# The International Corpus of Arabic: Compilation, Analysis and Evaluation

**Sameh Alansary**
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.
Sameh.alansary@bibalex.org

**Magdy Nagi**
Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.
Computer and System Engineering Dept. Faculty of Engineering, Alexandria University,
Alexandria Egypt.
Magdy.nagi@bibalex.org

## Abstract

This paper focuses on a project for building the first International Corpus of Arabic (ICA). It is planned to contain 100 million analyzed tokens with an interface which allows users to interact with the corpus data in a number of ways [ICA website]. ICA is a representative corpus of Arabic that has been initiated in 2006, it is intended to cover the Modern Standard Arabic (MSA) language as being used all over the Arab world. ICA has been analyzed by Bibliotheca Alexandrina Morphological Analysis Enhancer (BAMAE). BAMAE is based on Buckwalter Arabic Morphological Analyzer (BAMA). Precision and Recall are the evaluation measures used to evaluate the BAMAE system. At this point, Precision measurement ranges from 95%-92% while recall measurement was 92%-89%. This depends on the number of qualifiers retrieved for every word. The percentages are expected to rise by implementing the improvements while working on larger amounts of data.

## 1. Introduction

Arabic is the largest member of the Semitic language family, most closely related to Aramaic, Hebrew, Ugaritic and Phoenician. Arabic is one of the six official languages of the United Nations[1] and it is the main language of most of the Middle East countries. Arabic ranks fifth in the world's league table of languages, with an estimated 206 million native speakers, 24 million as 2nd language speakers to add up to total of 233 million and World Almanac estimates the total speakers as 255 million. Arabic language is the official language in all Arab nations as Egypt, Saudi Arabia and Algeria. Moreover, it is also an official language in non-Arab countries as Israel, Chad and Eritrea. It is also spoken as a 2nd language in other non-Arab countries as Mali and Turkey[2].

The formal Arabic language, known as Classical Arabic is the language in which the Qur'an is written and is considered to be the base of the syntactic and grammatical norms of the Arabic language. However, today it is considered more of a written language than a spoken one. Modern Standard Arabic (MSA) is similar to Classical Arabic, but it is an easier form. It is understood across the Arab world and it is used by television presenters and politicians, it is the form used to teach Arabic as a foreign language. There are different MSA varieties as the rate of similarity between every Arab country version of MSA and Classical Arabic differs. This is one of the issues that this paper will present.

Due to the fact that the need for Arabic corpus is increasing as well as the fact that the trials to build an Arabic corpus in the last few years were not enough to consider that the Arabic language has a real, representative and reliable corpus, it was necessary to build such an Arabic corpus that is capable of supporting various linguistic research in Arabic. Thus, ICA was inspired by the difficulties that encountered Arabic Language researches as a result of the lack of publicly available Arabic corpora.

Bibliotheca Alexandrina (BA) has initiated a big project to build the "International Corpus of Arabic (ICA)", a real trial to build a representative Arabic corpus as being used all over the Arab world to support research in Arabic. The International Corpus of Arabic is planned to contain 100 million words. The collection of samples is limited to written Modern Standard Arabic, selected from a wide range of sources and designed to represent a wide cross-section of Arabic; it is stimulating the first systematic investi-

[1] http://www.un.org/en/aboutun/languages.shtml

[2] http://www.vistawide.com/languages/top_30_languages.htm

*Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 8–17,
October 25, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

gation of the national varieties as being used all over the Arab world (Alansary, et al. 2007).

There were some trials for building Arabic corpora. Some of them were annotated corpora and others were raw texts corpora. Annotated corpora trails as Penn Arabic Treebank (PATB). The LDC was sponsored to develop an Arabic POS and Treebank of only 1,000,000 words. This corpus doesn't contain any branched genres except 600 stories from the ANNAHAR News Agency. The POS only annotated version of this ANNAHAR corpus was released in 2004[3.] The output from Buckwalter's Arabic Morphological Analyzer is used as the starting point for the morphological annotation and POS tagging of Arabic newswire text (Maamouri M., 2004).

Arabic Gigaword Corpus is an archive of newswire text data that depends on press only; it has been compiled from Arabic news sources by LDC[4]. The data coverage is limited, it was compiled from Egypt, Lebanon, Tunisia, Saudi Arabia and from outside the Arab world such as England. NEMLAR Annotated Written Corpus consists of about only 500, 000 words of Arabic text from 13 different categories, aiming to achieve a well-balanced corpus that offers a representation of the variety in syntactic, semantic and pragmatic features of modern Arabic language[5]. The accuracy of the automatic analysis is around 95% (Atiyya M. et al, 2005). Its analysis features are limited, moreover its use is restricted; it is not accessible for commercial use[6.]

KALIMAT is a free multipurpose Arabic corpus, consists of 18,167,183 annotated words representing 20,291 Arabic articles collected only from the Omani newspaper Alwatan. A morphological analysis process on the data collection using AL Khalil[7] morphological analyser was conducted to reach an accuracy of 96% (El-Haj M., 2013). Prague Arabic Dependency Treebank (PADT) version 1.0 distribution comprises over 113,500 tokens of data annotated analytically and provided with the disambiguating morphological information. In addition, the release includes complete annotations of MorphoTrees resulting in more than 148,000 tokens, 49,000 of which have received the analytical processing[8].

