage requires 14,000 words, a number that is away less than what would be needed for a similar coverage in English. This quick comparison would lead to the

conclusion that fewer words are needed for written or oral expression in Arabic than in English.

But there is more than a reason to doubt this conclusion. To begin with, the Leeds Corpus of Internet Arabic includes a lot of words of non-Arabic origin, a fact which can easily be proven by querying some of these words in the website mentioned above. For example, if the word "اث" occurring in the 3822nd rank is queried in the concordancer, most of the results of the query will turn out to be non-Arabic, though written in Arabic script. As explained by Sharoff (2006), internet corpora are harvested from open access webpages at a first stage, and are later on sifted to eliminate, among other things, texts that are not in the target language. It seems, however, that undesired texts do manage to slip into the final corpus against the compilers' wishes. Second, the Leeds corpus is composed not only of Standard Arabic texts, but also of colloquial texts pertaining to various regional dialects. To give an example, the verb "*šuf*" (to see) ranks 2134th in the order of frequency, though it is undoubtedly of colloquial nature. In fact, its frequency could become higher if other inflected forms of the verb are included. This remark paves the way for the third and major weakness of the list used by Masrai and Milton (2016), which has to do with the notion of lemma assumed there. As a matter fact, the list does not seem to use any recognizable definition of the lemma, though the website claims that it is a list of lemmatized words. As has just been hinted at, many forms of the verb "*šuf*" are listed as distinct words, including "*šufit*", "*šūfū*", "*šūfī*", among others. The same remark holds for most of the items in the list. In fact, there are many instances in which the same form is listed many times. In brief, a close scrutiny indicates that the list is basically a list of word types and, even as such, it is still in need of trimming.

In comparison, Backwalter and Parkinson (2011) is a much more systematic list. The list was extracted from a thirty million word corpus including both written texts and transcriptions of oral conversations. The problem with the list, however, is that it includes only 5000 items among the most frequent words in the corpus. Therefore, coverage beyond this level will remain unknown. Besides, their conception of the Arabic lemma may be controversial in some respects. For instance, the definite article "ال" is considered a separate lemma, as suggested in Table 1 above, a decision which will certainly affect the overall

coverage because of the article's extremely high frequency. But the most controversial aspect of the corpus and the list extracted from it is probably the inclusion of colloquial conversations. The issue as to whether Standard and colloquial Arabic form one language or two is still debatable, but native speakers seem to regard the standard as a separate variety which should be kept unaffected by the colloquial. In any case, the relation between the two varieties is not in any way near the relation between standard and non-standard varieties in non-diglossic situations.

For the purposes of this study, and in order to overcome the weaknesses of the existing frequency lists, a choice has been made to use the list in Landau (1959). Though this list is undoubtedly outdated and the corpus from which it was extracted rather small (272,178 tokens), it includes a set of 12,400 words of Modern Standard Arabic words and expressions only. The corpus was also various, with texts from newspapers and six other genres. Half of the corpus, however, was in the form of newspaper articles, a fact which makes it rather unbalanced. Since the list is published in book format, the first step was to convert it into an Excel spreadsheet in order to compute the coverage of items individually and cumulatively[1]. Some changes have also been introduced to maintain a consistent definition of the lemma adopted by Landau himself. In particular, the list included expressions of more than one word because Landau intended it to be used for pedagogical purposes, as explained in the book preface. So, those expressions were broken down into their constituent words, and the words' frequency added to their corresponding lemmas. If no corresponding lemma is already available, a new lemma is added to the list. When all this had been done, the result was a list of 12,011 items, which were re-ranked on the basis of their frequency.

A note about the notion of lemma adopted in the new list is in order. Landau (1959) took the decision to ignore all affixes and list only the base. Among the affixes excluded are affixal pronouns and the definite article. The conjunctions "*wa*" and "*fa*" were also discarded, though the reason behind that is not clear; but affixal prepositions were considered

[1] Special thanks go Intissar Louah, Nihal Bouabida and Omar Benjilali for their help in this time-consuming and rather tedious chore.

distinct lemmas. By opposition, some nouns in the accusative case were treated separately basically because they tended to be used as such in special pragmatic contexts. Examples of such nouns include "*tʿabʕan*" (of course) and "*šukran*" (thanks). For the purposes of consistency, however, the frequencies of these words were added to the corresponding bases. Some other minor changes were also introduced the details of which need not be of much concern to the reader.

