Towards Analyzing the International Corpus of Arabic (ICA):
Progress of Morphological Stage

Abstract:

This paper sheds light on four axes. The first axis deals with the levels of corpus analysis e.g. morphological analysis, lexical analysis, syntactic analysis and semantic analysis. The second axis captures some attempts of Arabic corpora analysis. The third axis demonstrates different available tools for Arabic morphological analysis (Xerox, Tim Buckwalter, Sakhr and RDI). The fourth axis is the basic section in the paper; it deals with the morphological analysis of ICA. It includes: selecting and describing the model of analysis, pre-analysis stage and full text analysis stages.

1. Introduction:

It can be said that corpus analysis highly depends on the availability of previous history of the analysis, because information with decisive solutions in one stage, are used in the next stages of the analysis. The major difference between creating and analyzing a corpus is that while the creator of a corpus has the option of adjusting what is included in the corpus to compensate for any complications that arise during the creation of the corpus, the corpus analyst is confronted with a fixed corpus, and has to decide whether to continue with the analysis, even if the corpus is not entirely suitable for analysis, or find a new corpus altogether (Meyer, 2002).

It is important, first of all, to begin the process with a very clear goal in mind; that the analysis should involve more than a simple (count) of linguistic features. Also, it is necessary to select the appropriate corpus for analysis: to make sure, for instance, that it contains the right types of texts for the analysis and that the samples to be examined are lengthy enough. Also, if more than one corpus is to be compared, the corpora must be comparable, or else the analysis will not be valid. After these preparations are made, the analyst must find the appropriate software tools to conduct the study, code the results, and finally subject these results to the appropriate statistical tests. If all of these steps are followed, the analyst can rest assured that the results obtained are valid and the generalizations that are made have a solid linguistic basis (Meyer, 2002).
2. Levels of corpus analysis:

Linguistic analysis has more than one level of analysis such as morphological analysis, lexical analysis, syntactic analysis (parsing) and semantic analysis. The focus of corpus analysis is empirical, whereas the interpretation can be either qualitative or quantitative.

**Morphological analysis** is the most basic type of linguistic corpus analysis because it forms the essential foundation for further types of analysis (such as syntactic parsing and semantic field annotation), and because it is a task that can be carried out with a high degree of accuracy by a computer. The aim of morphological analysis of corpora is not only to assign to each lexical unit in the text a code indicating its part of speech, but also to indicate other morphological information. There are many morphological dimensions for describing verbs, nouns and particles. Consequently, the morphological tag can either be extended to include all morphological features (including additional features such as transitivity, perfectness and voice for verbs, number, gender and derivation for nouns and agglutination for particles), or contracted to include only the main morphological tags and other morphological features are indicated separately (see Al-Sulaiti & Atwell, 2001).

There are two main approaches in morphological generation and analysis; namely, the Two-level approach (Non-concatenative approach) and the Concatenative approach. The two-level approach defines two levels of strings; lexical strings which represent morphemes, and surface strings which represent surface forms.

The two-level approach views the Arabic word vertically, as a composition of two layers; root and pattern. In Arabic, for instance, there is a clear sense that the forms in table 1 are morphologically related to one another, although they do not share isolable strings of segments in concatenated morphemes:

<table>
<thead>
<tr>
<th>Word</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>(kataba)</td>
<td>كتب He write</td>
</tr>
<tr>
<td>(makotuwb)</td>
<td>مكتوب Written</td>
</tr>
<tr>
<td>(kutub)</td>
<td>كتب Books</td>
</tr>
<tr>
<td>(kutiba)</td>
<td>كتب Be written</td>
</tr>
<tr>
<td>(kitab)</td>
<td>كتاب Book</td>
</tr>
<tr>
<td>(kut-Ab)</td>
<td>كتاب Writers/Quran school</td>
</tr>
<tr>
<td>(kAtib)</td>
<td>كاتب Writer</td>
</tr>
</tbody>
</table>

Table 1: variant words related to each other.

The Concatenative morphology, which appears almost exclusively in the more familiar languages, involves prefixation or suffixation only. In other words, morphemes are discrete elements linearly concatenated at the right or the left end of the base of the morphological operation (Hockett, 1947). Although the concatenative approach cannot predict the word-pattern automatically, it compensates for this by keeping a large database of Arabic lexemes with their related information including word-patterns.
Hence, the input word passes through less complicated processing than in the two-level approach.