The raw text corpora trails as (KACST) King Abdul-Aziz City for Science and Technology Corpus[9] contains 732,780,509 words representing 869,800 text files and 7,464,396 distinct words. It contains a lot of classical Arabic texts; however, it is neither analyzed nor well planned. AraбiCorpus[10] is a corpus that was developed by Dilworth Parkinson. It is a large corpus that could be accessed, but it is not analyzed. Words can be searched for in Arabic or Latin script. The website provides detailed instructions on the search. It contains 173,600,000 words in five main categories or genres: Newspapers, Modern Literature, Nonfiction, Egyptian Colloquial, and Premodern.

In what follows, Section 2 reviews the ICA data design, how it is compiled, discuss the copyrights issue and what is the current ICA statistics. Section 3 describes the analysis stage of ICA, the tool that is used in the analysis, why was it chosen followed by ICA evaluation and a comparison with another morphological disambiguator. Section 4 gives a brief review on the ICA website for the researchers to query its data. Conclusions and suggestions for further work are given in section 5.

## 2. ICA Design & Compilation Stage

The ICA is similar to the International Corpus of English (ICE) in terms of concept rather than in design. They are similar in trying to include the varieties of the language; the Modern Standard Arabic (MSA) includes publications from every Arab country that uses Arabic as official language and it has been decided to include Arabic publications from outside the Arab nations. However, they are different in terms of corpus design criteria and data compilation. For example, on the one hand, Egyptian Modern Standard Arabic is the most widespread variety that is used to represent MSA in ICA corpus. On the other hand, in building ICE[11] a fixed size from each variation was taken from any country that uses English as official language (one million words); however, balance in size does not always mean fixing a number of words for each variation as will be clarified in the next section.

[3] https://catalog.ldc.upenn.edu/LDC2005T20

[4] https://catalog.ldc.upenn.edu/LDC2003T12

[5] http://catalog.elra.info/product_info.php?products_id=873

[6] http://catalog.elra.info/product_info.php?products_id=873

[7] http://alkhalil-morpho-sys.soft112.com/

[8] https://catalog.ldc.upenn.edu/LDC2004T23

[9] http://www.kacstac.org.sa/Pages/default.aspx

[10] http://arabicorpus.byu.edu/search.php

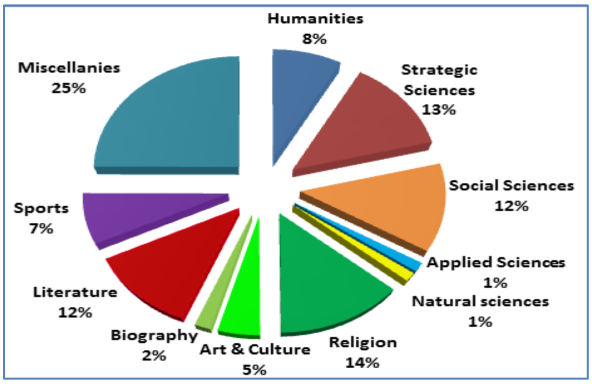
[11] http://ice-corpora.net/ICE/INDEX.HTM

It is important to realize that the creation of ICA is a "cyclical" process, requiring constant reevaluation during the corpus compilation. Consequently, we are willing to change our initial corpus design if there are any circumstances would arise that requires such changes.

## 2.1 ICA Design

ICA genre design relied on Dewey decimal classification of documents; however, this has been further classified to suit clear genre distinction rather than classifications for libraries. For example, Dewey decimal classification combines history and geography in one classification, while in ICA they are separated into two sub genres related to humanities genre. It has been designed to reflect a more or less real picture of how Arabic language exists in every field and in every country rather than relying on a theoretical image.

ICA is designed to include 11 genres, namely; Strategic Sciences, Social Sciences, Sports, Religion, Literature, Humanities, Natural Sciences, Applied Sciences, Art, Biography and Miscellaneous which are further classified into 24 sub-genres, namely; Politics, Law, Economy, Sociology, Islamic, Pros etc. Moreover, there are 4 sub-sub-genres, namely; Novels, Short Stories, Child Stories and plays. As shown in Figure 1.



“Figure 1: ICA Genres”

Planning of ICA data collection is based on some criteria related to corpus design such as representativeness, diversity, balance and size that were taken into the consideration. In collecting a corpus that represents the Arabic Language, the main focus was to cover the same genres from different sources and from all around the Arab nations. However, we decided to add Arabic data that belongs to the Arabic language even if they had been published outside as al-Hayat magazine which is published in London[12.]

Size criterion in the corpus design focuses on the number of words. However, issues of size are also related to the number of texts from different genres, the number of samples from each text, and the number of words in each sample. Such decisions were taken based on how common the genre or the source is. Balance in a corpus has not been addressed by having equal amounts of texts from different sources or genres. It has been addressed by the factual distribution of the language real use. For example, Literature genre represents 12% and biography genre represents 2% from the corpus data distribution.

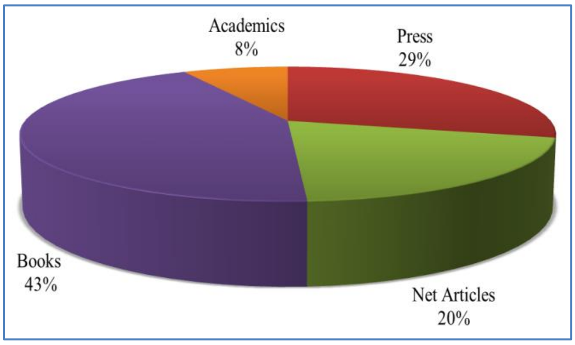
## 2.2 Text Compilation and Categorization

The International Corpus of Arabic has been compiled manually, and that enabled the corpus compilers to select all and only the MSA data rather than the colloquial Arabic data. Also, the ICA text categorization has been done manually according to the topic of the text and the distinct semantic features for each genre. These features keep the ICA data categorization objective rather than being subjective; depending on the compiler intuition. Accordingly, ICA texts can be considered as a good training data for text categorization system. ICA is planned to contain 100 million words. However, currently it is still around 80 million words.

