Analysis of Primary School Arabic Language Textbooks

B. Belkhouche¹, H. Harmain¹, H. Al Taha², L. Al Najjar², S. Tibi³
(1) Faculty of Information Technology (2) Faculty of Humanities (3) Faculty of Education
U.A.E. University, Al-Ain, U.A.E.

{ b.belkhouche,harmain, Huda.salem,lalnajja, tibi}@uaeu.ac.ae

Abstract

This paper reports on a preliminary analysis of a corpus consisting of Arabic language textbooks used in primary schools. The input to our process is a raw text extracted from Arabic textbooks used by the Emirates curriculum of grades 1 through 6. Various aspects, including word and root frequencies, parts-of-speech distribution, phonology, and themes, are investigated. A comparison among parts of speech, UAE grade 1 and Libya grade 1 is performed. Our analysis raises several issues on the criteria for selecting a word list that is appropriate for enriching the vocabulary of school children.

1. Introduction

Language learning has been greatly facilitated by the availability of electronic resources. In particular, electronic dictionaries and concordances have been instrumental in promoting language learning and vocabulary building [5]. However, electronic resources present some methodological challenges, not only for Arabic but even for languages that have been investigated for decades [10]. For example, the Japanese Electronic Dictionary [9] took over eight years to build an initial version and involved several organizations. Trends revolving around the building of electronic resources for the Arabic language fall more into the electronic duplication of hard copies. Even though a great deal of research and implementations has been done in many Indo-European languages, most of the results of such endeavors cannot be easily applied to Arabic. This can be attributed to the differences in language structures and the complexity of the Arabic morphology. However, some of the proposed work and knowledge representation schemes (e.g. FrameNet, WordNet, Treebank) provide an abstract conceptual basis that is language-independent [8], and thus, can be used effectively in the Arabic language context.

Our primary concern in this paper is to shed some light on the issues of building an Arabic eLearning dictionary with the main goal to sketch a methodology for building Arabic Learning resources. Such a task presents numerous and unique challenges, mainly because of the complex morphology of the Arabic language and the appropriateness of the vocabulary. Current Arabic printed dictionaries do not lend themselves to a straightforward categorization. Users of such dictionaries face great difficulties searching for words, because they are expected to master a substantial set of morphological derivation rules. For the young learner, this may be not just a challenging requirement, but a deterrent.

We can argue that the need for an Arabic electronic dictionary has become more urgent than ever before. The available Arabic electronic dictionaries are no more than electronic version of the printed dictionaries with the only advantage of providing an interface to search for words quickly. These dictionaries do not help in searching for inflectional words that do not need to be in the dictionary as separate entries (e.g. if you search for Taktub (كتب) and the dictionary entry is Yaktub (كتب), your search will yield no results). This is just one example of shortcoming in the Arabic electronic readable dictionaries. The Arabic electronic learners’ dictionary we propose should at least satisfy the following requirements:

1. Target primary grade school students.
2. Immerse the children into an interactive and exploratory mode.
3. Accommodate the development of word meanings to satisfy the current language usage (take for example the meaning for مسارة).
4. Provide more grammatical information that will enrich the dictionary and help the learners develop their linguistic skills, by including information such as cases, dual, plural, gender, and broken plurals.
5. Display the word meaning in an age-appropriate way and include relevant examples with illustrations.
6. Capture the relations of the word with the root and its derivatives.
7. Contain a morphological module to help search for words in their plain format and return all possible forms of a word.
8. Support cross-referencing and navigation.