**Lexical analysis** is the process of taking an input string of characters and producing a sequence of symbols called "lexical tokens", which may be handled easily by lexical analyzers (parsers, programs of lexical analysis). These analyzers have two phases of analysis; i.e. the scanning phase and tokenization phase, the process of determining and classifying a clause into tokens.

In **Syntactic analysis** the linear sequence of tokens is replaced by a tree structure through building a parse tree in order to define the language’s syntax according to the rules of formal grammar, and generate, or transform the parse tree. Parsing is also crucial in various applications in natural language processing, including text-to-speech synthesis, and machine translation (Patten, 1992).

**Semantic analysis** is one of the most important levels of analysis. In this level, the semantic information is added into the parse tree, the symbol table is built, and finally semantic checks are performed. Logically, semantic analysis intermediates the parsing phase and the code generation phase because it requires a complete parse tree. In machine learning, the semantic analysis of a corpus is the task of building structures that capture concepts from a large set of documents. It does not generally involve prior semantic understanding of the documents.

### 3. Some attempts of Arabic corpora analysis:

**CLARA (Corpus Linguae Arabicae):** The ultimate goal of this project is building a balanced and annotated corpus. The annotation should be done for morphological boundaries and Part Of Speech (POS). Some tools and databases are built for the sake of the analysis stage; for instance, a training corpus with marked morphological boundaries consisting of 100,000 words, a database of strings with marked morphological boundaries and another training corpus with annotation of parts of speech. Currently, the analyzed size of this corpus is about 15,000 words. The parts of speech tagset is based on the EAGLES recommendations.

**The Penn Arabic Treebank:** is a corpus of one million words of Arabic. Treebank is designed to support the development of data-driven approaches to natural language processing (NLP), human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research on Modern Standard Arabic (MSA) in general. There are two distinct phases of analysis in the Penn Arabic Treebank; namely, Part-of-Speech (POS) tagging, and Arabic Treebanking (ArabicTB) (Abdelali, 2004).

**Prague Arabic Dependency Treebank:** is a project of analyzing large amounts of linguistic data in Modern Written Arabic in terms of the formal representation of language that originates in the Functional Generative Description (Šgall et al. 1986, Šgall & Hajičůvá 2003). Prague Arabic Dependency Treebank (PADT) does not only

---

1 [http://www.ilc.pi.cnr.it/](http://www.ilc.pi.cnr.it/)
consist of multi-level linguistic annotations of the Modern Standard Arabic, but it even has a variety of unique software implementations, designed for general use in Natural Language Processing (NLP).

The linguistic analysis takes place in three stages: the morphological level (inflection of lexemes), the analytical level (surface syntax), and the tectogrammatical level (underlying syntax) (Smrž, 2004). The morphological level of PADT has for long been the same as that available in Penn Arabic Treebank, Part 2. However, PADT has adopted the way of Buckwalter Arabic Morphological Analyzer.

4. Existing Arabic Morphological analyzers:

There are many morphological analyzers for Arabic, some of them are available for research and evaluation while the others are proprietary commercial applications. Among those known in the literature are Xerox Arabic Morphological Analysis and Generation (Beesley, 1998a,2001), Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002), Sakhr and RDI Arabic Morphological Analyzer.

**Xerox Morphology:** is “based on solid and innovative finite-state technology” (Dichy & Fargaly, 2003). It adopts the root-and-pattern approach and includes 4,930 roots and 400 patterns, effectively generating 90,000 stems. Its main advantage is that it is rule based with wide coverage. It also reconstructs vowel marks and provides an English glossary for each word. At Xerox, the treatment of Arabic starts with a lexic grammar where prefixes and suffixes concatenate to stems in the usual way, and where stems are, similarly, represented as a concatenation of a root and a pattern (Beesley, 1998a & b).

The system includes more classical entries, and lacks more grammar-lexis specifications. Additional disadvantages of Xerox morphology are:

1. Overgeneration in word derivation. The distribution of patterns for roots is not even, and although each root was hand-coded in the system to select from among the 400 patterns, the task is understandably tedious and prone to mistakes as shown in table 2.