ICA data is composed of Modern Standard Arabic (MSA) written texts. There are different resources for compiling the data. It has been decided to compile all available Arabic data written in MSA. ICA will be composed of four sources, namely; 1. Press source which is divided into three sub-sources, namely; (a) Newspapers, (b) Magazines which had been compiled from the official magazines along with newspapers that are written in MSA such as Al Ahram from Egypt, Addstour from Jordan, Al Hayat from Lebanon … etc. finally the publications that have a printed copy as well as a soft electronic copy through world wide web such as (http://www.ahram.org.eg/), and (c) Electronic Press which had been compiled from magazines and newspapers that are written in MSA and have only soft electronic copy through world wide web. (2) Net articles which were compiled from forums and blogs that are also written in MSA. (3) Books which had been compiled from

[12] http://alhayat.com/AboutWebsite

all available books that are written in MSA and have a soft copy. (4) Academics which had been compiled from the scientific papers, researchers thesis, PhDs etc..



“Figure 2: ICA Sources”

## 2.3 Metadata

Each compiled text has its own text encoding. This coding process for the text file names will customize the search scope at which level of the corpus this file belongs. For example, the following filename coding [AH10-A1.1.1_140207] can be clarified as shown in Table1:

| | |
|---|---|
| **AH10** | AH: Indicate the source of the text which is Ahram newspaper.<br>10: This attached number that indicates that this file is the 10th article in that newspaper with the same genre, subgenre and date. |
| **A1.1.1** | Contains three pieces of information: Newspaper source (A1), Strategic science "genre" (A1.1) and Politics "sub-genre" (A1.1.1). |
| **140207** | Contains three pieces of issuing information: The day (14), the month (02) and the year (2007). |

“Table 1: An example of filenames coding”

ICA Metadata covers the needed information related to Corpus for each compiled text as data source providers, Text code name, Text size, Website, date of publishing, publisher (name and country), writer (name, gender, age, nationality and educational level) and Collection/Annotation Specifications.

## 2.4 Copyrights

One of the serious constraints on developing large corpora and their widespread use is national and international copyright legalizations. According to copyright laws, it is necessary and sensible to protect the authors as well as the publishers rights of the texts that they had produced. ICA data Copy rights and publishing issues are in progress by Bibliotheca Alexandrina Legal Affairs. For that reason, the ICA data is not available to be downloaded but the researchers can search the ICA data via the ICA website[13.]

## 2.5 ICA statistics

Corpus analysis is both qualitative and quantitative. One of the advantages of corpora is that they can readily provide quantitative data which intuitions cannot provide reliably. The use of quantification in corpus linguistics typically goes well beyond simple counting.

Table 2 shows some of the numbers of ICA data coverage. It must be noted that total number of “Tokens” refers to all word forms except numbers, foreign words and punctuations to reflect the real size of the used word forms before the analysis stage. Coverage interval starts from 1993 up to 2014; however, there is a compilation problems as result of the data availability since the size of the data was not equal throughout the years. Balance is considered as an issue for the ICA current situation. It deals with the coverage of texts over the years rather than balance according to time span and that will remain as issue in the future.

| Statistics | Total Number |
|---|---|
| No. of texts | 70,022 |
| No. of words | 79,569,384 |
| No. of Tokens | 76,199,414 |
| No. of unique words | 1,272,766 |
| No. of ICA sources | 4 |
| No. of sub sources | 3 |
| No. of genres | 11 |
| No. of sub genres | 24 |
| No. of sub sub-genres | 4 |
| No. of countries | 20 |
| No. of covered years | 22 |
| No. of writers | 1021 |

“Table 2 : Shows qualitative linguistic analysis for ICA statistics”

[13] http://www.bibalex.org/ica/ar/

## 3. ICA Analysis stage

The first stage of linguistic analysis of the International corpus of Arabic is to analyze the 100 million words morphologically.

The stem-based approach "concatenative approach" has been adopted as the linguistic approach. There are many morphological analyzers for Arabic; some of them are available for research and evaluation while the rest are proprietary commercial applications. Buckwalter Arabic Morphological Analyzer (Buckwalter, 2004) is a well-known analyzer in the field`s literature and has even been considered as the "most respected lexical resource of its kind" (Hajič et al, 2005). It is used in LDC Arabic POS-tagger, Penn Arabic Dependency Treebank, and the Prague Arabic Dependency Treebank. It is designed to consist of a main database of word forms that interact with other concatenation databases. Every word form is entered separately, and the stem is used as the base form. The word is viewed as to be composed of a basic unit that can be combined with morphemes governed by morph tactic rules. It makes use of three lexicons: a Prefixes lexicon, a Stem lexicon, and a Suffixes lexicon.

Buckwalter Arabic Morphological Analyzer (BAMA) has been selected since it was the most suitable lexical resource to our approach. (Alansary, et al. 2008).

Although it has many advantages including its ability to provide a lot of information such as Lemma, Vocalization, Part of Speech (POS), Gloss, Prefix(s), Stem, Word class, Suffix(s), Number, Gender, Definiteness and Case or Mood, it does not always provide all the information that the ICA requires, and in some cases, the provided analyses would need some modification. Its results may give the right solution for the Arabic input word, provide more than one result that needs to be disambiguated to reach the best solution, provide many solutions but none of them is right, segment the input words wrongly without taking the segmentation rules in consideration or provide no solutions. Consequently, solutions enhancement is needed in these situations.