The rest of the paper is organized as follows: Section 2 discusses our concerns about methodological issues. Section 3 describes our text processing tools. Section 4 gives details of our corpus. Results are discussed in Section 5 and conclusions are given in Section 6.
2. Methodological Issues

The use of grade 1-6 language arts textbooks as a corpus to identify the vocabulary for an Arabic e-dictionary raises several methodological issues. The first difficulty of pragmatic import is the availability of these resources in electronic form. A fundamental question is whether these textbooks can be considered as “canonical”, that is, do they offer a rich and representative enough corpus from which to identify our vocabulary word list? This has implications on the methodology used to develop textbooks, if indeed a methodology exists within and among Arab countries. Moreover, as Tables 1 and 2 reveal, these textbooks cannot be analyzed as a homogeneous group as the data they contain do not show similar patterns in terms of sizes, frequencies, and word similarities.

The traditional approach for building word lists relies on large corpora. The selection of words is then based on their frequencies. A major assumption of this approach is that these corpora are highly representative of the word usage. For many reasons, and among them diglossia, this assumption does not hold in the Arab world context. Moreover, recent research has shown that raw word frequencies may not be good measures for building word lists [1, 3]. Thus, rather than analyzing the corpora in a quantitative fashion, our approach is to concentrate on a comparative analysis of the distributions of some general linguistic principles, such as parts-of-speech, themes, and sounds.

3. Text Processing Tool

Figure 1 shows a snapshot from AlKhaleel, a corpus processing tool specifically designed to analyze Arabic texts [6]. AlKhaleel accepts texts in different formats and can be used to perform simple text analysis such as calculating word frequencies and concordances, or to perform more sophisticated analysis using Natural Language Processing techniques for morphological analysis.

AlKhaleel is a modular system designed to integrate several language analysis tools to support the study and analysis of Arabic Language. AlKhaleel is a menu-driven Windows application implemented in Visual C#. The main modules of the system are discussed in this section.

3.1 Corpus Loader

The Corpus Loader is used to load the text document(s) into the system. AlKhaleel accepts texts in different formats including HTML, XML, plain text, and MS Word documents. The corpus can be stored in one or multiple files. Two text encodings currently supported for Arabic texts are UTF-8 and MS Windows CP-1256. These are the two main and widely used Arabic text encodings. The Corpus Loader extracts Arabic words from the loaded texts and builds an internal data structure. The data structure consists of a unique word index, the word in its surface form, the word without short vowels, and the file name and line number in which the word appeared. Punctuation marks, HTML, and XML tags are ignored.

3.2 Word Frequency Analyzer

The Word Frequency Analyzer reads the data structure built by the Corpus Loader and builds a word frequency list. This list consists of the word, its frequency number, and a set of pointers to all its occurrences in the text. The word frequency list is stored in a data store and displayed to the user in a grid format as shown in Fig. 1. The frequency list can be sorted alphabetically or based on the word frequencies.

3.3 Concordancer

We use a simple Keyword in Context (KWIC) technique to build concordances. The concordances are displayed right-to-left and can be exported to MS Excel file. The context of the keywords is given in its original format with all diacritics and punctuation marks included. The user can display a larger context of the keyword with several lines before and after it.

3.4 Dictionary Lookup

AlKhaleel includes an Arabic electronic dictionary. This is our own implementation of Al-Waseet Dictionary [9]. We have implemented a simple look-up algorithm to search to the word in the dictionary. First the given word is searched in the keyword entries of the dictionary. If it is not found we search for it in the sub-entry forms. If it is not found then we perform some morphological analysis by deleting the prefixes and/or suffixes if any and search again.

3.5 Morphological Analyzer

The Morphological Analyzer is a re-implementation of
Buckwalter’s morphological analyzer [10]. We used the same data dictionaries given in Buckwalter Analyzer v.1.0 and re-implemented the algorithms in C#. The results of analysis are presented in a table which includes the English translation of the word (see Fig. 4). This can be of great help to the Arabic second language learners.

### 3.6 Output Generator

To better support its users (teachers, learners, and researchers) AlKhaleel allows for all of its results to be exported to MS Excel files. These Excel files can be used for printing handouts or for performing extra statistical analysis.

For the purpose of this paper we used the Word Frequency Analyzer, Morphological Analyzer, and the Output Generator.