<table>
<thead>
<tr>
<th>Word</th>
<th>Transliteration</th>
<th>Root</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>قال</td>
<td>qaal</td>
<td>qwl</td>
<td>Say (verb)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>qlw</td>
<td>Fry (active participle)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>qll</td>
<td>decrease (active participle)</td>
</tr>
</tbody>
</table>

Table 2: Example of over generation.

The first root analysis is valid, while the other two are illegal derivations that have no place in the Arabic language, and not mentioned in classical dictionaries.

2. Underspecification: in POS classification, which makes it unsuited for serving a syntactic parser. Words are only classified into: (verbs, nouns which include adjectives and adverbs, participles and function words which, in turn, include prepositions, conjunctions, subordinating conjunctions, articles, negative particles…etc).
3. Increased rate of ambiguity: due to the above-mentioned factors, the system suffers from a very high level of ambiguity, as it provides so many analyses (many of them spurious) for most words (Attia, 2006).

**Buckwalter Arabic Morphological Analyzer:** It uses a concatenative lexicon-driven approach where morphotactics and orthographic rules are built directly into the lexicon itself instead of being specified in terms of general rules that interact to realize the output (Buckwalter, 2002). Buckwalter Morphology contains of 38,600 lemmas, and is used in LDC Arabic POS-tagger, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms and it interacts with other concatenation databases. Every word form is entered separately, Buckwalter’s morphology reconstructs vowel marks and provides English glossary. It takes the stem as the base form and root information is provided (Attia, 2000). In Buckwalter analyzer, Arabic words are segmented into prefix, stem and suffix strings according to the following rules:

- the prefix can be 0 to 4 characters long.
- the stem can be 1 to infinite characters long.
- the suffix can be 0 to 6 characters long.

**Sakhr Arabic Morphological Processor:** It is a morphological analyzer-synthesizer that provides basic analysis for a single Arabic word, covering the whole range of modern and classical Arabic. The analyzer identifies all possible stem forms of a word; i.e. extracting its basic form stripped from the affixes, the morphological data such as root, the Morphological Pattern (MP), and its part of speech. The synthesizer works in a reverse mode to regenerate the word from its morphological forms (stem, root, morphological pattern, part of speech and/or affixes). Sakhr has designed the Morphological Processor to produce word level analysis through regeneration and comparison.

In Sakhr morphological processor each regular derivative root is allowed to be combined with a selected set of forms or patterns to produce words that can be found in standard Arabic dictionaries. Sakhr did not publish any technical documents about its Arabic morphological analyzer; no one knows how its model of Arabic morphology looks like. (Attia, 2000).

**RDI Arabic Morphological Analyzer:** The main RDI’s NLP core engine is the basis of Arabic morphological analysis, Arabic POS tagging, and Arabic Lexical Semantic Analysis. ArabMorpho is a morpheme-based lexical analyzer/synthesizer which distinguishes it from its vocabulary-based rivals and boosts its flexibility. After morphological rules are exhausted, deep-horizon dynamic statistical analysis is employed to realize disambiguation; hence, word accuracy can reach up to 96%. In RDI analyzer each regular derivative root is allowed to combine freely with any form.

---

2 http://www.ldc.upenn.edu/Catalog/docs/LDC2004L02/readme.txt
4 http://www.rdi-eg.com/rdi/technologies/arabic_nlp.htm
as long as this combination is morphologically allowed. This allows the system to deal with all the possible Arabic words and eradicates the need to be tied to a fixed vocabulary (Attia, 2000).

5. The International Corpus of Arabic (ICA) “Analysis stage”:

Alansary et al. (2007) surveyed the compilation of ICA, its design and the preliminary software used in interrogating the compiled corpus. This attempt can be considered one of the most successful approaches for building a representative corpus for MSA. It is important to realize that the creation of ICA is a "cyclical" process, requiring constant re-evaluation as the corpus is being compiled. Once the process of collecting and computerizing texts is completed, texts will be ready for the final stage of preparation; mark up, from there, it is easy to deal with texts in the analysis stage.