Number, gender and definiteness need to be modified according to their morphosyntactic properties. Some tags had been added to Buckwalter's analyzer lexicon, some lemmas, glossaries had been modified and others had been added. In addition, new analysis and qualifiers had been added as root, stem pattern and name entities. (Alansary, et al. 2008)

Due to all these modifications, there are some clear differences between the tool adopted by ICA and BAMA 2.0 as:

- There are 44,756 distinct lemmas in ICA lexicon while they are 40,654 in BAMA 2.0.
- The root feature has been added to ICA lexicon representing 3,451 distinct roots, the pattern feature has been added to ICA lexicon representing 782 distinct stem patterns and they will be increased to cover all Arabic roots.
- There are 191 distinct tags in ICA while they are 167 in BAMA 2.0. Table 3 shows some tags that have been added to ICA lexicon that are not found in BAMA:

| Tag | Description |
|---|---|
| NOUN(ADV_M) | Adverb of Manner |
| NOUN(ADV_T) | Adverb of Time |
| NOUN(ADV_P) | Adverb of Place |
| NOUN(VERBAL) | Verbal noun |
| NOUN_PROP(ADV_T) | Proper nouns that refer to adverb of time |
| NOUN(INTERJ) | The vocative nouns |

"Table 3: Added Tags in ICA lexicon"

- Table 4 shows some tags that are added to prefixes and suffixes:

| Sample of Added Prefixes and suffixes | |
|---|---|
| CV_SUBJ:2FP | Prefixes |
| CV_SUBJ:2FS | |
| CV_SUBJ:2MP | |
| CV_SUBJ:2MS | |
| wa/PREP | |
| la/PREP | |
| >a/INTERROG_PART | |
| hAt/NSUFF | Suffixes |
| NSUFF_SUBJ:2MS | |
| CVSUFF_SUBJ:2MD | |
| CVSUFF_SUBJ:2FP | |
| CVSUFF_DO:3FS | |
| CVSUFF_DO:3FS | |

"Table 4: Sample of added prefixes and suffixes."

Moreover, new features have been added in number as well as in definiteness qualifiers as the plural broken (PL_BR) and the EDAFAH features.

These modifications and other new features were used in disambiguating two million words to be used as a training data extracted from the

ICA corpus to represent a sample of Arabic texts. After disambiguating the training date, some linguistic rules had been extracted, depending on the contexts, to help in the automatic disambiguation process of Bibliotheca Alexandrina Morphological Analysis Enhancer (BAMAE) as will be discusses in the next section.

After solving the BAMA's problems and disambiguating the data according to its context, the BAMA enhanced output along with the training data will be ready to be used in the next phase of analysis.

In the ICA, There are 5 tag sets categories of the stem which are divided into 26 tag types:

1. Verbal category: it contains 5 tag types; Command Verb, Imperfect Verb, Imperfect Passive Verb, Past Verb and Past Passive Verb.
2. Nominal category: it contains 9 tag types; Adjective, Noun, Adverb of Manner, Adverb of Place, Adverb of Time, Verbal Noun, Proper Noun, Proper Noun (Adverb of Time) and Proper Noun (Interjection).
3. Pronouns category: it contains 3 tag types; Demonstrative Pronoun, Pronoun and Relative Pronoun.
4. Particles category: it contains 7 tags; Focus Particle, Future Particle, Interrogative Particle, Negative Particle, Particle, Verbal Particle and Exception Particle.
5. Conjunctions category: it contains 2 tags; Conjunctions and Sub Conjunctions.

In addition, there are 2 tags that are not divided into any types; Preposition and Interjection tags.

Some words were found to have no solution for one of three reasons. First, some words are not analyzed altogether by BAMA; second, some words are analyzed, but none of the provided solutions is suitable to their contexts in the text; third, some words are wrongly segmented by BAMA. Consequently, 15,605 words have been analyzed manually in the same manner they would have been analyzed automatically.

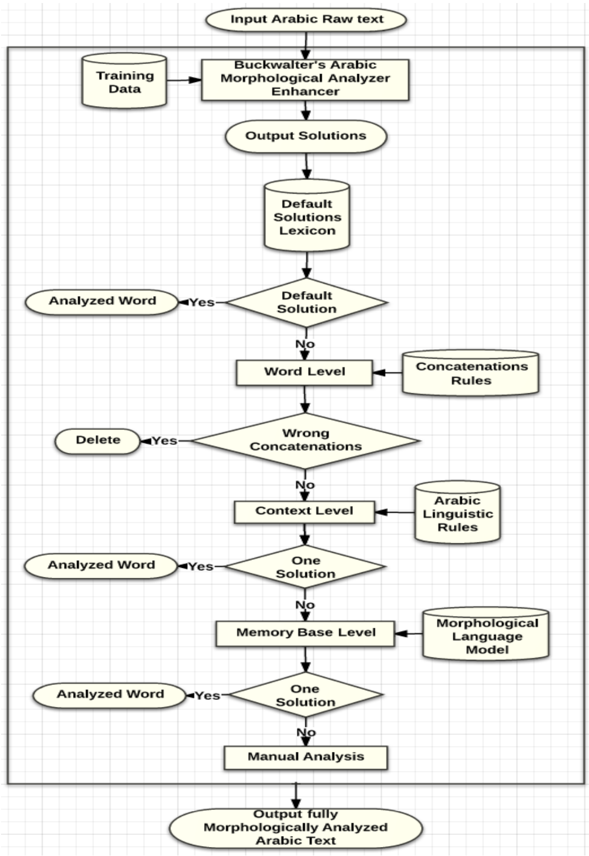
### 3.1 Bibliotheca Alexandrina Morphological Analysis Enhancer (BAMAE)

It is a system that has been built to morphologically analyze and disambiguate the Arabic texts depending on BAMA's enhanced output of the ICA. It was preferred to use BAMA's enhanced output of the ICA since it contains more information than any other system of BAMA's output. This is the reason that made the members of the ICA team aim to build their own morphological disambiguator (BAMAE).