### 4. The Corpus

When selecting a corpus for use in any linguistic task, two main issues must be taken into consideration: the corpus size and its representativeness. Unfortunately, for the Arabic language there are no available corpora like the British National Corpus or the Bank of English [2]; therefore we must rely on other sources to build a new corpus.

### 5. Analysis of Results

The first major issue to confront is finding an appropriate methodology for building an Arabic dictionary for grammar school children. Studies in other languages may give some hints, but their assumptions are not transferrable to the Arabic context. Fundamental studies related to Arabic dictionaries for children are unknown to us. The second major issue is the notion of fundamental vocabulary and its extent. Determining reliable sources and filters to extract such a vocabulary is not only complex, but it may not have a positive answer. Thus, defining a suitable word list becomes a challenging task.

The intent of the analysis is to quantify and categorize the various aspects of the use of the vocabulary in grammar school in order to build a dictionary that responds to the needs of the pupils. Based on the available material from public schools in the UAE, we performed various statistical analyses of which we present an overview related to data in grade 1. Our analysis includes: (1) parts-of-speech distribution; (2) verb frequency; (3) theme frequency; and (4) sound frequency.

#### 5.1 Parts-of-Speech Distribution

Major grammatical categories (parts of speech, POS) were identified. Figure 2 shows the distribution of the different parts of speech categorized as noun, verb, particle, and number. It shows a substantial use of nouns, whereas verbs and particles constitute half of the usage of words. The use of adjectives seems to be very limited. The number category may not be relevant, since the textbook deal with language arts. In the absence of sound studies on vocabulary requirements, or at least a suitable vocabulary, it is difficult to provide an interpretation of this distribution.
5.2 Verb and Verb Root Frequencies

Since roots play a major role in Arabic, we present here frequencies of verbs in word and root forms. Figure 3 summarizes the frequency of verbs within the corpus. It indicates that instances of verbs do not appear repeatedly. Indeed, it shows that 77% of the verbs are used only once and 14% used only twice.

Reducing the verbs to their roots (Figure 4) results in an increase of instance repetition. Still, the highest frequency (8 repetitions) constitutes only 22% of the roots. In any case, one would question the impact of such low frequencies on vocabulary development. A raw frequency of 100 is suggested in [11].

Figure 5 shows the distribution of sounds within verbs. An issue that we were trying to elucidate is whether difficult sounds (e.g., letters such as $\check{\imath}$) would imply their infrequent use. The analysis according to the major categories does not support this assumption. A further analysis of the letter frequencies confirms the distribution in the Figure 5.

That is, there is no pattern differentiating easy and difficult sounds. Figure 6 shows the distribution of verbs according to themes as defined in Roget’s thesaurus. Over 90% of the verbs are action verbs. Even though excessively imbalanced, the bias towards action verbs seems appropriate for the age group under consideration.

To further our analysis, we began an investigation of grades 1 to 6 textbooks from Libya. For comparison
purposes, we followed the same pattern used for the UAE textbooks. Figure 7 shows the POS distribution.

Figure 7 POS Distribution (Libya)

We also compared the POS distribution in an English dictionary for children (1000 words), the UAE grade 1 language textbooks, and the Libyan grade 1 textbook. Figure 8 summarizes our comparison.

Figure 8 POS Comparison

6. Conclusion

Our results shed some light on the data and the tasks needed to develop a grammar school dictionary. The use of grade 1-6 language arts textbooks as a corpus to identify the vocabulary for an Arabic e-dictionary raised several methodological issues. A fundamental question is whether these textbooks can be considered as “canonical”, that is, do they offer a rich and representative enough corpus from which to identify our vocabulary word list? This has implications on the methodology used to develop textbooks, if indeed, a methodology exists. Noting that a grammar school dictionary is not just another general dictionary, an underlying question is to elaborate the distributions of parts-of-speech, themes, and sounds. The methodology used in [7] provides some solid guidelines on how to extend our work.

References