The process of analyzing a corpus is in many respects similar to the process of creating a corpus. Like the compiler, the corpus analyst needs to consider some factors such as: whether the corpus to be analyzed is lengthy enough for the particular linguistic study being undertaken and whether the samples in the corpus are balanced and representative (Meyer, 2002).

This section is devoted to describing the process of analyzing the ICA corpus. It will focus on selecting and describing the model of analysis, pre-analysis stage (data processing), full text analysis stages, adding root information and current state of ICA.

5.1 Selecting and describing the model of analysis:

According to our adopted model in the morphological analysis, the word is viewed as composed of a basic unit that can be combined with morphemes governed by morphotactic rules. Therefore, the stem-based approach (concatenative approach) is adopted as a linguistic approach to analyze the ICA. According to this linguistic approach, it was expected that a feature based on the right and left stems would lead to improvement in system accuracy. The Arabic Morphology module uses a simple approach of dividing the Arabic word into three parts:

Prefix: consist of as many as three concatenated prefixes, or could be null.
Stem: it is composed of root and pattern morphemes.
Suffix: consist of as many as two concatenated suffixes, or could be null.

The three-part approach entails the use of three lexicons: Prefixes lexicon, Stem lexicon, and Suffixes lexicon. For a word to be analyzed, its parts must have an entry in each lexicon, assuming that a null prefix or a null suffix are both possible. Table 3 shows example of valid word forms:

---

Table 3: valid word forms.

Not every Prefix-Stem-Suffix combination is necessarily a valid or a legal word. To confirm that the Prefix-Stem-Suffix composition is a valid Arabic word, morphological categories are assigned to each entry in the lexicons.

When trying to select the morphological analyzer system to be used in analyzing the ICA, Buckwalter morphological analyzer has been selected to analyze the ICA as it was found to be the most suitable lexical resource to our approach.

The Buckwalter’s morphological analyzer has many advantages such as its ability to provide a lot of information like Lemma, Vocalization, Part of Speech (POS) and Gloss. Also, Buckwalter is capable of supplying other information such as prefix(s), stem, word class, suffix(s), number, gender, definiteness and case. The output of Buckwalter appears in XML format.

A single word may belong to more than one word class. For example the word “كتب” appears in Buckwalter output as noun or verb as shown in figure 1:

```
<variant>
  <kitb>
    <solution>
      <lemmaId=kotab-u_1</lemmaId>
      <pos>kotab</pos>
      <gloss>he, it verb</gloss>
    </solution>
    <solution>
      <lemmaId=kotab-u_1</lemmaId>
      <pos>kotab/PV/a/PV/SUFF_SUBJ=3MS</pos>
      <gloss>write, he, it verb</gloss>
    </solution>
    <solution>
      <lemmaId=kotab-u_1</lemmaId>
      <pos>kotab/PV_PASS/A/PV/SUFF_SUBJ=3MS</pos>
      <gloss>be written, be fated, be destined, he, it verb</gloss>
    </solution>
    <solution>
      <lemmaId=kitAb_1</lemmaId>
      <pos>kutub</pos>
      <gloss>books</gloss>
    </solution>
    <solution>
      <lemmaId=kitAb_1</lemmaId>
      <pos>kutub/NOUN</pos>
      <gloss>books</gloss>
    </solution>
  </kitb>
</variant>
```

Figure 1: The word classes of “كتب”
The word "من" appears in Buckwalter output as a Noun, verb, Preposition, Relative Pronoun or Interrogative part as shown in figure 2:

Buckwalter’s morphological analyzer can also determine the number of prefixes and suffixes in each word. For example the word "لوبسیبلنها" has three prefixes and two suffixes as shown in figure 3:
Additionally, a single Arabic word may have more than one meaning according to its context. Buckwalter has the ability to indicate this feature by showing different glosses for the same word with the same word class. For example, the word صدور when classified as a noun it may have more than one gloss as shown in figure 4:

Figure 3: The prefixes and suffixes of "وسيلغونها".

Figure 4: The prefixes and suffixes of "صدر".
5.2 Pre-analysis stage:

The basic idea behind the rule-based approach to parts-of-speech tagging is to provide the analyzer software with three lexicons (a prefix lexicon, a stem lexicon and a suffix lexicon) and some sorts of internal grammar which use grammatical rules to disambiguate words.

Surely there must be some objective criteria that enable the analyst to decide to which class a word belongs in order to assign the part-of-speech class. Hence, if one word can be assigned to more than one class, this must be mentioned in the lexicon of the analysis system.

There is a number of general considerations to bear in mind when beginning the process of analyzing the ICA corpus. The pre-analysis stage is an important stage that includes:

A. Handling Buckwalter’s output: When dealing with texts and Buckwalter’s output it was preferred to use a database format because it helps in capturing, editing and changing any part of the information easily. The conversion to database format caused a problem because Buckwalter’s output is divided into three tables: A table for analyzed words with all possible solutions, a table for unanalyzed words that do not exist in the analyzer’s lexicon and a third for punctuation marks found in the text being analyzed. However, this process results in the loss of the context of the text to be analyzed.

B. Handling texts: This stage includes transferring texts from ‘plain text’ horizontal format to database vertical format (from text to list). This process of handling texts helps in keeping the context of words in each text file to be analyzed in one hand, and enabling a list of features to be inserted horizontally besides each word in the list on the other hand.

C. Mapping between Buckwalter’s solutions and word list: In this stage each word in the word list will be mapped with its suitable morphological solutions according to Buckwalter’s output.

An interface has been used to map between Buckwalter’s solutions and the word list. It leads to have a table containing 16 columns of information as follows: Word, Lemma, Vocalization, Gloss, Prefix1, Prefix2, Prefix3, Stem, word class, Suffix1, Suffix2, number, gender, definiteness, Arabic stem and case. Figure 5 shows the following:

- Each solution appears in a separate row.
- Each solution has 16 types of information separated in an independent column.
5.3 Full text analysis stages:

The full text analysis stage includes: disambiguation of words that may have multiple solutions, modifying and adding extra linguistic information and manual analysis of unanalyzed words.

5.3.1 Disambiguating words:

The suitable analysis for each word is chosen according to its context. An interface is used to select the correct analysis solution. Figure 6 shows an example of disambiguating the word "كتّب".

Figure 6: An example of the disambiguation process.
Some information in the output of Buckwalter’s analyzer such as number, gender and definiteness needed modifications according to their morphosyntactic properties. These features can be explained as follows:

- **Gender**: Buckwalter’s analyzer does not identify the gender of Arabic words in two case. The first, if a masculine word or a broken plural ends in “س” e.g. “اسماء” and “اسمادة”, it considers both of them as feminine. The second, if a feminine word or a broken plural does not end in “س” e.g. “أبناء”, “أبواب”, the analyzer does not identify the gender and assigns “NULL” to the words under identification. In both cases, a manual intervention is used to fix the gender.

- **Number**: It has been noted that Buckwalter’s analyzer has a problem with broken plurals; it deals with some of these words as singular, e.g. “أبن” and “أبواب”, and deals with others by assigning them (NULL), e.g. “أبناء”, “أبواب”. This type of plural is given “PL_BR” for number manually. In addition all other nouns that do not end in any morpheme the denotes gender e.g. “أبن”, “أبواب”, have been assigned “NULL”. All number problems have been fixed manually.
**Definiteness**: Buckwalter could detect the suitable definiteness for most words, however, there are some indefinite words that Buckwalter identified as definite words such as “المهامات"، "الالتزام، التحاق، الإتفاق"، these words have been modified to be indefinite. In addition, the analyst added a new value for the feature of definiteness (DEF_EDAFAH), e.g. as in "مهماته"، in order to make the feature of definiteness more expressive.

Figure 8 shows the new modifications for Gender, Number and Definiteness according to their contexts:

<table>
<thead>
<tr>
<th>word</th>
<th>voc</th>
<th>gen</th>
<th>num</th>
<th>def</th>
</tr>
</thead>
<tbody>
<tr>
<td>/D</td>
<td>BOF_Doc</td>
<td>BOF_Doc</td>
<td>BOF_Doc</td>
<td>BOF_Doc</td>
</tr>
<tr>
<td>/t</td>
<td>BOF_Tit</td>
<td>BOF_Tit</td>
<td>BOF_Tit</td>
<td>BOF_Tit</td>
</tr>
<tr>
<td>في</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>استثناء</td>
<td>[isotifotA'M]</td>
<td>MASC</td>
<td>SG</td>
<td>INDEF</td>
</tr>
<tr>
<td>شخصي</td>
<td>DimoniY- 'K'</td>
<td>MASC</td>
<td>SG</td>
<td>INDEF</td>
</tr>
<tr>
<td>على</td>
<td>EalaY</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>رأس</td>
<td>ri'Asapi</td>
<td>FEM</td>
<td>SG</td>
<td>DEF</td>
</tr>
<tr>
<td>بوش</td>
<td>buw$</td>
<td>MASC</td>
<td>SG</td>
<td>DEF</td>
</tr>
<tr>
<td>:</td>
<td>Punc</td>
<td>Punc</td>
<td>Punc</td>
<td>Punc</td>
</tr>
<tr>
<td>توافقات</td>
<td>tawaq- 'u'EaN</td>
<td>FEM</td>
<td>PL</td>
<td>INDEF</td>
</tr>
<tr>
<td>استعادة</td>
<td>bi[isotifAdapi]</td>
<td>FEM</td>
<td>NULL</td>
<td>DEF (EDAFAH)</td>
</tr>
<tr>
<td>الديمقراطية</td>
<td>Ald- 'yimuqrATby- 'lyna</td>
<td>MASC</td>
<td>PL</td>
<td>DEF</td>
</tr>
<tr>
<td>السلطة</td>
<td>Als- ayoTarapa</td>
<td>FEM</td>
<td>SG</td>
<td>DEF</td>
</tr>
<tr>
<td>على</td>
<td>EalaY</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>مجلس</td>
<td>majoliS</td>
<td>MASC</td>
<td>NULL</td>
<td>DEF (EDAFAH)</td>
</tr>
<tr>
<td>النواب</td>
<td>Aln-'uw-'Abi</td>
<td>MASC</td>
<td>PL_BR</td>
<td>DEF</td>
</tr>
<tr>
<td>بعد</td>
<td>baEoda</td>
<td>NULL</td>
<td>NULL</td>
<td>DEF (EDAFAH)</td>
</tr>
<tr>
<td>12</td>
<td>Num</td>
<td>Num</td>
<td>Num</td>
<td>Num</td>
</tr>
<tr>
<td>عامة</td>
<td>EAmAF</td>
<td>MASC</td>
<td>NULL</td>
<td>INDEF</td>
</tr>
<tr>
<td>T/</td>
<td>EOF_Tit</td>
<td>EOF_Tit</td>
<td>EOF_Tit</td>
<td>EOF_Tit</td>
</tr>
<tr>
<td>/P</td>
<td>BOF_Prg</td>
<td>BOF_Prg</td>
<td>BOF_Prg</td>
<td>BOF_Prg</td>
</tr>
<tr>
<td>والوطن</td>
<td>wASinoTun</td>
<td>FEM</td>
<td>SG</td>
<td>DEF</td>
</tr>
<tr>
<td>.</td>
<td>Punc</td>
<td>Punc</td>
<td>Punc</td>
<td>Punc</td>
</tr>
<tr>
<td>وكالات</td>
<td>wakAIAtu</td>
<td>FEM</td>
<td>PL</td>
<td>DEF (EDAFAH)</td>
</tr>
<tr>
<td>الإناث</td>
<td>Al'banobA'l</td>
<td>FEM</td>
<td>PL_BR</td>
<td>DEF</td>
</tr>
</tbody>
</table>

In order to make the morphological analysis more expressive, we have seen that the following extra information that exceed the scope of Buckwalter’s analyzer should be added:

A. **Name entities**: name entities are words that represent the title of an institute, ministry, association, compound name of a country, book, film, company or conference. Analysts identified these names by adding the feature (NOUN_PROP) right after the basic word class of these words. For example "الولايات المتحدة الأميركية" appears in analysis as shown in table 4:
By adding the name entity feature, researchers can capture name entities easily in addition to capturing the word with respect to the part of speech. Figure 9 shows some examples of name entities within their contexts:

B. One of the disadvantages of the Buckwalter’s morphological analyzer is that it determines the word class of Arabic words according to their counterparts in English. For example, Buckwalter’s has classified some adverbs in Arabic as prepositions. Figure 10 shows Buckwalter’s analysis of "بين" which should be analyzed as an adverb.
According to Buckwalter’s analysis of adverbs (figure 10), four observations can be noticed. First, the word "بين" should be analyzed as an adverb; it can be used to describe either a place, as in "بين الساعة الخامسة والخامسة والنصف" or a time as in "بين الأشجار واثنين". Second, Some adverbs are nominalized (no longer adverbs) if they occur after a preposition; in this case their case is genitive as shown in example (1):

(1) "ما زال تنظيم الأسرة من بين التحديات التي تواجه المجتمع" (bayon/NOUN+i/CASE_DEF_GEN)

However, when Buckwalter’s analyzer dealt with "بين" as a noun it gave out three possible cases, namely: nominative, accusative, and genitive (u/NOM, a/ACC, i/GEN, N/NOM and K/GEN), which is not correct. Third, Buckwalter’s analyzer mistakenly analyzed some adverbs not only as prepositions but also as sub conjunctions (SUB_CONJ) as shown in figure 11.

Forth, adverbs in Arabic are tagged with respect to two classes: adverbs which describe time (ADV_T) and adverbs which describe place (ADV_P). The same adverb may describe both time and place in different contexts. Buckwalter’s analyzer can analyze some words as adverbs without determining the manner of that adverb (time or place) as shown in figure 12.
In retagging adverbs two criteria have been taken into account:

1. Separating the case tag from the stem; when Buckwalter analyzes the adverbs it considers the case as a part of the stem and consequently a part of a lemma; for example, the stem of "هنالك" is (hunAka/ADV) and the lemma is “hunAka”. So the case should be separated from stem and lemma.

2. In Arabic adverbs are nouns. Accordingly this has been tagged to every adverb. Consequently, the analysis of adverbs should contain three pieces of information: noun, adverb and time or place (T/P) as table 5 shows.

<table>
<thead>
<tr>
<th>Word</th>
<th>Buckwalter analysis</th>
<th>New analysis</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>عند</td>
<td>Einoda/PREP</td>
<td>Einod/NOUN(ADV_T)</td>
<td>يرجى الاتصال عند حدوث أي مشكلة.</td>
</tr>
<tr>
<td>بعد</td>
<td>baEoda/PREP</td>
<td>baEod/NOUN(ADV_T)</td>
<td>سيتم تشغيلهما بعد الحصول على الترخيص.</td>
</tr>
<tr>
<td>بين</td>
<td>bayona/PREP</td>
<td>bayon/NOUN(ADV_T)</td>
<td>الفترة ما بين العامين الماضيين حلت من التطور.</td>
</tr>
<tr>
<td>أمام</td>
<td>&gt;amAma/PREP</td>
<td>&gt;amAm/NOUN(ADV_P)</td>
<td>إن امام قضية خطيرة.</td>
</tr>
<tr>
<td>عبر</td>
<td>Eabora/PREP</td>
<td>Eabor/NOUN(ADV_P)</td>
<td>تم أرسال البيانات عبر شبكة المعلومات.</td>
</tr>
<tr>
<td>قبل</td>
<td>qabola/PREP</td>
<td>qabol/NOUN(ADV_T)</td>
<td>المبادرة التي اتخذها قبل بضعة أشهر.</td>
</tr>
<tr>
<td>فور</td>
<td>fawora/PREP</td>
<td>favor/NOUN(ADV_T)</td>
<td>مستعد إلى الفعارة قهر انتهاء أعمالها.</td>
</tr>
</tbody>
</table>

Table 5: Example for adverbs.
Figure 13 shows the analysis of some adverbs which have been found in the ICA analyzed corpus:

**NOUN(ADV_M):** This type of adverbs needs the context to be detected, but Buckwalter’s did not identify this type of adverbs As shown in example (2):

(2)

![Example Image](image_url)

Figure 14 shows an example of NOUN(ADV_M) within its context:

![Example Image](image_url)
C. For more accuracy, analysts added new information that Buckwalter’s analyzer does not provide; namely, root information.

The root of each word was detected according to its lemma. It was noted that some words have no root like "... أدا، إفريقيا، أسفلت،...". Analysts gave such words the root “NONE”. Also some foreign words were found in Arabic orthography such as " إنترناشونال، سوسيتي، شارون،..." , analysts gave these words the root “FOREIGN”.

In addition, some words may have two roots as shown in table 6:

<table>
<thead>
<tr>
<th>Word</th>
<th>Lemma</th>
<th>Root</th>
</tr>
</thead>
<tbody>
<tr>
<td>إبناه</td>
<td>{ibon}</td>
<td>bnw/bny</td>
</tr>
<tr>
<td>أزال</td>
<td>&gt;azAl</td>
<td>zwl/zyl</td>
</tr>
<tr>
<td>تنمية</td>
<td>tanomiyap</td>
<td>nmw/nmy</td>
</tr>
</tbody>
</table>

Table 6: example of words may take two roots.

Figure 15 shows each word, lemma and its detected root:

![Figure 15: Examples of root table.](image)

5.3.3 Manual analysis of unanalyzed words:

After choosing the suitable analysis for each word according to the context, some words were found to have no solution for one of two reasons. The First, some words have no analysis according to Buckwalter’s analyzer. The Second, some words can be analyzed but no suitable analysis can be selected according to their context in the text. Therefore, these words have been analyzed manually according to their contexts as if they have been analyzed automatically.
It has been noted that not all unanalyzed words were MSA Arabic words some of them are:

A. Colloquial words like "..." which analysts tagged as (Colloquial).
B. Loan words like "...". These words have no counterpart in Arabic language and therefore have been tagged (Loan).
C. Non Arabic words that are used commonly like "..." and also English words. These words have been tagged as (Not_Arabic).

5.4 ICA: A final analyzed view:

The current state of ICA analyzed corpus helps in interrogating a lot of phenomena since there is one database containing all analyzed words in their context and with their Meta data information. Each word has 17 pieces of information namely: Word, Lemma, Vocalization, Gloss, Prefix1, Prefix2, Prefix3, Stem, word class, Suffix1, Suffix2, number, gender, definiteness, Arabic stem, case and root as shown in figure 16.

Through the analyzed ICA sample the analysts can capture any information easily. For example the analysts can capture all the imperative verbs whether in their contexts or
without context as shown in figure 17 & 18. This can help in building a good search engine tool.

Figure 17: CV within context.

Figure 18: CV without context.
6. Conclusion:

This paper presented a road map of a trial for Arabic corpus analysis. The analysts followed a stem-based approach to be used in analyzing ICA. Buckwalter Morphological analyzer is the most suitable available lexical resource for our approach. The paper discussed a number of general considerations to bear in mind when beginning the process of analyzing the ICA corpus. This trial can be considered one of the most successful approaches for analyzing modern standard Arabic (MSA) in comparison with other trials of Arabic analyzed corpora.

This analyzed sample will be developed to be used as a training corpus to analyze the target size of ICA (100 million words). The ICA software will be developed to interrogate the analyzed version to help researchers to capture powerful textual search.

7. References:

Attia M. (2000), A large-scale computational processor of the Arabic morphology and applications, Faculty of engineering, Cairo university.
Beesley K. (1998a.), Arabic morphology using only finite-state operations, Computational Approaches to Semitic Languages, Proceedings of the Workshop, pages 50–57, Montréal, Quécbec, August 16. Université de Montréal.
Habash N. & TALN J. (2004), Scale Lexeme Based Arabic Morphological Generation, Session Traitement Automatique de l’Arabe, Institute for Advanced Computer Studies, University of Maryland College Park College Park, Maryland, 20742.


Hilbert D. & Krenn B. (2006), Computational Approaches to Collocations, UCS toolkit v0.5 pre-release version fixes some compatibility issues (11-01).


Krauwer S. (2003), The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap, Proceedings of 2nd International Conference on Speech and computer.


Resnik P. (1998), Statistical Methods in NLP, July 8-10, Short Course.

Ritchey T. (2002-2006), General Morphological Analysis, A general method for non-quantified modeling. Downloaded from the Swedish Morphological Society, Adapted from the paper "Fritz Zwicky, Morphology and Policy Analysis”.