In order to reach the best solution for the input word, BAMAE preforms automatic disambiguation process carried on three levels that depends primarily on the basic POS information (Prefix(s), Stem, Tag and Suffixes) that is obtained from the enhanced BAMA's output. (Alansary, 2012):

- Word level which avoids or eliminates the impossible solutions that Buckwalter provides due to the wrong concatenations of prefix(s), stem and suffix(s).
- Context level where some linguistic rules have been extracted from the training data to help in disambiguating words depending on their context.
- Memory based level which is not applicable in all cases; it is only applicable when all the previous levels fail to decide the best solution for the Arabic input word.

Figure 3 shows BAMAE architecture starting from the input text and the numerous solutions for each word in order to predict the best POS solution for each word.

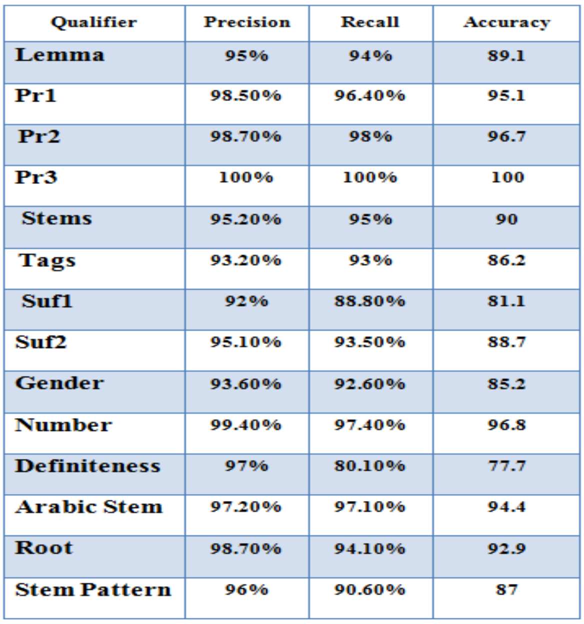


"Figure 3: BAMAE Architecture."

After selecting the best POS solution for each word, BAMAE detects the rest of information accordingly. It detects the lemmas, roots (depending primarily on the lemmas), stem patterns (depending on stems, roots and lemmas), number (depending on basic POS and stem patterns), gender (depending also on basic POS, stem patterns and sometimes depending on number), definiteness (depending on POS or their sequences), case (depending on definiteness and sequences of POS) and finally it detects the vocalization of each word.

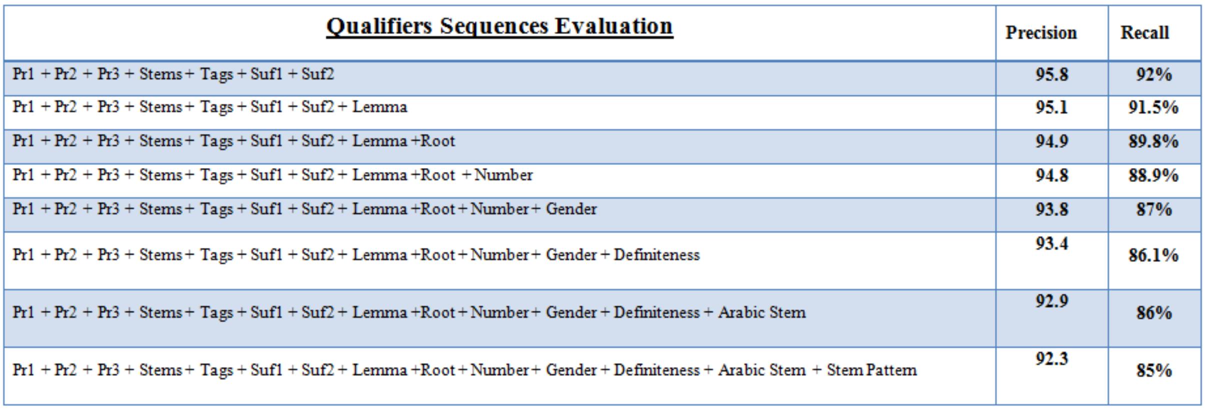
### 3.2 ICA Analysis Evaluation

The testing data has been evaluated based on the rules extracted from the manually disambiguated training data in order to determine the strengths and weaknesses of the enhancer module in reaching the best solution. The testing data set will contain 1,000,000 representative words that were manually analyzed specially for the testing stage. Precision and Recall are the evaluation measures used to evaluate the BAMAE system. Precision is a measure of the ability of a system to present only relevant results. Recall is a measure of the ability of a system to present all relevant results. The evaluation has been conducted on two levels; the first level includes the precision, recall and accuracy for each qualifier separately as shown in table 5. The second level includes the basic POS in addition to adding a new qualifier each time to investigative how it would affect the accuracy as shown in table 6.