With a little more than 12,000 lemmas, lexical coverage in Landau's corpus could be representative of coverage in Arabic. The table below shows the cumulative percentage of sets of items in parallel to those in Table 2 above so that comparison between the two languages can be made:

Table 4: Frequency and coverage in Landau's corpus

Coverage (%)	Number of words
19,45	10
38,62	100
70,96	1,000
82,62	2,000
88,7	3,000
92,33	4,000
94,71	5,000
96,33	6,000
97,74	7,000
98,25	8,000
98,89	9,000
99,26	10,000
100	12,011

As can be noticed, the first band accounts for almost 71% of the corpus, with three points less than what the same number of words cover in Brown Corpus. But the second band increases coverage to more than 82%, thus exceeding what is recorded for the English corpus. With 5000 words, more than 96% of the Arabic corpus would be familiar to the reader while the same amount of vocabulary makes only 89% of the English corpus accessible. Similarly, 12,000 words accounts for almost 100% of the Arabic corpus but only 95% of the English corpus. Obviously, some of the differences between these percentages are due

to the number of lemmas extracted from the two corpora. The list extracted from Landau's corpus is composed of around 12,000 lemmas only whereas that extracted from Brow Corpus includes 87,000 lemmas. Therefore, coverage in the two languages would seem incommensurable through these lists.

However, there could be more to the difference than just the number of lemmas. For one thing, many grammatical meanings are expressed by independent words in English but by affixes only in Arabic. Examples of such words are the definite and the indefinite articles, as was explained earlier. Since these are highly frequent in English, they are likely to affect coverage in a significant way. In addition, pronouns appear as independent words in English but in Arabic, they are most of the time attached to the verb and, consequently, are considered to form distinct words. To limit the discussion to these couple of remarks, it is only natural that Arabic uses fewer words than English, according to the definition of word adopted here. Thus, an English sentence like "*the man has finished his lunch*" is composed of six words while its Arabic equivalent "أنهى الرجل غذاءه" includes only half that number.

That being said, a deeper understanding of lexical coverage in the two languages can only be achieved when corpora of similar size are available. It might be the case that such a comparison will require the use of vocabulary units other than lemma or word family.

4. Conclusion: Some Pedagogical A/Implications

The above discussion can have a number of pedagogical implications and applications. The measurement of vocabulary knowledge has become an established practice in foreign language testing, using word frequency as a major factor in designing vocabulary tests. In research as well, the relation between vocabulary size and other aspects of language proficiency has witnessed an increasing interest on the part of researchers working on English and a few other languages. The problem, however, is that this kind of research has not been extended to languages like Arabic yet, and that could be due mainly to the lack of adequate tools to explore the existing corpora and extract frequency lists, collocations and collostructions, among other things. Therefore, there is an extreme urgency about the development of Arabic corpus linguistics not only because Arabic is being learned as a foreign

language by thousands of students around the world, but also because it functions as a second language of some sort to its speakers in Arab countries. It is no surprise that educators and other stake-holders continue to complain about its poor mastery by learners in these countries even after many years of instruction. As researchers, our immediate objective should be to observe closely its acquisition at different stages and suggest better ways of its instruction accordingly. One way to do that is to focus on the acquisition of its vocabulary and the current practices of teaching it, in addition to sharing up-to-date knowledge with the community of Arabic teachers. This paper has shown that some words are highly frequent in the language and, on the account; learners' attention should be directed to them in order to speed up the development of the four major skills. Research has demonstrated, for example, that without knowledge of the 2000 most frequent lemmas in English, very little progress can be achieved in reading (Milton 2009, Schmidt 2010). Therefore, any means possible should be deployed to teach those items.

References

Buckwalter, T., & D. Parkinson (2011). *A frequency dictionary of Arabic*: *Core vocabulary for learners*. London/New York: Routledge

Chomsky, N. (1959). A review of Skinner's verbal behavior. *Language,* 35(1), 26-58.

Crystal, D. (1980/2008). *A dictionary of linguistics and phonetics*. Blackwell Publishing

Harris, Z. (1951). *Methods in structural linguistics*. Chicago: Chicago University Press

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

Landau, Y. (1959). *A word count of modern Arabic prose*. New York: American Council of Learned Societies.

Masrai, A., & Milton, J. (2016). How different is Arabic from other languages? The relationship between word frequency and lexical coverage. *Journal of Applied Linguistics and Language Research, 3*(1), 15-35.

McEnery, T., A. Hardie, N. Younis (Eds.). (2019). *Arabic corpus linguistics*. Edinburgh: Edinburgh University Press

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

Piantadosi, S.T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21, 1112-1130

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

Schmidt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave: Macmillan

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.

Zaghouani, W. (2017). Critical survey of freely available Arabic corpora. *Computer Science. https://arxiv.org/ftp/arxiv/papers/1702/1702.07835.pdf.*

Zipf, G. (1936). *The psychobiology of language*. London: Routledge.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.