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

Sameh Alansary[*†]
Sameh.alansary@bibalex.org

Magdy Nagi[*††]
magdy.nagi@bibalex.org

Noha Adly[*††]
noha.adly@bibalex.org

[*] Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.
[†] Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University , El Shatby, Alexandria, Egypt.
[††] Computer and System Engineering Dept. Faculty of Engineering, Alexandria University, Alexandria Egypt.

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:

Word	Gloss
(kataba) كَتب	He write
(makotuwb) مكتوب	Written
(kutub) كُتب	Books
(kutiba) كُتب	Be written
(kitab) كتاب	Book
(kut~Ab) كُتاب	Writers/Quran school
(kAtib) كاتب	Writer

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[1].

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 (Sgall et al. 1986, Sgall & Hajičová 2003). Prague Arabic Dependency Treebank (PADT) does not only

[1] 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 lexc 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.

Word	Transliteration	Root	Meaning
قال	qaal	qwl	Say (verb)
		qlw	Fry (active participle)
		qll	decrease (active participle)

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[2]:

- 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[3].

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%4. 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
[3] http://www.sakhr.com/Technology/Morphology/Default.aspx?sec=Technology&item=Morphology
[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].

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:

[5] http://www.rdi-eg.com/rdi/Downloads/Scientific%20Papers/M_Atiyya_MScThesis2000.pdf

Suffix	Stem	Prefix
xxx	كتاب	الــ
ان	كتاب	xxx
ين	كتاب	والــ
xxx	كتب	يــ
xxx	كتب	xxx
ين	كتب	تــ

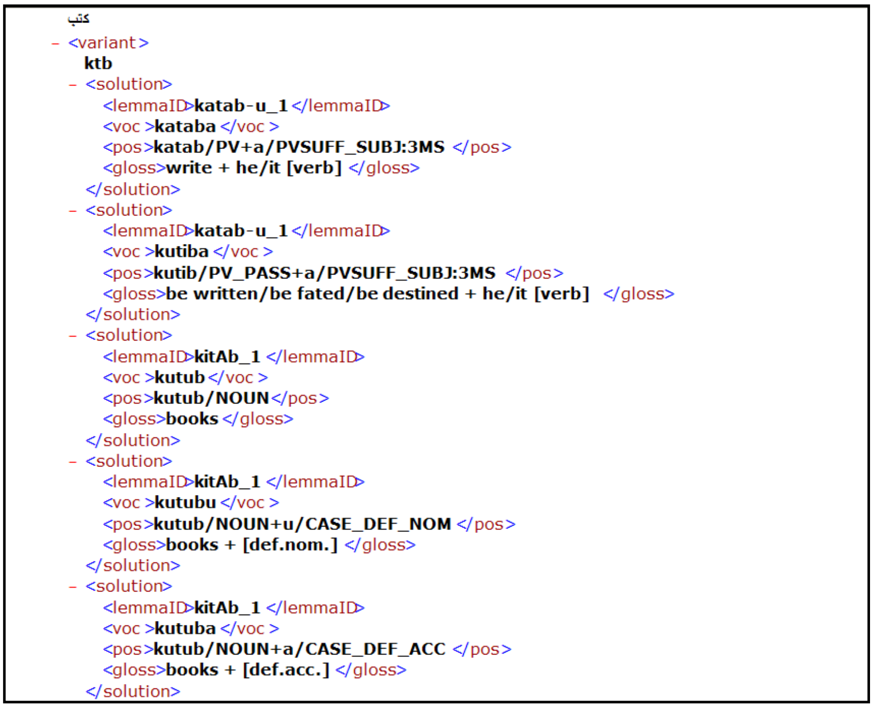
Table 3: valid word forms.

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 that 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.

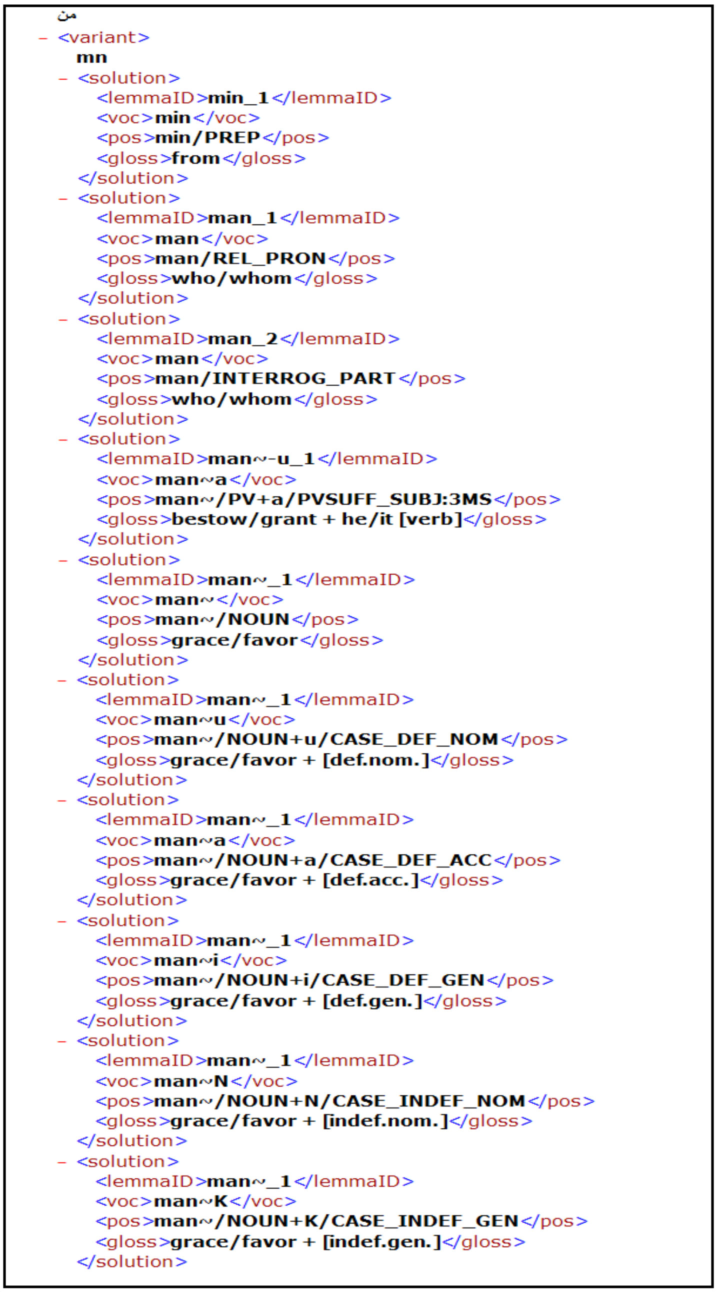
A single word may belong to more than one word class. For example the word "كتب" appears in Buckwalter output as noun or verb as shown in figure 1:



```
   كتب
 - <variant>
     ktb
   - <solution>
       <lemmaID>katab-u_1</lemmaID>
       <voc >kataba </voc >
       <pos>katab/PV+a/PVSUFF_SUBJ:3MS </pos>
       <gloss>write + he/it [verb] </gloss>
     </solution>
   - <solution>
       <lemmaID>katab-u_1</lemmaID>
       <voc >kutiba </voc >
       <pos>kutib/PV_PASS+a/PVSUFF_SUBJ:3MS </pos>
       <gloss>be written/be fated/be destined + he/it [verb] </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc >kutub </voc >
       <pos>kutub/NOUN</pos>
       <gloss>books </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc >kutubu </voc >
       <pos>kutub/NOUN+u/CASE_DEF_NOM </pos>
       <gloss>books + [def.nom.] </gloss>
     </solution>
   - <solution>
       <lemmaID>kitAb_1 </lemmaID>
       <voc >kutuba </voc >
       <pos>kutub/NOUN+a/CASE_DEF_ACC </pos>
       <gloss>books + [def.acc.] </gloss>
     </solution>
```

Figure 1: The word classes of "كتب"

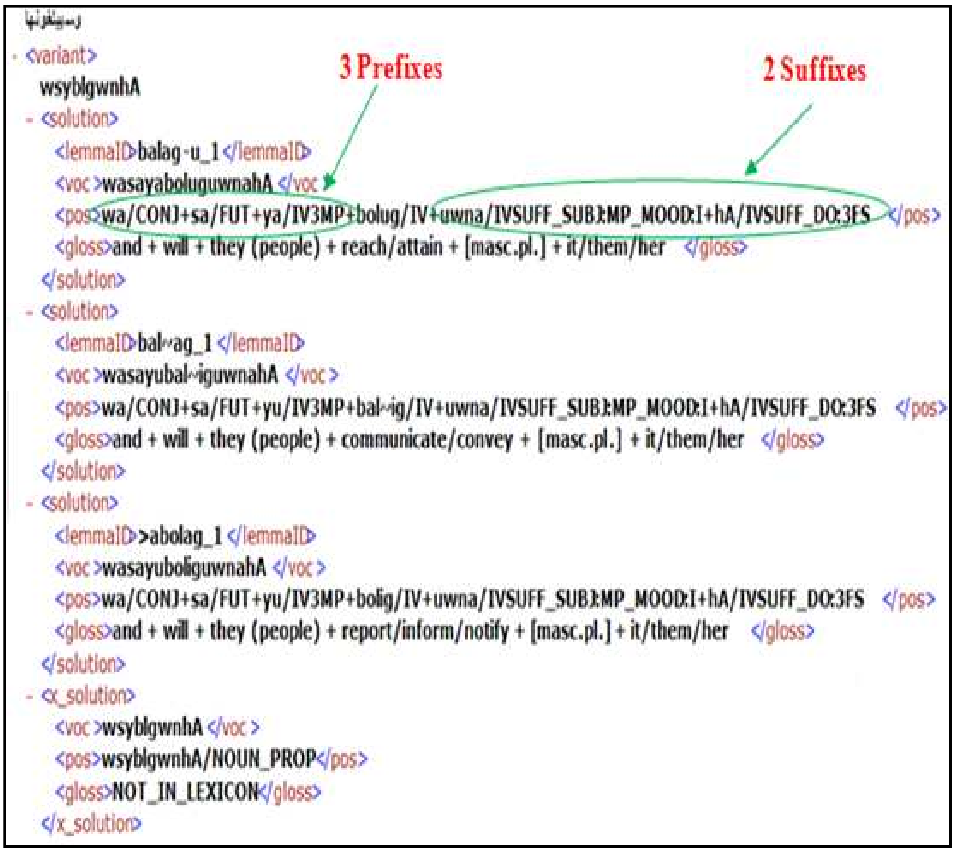
The word "من" appears in Buckwalter output as a Noun, verb, Preposition, Relative Pronoun or Interrogative part as shown in figure 2:



```
    من
 - <variant>
     mn
   - <solution>
       <lemmaID>min_1</lemmaID>
       <voc>min</voc>
       <pos>min/PREP</pos>
       <gloss>from</gloss>
     </solution>
   - <solution>
       <lemmaID>man_1</lemmaID>
       <voc>man</voc>
       <pos>man/REL_PRON</pos>
       <gloss>who/whom</gloss>
     </solution>
   - <solution>
       <lemmaID>man_2</lemmaID>
       <voc>man</voc>
       <pos>man/INTERROG_PART</pos>
       <gloss>who/whom</gloss>
     </solution>
   - <solution>
       <lemmaID>man~-u_1</lemmaID>
       <voc>man~a</voc>
       <pos>man~/PV+a/PVSUFF_SUBJ:3MS</pos>
       <gloss>bestow/grant + he/it [verb]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~</voc>
       <pos>man~/NOUN</pos>
       <gloss>grace/favor</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~u</voc>
       <pos>man~/NOUN+u/CASE_DEF_NOM</pos>
       <gloss>grace/favor + [def.nom.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~a</voc>
       <pos>man~/NOUN+a/CASE_DEF_ACC</pos>
       <gloss>grace/favor + [def.acc.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~i</voc>
       <pos>man~/NOUN+i/CASE_DEF_GEN</pos>
       <gloss>grace/favor + [def.gen.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~N</voc>
       <pos>man~/NOUN+N/CASE_INDEF_NOM</pos>
       <gloss>grace/favor + [indef.nom.]</gloss>
     </solution>
   - <solution>
       <lemmaID>man~_1</lemmaID>
       <voc>man~K</voc>
       <pos>man~/NOUN+K/CASE_INDEF_GEN</pos>
       <gloss>grace/favor + [indef.gen.]</gloss>
     </solution>
```

Figure 2: The word classes of “من”.

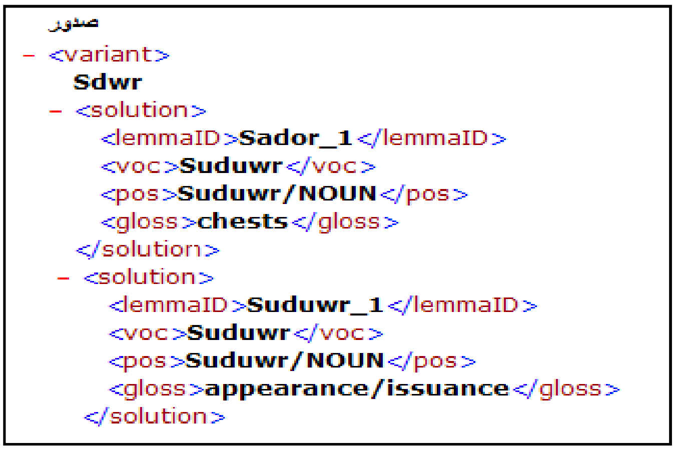
Buckwalter’s morphological analyzer can also determine the number of prefixes and suffixes in each word. For example the word **“وسيبلغونها”** has three prefixes and two suffixes as shown in figure 3:



```
وسيبلغونها
- <variant>
  wsyblgwnhA
 - <solution>
    <lemmaID>balag-u_1</lemmaID>
    <voc >wasayaboluguwnahA </voc >
    <pos>wa/CONJ+sa/FUT+ya/IV3MP+bolug/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS </pos>
    <gloss>and + will + they (people) + reach/attain + [masc.pl.] + it/them/her  </gloss>
  </solution>
 - <solution>
    <lemmaID>bal~ag_1 </lemmaID>
    <voc >wasayubal~iguwnahA </voc >
    <pos>wa/CONJ+sa/FUT+yu/IV3MP+bal~ig/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS  </pos>
    <gloss>and + will + they (people) + communicate/convey + [masc.pl.] + it/them/her  </gloss>
  </solution>
 - <solution>
    <lemmaID>>abolag_1 </lemmaID>
    <voc >wasayuboliguwnahA </voc >
    <pos>wa/CONJ+sa/FUT+yu/IV3MP+bolig/IV+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS  </pos>
    <gloss>and + will + they (people) + report/inform/notify + [masc.pl.] + it/them/her  </gloss>
  </solution>
 - <x_solution>
    <voc >wsyblgwnhA </voc >
    <pos>wsyblgwnhA/NOUN_PROP</pos>
    <gloss>NOT_IN_LEXICON</gloss>
  </x_solution>
```

Figure 3:The prefixes and suffixes of “وسيبلغونها”

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:



```
  صدور
- <variant>
    Sdwr
  - <solution>
      <lemmaID>Sador_1</lemmaID>
      <voc>Suduwr</voc>
      <pos>Suduwr/NOUN</pos>
      <gloss>chests</gloss>
    </solution>
  - <solution>
      <lemmaID>Suduwr_1</lemmaID>
      <voc>Suduwr</voc>
      <pos>Suduwr/NOUN</pos>
      <gloss>appearance/issuance</gloss>
    </solution>
```

Figure 4: The prefixes and suffixes of “صدور”.

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:

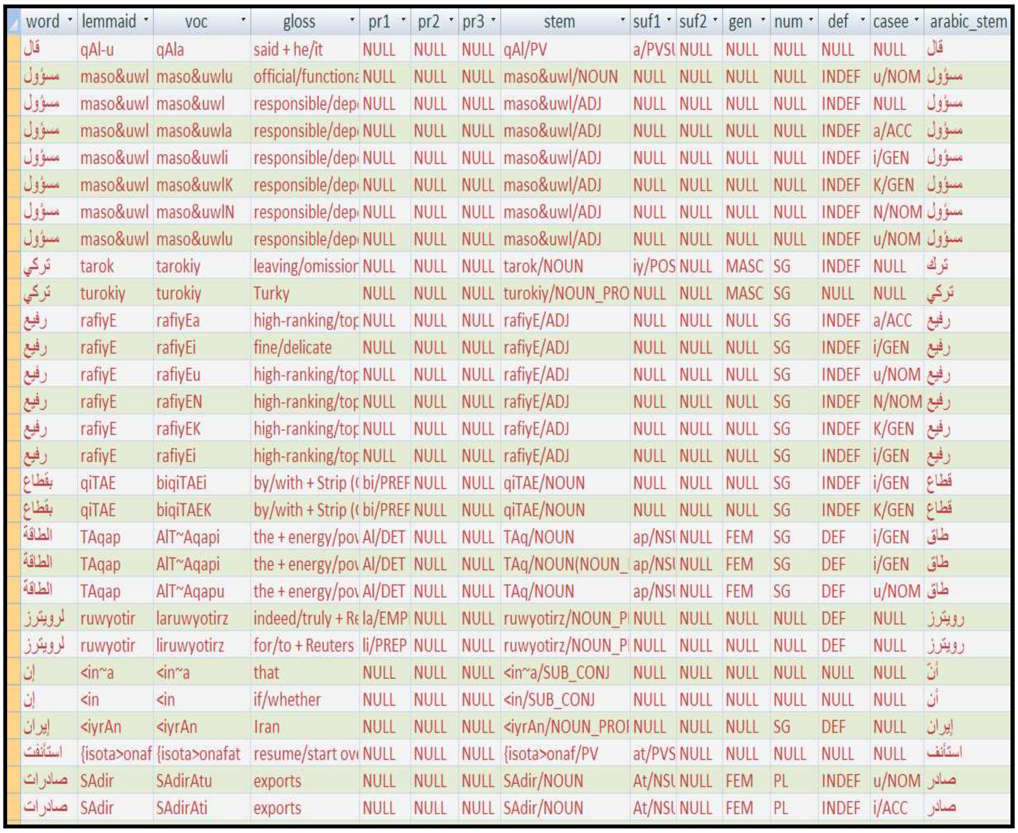
A. **Handling Buckwalter's output:** When dealing with texts and Buckwalter's output it was preferred to use a database format because it helps in capturing, editing and changing any part of the information easily. The conversion to database format caused a problem because Buckwalter's output is divided 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.



word	lemmaid	voc	gloss	pr1	pr2	pr3	stem	suf1	suf2	gen	num	def	casee	arabic_stem
قال	qAl-u	qAla	said + he/it	NULL	NULL	NULL	qAl/PV	a/PVSl	NULL	NULL	NULL	NULL	NULL	قال
مسؤول	maso&uwl	maso&uwlu	official/functiona	NULL	NULL	NULL	maso&uwl/NOUN	NULL	NULL	NULL	NULL	INDEF	u/NOM	مسؤول
مسؤول	maso&uwl	maso&uwl	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	NULL	مسؤول
مسؤول	maso&uwl	maso&uwla	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	a/ACC	مسؤول
مسؤول	maso&uwl	maso&uwli	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	i/GEN	مسؤول
مسؤول	maso&uwl	maso&uwlK	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	K/GEN	مسؤول
مسؤول	maso&uwl	maso&uwlN	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	N/NOM	مسؤول
مسؤول	maso&uwl	maso&uwlu	responsible/dep	NULL	NULL	NULL	maso&uwl/ADJ	NULL	NULL	NULL	NULL	INDEF	u/NOM	مسؤول
تركي	tarok	tarokiy	leaving/omissior	NULL	NULL	NULL	tarok/NOUN	iy/POS	NULL	MASC	SG	INDEF	NULL	ترك
تركي	turokiy	turokiy	Turky	NULL	NULL	NULL	turokiy/NOUN_PRO	NULL	NULL	MASC	SG	NULL	NULL	تركي
رفيع	rafiyE	rafiyEa	high-ranking/top	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	a/ACC	رفيع
رفيع	rafiyE	rafiyEi	fine/delicate	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	i/GEN	رفيع
رفيع	rafiyE	rafiyEu	high-ranking/top	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	u/NOM	رفيع
رفيع	rafiyE	rafiyEN	high-ranking/top	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	N/NOM	رفيع
رفيع	rafiyE	rafiyEK	high-ranking/top	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	K/GEN	رفيع
رفيع	rafiyE	rafiyEi	high-ranking/top	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	NULL	SG	INDEF	i/GEN	رفيع
بقطاع	qiTAE	biqiTAEi	by/with + Strip ((	bi/PREF	NULL	NULL	qiTAE/NOUN	NULL	NULL	NULL	SG	INDEF	i/GEN	قطاع
بقطاع	qiTAE	biqiTAEK	by/with + Strip ((	bi/PREF	NULL	NULL	qiTAE/NOUN	NULL	NULL	NULL	SG	INDEF	K/GEN	قطاع
الطاقة	TAqap	AlT~Aqapi	the + energy/pov	Al/DET	NULL	NULL	TAq/NOUN	ap/NSl	NULL	FEM	SG	DEF	i/GEN	طاق
الطاقة	TAqap	AlT~Aqapi	the + energy/pov	Al/DET	NULL	NULL	TAq/NOUN(NOUN_	ap/NSl	NULL	FEM	SG	DEF	i/GEN	طاق
الطاقة	TAqap	AlT~Aqapu	the + energy/pov	Al/DET	NULL	NULL	TAq/NOUN	ap/NSl	NULL	FEM	SG	DEF	u/NOM	طاق
لرويترز	ruwyotir	laruwyotirz	indeed/truly + Re	la/EMP	NULL	NULL	ruwyotirz/NOUN_P	NULL	NULL	NULL	NULL	DEF	NULL	رويترز
لرويترز	ruwyotir	liruwyotirz	for/to + Reuters	li/PREP	NULL	NULL	ruwyotirz/NOUN_P	NULL	NULL	NULL	NULL	DEF	NULL	رويترز
إن	<in~a	<in~a	that	NULL	NULL	NULL	<in~a/SUB_CONJ	NULL	NULL	NULL	NULL	NULL	NULL	أنّ
إن	<in	<in	if/whether	NULL	NULL	NULL	<in/SUB_CONJ	NULL	NULL	NULL	NULL	NULL	NULL	أن
إيران	<iyrAn	<iyrAn	Iran	NULL	NULL	NULL	<iyrAn/NOUN_PROI	NULL	NULL	NULL	SG	DEF	NULL	إيران
استأنفت	{isota>onaf	{isota>onafat	resume/start ov	NULL	NULL	NULL	{isota>onaf/PV	at/PVS	NULL	NULL	NULL	NULL	NULL	استأنف
صادرات	SAdir	SAdirAtu	exports	NULL	NULL	NULL	SAdir/NOUN	At/NSl	NULL	FEM	PL	INDEF	u/NOM	صادر
صادرات	SAdir	SAdirAti	exports	NULL	NULL	NULL	SAdir/NOUN	At/NSl	NULL	FEM	PL	INDEF	i/ACC	صادر

Figure 5: The database after mapping word list with Buckwalter's solutions.

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 "كتب".

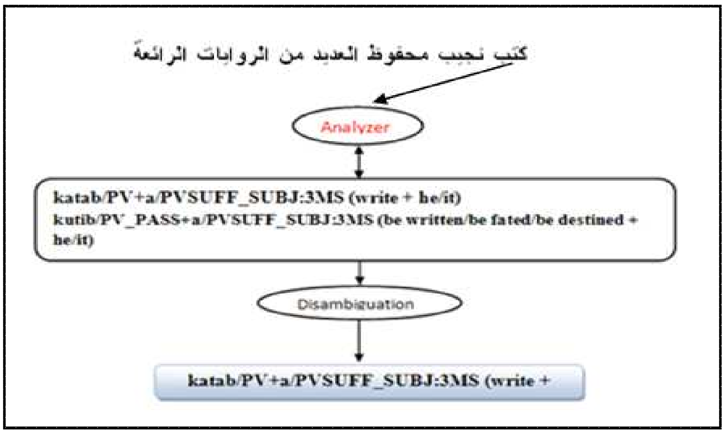
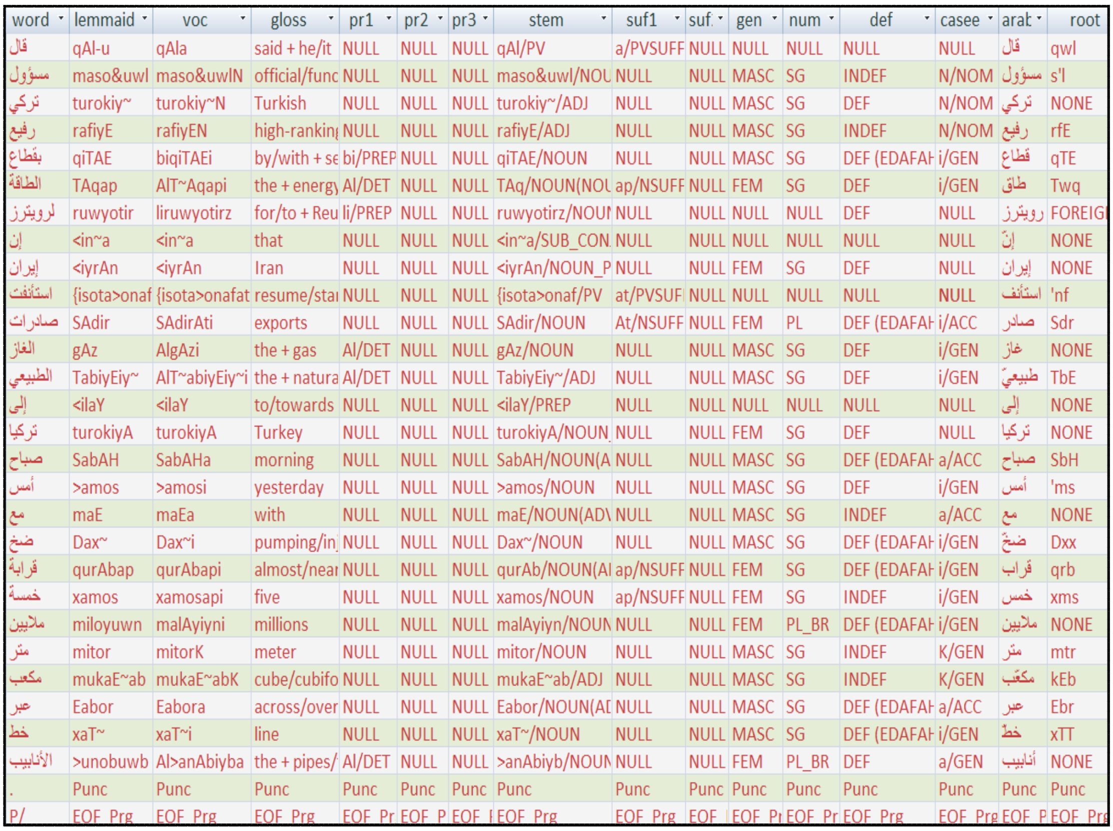

كتب نجيب محفوظ العديد من الروايات الرائعة
Analyzer
katab/PV+a/PVSUFF_SUBJ:3MS (write + he/it)
kutib/PV_PASS+a/PVSUFF_SUBJ:3MS (be written/be fated/be destined + he/it)
Disambiguation
katab/PV+a/PVSUFF_SUBJ:3MS (write +

Figure 6: An example of the disambiguation process.

Figure 7 shows one text after it was disambiguated:



word	lemmaid	voc	gloss	pr1	pr2	pr3	stem	suf1	suf	gen	num	def	casee	arab	root
قال	qAl-u	qAla	said + he/it	NULL	NULL	NULL	qAl/PV	a/PVSUFF	NULL	NULL	NULL	NULL	NULL	قال	qwl
مسؤول	maso&uwl	maso&uwlN	official/func	NULL	NULL	NULL	maso&uwl/NOU	NULL	NULL	MASC	SG	INDEF	N/NOM	مسؤول	s'l
تركي	turokiy~	turokiy~N	Turkish	NULL	NULL	NULL	turokiy~/ADJ	NULL	NULL	MASC	SG	DEF	N/NOM	تركي	NONE
رفيع	rafiyE	rafiyEN	high-rankin	NULL	NULL	NULL	rafiyE/ADJ	NULL	NULL	MASC	SG	INDEF	N/NOM	رفيع	rfE
بقطاع	qiTAE	biqiTAEi	by/with + se	bi/PREP	NULL	NULL	qiTAE/NOUN	NULL	NULL	MASC	SG	DEF (EDAFAH	i/GEN	قطاع	qTE
الطاقة	TAqap	AlT~Aqapi	the + energy	Al/DET	NULL	NULL	TAq/NOUN(NOU	ap/NSUFF	NULL	FEM	SG	DEF	i/GEN	طاق	Twq
لرويترز	ruwyotir	liruwyotirz	for/to + Reu	li/PREP	NULL	NULL	ruwyotirz/NOUN	NULL	NULL	NULL	NULL	DEF	NULL	رويترز	FOREIG
إن	<in~a	<in~a	that	NULL	NULL	NULL	<in~a/SUB_CON	NULL	NULL	NULL	NULL	NULL	NULL	إنّ	NONE
إيران	<iyrAn	<iyrAn	Iran	NULL	NULL	NULL	<iyrAn/NOUN_P	NULL	NULL	FEM	SG	DEF	NULL	إيران	NONE
استأنفت	{isota>onaf	{isota>onafat	resume/star	NULL	NULL	NULL	{isota>onaf/PV	at/PVSUF	NULL	NULL	NULL	NULL	NULL	استأنف	'nf
صادرات	SAdir	SAdirAti	exports	NULL	NULL	NULL	SAdir/NOUN	At/NSUFF	NULL	FEM	PL	DEF (EDAFAH	i/ACC	صادر	Sdr
الغاز	gAz	AlgAzi	the + gas	Al/DET	NULL	NULL	gAz/NOUN	NULL	NULL	MASC	SG	DEF	i/GEN	غاز	NONE
الطبيعي	TabiyEiy~	AlT~abiyEiy~i	the + natura	Al/DET	NULL	NULL	TabiyEiy~/ADJ	NULL	NULL	MASC	SG	DEF	i/GEN	طبيعيّ	TbE
إلى	<ilaY	<ilaY	to/towards	NULL	NULL	NULL	<ilaY/PREP	NULL	NULL	NULL	NULL	NULL	NULL	إلى	NONE
تركيا	turokiyA	turokiyA	Turkey	NULL	NULL	NULL	turokiyA/NOUN	NULL	NULL	FEM	SG	DEF	NULL	تركيا	NONE
صباح	SabAH	SabAHa	morning	NULL	NULL	NULL	SabAH/NOUN(A	NULL	NULL	MASC	SG	DEF (EDAFAH	a/ACC	صباح	SbH
أمس	>amos	>amosi	yesterday	NULL	NULL	NULL	>amos/NOUN	NULL	NULL	MASC	SG	DEF	i/GEN	أمس	'ms
مع	maE	maEa	with	NULL	NULL	NULL	maE/NOUN(ADV	NULL	NULL	MASC	SG	INDEF	a/ACC	مع	NONE
ضخ	Dax~	Dax~i	pumping/in	NULL	NULL	NULL	Dax~/NOUN	NULL	NULL	MASC	SG	DEF (EDAFAH	i/GEN	ضخّ	Dxx
قرابة	qurAbap	qurAbapi	almost/near	NULL	NULL	NULL	qurAb/NOUN(A	ap/NSUFF	NULL	FEM	SG	DEF (EDAFAH	i/GEN	قراب	qrb
خمسة	xamos	xamosapi	five	NULL	NULL	NULL	xamos/NOUN	ap/NSUFF	NULL	FEM	SG	INDEF	i/GEN	خمس	xms
ملايين	miloyuwn	malAyiyni	millions	NULL	NULL	NULL	malAyiyn/NOUN	NULL	NULL	FEM	PL_BR	DEF (EDAFAH	i/GEN	ملايين	NONE
متر	mitor	mitorK	meter	NULL	NULL	NULL	mitor/NOUN	NULL	NULL	MASC	SG	INDEF	K/GEN	متر	mtr
مكعب	mukaE~ab	mukaE~abK	cube/cubifo	NULL	NULL	NULL	mukaE~ab/ADJ	NULL	NULL	MASC	SG	INDEF	K/GEN	مكعّب	kEb
عبر	Eabor	Eabora	across/over	NULL	NULL	NULL	Eabor/NOUN(A	NULL	NULL	MASC	SG	DEF (EDAFAH	a/ACC	عبر	Ebr
خط	xaT~	xaT~i	line	NULL	NULL	NULL	xaT~/NOUN	NULL	NULL	MASC	SG	DEF (EDAFAH	i/GEN	خطّ	xTT
الأنابيب	>unobuwb	Al>anAbiyba	the + pipes/	Al/DET	NULL	NULL	>anAbiyb/NOUN	NULL	NULL	FEM	PL_BR	DEF	a/GEN	أنابيب	NONE
.	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc
P/	EOF_Prg	EOF_Prg	EOF_Prg	EOF_Pr	EOF_P	EOF_	EOF_Prg	EOF_Prg	EOF_	EOF_Pr	EOF_Pr	EOF_Prg	EOF_Prg	EOF_P	EOF_Prg

Figure 7: One of disambiguated texts.

5.3.2 Modifying and adding some linguistic information:

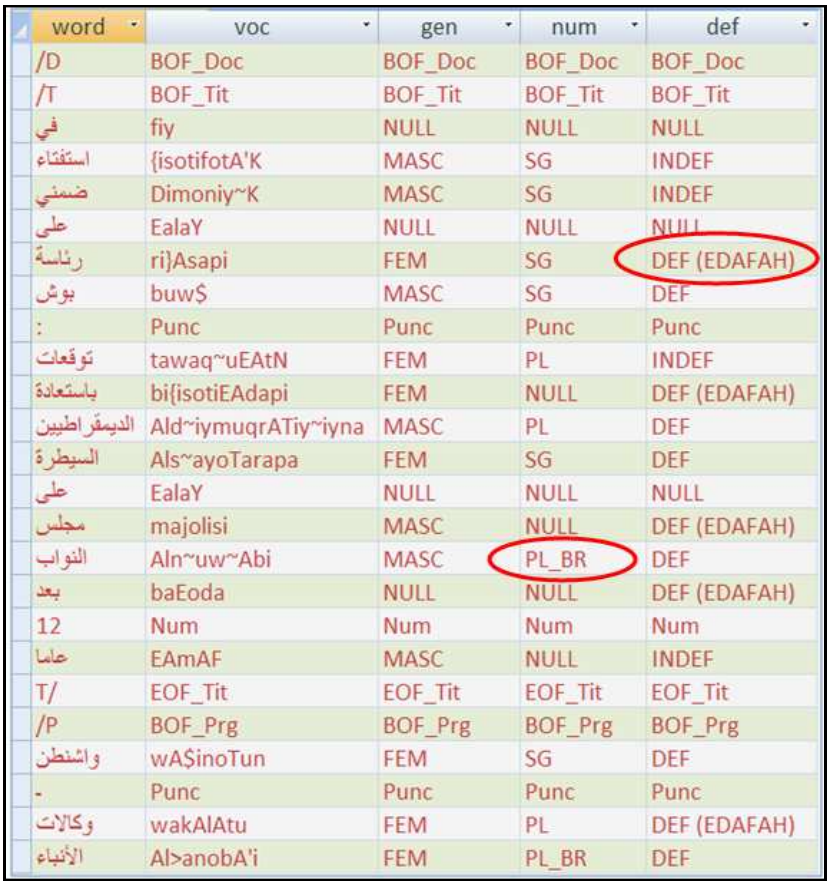
Some information in the output of Buckwalter's analyzer such as number, gender and definiteness needed modifications according to their morphosyntactic properties. These features can be explained as follows:

- ***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 "ة" e.g. "أسامة" and "أساتذة", it considers both of them as feminine. The second, if a feminine word or a broken plural does not end in "ة" e.g. "نساء ، أملاك ، 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. "أبخرة ، أحذية", and deals with others by assigning them (NULL), e.g. "أبواب ، أحوال ، أنحاء". 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. ,"أسمنت ، أبلغ ، أكبر", 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 “التفاف ، التحاق ، التزام”, 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 “مهاراته”, in order to make the feature o definiteness more expressive. " مهاراته .

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



word	voc	gen	num	def
/D	BOF_Doc	BOF_Doc	BOF_Doc	BOF_Doc
/T	BOF_Tit	BOF_Tit	BOF_Tit	BOF_Tit
في	fiy	NULL	NULL	NULL
استفتاء	{isotifotA'K	MASC	SG	INDEF
ضمني	Dimoniy~K	MASC	SG	INDEF
على	EalaY	NULL	NULL	NULL
رئاسة	ri}Asapi	FEM	SG	DEF (EDAFAH)
بوش	buw$	MASC	SG	DEF
:	Punc	Punc	Punc	Punc
توقعات	tawaq~uEAtN	FEM	PL	INDEF
باستعادة	bi{isotiEAdapi	FEM	NULL	DEF (EDAFAH)
الديمقراطيين	Ald~iymuqrATiy~iyna	MASC	PL	DEF
السيطرة	Als~ayoTarapa	FEM	SG	DEF
على	EalaY	NULL	NULL	NULL
مجلس	majolisi	MASC	NULL	DEF (EDAFAH)
النواب	Aln~uw~Abi	MASC	PL_BR	DEF
بعد	baEoda	NULL	NULL	DEF (EDAFAH)
12	Num	Num	Num	Num
عاما	EAmAF	MASC	NULL	INDEF
T/	EOF_Tit	EOF_Tit	EOF_Tit	EOF_Tit
/P	BOF_Prg	BOF_Prg	BOF_Prg	BOF_Prg
واشنطن	wA$inoTun	FEM	SG	DEF
-	Punc	Punc	Punc	Punc
وكالات	wakAlAtu	FEM	PL	DEF (EDAFAH)
الأنباء	Al>anobA'i	FEM	PL_BR	DEF

Figure 8 : Gender, Number and Definiteness.

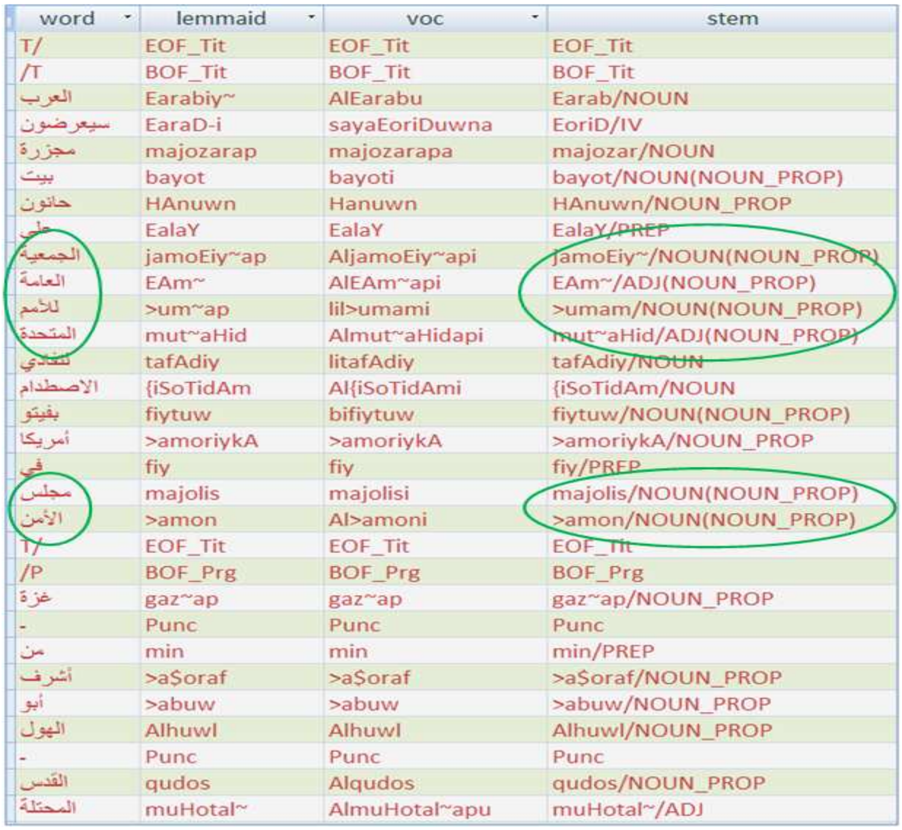
In order to make the morphological analysis more expressive, we have seen that the following extra information that exceed the scope of Buckwalter’s analyzer should be added:

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:

Word	Word Class
الولايات	NOUN(NOUN_PROP)
المتحدة	ADJ(NOUN_PROP)
الأميركية	ADJ(NOUN_PROP)

Table 4 : An example of a name entity.

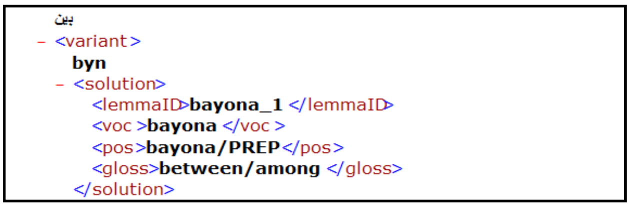
By adding the name entity feature, researchers can capture name entities easily in addition to capturing the word with respect to the part of speech. Figure 9 shows some examples of name entities within their contexts:



word	lemmaid	voc	stem
T/	EOF_Tit	EOF_Tit	EOF_Tit
/T	BOF_Tit	BOF_Tit	BOF_Tit
العرب	Earabiy~	AlEarabu	Earab/NOUN
سيعرضون	EaraD-i	sayaEoriDuwna	EoriD/IV
مجزرة	majozarap	majozarapa	majozar/NOUN
بيت	bayot	bayoti	bayot/NOUN(NOUN_PROP)
حانون	HAnuwn	Hanuwn	HAnuwn/NOUN_PROP
على	EalaY	EalaY	EalaY/PREP
الجمعية	jamoEiy~ap	AljamoEiy~api	jamoEiy~/NOUN(NOUN_PROP)
العامة	EAm~	AlEAm~api	EAm~/ADJ(NOUN_PROP)
للأمم	>um~ap	lil>umami	>umam/NOUN(NOUN_PROP)
المتحدة	mut~aHid	Almut~aHidapi	mut~aHid/ADJ(NOUN_PROP)
لتفادي	tafAdiy	litafAdiy	tafAdiy/NOUN
الاصطدام	{iSoTidAm	Al{iSoTidAmi	{iSoTidAm/NOUN
بفيتو	fiytuw	bifiytuw	fiytuw/NOUN(NOUN_PROP)
أمريكا	>amoriykA	>amoriykA	>amoriykA/NOUN_PROP
في	fiy	fiy	fiy/PREP
مجلس	majolis	majolisi	majolis/NOUN(NOUN_PROP)
الأمن	>amon	Al>amoni	>amon/NOUN(NOUN_PROP)
T/	EOF_Tit	EOF_Tit	EOF_Tit
/P	BOF_Prg	BOF_Prg	BOF_Prg
غزة	gaz~ap	gaz~ap	gaz~ap/NOUN_PROP
-	Punc	Punc	Punc
من	min	min	min/PREP
أشرف	>a$oraf	>a$oraf	>a$oraf/NOUN_PROP
أبو	>abuw	>abuw	>abuw/NOUN_PROP
الهول	Alhuwl	Alhuwl	Alhuwl/NOUN_PROP
-	Punc	Punc	Punc
القدس	qudos	Alqudos	qudos/NOUN_PROP
المحتلة	muHotal~	AlmuHotal~apu	muHotal~/ADJ

Figure 9: Some name entities according to context.

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.



```
  بين
- <variant>
    byn
  - <solution>
      <lemmaID>bayona_1 </lemmaID>
      <voc >bayona </voc >
      <pos>bayona/PREP</pos>
      <gloss>between/among </gloss>
    </solution>
```

Figure 10: The word "بين" as preposition.

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 الأشجار", or a time as in "بين الساعة الخامسة والخامسة والنصف". 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):

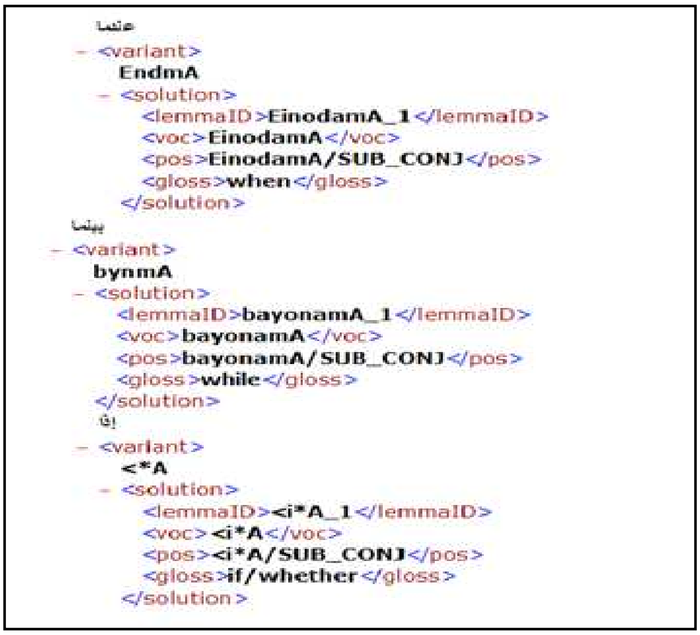
(1)

"ما زال تنظيم الأسرة من **بين** التحديات التي تواجه المجتمع"

↓

(bayon/NOUN+i/CASE_DEF_GEN)

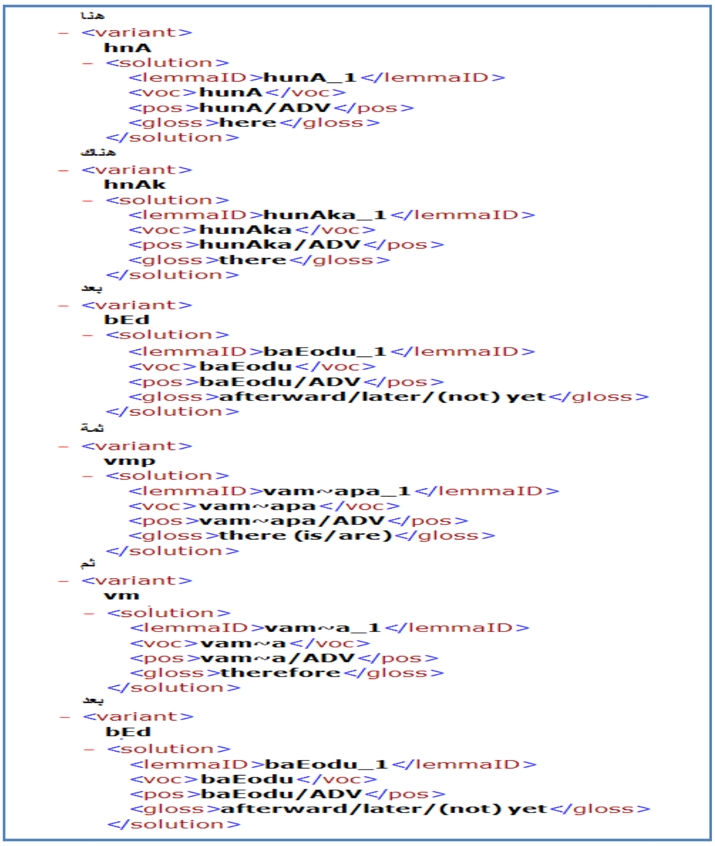
However, when Buckwalter’s analyzer dealt with "بين" as a noun it gave out three possible cases, namely: nominative, accusative, and genitive (u/NOM, a/ACC, i/GEN, N/NOM and K/GEN), which is not correct. Third**,** Buckwalter's analyzer mistakenly analyzed some adverbs not only as prepositions but also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



```
    عندما
  - <variant>
      EndmA
    - <solution>
        <lemmaID>EinodamA_1</lemmaID>
        <voc>EinodamA</voc>
        <pos>EinodamA/SUB_CONJ</pos>
        <gloss>when</gloss>
      </solution>
  بينما
- <variant>
    bynmA
  - <solution>
      <lemmaID>bayonamA_1</lemmaID>
      <voc>bayonamA</voc>
      <pos>bayonamA/SUB_CONJ</pos>
      <gloss>while</gloss>
    </solution>
    إذا
  - <variant>
      <*A
    - <solution>
        <lemmaID><i*A_1</lemmaID>
        <voc><i*A</voc>
        <pos><i*A/SUB_CONJ</pos>
        <gloss>if/whether</gloss>
      </solution>
```

Figure 11: Example of Buckwalter output.

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.



```
هنا
– <variant>
    hnA
  – <solution>
      <lemmaID>hunA_1</lemmaID>
      <voc>hunA</voc>
      <pos>hunA/ADV</pos>
      <gloss>here</gloss>
    </solution>
هناك
– <variant>
    hnAk
  – <solution>
      <lemmaID>hunAka_1</lemmaID>
      <voc>hunAka</voc>
      <pos>hunAka/ADV</pos>
      <gloss>there</gloss>
    </solution>
بعد
– <variant>
    bEd
  – <solution>
      <lemmaID>baEodu_1</lemmaID>
      <voc>baEodu</voc>
      <pos>baEodu/ADV</pos>
      <gloss>afterward/later/(not) yet</gloss>
    </solution>
ثمة
– <variant>
    vmp
  – <solution>
      <lemmaID>vam~apa_1</lemmaID>
      <voc>vam~apa</voc>
      <pos>vam~apa/ADV</pos>
      <gloss>there (is/are)</gloss>
    </solution>
ثم
– <variant>
    vm
  – <solution>
      <lemmaID>vam~a_1</lemmaID>
      <voc>vam~a</voc>
      <pos>vam~a/ADV</pos>
      <gloss>therefore</gloss>
    </solution>
بعد
– <variant>
    bEd
  – <solution>
      <lemmaID>baEodu_1</lemmaID>
      <voc>baEodu</voc>
      <pos>baEodu/ADV</pos>
      <gloss>afterward/later/(not) yet</gloss>
    </solution>
```

Figure 12: Buckwalter Adverbs analysis.

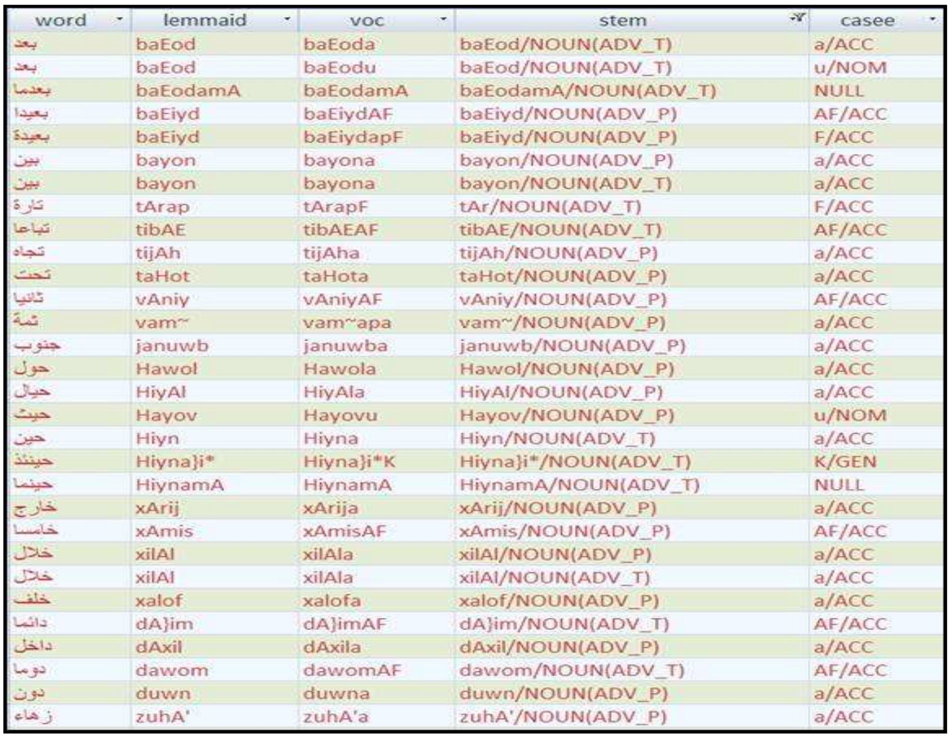
In retagging adverbs two criteria have been taken into account:

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 lamma; for example, the stem of "هناك" 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.

Word	Buckwalter analysis	New analysis	Example
عند	Einoda/PREP	Einod/NOUN(ADV_T) Einod/NOUN(ADV_P)	يرجى الاتصال **عند** حدوث أي مشكلة. يلزم بناء سد **عند** مدخل الفيوم.
بعد	baEoda/PREP	baEod/NOUN(ADV_T) baEod/NOUN(ADV_P)	سيتم تشغيلها **بعد** الحصول على الترخيص. الشريك التجاري الثاني **بعد** تركيا.
بين	bayona/PREP	bayon/NOUN(ADV_T) bayon/NOUN(ADV_P)	الفترة ما **بين** العامين الماضيين خلت من التطور. إن التنسيق **بين** مصر وسوريا منتظم.
أمام	>amAma/PREP	>amAm/NOUN(ADV_P)	إننا **أمام** قضية خطيرة.
عبر	Eabora/PREP	Eabor/NOUN(ADV_P)	تم إرسال البيانات **عبر** شبكة المعلومات.
قبل	qabola/PREP	qabol/NOUN(ADV_T)	المبادرة التي اتخذها **قبل** بضعة أشهر.
فور	fawora/PREP	fawor/NOUN(ADV_T)	ستعود إلى القاهرة **فور** انتهاء أعمالها.

Table 5: Example for adverbs.

Figure 13 shows the analysis of some adverbs which have been found in the ICA analyzed corpus:



word	lemmaid	voc	stem	casee
بعد	baEod	baEoda	baEod/NOUN(ADV_T)	a/ACC
بعد	baEod	baEodu	baEod/NOUN(ADV_T)	u/NOM
بعدما	baEodamA	baEodamA	baEodamA/NOUN(ADV_T)	NULL
بعيدا	baEiyd	baEiydAF	baEiyd/NOUN(ADV_P)	AF/ACC
بعيدة	baEiyd	baEiydapF	baEiyd/NOUN(ADV_P)	F/ACC
بين	bayon	bayona	bayon/NOUN(ADV_P)	a/ACC
بين	bayon	bayona	bayon/NOUN(ADV_T)	a/ACC
تارة	tArap	tArapF	tAr/NOUN(ADV_T)	F/ACC
تباعا	tibAE	tibAEAF	tibAE/NOUN(ADV_T)	AF/ACC
تجاه	tijAh	tijAha	tijAh/NOUN(ADV_P)	a/ACC
تحت	taHot	taHota	taHot/NOUN(ADV_P)	a/ACC
ثانيا	vAniy	vAniyAF	vAniy/NOUN(ADV_P)	AF/ACC
ثمة	vam~	vam~apa	vam~/NOUN(ADV_P)	a/ACC
جنوب	januwb	januwba	januwb/NOUN(ADV_P)	a/ACC
حول	Hawol	Hawola	Hawol/NOUN(ADV_P)	a/ACC
حيال	HiyAl	HiyAla	HiyAl/NOUN(ADV_P)	a/ACC
حيث	Hayov	Hayovu	Hayov/NOUN(ADV_P)	u/NOM
حين	Hiyn	Hiyna	Hiyn/NOUN(ADV_T)	a/ACC
حينئذ	Hiyna}i*	Hiyna}i*K	Hiyna}i*/NOUN(ADV_T)	K/GEN
حينما	HiynamA	HiynamA	HiynamA/NOUN(ADV_T)	NULL
خارج	xArij	xArija	xArij/NOUN(ADV_P)	a/ACC
خامسا	xAmis	xAmisAF	xAmis/NOUN(ADV_P)	AF/ACC
خلال	xilAl	xilAla	xilAl/NOUN(ADV_P)	a/ACC
خلال	xilAl	xilAla	xilAl/NOUN(ADV_T)	a/ACC
خلف	xalof	xalofa	xalof/NOUN(ADV_P)	a/ACC
دائما	dA}im	dA}imAF	dA}im/NOUN(ADV_T)	AF/ACC
داخل	dAxil	dAxila	dAxil/NOUN(ADV_P)	a/ACC
دوما	dawom	dawomAF	dawom/NOUN(ADV_T)	AF/ACC
دون	duwn	duwna	duwn/NOUN(ADV_P)	a/ACC
زهاء	zuhA'	zuhA'a	zuhA'/NOUN(ADV_P)	a/ACC

Figure 13: Some adverbs in the ICA analyzed corpus.

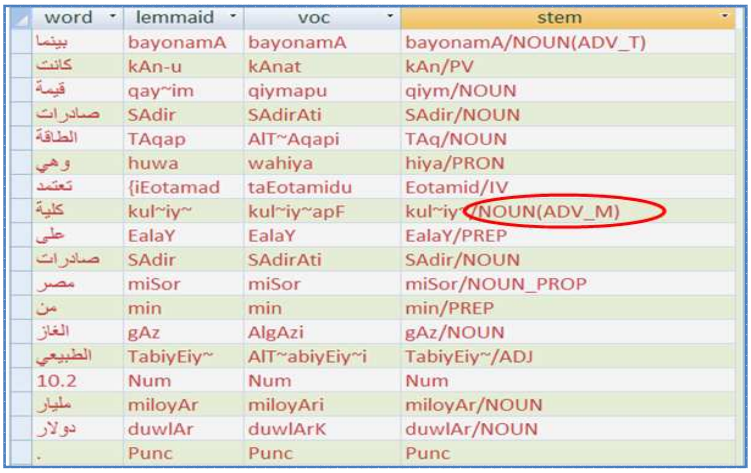
NOUN(ADV_M): This type of adverbs needs the context to be detected, but Buckwalter's did not identify this type of adverbs As shown in example (2):

(2)

جاء الولد مسرعا

NOUN(ADV_M)

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



word	lemmaid	voc	stem
بينما	bayonamA	bayonamA	bayonamA/NOUN(ADV_T)
كانت	kAn-u	kAnat	kAn/PV
قيمة	qay~im	qiymapu	qiym/NOUN
صادرات	SAdir	SAdirAti	SAdir/NOUN
الطاقة	TAqap	AlT~Aqapi	TAq/NOUN
وهي	huwa	wahiya	hiya/PRON
تعتمد	{iEotamad	taEotamidu	Eotamid/IV
كلية	kul~iy~	kul~iy~apF	kul~iy~/NOUN(ADV_M)
على	EalaY	EalaY	EalaY/PREP
صادرات	SAdir	SAdirAti	SAdir/NOUN
مصر	miSor	miSor	miSor/NOUN_PROP
من	min	min	min/PREP
الغاز	gAz	AlgAzi	gAz/NOUN
الطبيعي	TabiyEiy~	AlT~abiyEiy~i	TabiyEiy~/ADJ
10.2	Num	Num	Num
مليار	miloyAr	miloyAri	miloyAr/NOUN
دولار	duwlAr	duwlArK	duwlAr/NOUN
.	Punc	Punc	Punc

Figure 14: An example of NOUN(ADV_M) within context.

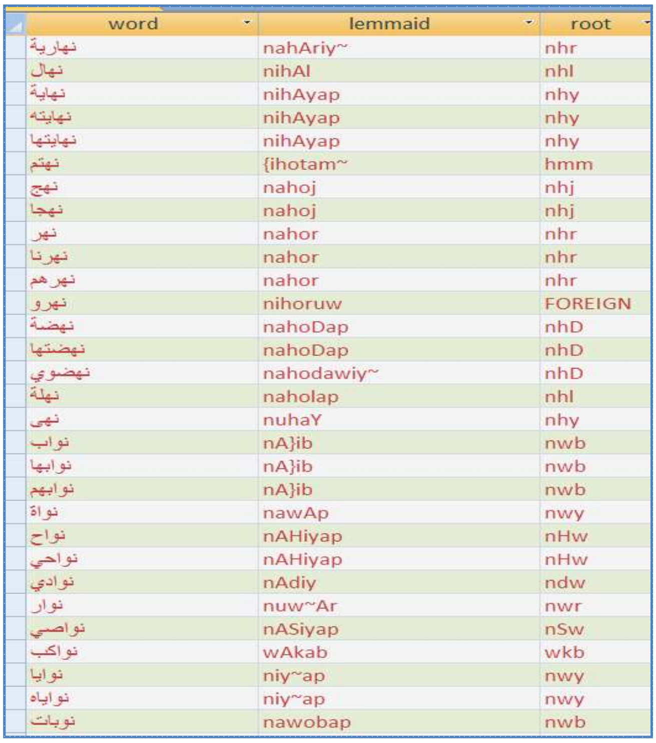
C. For more accuracy, analysts added new information that Buckwalter’s analyzer does not provide; namely, root information.

The root of each word was detected according to its lemma. It was noted that some words have no root like "إذا، إفريقيا، أسفلت، ..." . Analysts gave such words the root “NONE”. Also some foreign words were found in Arabic orthography such as, "إنترناشونال، سوستيه، شارون، ..." , analysts gave these words the root “FOREIGN”. In addition, some words may have two roots as shown in table 6:

Word	Lemma	Root
أبناء	{ibon	bnw/bny
أزال	>azAl	zwl/zyl
تنمية	tanomiyap	nmw/nmy

Table 6: example of words may take two roots.

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



word	lemmaid	root
نهارية	nahAriy~	nhr
نهال	nihAl	nhl
نهاية	nihAyap	nhy
نهايته	nihAyap	nhy
نهايتها	nihAyap	nhy
نهتم	{ihotam~	hmm
نهج	nahoj	nhj
نهجا	nahoj	nhj
نهر	nahor	nhr
نهرنا	nahor	nhr
نهرهم	nahor	nhr
نهرو	nihoruw	FOREIGN
نهضة	nahoDap	nhD
نهضتها	nahoDap	nhD
نهضوي	nahodawiy~	nhD
نهلة	naholap	nhl
نهى	nuhaY	nhy
نواب	nA}ib	nwb
نوابها	nA}ib	nwb
نوابهم	nA}ib	nwb
نواة	nawAp	nwy
نواح	nAHiyap	nHw
نواحي	nAHiyap	nHw
نوادي	nAdiy	ndw
نوار	nuw~Ar	nwr
نواصي	nASiyap	nSw
نواكب	wAkab	wkb
نوايا	niy~ap	nwy
نواياه	niy~ap	nwy
نوبات	nawobap	nwb

Figure 15: Examples of root table.

5.3.3 Manual analysis of unanalyzed words:

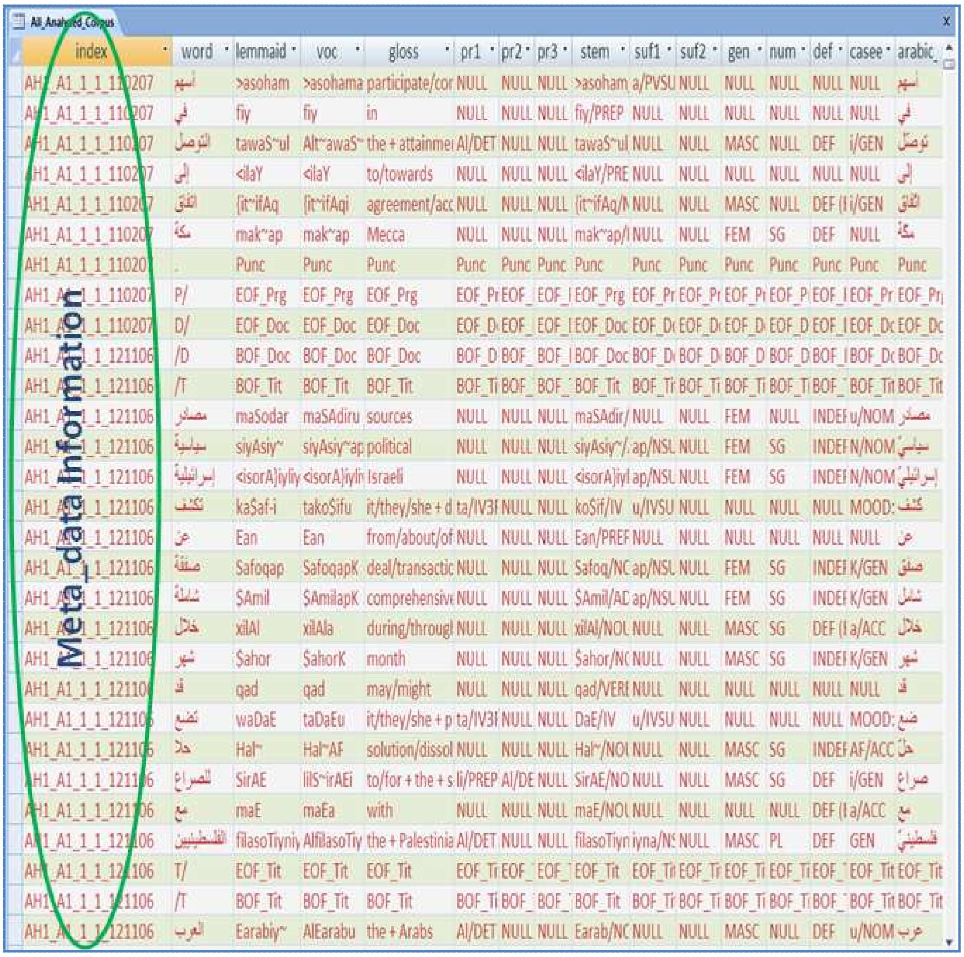
After choosing the suitable analysis for each word according to the context, some words were found to have no solution for one of two reasons. The First, some words have no analysis according to Buckwalter’s analyzer. The Second, some words can be analyzed but no suitable analysis can be selected according 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 "... جواهرجي – بتحبك – حنشوف – إزاي" which analysts tagged as (Colloquial).
B. Loan words like "... بلودوج – البرجماتية – تكنوكراتي ". These words have no counterpart in Arabic language and therefore have been tagged (Loan).
C. Non Arabic words that are used commonly like "... سنجل – ديكشنري " and also English words. These words have been tagged as (Not_Arabic).

5.4 ICA: A final analyzed view:

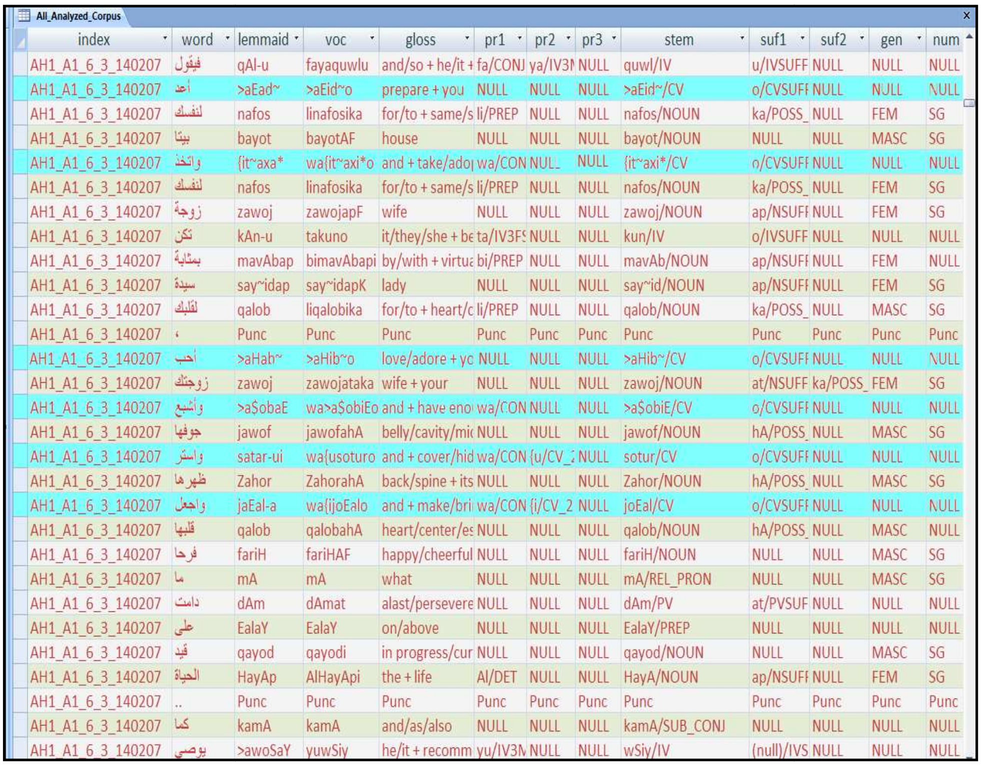
The current state of ICA analyzed corpus helps in interrogating a lot of phenomena since there is one database containing all analyzed words in their context and with their Meta data information. Each word has 17 pieces of information namely: Word, Lemma, Vocalization, Gloss, Prefix1, Prefix2, Prefix3,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



index	word	lemmaid	voc	gloss	pr1	pr2	pr3	stem	suf1	suf2	gen	num	def	casee	arabic_
AH1_A1_1_1_110207	أسهم	>asoham	>asohama	participate/cor	NULL	NULL	NULL	>asoham	a/PVSU	NULL	NULL	NULL	NULL	NULL	أسهم
AH1_A1_1_1_110207	في	fiy	fiy	in	NULL	NULL	NULL	fiy/PREP	NULL	NULL	NULL	NULL	NULL	NULL	في
AH1_A1_1_1_110207	التوصل	tawaS~ul	Alt~awaS~	the + attainmei	Al/DET	NULL	NULL	tawaS~ul	NULL	NULL	MASC	NULL	DEF	i/GEN	توصّل
AH1_A1_1_1_110207	إلى	<ilaY	<ilaY	to/towards	NULL	NULL	NULL	<ilaY/PRE	NULL	NULL	NULL	NULL	NULL	NULL	إلى
AH1_A1_1_1_110207	اتفاق	{it~ifAq	{it~ifAqi	agreement/acc	NULL	NULL	NULL	{it~ifAq/N	NULL	NULL	MASC	NULL	DEF (I	i/GEN	اتفاق
AH1_A1_1_1_110207	مكة	mak~ap	mak~ap	Mecca	NULL	NULL	NULL	mak~ap/I	NULL	NULL	FEM	SG	DEF	NULL	مكة
AH1_A1_1_1_110207	.	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc
AH1_A1_1_1_110207	P/	EOF_Prg	EOF_Prg	EOF_Prg	EOF_Pr	EOF_	EOF_I	EOF_Prg	EOF_Pr	EOF_Pr	EOF_Pr	EOF_P	EOF_I	EOF_Pr	EOF_Pr
AH1_A1_1_1_110207	D/	EOF_Doc	EOF_Doc	EOF_Doc	EOF_D	EOF_	EOF_I	EOF_Doc	EOF_D	EOF_D	EOF_D	EOF_D	EOF_I	EOF_Dc	EOF_Dc
AH1_A1_1_1_121106	/D	BOF_Doc	BOF_Doc	BOF_Doc	BOF_D	BOF_	BOF_I	BOF_Doc	BOF_D	BOF_D	BOF_D	BOF_D	BOF_I	BOF_Dc	BOF_Dc
AH1_A1_1_1_121106	/T	BOF_Tit	BOF_Tit	BOF_Tit	BOF_Ti	BOF_	BOF_	BOF_Tit	BOF_Ti	BOF_Ti	BOF_Ti	BOF_Ti	BOF_	BOF_Tit	BOF_Tit
AH1_A1_1_1_121106	مصادر	maSodar	maSAdiru	sources	NULL	NULL	NULL	maSAdir/	NULL	NULL	FEM	NULL	INDEF	u/NOM	مصادر
AH1_A1_1_1_121106	سياسية	siyAsiy~	siyAsiy~ap	political	NULL	NULL	NULL	siyAsiy~/	ap/NSU	NULL	FEM	SG	INDEF	N/NOM	سياسي
AH1_A1_1_1_121106	إسرائيلية	<isorA}iyliy	<isorA}iylr	Israeli	NULL	NULL	NULL	<isorA}iyl	ap/NSU	NULL	FEM	SG	INDEF	N/NOM	إسرائيلي
AH1_A1_1_1_121106	تكشف	ka$af-i	tako$ifu	it/they/she + d	ta/IV3I	NULL	NULL	ko$if/IV	u/IVSU	NULL	NULL	NULL	NULL	MOOD:	كشف
AH1_A1_1_1_121106	عن	Ean	Ean	from/about/of	NULL	NULL	NULL	Ean/PREF	NULL	NULL	NULL	NULL	NULL	NULL	عن
AH1_A1_1_1_121106	صفقة	Safoqap	SafoqapK	deal/transactic	NULL	NULL	NULL	Safoq/NC	ap/NSU	NULL	FEM	SG	INDEF	K/GEN	صفق
AH1_A1_1_1_121106	شاملة	$Amil	$AmilapK	comprehensive	NULL	NULL	NULL	$Amil/AE	ap/NSU	NULL	FEM	SG	INDEF	K/GEN	شامل
AH1_A1_1_1_121106	خلال	xilAl	xilAla	during/througl	NULL	NULL	NULL	xilAl/NOU	NULL	NULL	MASC	SG	DEF (I	a/ACC	خلال
AH1_A1_1_1_121106	شهر	$ahor	$ahorK	month	NULL	NULL	NULL	$ahor/NC	NULL	NULL	MASC	SG	INDEF	K/GEN	شهر
AH1_A1_1_1_121106	قد	qad	qad	may/might	NULL	NULL	NULL	qad/VERI	NULL	NULL	NULL	NULL	NULL	NULL	قد
AH1_A1_1_1_121106	تضع	waDaE	taDaEu	it/they/she + p	ta/IV3I	NULL	NULL	DaE/IV	u/IVSU	NULL	NULL	NULL	NULL	MOOD:	ضع
AH1_A1_1_1_121106	حلا	Hal~	Hal~AF	solution/dissol	NULL	NULL	NULL	Hal~/NOI	NULL	NULL	MASC	SG	INDEF	AF/ACC	حلّ
AH1_A1_1_1_121106	للصراع	SirAE	lilS~irAEi	to/for + the + s	li/PREP	Al/DE	NULL	SirAE/NO	NULL	NULL	MASC	SG	DEF	i/GEN	صراع
AH1_A1_1_1_121106	مع	maE	maEa	with	NULL	NULL	NULL	maE/NOU	NULL	NULL	NULL	NULL	DEF (I	a/ACC	مع
AH1_A1_1_1_121106	الفلسطينيين	filasoTiyniy	AlfilasoTiy	the + Palestinia	Al/DET	NULL	NULL	filasoTiyn	iyna/N:	NULL	MASC	PL	DEF	GEN	فلسطيني
AH1_A1_1_1_121106	T/	EOF_Tit	EOF_Tit	EOF_Tit	EOF_Ti	EOF_	EOF_	EOF_Tit	EOF_Ti	EOF_Ti	EOF_Ti	EOF_Ti	EOF_	EOF_Tit	EOF_Tit
AH1_A1_1_1_121106	/T	BOF_Tit	BOF_Tit	BOF_Tit	BOF_Ti	BOF_	BOF_	BOF_Tit	BOF_Ti	BOF_Ti	BOF_Ti	BOF_Ti	BOF_	BOF_Tit	BOF_Tit
AH1_A1_1_1_121106	العرب	Earabiy~	AlEarabu	the + Arabs	Al/DET	NULL	NULL	Earab/NC	NULL	NULL	MASC	NULL	DEF	u/NOM	عرب

Figure 16: Final view of ICA analyzed corpus.

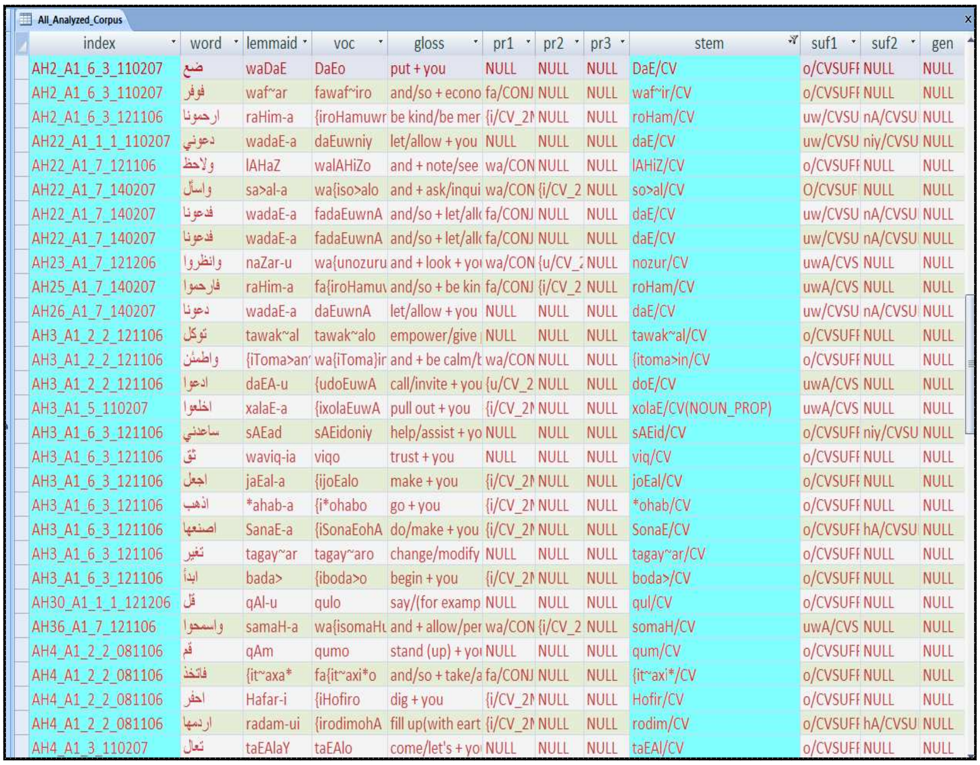
Through the analyzed ICA sample the analysts can capture any information easily. For example the analysts can capture all the imperative verbs whether in their contexts or without context as shown in figure 17 & 18. This can help in building a good search engine tool.



All_Analyzed_Corpus

index	word	lemmaid	voc	gloss	pr1	pr2	pr3	stem	suf1	suf2	gen	num
AH1_A1_6_3_140207	فيقول	qAl-u	fayaquwlu	and/so + he/it +	fa/CONJ	ya/IV3M	NULL	quwl/IV	u/IVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	أعد	>aEad~	>aEid~o	prepare + you	NULL	NULL	NULL	>aEid~/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	لنفسك	nafos	linafosika	for/to + same/s	li/PREP	NULL	NULL	nafos/NOUN	ka/POSS_	NULL	FEM	SG
AH1_A1_6_3_140207	بيتا	bayot	bayotAF	house	NULL	NULL	NULL	bayot/NOUN	NULL	NULL	MASC	SG
AH1_A1_6_3_140207	واتخذ	{it~axa*	wa{it~axi*o	and + take/adoj	wa/CON	NULL	NULL	{it~axi*/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	لنفسك	nafos	linafosika	for/to + same/s	li/PREP	NULL	NULL	nafos/NOUN	ka/POSS_	NULL	FEM	SG
AH1_A1_6_3_140207	زوجة	zawoj	zawojapF	wife	NULL	NULL	NULL	zawoj/NOUN	ap/NSUFF	NULL	FEM	SG
AH1_A1_6_3_140207	تكن	kAn-u	takuno	it/they/she + be	ta/IV3FS	NULL	NULL	kun/IV	o/IVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	بمثابة	mavAbap	bimavAbapi	by/with + virtua	bi/PREP	NULL	NULL	mavAb/NOUN	ap/NSUFF	NULL	FEM	NULL
AH1_A1_6_3_140207	سيدة	say~idap	say~idapK	lady	NULL	NULL	NULL	say~id/NOUN	ap/NSUFF	NULL	FEM	SG
AH1_A1_6_3_140207	لقلبك	qalob	liqalobika	for/to + heart/c	li/PREP	NULL	NULL	qalob/NOUN	ka/POSS_	NULL	MASC	SG
AH1_A1_6_3_140207	،	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc
AH1_A1_6_3_140207	أحب	>aHab~	>aHib~o	love/adore + yo	NULL	NULL	NULL	>aHib~/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	زوجتك	zawoj	zawojataka	wife + your	NULL	NULL	NULL	zawoj/NOUN	at/NSUFF	ka/POSS_	FEM	SG
AH1_A1_6_3_140207	وأشبع	>a$obaE	wa>a$obiEo	and + have eno	wa/CON	NULL	NULL	>a$obiE/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	جوفها	jawof	jawofahA	belly/cavity/mi	NULL	NULL	NULL	jawof/NOUN	hA/POSS_	NULL	MASC	SG
AH1_A1_6_3_140207	واستر	satar-ui	wa{usoturo	and + cover/hid	wa/CON	{u/CV_2	NULL	sotur/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	ظهرها	Zahor	ZahorahA	back/spine + its	NULL	NULL	NULL	Zahor/NOUN	hA/POSS_	NULL	MASC	SG
AH1_A1_6_3_140207	واجعل	jaEal-a	wa{ijoEalo	and + make/bri	wa/CON	{i/CV_2	NULL	joEal/CV	o/CVSUFF	NULL	NULL	NULL
AH1_A1_6_3_140207	قلبها	qalob	qalobahA	heart/center/es	NULL	NULL	NULL	qalob/NOUN	hA/POSS_	NULL	MASC	NULL
AH1_A1_6_3_140207	فرحا	fariH	fariHAF	happy/cheerful	NULL	NULL	NULL	fariH/NOUN	NULL	NULL	MASC	SG
AH1_A1_6_3_140207	ما	mA	mA	what	NULL	NULL	NULL	mA/REL_PRON	NULL	NULL	MASC	SG
AH1_A1_6_3_140207	دامت	dAm	dAmat	alast/persevere	NULL	NULL	NULL	dAm/PV	at/PVSUF	NULL	NULL	NULL
AH1_A1_6_3_140207	على	EalaY	EalaY	on/above	NULL	NULL	NULL	EalaY/PREP	NULL	NULL	NULL	NULL
AH1_A1_6_3_140207	قيد	qayod	qayodi	in progress/cur	NULL	NULL	NULL	qayod/NOUN	NULL	NULL	MASC	SG
AH1_A1_6_3_140207	الحياة	HayAp	AlHayApi	the + life	Al/DET	NULL	NULL	HayA/NOUN	ap/NSUFF	NULL	FEM	SG
AH1_A1_6_3_140207	..	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc
AH1_A1_6_3_140207	كما	kamA	kamA	and/as/also	NULL	NULL	NULL	kamA/SUB_CONJ	NULL	NULL	NULL	NULL
AH1_A1_6_3_140207	يوصي	>awoSaY	yuwSiy	he/it + recomm	yu/IV3M	NULL	NULL	wSiy/IV	(null)/IVS	NULL	NULL	NULL

Figure 17: CV within context.



All_Analyzed_Corpus

index	word	lemmaid	voc	gloss	pr1	pr2	pr3	stem	suf1	suf2	gen
AH2_A1_6_3_110207	ضع	waDaE	DaEo	put + you	NULL	NULL	NULL	DaE/CV	o/CVSUFF	NULL	NULL
AH2_A1_6_3_110207	فوفر	waf~ar	fawaf~iro	and/so + econo	fa/CONJ	NULL	NULL	waf~ir/CV	o/CVSUFF	NULL	NULL
AH2_A1_6_3_121106	ارحمونا	raHim-a	{iroHamuwn	be kind/be mer	{i/CV_2M	NULL	NULL	roHam/CV	uw/CVSU	nA/CVSU	NULL
AH22_A1_1_1_110207	دعوني	wadaE-a	daEuwniy	let/allow + you	NULL	NULL	NULL	daE/CV	uw/CVSU	niy/CVSU	NULL
AH22_A1_7_121106	ولاحظ	lAHaZ	walAHiZo	and + note/see	wa/CON	NULL	NULL	lAHiZ/CV	o/CVSUFF	NULL	NULL
AH22_A1_7_140207	واسأل	sa>al-a	wa{iso>alo	and + ask/inqui	wa/CON	{i/CV_2	NULL	so>al/CV	O/CVSUF	NULL	NULL
AH22_A1_7_140207	فدعونا	wadaE-a	fadaEuwnA	and/so + let/all	fa/CONJ	NULL	NULL	daE/CV	uw/CVSU	nA/CVSU	NULL
AH22_A1_7_140207	فدعونا	wadaE-a	fadaEuwnA	and/so + let/all	fa/CONJ	NULL	NULL	daE/CV	uw/CVSU	nA/CVSU	NULL
AH23_A1_7_121206	وانظروا	naZar-u	wa{unozuru	and + look + you	wa/CON	{u/CV_2	NULL	nozur/CV	uwA/CVS	NULL	NULL
AH25_A1_7_140207	فارحموا	raHim-a	fa{iroHamuw	and/so + be kin	fa/CONJ	{i/CV_2	NULL	roHam/CV	uwA/CVS	NULL	NULL
AH26_A1_7_140207	دعونا	wadaE-a	daEuwnA	let/allow + you	NULL	NULL	NULL	daE/CV	uw/CVSU	nA/CVSU	NULL
AH3_A1_2_2_121106	توكل	tawak~al	tawak~alo	empower/give	NULL	NULL	NULL	tawak~al/CV	o/CVSUFF	NULL	NULL
AH3_A1_2_2_121106	واطمئن	{iToma>an~	wa{iToma}ir	and + be calm/t	wa/CON	NULL	NULL	{itoma>in/CV	o/CVSUFF	NULL	NULL
AH3_A1_2_2_121106	ادعوا	daEA-u	{udoEuwA	call/invite + you	{u/CV_2	NULL	NULL	doE/CV	uwA/CVS	NULL	NULL
AH3_A1_5_110207	اخلعوا	xalaE-a	{ixolaEuwA	pull out + you	{i/CV_2M	NULL	NULL	xolaE/CV(NOUN_PROP)	uwA/CVS	NULL	NULL
AH3_A1_6_3_121106	ساعدني	sAEad	sAEidoniy	help/assist + yo	NULL	NULL	NULL	sAEid/CV	o/CVSUFF	niy/CVSU	NULL
AH3_A1_6_3_121106	ثق	waviq-ia	viqo	trust + you	NULL	NULL	NULL	viq/CV	o/CVSUFF	NULL	NULL
AH3_A1_6_3_121106	اجعل	jaEal-a	{ijoEalo	make + you	{i/CV_2M	NULL	NULL	joEal/CV	o/CVSUFF	NULL	NULL
AH3_A1_6_3_121106	اذهب	*ahab-a	{i*ohabo	go + you	{i/CV_2M	NULL	NULL	*ohab/CV	o/CVSUFF	NULL	NULL
AH3_A1_6_3_121106	اصنعها	SanaE-a	{iSonaEohA	do/make + you	{i/CV_2M	NULL	NULL	SonaE/CV	o/CVSUFF	hA/CVSU	NULL
AH3_A1_6_3_121106	تغير	tagay~ar	tagay~aro	change/modify	NULL	NULL	NULL	tagay~ar/CV	o/CVSUFF	NULL	NULL
AH3_A1_6_3_121106	ابدأ	bada>	{iboda>o	begin + you	{i/CV_2M	NULL	NULL	boda>/CV	o/CVSUFF	NULL	NULL
AH30_A1_1_1_121206	قل	qAl-u	qulo	say/(for examp	NULL	NULL	NULL	qul/CV	o/CVSUFF	NULL	NULL
AH36_A1_7_121106	واسمحوا	samaH-a	wa{isomaHu	and + allow/per	wa/CON	{i/CV_2	NULL	somaH/CV	uwA/CVS	NULL	NULL
AH4_A1_2_2_081106	قم	qAm	qumo	stand (up) + you	NULL	NULL	NULL	qum/CV	o/CVSUFF	NULL	NULL
AH4_A1_2_2_081106	فاتخذ	{it~axa*	fa{it~axi*o	and/so + take/a	fa/CONJ	NULL	NULL	{it~axi*/CV	o/CVSUFF	NULL	NULL
AH4_A1_2_2_081106	احفر	Hafar-i	{iHofiro	dig + you	{i/CV_2M	NULL	NULL	Hofir/CV	o/CVSUFF	NULL	NULL
AH4_A1_2_2_081106	اردمها	radam-ui	{irodimohA	fill up(with eart	{i/CV_2M	NULL	NULL	rodim/CV	o/CVSUFF	hA/CVSU	NULL
AH4_A1_3_110207	تعال	taEAlaY	taEAlo	come/let's + you	NULL	NULL	NULL	taEAl/CV	o/CVSUFF	NULL	NULL

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:

Abdelali A. (2004), **Localization in Modern Standard Arabic**, Journal of the American Society for Information Science and technology (JASIST), Volume 55, Number 1, 2004. pp. 23-28.

Al-Sulaiti L. & Atwell E. (2001), **Extending the Corpus of Contemporary Arabic**, School of Computing, University of Leeds.

Attia M. (2000), **A large-scale computational processor of the Arabic morphology and applications**, Faculty of engineering, Cairo university.

Attia M. (2006), **An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks** , School of Informatics, The University of Manchester.

Beesley K. (1996), **Arabic finite-state morphological analysis and generation**, In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pages volume 1, 89–94, Copenhagen, Denmark.

Beesley K. (1998a.), **Arabic morphology using only finite-state operations**, **Computational Approaches to Semitic Languages**, Proceedings of the Workshop, pages 50–57, Montr´eal, Qu´ebec, August 16. Universit´e de Montr´eal.

Beesley K. (1998b.), **Arabic Linguistic Society**, Paper presented at the 12th Symposium on Arabic Linguistics, 6-7 March, Champaign, IL.

Buckwalter T. (2002), **Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium**, University of Pennsylvania, LDC Catalog No.: LDC2002L49.

Choukri K., Krawner S. (2004), **Arabic Language Resources and Tools**, Nemlar.

Choukri K., Krawner S., Maegaard B., The BLARK (2006), **concept and BLARK for Arabic**, Proceedings of the 5th International Conference on Language Resources and Evaluation. Genova.

Darwish K. (2002), **Building a Shallow Morphological Analyzer in One Day,** In Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA.

Dichy J. & Fargaly A. (2003), **Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?**, Proceedings of the MTSummit IX workshop on Machine Translation for Semitic Languages, New-Orleans.

Eriksson T. & Ritchey T. (2002), **Scenario Development using Computerised Morphological analysis**, Presented at the Winchester International OR Conference, England.

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.

Hajič O. & et al (2006), **THE CHALLENGE OF ARABIC FOR NLP/MT, Tips and Tricks of the Prague Arabic Dependency Treebank**, International Conference at The British Computer Society (BCS), 23 October, London.

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

Hockett C., 1947, **problems of morphemic analysis** , Linguistic Society of America, Language, Vol. 23, No. 4 (Oct. - Dec., 1947), pp. 321-343.

Hulstijg J. (1992), **Retention of inferred and given word meanings: experiments in incidental vocabulary learning**, In P.J.L Arnaud and H.bejoint (eds), vocabulary and applied linguistics. London: Macmillan, 113-25.

Kaplan J. & Holland V. (1995), **Natural language processing techniques in computer assisted language learning: status and instructional issues**, Springer, Instructional Science. 23,351-80.

Karttunen L. (2005), **Twenty-five years of finite-state morphology**, CSLI Publications.

Karttunen, Kaplan R., & Zaenen A. (1992), **Two-level morphology with composition**, In Proceedings of Fourteenth International Conference on Computational Linguistics (COLING-92), pages 141–148, Nantes, July 20–28, France.

Kiraz G.(1994), **Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology**, In Proceedings of Fifteenth International Conference on Computational Linguistics (COLING-94), pages 180–186, Kyoto, Japan.

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.

Landauer T., Foltz P., & Laham D. (1998), **Introduction to Latent Semantic Analysis**., Discourse Processes, 25, 259-284.

Lee Y. (2004), **Morphological Analysis for Statistical Machine Translation**, IBM T. J. Watson Research Center, Yorktown Heights, NY-10598.

Maamouri M., Bies A., Buckwalter T. & Mekki W. (2004), **The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus**, NEMLAR Conference on Arabic Language Resources and Tools.

Manning C. & Schütze H. (1999), **Foundations of Statistical Natural Language Processing**, MIT Press, Cambridge, Massachusetts.

Meyer C. (2002), **English corpus linguistics, an introduction**, Cambridge University Press.

Nerbonne J., Jager S. & Essen A. (1997), **Language Teaching and Language Technology**, the University of Groningen, April 28-29, 1997.

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 Societ, Adapted from the paper "Fritz Zwicky, Morphology and Policy Analysis".

Ritchey T. (2005-2008), **Wicked Problems, Structuring Social Messes with Morphological Analysis**, Swedish Morphological Society.

Ritchey, T. (1998), **General Morphological Analysis, A general method for non-quantified modeling**, "Fritz Zwicky, **'Morphologie' and Policy Analysis**", Presented at the 16th Euro Conference on Operational Analysis, Brussels.

Soudi A., Cavalli-Sforza V., & Jamari A. (2001), **A Computational Lexeme-Based Treatment of Arabic Morphology**, Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001), Jul 6, Toulouse, France.

Soudi A., Bosch A. & Neumann G. (2007), **Arabic Computational Morphology, Knowledge-based and Empirical Methods**, Springer.

Swaab T. & Kaan E. (2003), **Repair, Revision, and, Complexity in Syntactic Analysis: An Electrophysiological Differentiation**, The MIT Press, Journal of Cognitive Neuroscience.

ZAUGUAGE S, Varga D. (1955), **Syntactic analysis in the case of highly inflecting languages, international conference on computational linguistics**, Computing Centre of the Hungarian Academy of Sciences, 53, Uri u., Budapest I., Hungary.A1

Zemanek P. (2001), **CLARA (Corpus Linguae Arabicae): An Overview**, Proceedings of ACL/EACL Workshop on Arabic Language.