| Qualifier | Precision | Recall | Accuracy |
|---|---|---|---|
| Lemma | 95% | 94% | 89.1 |
| Pr1 | 98.50% | 96.40% | 95.1 |
| Pr2 | 98.70% | 98% | 96.7 |
| Pr3 | 100% | 100% | 100 |
| Stems | 95.20% | 95% | 90 |
| Tags | 93.20% | 93% | 86.2 |
| Suf1 | 92% | 88.80% | 81.1 |
| Suf2 | 95.10% | 93.50% | 88.7 |
| Gender | 93.60% | 92.60% | 85.2 |
| Number | 99.40% | 97.40% | 96.8 |
| Definiteness | 97% | 80.10% | 77.7 |
| Arabic Stem | 97.20% | 97.10% | 94.4 |
| Root | 98.70% | 94.10% | 92.9 |
| Stem Pattern | 96% | 90.60% | 87 |

"Table 5: Precision, Recall and Accuracy for each qualifier"



| Qualifiers Sequences Evaluation | Precision | Recall |
|---|---|---|
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 | 95.8 | 92% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma | 95.1 | 91.5% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root | 94.9 | 89.8% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root + Number | 94.8 | 88.9% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root + Number + Gender | 93.8 | 87% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root + Number + Gender + Definiteness | 93.4 | 86.1% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root + Number + Gender + Definiteness + Arabic Stem | 92.9 | 86% |
| Pr1 + Pr2 + Pr3 + Stems + Tags + Suf1 + Suf2 + Lemma +Root + Number + Gender + Definiteness + Arabic Stem + Stem Pattern | 92.3 | 85% |

"Table 6: Accuracy decreasing as a result of adding new qualifier each time to the main POS Tag"

### 3.3 Comparing BAMAE with MADA

MADA (Morphological Analysis and Disambiguation for Arabic) is selected to be compared with BAMAE since both of them uses Buckwalter's output analyses to help in disambiguating the Arabic texts. The primary purpose of MADA 3.2 is to extract linguistic information as much as possible about each word in the text, from given raw Arabic text, in order to reduce or eliminate any ambiguity concerning the word. MADA does this by using ALMORGEANA[14] (an Arabic lexeme-based morphology analyzer) to generate every possible interpretation of each input word. Then, MADA  applies a number of language models to determine which analysis is the most probable for each word, given the word's context.

MADA makes use of up to 19 orthogonal features to select, for each word, a proper analysis from a list of potential analyses that are provided

[14] http://clipdemos.umiacs.umd.edu/ALMORGEANA/

by the Buckwalter Arabic Morphological Analyzer (BAMA; Buckwalter 2004). The BAMA analysis that most closely matches the collection of weighted, predicted features, is chosen. The 19 features include 14 morphological features that MADA predicts using 14 distinct Support Vector Machines (SVMs) trained on the PATB. In addition, MADA uses five features that capture information such as spelling variations and n-gram statistics.

Since MADA selects a complete analysis from BAMA, all decisions regarding morphological ambiguity, lexical ambiguity, tokenization, diacritization and POS tagging in any possible POS tag set are made in one fell swoop (Habash and Rambow, 2005; Habash and Rambow 2007; Roth et al, 2008). The choices are ranked in terms of their score. MADA has over 96% accuracy on basic morphological choice (including tokenization but excluding case, mood, and nunation) and on lemmatization. MADA has over 86% accuracy in predicting full diacritization (including case and mood). Detailed comparative evaluations are provided in the following publications: (Habash and Rambow, 2005; Habash and Rambow 2007; Roth et al, 2008).

In order to compare between BAMAE and MADA, the selected text, to be run on both systems, was selected from the ICA training data to facilitate the comparing process. To make the comparing process more accurate, some justifications were needed in MADA to be compatible with BAMAE format. For example, in number qualifier the feature of singular (s) was handled to be (SG), in case qualifier the feature of nominative (u) was handled to be (NOM), in tags qualifier the verbs were handled with relation to aspect and stem category. The comparing process will be done in terms of some qualifiers; diacritization, tags, stems, number, gender and definiteness including Arabic words only as shown in Table 7:

| Qualifier | BAMAE | MADA |
|---|---|---|
| **Diacritization** | 89.61% | 78.78% |
| **Tags** | 93.94% | 85.28% |
| **Stems** | 96.97% | 91.34% |
| **Number** | 96.10% | 64.93% |
| **Gender** | 96.53% | 66.67% |
| **Definiteness** | 96.53% | 60.61% |

“Table 7: Comparing between MADA and BAMAE.”

There are some notes that must be taken into consideration:

- The problems of detecting the diacritization in BAMAE are related to either predicting the case ending wrongly or predicting the whole solution wrongly.
- The problems of detecting the diacritization in MADA are related to predicting the case ending wrongly, predicting the whole solution wrongly, missing some diacritics in some words, or missing all diacritics in some words.
- The problems of detecting the tags in MADA are related to either predicting the tags wrongly or the differences in some tags from those of BAMAE. For example the adverbs of time or place in BAMAE are assigned with ‘NOUN (ADV_T)’ or ‘NOUN (ADV_P)’ in BAMAE while they are assigned with ‘NOUN’, sub conjunction ‘SUB_CONJ’, and preposition ‘PREP’. This happens as a result of using BAMA’s output without enhancing such tags. In addition the wrong concatenations of BAMA’s output cause problems in detecting some tags.
- The problems of detecting stems in both BAMAE and MADA are related to predicting the solution wrongly.
- The problem of detecting number, gender and definiteness in MADA are related to using BAMA’s output without regarding morphosyntactic properties.
- The comparison between cases in BAMAE and MADA can’t be done since MADA assigns case without regarding the diacritics of this case. For example, it assigns the accusative case ‘ACC’ for both ‘a/ACC’ and ‘i/ACC’ in BAMAE.
- There are some qualifiers in BAMAE which are not found in MADA; Root and Stem Pattern. The root qualifier has been assigned with accuracy 99.45% while the stem pattern qualifier has been assigned with accuracy 94.34%.
- The lemma qualifier has been assigned in BAMAE with accuracy 96.54%, while it is does not existed in MADA.

## 4. ICA Website[15]

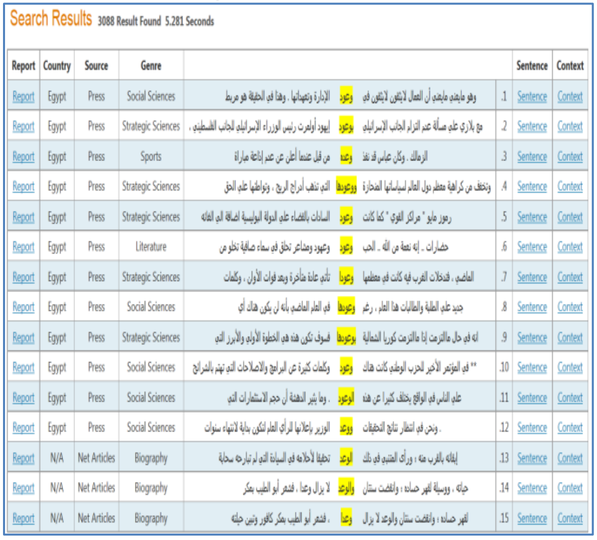
It is an interface that allows users to interact with the corpus data in a number of ways. The interface provides four options of searching the corpus content; namely, Exact Match Search, Lemma Based Search, Root Based Search and Stem Based Search.

More search options are available; namely, Word Class and Sub Class, Stem Pattern, Num-

[15] http://www.bibalex.org/ica/en/

ber, Definiteness, Gender, Country (Advanced search). Moreover, the scope of search may include the whole corpus, Source(s), Sub-Source(s), Genre(s), Sub-Sub-Genre(s) or Sub-Genre(s).

Figure 4 presents an example of a query of the analyzed data that states: when the word ‘وعد’ is searched for using a Lemma-Based search option, the system will highlight all possible lemmas that the word may have, since Arabic is orthographically ambiguous. In this example, the system will highlight several possible lemmas; ‘waʕada’ ‘to promise’, ‘waʕd’ ‘Promise’ and ‘ʕaada’ ‘return’. If the lemma ‘waʕd’ ‘Promise’ is chosen the output search in this case will include all words that have this lemma such as ‘وعود’ ‘Promises’, ‘alwaʕd’…etc. with all possible word forms together with concordance lines.



Search Results 3088 Result Found 5.281 Seconds

| Report | Country | Source | Genre | | | | | Sentence | Context |
|---|---|---|---|---|---|---|---|---|---|
| Report | Egypt | Press | Social Sciences | الإدارة وتعهداتها . وهنا في الحقيقة هو مربط | وعود | وهو مايعني مايعني أن العمال لايثقون لايثقون في | .1 | Sentence | Context |
| Report | Egypt | Press | Strategic Sciences | إيهود أولمرت رئيس الوزراء الإسرائيلي للجانب الفلسطيني . | بوعود | مع بلازي على مسألة عدم التزام الجانب الإسرائيلي | .2 | Sentence | Context |
| Report | Egypt | Press | Sports | من قبل عندما أعلن عن عدم إذاعة مباراة | وعده | الزمالك . وكان عباس قد نفذ | .3 | Sentence | Context |
| Report | Egypt | Press | Strategic Sciences | التي تذهب أدراج الريح ، وتواطئها على الحق | ووعودها | وتخفف من كراهية معظم دول العالم لسياساتها المنحازة | .4 | Sentence | Context |
| Report | Egypt | Press | Strategic Sciences | السادات بالقضاء على الدولة البوليسية اضافة الى الغائه | وعود | رموز مايو " مراكز القوى " كما كانت | .5 | Sentence | Context |
| Report | Egypt | Press | Literature | وعهود ومشاعر تحلق في سماء صافية تخلو من | وعود | حضارات .. إنه نعمة من الله .. الحب | .6 | Sentence | Context |
| Report | Egypt | Press | Strategic Sciences | تأتي عادة متأخرة وبعد فوات الأوان ، وكلمات | وعودا | الماضي . فتدخلات الغرب فيه كانت في معظمها | .7 | Sentence | Context |
| Report | Egypt | Press | Social Sciences | في العام الماضي بأنه لن يكون هناك أي | وعودها | جديد على الطلبة والطالبات هذا العام ، رغم | .8 | Sentence | Context |
| Report | Egypt | Press | Strategic Sciences | فسوف تكون هذه هي الخطوة الأولى والأبرز التي | بوعودها | انه في حال مالتزمت إذا مالتزمت كوريا الشمالية | .9 | Sentence | Context |
| Report | Egypt | Press | Social Sciences | وكلمات كثيرة عن البرامج والاصلاحات التي تهتم بالشرائح | وعود | ** في المؤتمر الأخير للحزب الوطني كانت هناك | .10 | Sentence | Context |
| Report | Egypt | Press | Social Sciences | . وما يثير الدهشة أن حجم الاستثمارات التي | الوعود | على الناس في الواقع يختلف كثيرا عن هذه | .11 | Sentence | Context |
| Report | Egypt | Press | Social Sciences | الوزير بإعلانها للرأي العام لتكون بداية لانتهاء سنوات | ووعد | . ونحن في انتظار نتائج التحقيقات | .12 | Sentence | Context |
| Report | N/A | Net Articles | Biography | تحقيقا لأحلامه في السيادة التي لم تبارحه سحابة | الوعد | إبقائه بالقرب منه ؛ ورأى المتنبي في ذلك | .13 | Sentence | Context |
| Report | N/A | Net Articles | Biography | لا يزال وعدا . فشعر أبو الطيب بمكر | والوعد | حيله . ووسيلة لقهر حساده ؛ وانقضت سنتان | .14 | Sentence | Context |
| Report | N/A | Net Articles | Biography | . فشعر أبو الطيب بمكر كافور وتبين حيله | وعدا | لقهر حساده ؛ وانقضت سنتان والوعد لا يزال | .15 | Sentence | Context |

“Figure 4: The lemma ‘waʕd’ ‘Promise’ output search.”

In the search output information about the number of search result, country, source, genre, sentence and context are also available. This is phase one of ICA website and more enhancements are expected in later phases. The current phase of ICA application does not represent the final release as we are still receiving users comments and reports till all of them are implemented. However, The official phase of ICA application will give the opportunity for the researchers to save their query results.

## 5. Conclusion

The International Corpus of Arabic (ICA) is built, about 80 million words have been collected, covering all of the Arab world. About 2 million words have been disambiguated manually as a training data. About 50 million words have been disambiguated using (BAMAE). The evaluation has been done using precision and recall measurements for 1,000,000 words. At this point, Precision measurement ranges from 95%-92% while recall measurement was 92%-89%. The percentages are expected to rise by implementing the improvements while working on larger amounts of data. ICA website plays a role in overcoming the lack of Arabic resources. It is the 1st online freely available easy access query on 100,000,000 words which reflect the richness and variation of the ICA analyzed corpus to help the NLP community in specific and other researchers in general.

## References

Ahmed Abdelali, James Cowie&Hamdy S. Soliman. 2005, *Building a modern standard Arabic corpus*. In Proceedings of workshop on computational modeling of lexical acquisition. The split meeting. Croatia, (25-28 July).

CaminoR. Rizzo. 2010, *Getting on with corpus compilation: From theory to practice*. In ESP World Journal, Issue 1 (27), Volume 9.

Charles F. Meyer. 2002, *English Corpus Linguistics: An Introduction*, Cambridge University Press.

Hajic J., Smrz O., Zemánek P., Šnaidauf J., &Beška E. (2004, September), Prague Arabic Dependency Treebank: Development in Data and Tools. InProc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools (pp. 110-117).

Jan Hajic J., OtkarSmrz O., Petr Zemánek P., Jan Šnaidauf J., &Emanuel Beška E. 2004, *(2004, September), Prague Arabic Dependency Treebank: Development in Data and Tools*. In Proceedings of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools (pp. 110-117). (2004, September).

Jan Hajic, OtakarSmrz, Tim Buckwalter ,& Hubert Jin. September. 2005, *Feature-based tagger of approximations of functional Arabic morphology*. In Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, Spain.

John Sinclair. 1991, *Corpus, Concordance and Collocation (Describing English Language).* Oxford University Press.

Mahmoud El-Haj & Rim Koulali. 2013, *KALIMAT a Multipurpose Arabic Corpus*. In Proceedings of the

2nd Workshop on Arabic Corpus Linguistics (WACL-2) .

Mohamed Maamouri, Ann Bies, Tim Buckwalter &WegdanMekki. 2004, *The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus*, In Proceedings of NEMLAR conference on Arabic language resources and tools (pp. 102-109).

Muhammad Atiyya, Khalid Choukri& Mustafa Yaseen. (2005), *Specifcations of the Arabic Written Corpus.* NEMLAR Project. September 29th 2005.

Nizar Habash , Owen Rambow, & Ryan Roth. 2009, *MADA+ TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization*. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Nizar Habash and Owen Rambow. 2005, *Arabic Tokenization, Part-Of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. In Proceedings of ACL'05, Ann Arbor, MI, USA.

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

Piotr Pęzik. 2010, *New Ways to Language*. Chapter 21 (pp. 433-46), WydawnictwoUniwersytetuŁódzkiego.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008, *Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking*. In Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages (117–120), Columbus, Ohio, USA, June 2008.

Sameh Alansary, Magdy Nagi & Noha Adly. 2008, *Towards Analysing the International Corpus of Arabic (ICA): Progress of Morphological Stage*. In Proceedings of 8th International Conference on Language Engineering, Egypt.

Sameh Alansary. 2012, *BAMAE: Buckwalter Arabic Morphological Analyser Enhancer*. in Proceedings of 4th international conference on Arabic language processing, Mohamed Vth University Souissi, Rebate, Morocco, May 2-3 2012.

Sue Atkins S., Jeremy Clear J.& Nicholas Ostler N. (1992), *Corpus Design Criteria*, Literary and linguistic computing, 7(1), 1-16.

Tim Buckwalter. 2004, *Buckwalter Arabic Morphological Analyzer Version 2.0*. Linguistic Data Consortium (LDC) catalogue number LDC2004L02, ISBN 1-58563-324-0.

WajdiZaghouani. 2014, *Critical Survey of the Freely Available Arabic Corpora*. In Proceedings of Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools.