The Seventeenth Conference On Language Engineering

December 6-7, 2017, Cairo, Egypt
(ESOLEC'2017)

Organized by

Egyptian Society of Language Engineering (ESOLE)

Under the Auspices of

PROF. DR. ABD ELWAHAB AZZAT
President of Ain Shams University

PROF. DR. MOHAMED SHEBL EL KOMY
President of Misr International University

PROF. DR. MOHAMED AYMAN ASHOUR
Dean, Faculty of Engineering, Ain Shams University

PROF. DR. AYMAN BAHAA ELDIN
Dean, Faculty of Computer Science, MisrInternational University

Conference Chairperson
PROF. DR. M. A. R. GHONAIMY

Conference Co-Chairpersons
PROF. DR. SALWA ELRAMLY
PROF. DR. AYMAN BAHAA

Cairo, Egypt

http://esole-eg.org

Conference Chairman:
Prof. Dr. M. R. A.Ghonaimy

Technical Program Committee:
Prof. TaghridAnber, **ASU, Egypt**
Prof. I. Abdel Ghaffar, **CU, Egypt**
Prof. M. Ghaly, **CU, Egypt**
Prof. M. Z. Abdel Mageed, **Azhar U., Egypt**
Prof. Khalid Choukri, ELDA, **France**
Prof. Nadia Hegazy, **IRE, Egypt**
Prof. Christopher Ciri, LDC, **U.S.A**
Prof. Mona T. Diab, Stanford U., **U.S.A**
Prof. Ayman ElDossouki,**IRE,Egypt**
Prof. AfafAbdelFattah, CU, **Egypt**
Prof. Y. ElGamal**, AAST, Egypt**
Prof. M. Elhamalaway**, Azhar U., Egypt**
Prof. S. Elramly, **ASU, Egypt**
Prof. SamiaMashaly**, IRE, Egypt**
Prof. A. A. Fahmy**, Egypt**
Prof. I. Farag,**CU, Egypt**
Prof. MagdiFikry, **CU, Egypt**
Prof. WafaKamel, CU, **Egypt**
Prof. S. Krauwer, **Netherlands**
Prof. Dr. Sameh Alansary, **Alex. U.**
Prof. BenteMaegaard, CST, **Denmark**
Prof. M. Nagy**, Alex U., Egypt**
Prof. A. Rafae, **AUC, Egypt**
Prof. Mohsen Rashwan**, CU, Egypt**
Prof. S.I. Shaheen**, CU, Egypt**
Prof.H.M. AL-Barhamtoshy**, KAU, KSA**
Prof. M. F. Tolba, **ASU, Egypt**
Dr.Tarik F. Himdi**, KAU, KSA**

Organizing Committee

Prof. I. Farag
Prof. Ayman Bahaa
Prof. H. Korashy
Dr. Passant El-kafrawy
Dr. MonaZakaria
Dr.Ayman Nabil
Prof. Salwa Elramly
Prof. Hany Kamal
Prof. Mostafa Aref
Dr. Ashraf AbdelRaouf
Dr. Bassant A. Hamid
Dr.Ayman Ezzat

Conference Co-Chairpersons

Prof. Dr. SalwaElramly
Prof. Dr. Ayman Bahaa

Conference Sponsors

Scope of theConference:

- **Language analysis and comprehension**
- **Language generation**
- **Spoken language understanding**
- **Evaluation of natural language processing systems**
- **Large corpora**
- **Speech processing recognition and synthesis**
- **Natural language processing for information retrieval**
- **Machine translation**
- **Language engineering frameworks & methodologies**
- **Language engineering & artificial intelligence**
- **Automatic character recognition**
- **Semantic Web and Ontology Languages**
- **Mobile Web**
- **Social networks and contents development challenges**

The Seventeenth Conference on Language Engineering
Final Program

Wednesday,6 December 2017:(Misr International University)

9.00 - 10.00 Registration

10.00 - 10.30 Opening Session

10.30 - 11.15 **Session 1:Invited Paper**s: (Room 00A)
Chairman: Prof. Dr. I. Farag Eisaa

- **Recent Advances in Speech and Speaker Recognition Using Deep Learning Techniques**
  M. Affifi, *Microsoft Advanced Technology Lab Cairo, Egypt*

11:15 - 12:00

- **Ubiquitous Computing and Human Computer Interaction Research**
  Prof. Jiro Tanaka, Interactive Programming Laboratory (IPLAB), Graduate School of Information, Production and Systems, WASEDA University

12.00 - 12.30 Coffee break

12.30 - 13.30 **Session 2: Computational Linguistics I** (Room 00A)
Chairman: Prof. Dr. Wafa Kamel

1. **Automatic Arabic Diacritization for Text to Speech Systems**
   Sameh Alansary
   Bibliotheca Alexandrina, Alexandria, Egypt
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

2. **DiaVator: A Tool for Evaluating Arabic Diacritization Systems**
   Sameh Alansary
   Bibliotheca Alexandrina, Alexandria, Egypt
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

12.30 - 13.30 **Session 3: NLP for Information Retrieval: (**Room 00B**)**
Chairman: Prof. Dr. Ayman Bahaa

1. **Query Expansion for Arabic Information Retrieval Model: Performance Analysis and Modification**
   Ayat Elnahaas[*], Nawal Alfishawy[**], Mohamed Nour[*], Gamal Attiya[**], MahaTolba[**]
   **Electronics Research Institute, Cairo, Egypt*
   ***Faculty of Electronic Engineering, Menoufia University, Egypt*

2. دور السياق فى صياغة المعنى فى الترجمة
أسماء جعفر عبد الرسول
قسم اللغة الفرنسية، كلية الآداب، جامعة المنوفية

13.30 - 14.30 **Session 4: Automatic Character Recognition:(**Room 00B**)**
Chairman: Dr. Ashraf M. AbdelRaouf

1. **Arabic Handwritten Recognition using IoT Technology in Cloud Computing**
   Nada A. Shorim, Norhan M. Eltopgy, Sahar K. Mohamed, Shehab Salah, Taraggy M. Ghanim, Ashraf M. AbdelRaouf
   Faculty of Computer Science, Misr International University, Cairo.

13.30 - 14.30 **Session 5: Computational Linguistics II: (**Room 00A**)**
Chairman: Prof. Dr. Sameh Alansary

1. **Modern Standard Arabic Grammar Extraction from Penn Arabic Treebank Using Natural Language Toolkit**
   Amira Abdelhalim, Sameh Alansary
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

2. **A Formal Grammar for Describing Modern Standard Arabic Structures**
   Marwa Saber, Sameh Alansary
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

14:30 - 15:30 Lunch Break

15:30 - 16:30 **Session 6: Computational Linguistics III:** (Room 00A)
Chairman: Prof. Dr. M. Zaki Abdel Mageed

1. **How to Store a Syntactically Annotated Corpus in a Database?**
   Israa Elhosiny, Sameh Alansary
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

2. **Predicting Diacritics for Arabic Unknown Words**
   Amany Fashwan, Sameh Alansary
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

3. **Quantitative Analysis of Egyptian Aphorisms by Using R Language**
   Eman M. Yousri El-Gamal, Sameh Alansary
   Dept. of Phonetics & linguistics, Faculty of Arts, University of Alexandria. Alexandria, Egypt

15:30 - 16:30 **<u>Session 7:</u> Speech Analysis and Recognition(**Room 00B**)**
Chairman: **Prof. Dr.** M. Fahmy Tolba

1. **A High Efficient Automatic Speech Segmentation Algorithm Using A Hybrid Approach**
   Manal Shabaan Ismail, Amr Saleh, Ahmed Nashat, Amr Gody
   Electrical Engineering Department, Faculty of Engineering, Fayoum University El-Fayoum Egypt

2. **Arabic Dialect Identification from Speech Signals using i-vector and GMM-UBM**
   Mohsen Moftah[*], Waleed Fakhr[*], Salwa El Ramly[*]
   [*]*Electronics &Communications Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt*
   [**]*College of Computing, Arab Academy for Science and Technology, Egypt*

3. **Effect of Reducing the Number of Linear Predictive Coefficients on the Voice Quality of the CELP Vocoder using the Arabic Words**
   Nayra AbdElhalim, Noha Korany, Onsy Abdel Alim
   Electrical Department, Faculty of Engineering, Alexandria University, Egypt

Thursday, 7 December 2017 Conferences Centre, Faculty of Eng., Ain Shams University

09.00 - 14.30 **Session 8: Workshop: Deep Learning for NLP and Best Practices**

Time	Topic	Speaker
9.00	**Registration**	
10.00	**Welcome**	Prof. Salwa ElRamly
10.15	**Workshop Overview**	Prof. Aly Fahmy
10.30	**Natural Language Processing Tasks**	Prof. Aly Fahmy–Dr. Hanaa Bayomi
11.15	**Introduction to Deep Learning and Neural Networks**	Prof. Aly Fahmy–Dr. Hanaa Bayomi
12.00	**Coffee Break**	
12.30	**The amazing power of Word and Sentence Vectors**	Prof. Aly Fahmy–Dr. Wael Gomaa
1.15	**Keras Hands-on tutorial**	Prof. Aly Fahmy–Dr. Wael Gomaa
2.15	**Closing Remarks**	Prof.Aly Fahmy–Prof.Salwa ElRamly

14.30 - 15.30 Lunch Break

15.30 - 17.00 **Session 9:Text Mining(Room A)**
Chairman: Prof. Dr. Waleed Fakhr

1. Text Mining Mood Miner: Sentiment Mining of Financial Market
Hany Mohamed*, Ayman Atia**, and Mostafa Sami***
**Faculty of Computer Science, Helwan University, Egypt*
***Faculty of Computer Science, Misr International University, Helwan University, HCI- LAB, Egypt*
****Faculty of Computer Science, Helwan University, HCI-LAB, Egypt*

2. SimAll: A Flexible Tool for Text Similarity
Wael H. Gomaa*, Aly. A. Fahmy**
**Computer Science Department, Faculty of Computers and Information, Beni-Suef University, Egypt*
***Computer Science Department, Faculty of Computers and Information, Cairo University, Egypt*

3. A Survey on Mental Illness Detection using Language via Social Media Networks
EmanHamdi*, SherineRady**, Mostafa Aref*
**Computer Science Department, Faculty of Computer and Information Science, Ain Sham University, Cairo, Egypt*
***Information Systems Department, Faculty of Computer and*

Information Science, Ain Shams University

4. Cueing Conspiratorial Ideation in the Egyptian Tweets using Web-as-Corpus
Bacem A. Essam*, Mostafa M. Aref **
** English Language Department, Faculty of Al-Alsun, Ain Shams University*
*** Computer Science Department, Faculty of Computer Science and Information Sciences, Ain Shams University, Cairo, Egypt*

15.30 - 17.00 **Session 10:Students Workshop**: **(Room B)** (Posters)
Chairman: Prof. Dr. M. Younis Elhamalawy

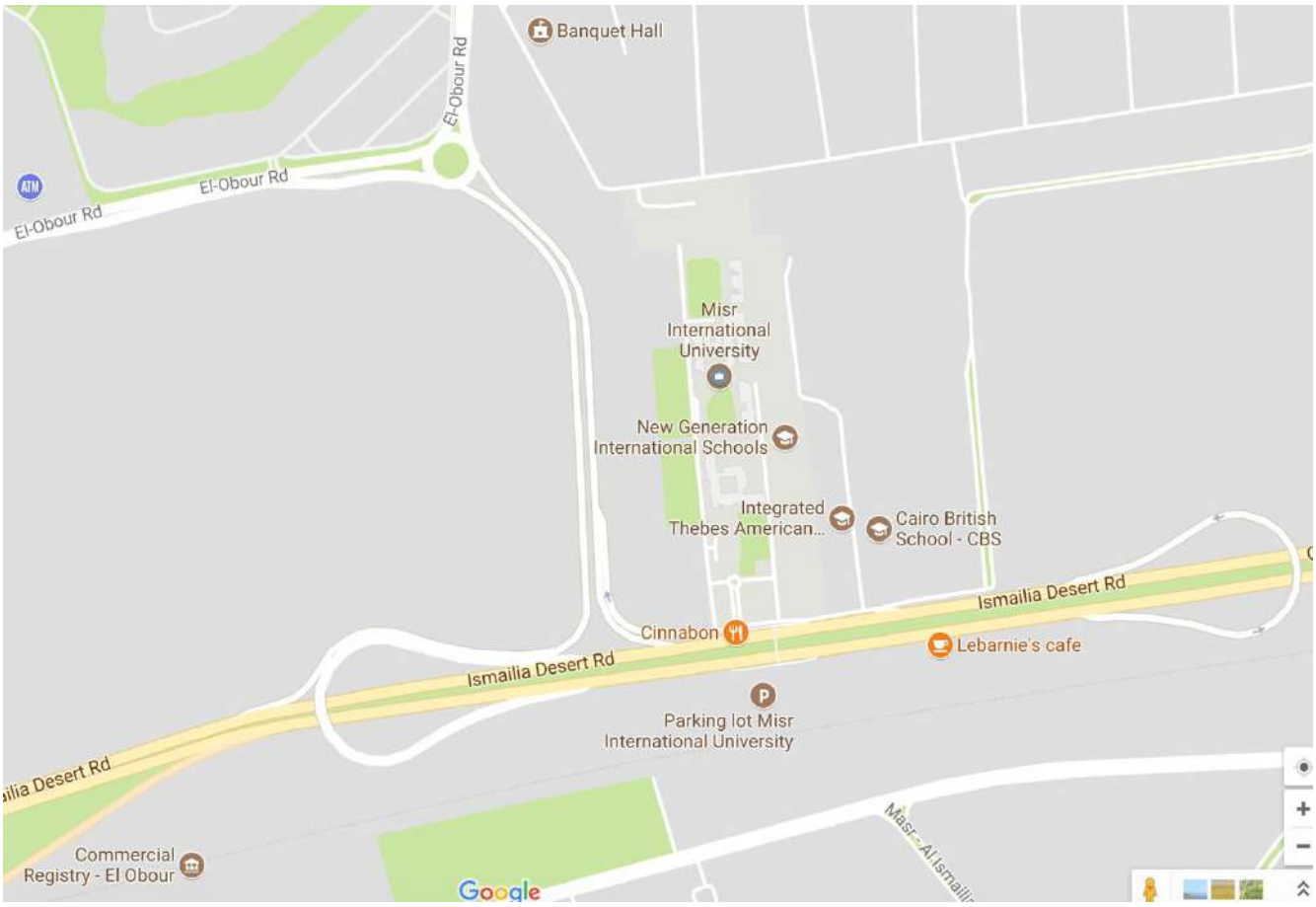
1. FARASA Named-Entity Handler
Amira Abdelhalim, Manar Hani, Miramar Shafiq
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

2. Lemmatization and Root Extraction in Arabic
Alaa Ayman, Doha Mahmoud, Mirna Shahin, Wafaa Muhammed
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

3. Aggregation in Arabic
Mai Hani, Heba Mahmoud
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University,Alexandria, Egypt

17.00 - 17.30 **Session 11:Closing Session (Room A)**
Chairman: Prof. Dr. SalwaElramly


El-Gereina
Rotana Phone
روتانا
Future Electronics - Egypt
Student Affairs
Greek Hospital St
El-Zahraa Hospital
El Zahraa University Hospital
مستشفى الزهراء الجامعى
Education Directorate - Cairo
مديرية التربية والتعليم - القاهرة
ASU - Engineering
Engineering Student Union
Toyota Egypt
تويوتا مصر
Sabel Al Khazndara
Sono Cairo
شركة صوت القاهرة للصوتيات والمرئيات


Banquet Hall
El-Obour Rd
El-Obour Rd
El-Obour Rd
ATM
Misr International University
New Generation International Schools
Integrated Thebes American...
Cairo British School - CBS
Ismailia Desert Rd
Cinnabon
Lebarnie's cafe
Ismailia Desert Rd
Parking lot Misr International University
Commercial Registry - El Obour
Google

أعضاء الجمعية من المؤسسات

١ –ركز نظم المعلومات – كلية الهندسة – جامعة عين شمس
٢ –معهد الدراسات والبحوث الإحصائية – جامعة القاهرة
٣ –مركز الحساب العلمي – جامعة عين شمس
٤ –الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحري
٥ –أكاديمية أخبار اليوم
٦ –معهد بحوث الإلكترونيات
٧ –معهد تكنولوجيا المعلومات
٨ –مكتبة الإسكندرية
٩ –المعهد القومي للاتصالات (NTI)
١٠ -الشركة الهندسية لتطوير نظم الحاسبات (RDI)
١١ -الهيئة القومية للاستشعار من بعد و علوم الفضاء
١٢ -كلية الحاسبات و المعلومات جامعة قناة السويس
١٣ -دار التأصيل للبحث و الترجمة

أهداف الجمعية

١ –الاهتمام بمجال هندسة اللغويات مع ا لتركيز على اللغة العربية بصفتها لغتنا القومية والتركيز على قواعد البيانات المعجمية وصرفها ونحوها ودلالتها بهدف الوصول إلى أنظمة ألية لترجمة النصوص من اللغات الأجنبية إلى اللغة العربية والعكس، وكذلك معالجة اللغة المنطوقة والتعرف عليها وتوليدها، ومعالجة الأنماط مع التركيز على اللغة المكتوبة بهدف إدخالها إلى الأجهزة الرقمية.
٢ –متابعة التطور فى العلوم والمجالات المختصة بهندسة اللغة.
٣ –التعاون مع الجمعيات العلمية المماثلة على المستوى المحلى والقومى والعالمى.
٤ –إنشاء قواعد بيانات عن البحوث التى سبق نشرها والنتائج التى تم التوصل إليها فى مجال هندسة اللغة بالإضافة إلى المراجع التى يمكن الرجوع إليها سواء فى اللغة العربية أو اللغات الأخرى.
٥ –إنشاء مجلة علمية دورية للجمعية ذات مستوى عال لنشر البحوث الخاصة بهندسة اللغة وكذلك بعض النشرات الدورية الإعلامية الأخرى بعد موافقة الجهات المختصة.
٦ –عقد ندوات لرفع الوعي في مجال هندسة اللغة.
٧ –تنظيم دورات تدريبية يستعان فيها بالمتخصصين وتتاح لكل من يهمه الموضوع . وذلك من أجل تحسين أداء المشتغلين فى البحث لخلق لغة مشتركة للتفاهم بين الأعضاء.
٨ –إنشاء مكتبة تتاح للمهتمين بالموضوع تشمل المراجع وأدوات البحث من برامج وخلافه.
٩ –خلق مجال للتعاون وتبادل المعلومات وذلك عن طريق تهيئة الفرصة لعمل بحوث مشتركة بين المشتغلين فى نفس الموضوعات.
١٠ تقييم المنتجات التجارية أو البحثية والتى تتعرض لعملية ميكنة اللغة.
١١ رصد الجوائز التشجيعية للجهود المتميزة فى مجالات هندسة اللغة.
١٢ إنشاء فروع للجمعية فى المحافظات.

المؤتمرالسابع عشر لهندسة اللغة

٦ – ٧ديسمبر ٢٠١٧

القاهرة - جمهورية مصر العربية

ينظم المؤتمر

الجمعية المصرية لهندسة اللغة

تحت رعاية

الأستاذ الدكتور/ عبد الوهاب محمد عزت
رئيس جامعة عين شمس

الأستاذ الدكتور/ محمد شبل الكومي
رئيس جامعة مصر الدولية

الأستاذ الدكتور/محمد أيمن عاشور
عميد كلية الهندسة -جامعة عين شمس

الأستاذ الدكتور/ أيمن محمد بهاء الدين
عميد كلية حاسباتمصر الدولية

رئيس المؤتمر

الأستاذ الدكتور/ محمد أديب رياض غنيمي

مقرر المؤتمر

الأستاذ الدكتور / سلوى حسينالرملي
كلية الهندسة - جامعة عين شمس

الأستاذ الدكتور/ أيمن محمد بهاء الدين
كلية الحاسبات – جامعة مصر الدولية

http:// www.esole-eg.org

المؤتمر السابع عشرلهندسة اللغة

برنامج المؤتمر

٦-٧ديسمبر ٢٠١٧

القاهرة - جمهورية مصر العربية

Table of Contents

Automatic Arabic Text Diacritization for Text-To-Speech Systems

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
sameh.alansary@bibalex.org

***Abstract*— Diacritization has a significant impact on Arabic NLP applications. Diacritized words are useful when developing Arabic Text-To-Speech Systems (TTS), since it could reduce the ambiguity. Text-To-Speech synthesis has been the focus of a large number of studies for quite some time now. In this paper, we present generative techniques for diacritizing the Arabic texts and present the evaluation results for these techniques on widely available data sets. The focus of this paper is the issue of generating diacritized numeric nouns accurately out of numeric values and diacrtizing them using Alserag diacritization system. The system scored less than 10% error rate in the generation of the diacritized numeric nouns.**

***Keywords*: Diacritization, Arabic, NLP, Alserag, computational linguistics, language engineering, TTS.**

1 INTRODUCTION

Arabic language is a diacritized language where the pronunciation of the text is not fully determined by its spelling characters. It needs specific marks to define the correct pronunciation and realize the meaning of the text. In addition, the processes of grapheme-to-phoneme conversion and syllabification need also a full diacritized text to be accomplished [1]. Diacritization is crucial for many NLP applications. In particular, it is extremely useful for text-to-speech (TTS) applications [2]. Text-to-speech system (TTS) is also known as speech synthesizer [1]. One of the most TTS challenges of the Arabic language is the lack of diacritics and punctuation. The importance of diacritization for speech synthesis is much more than for speech recognition since it is urgent for TTS where the correct vowels are pronounced – an incorrect vowel may completely alter the meaning of the utterance [3].
Text to speech engine converts Arabic text into a natural, human-sounding voice output. The lack of diacritics leads to problems for both acoustic and language modeling, because it is difficult to train accurate acoustic models for short vowels if their identity and location in the signal is unknown and the absence of diacritics leads to a large set of linguistic context for a given word form [4].

With the increasing number of users of text to speech applications, high quality speech synthesis is required. Therefore, diacritization system is proposed to resolve this problem [5].Diacritization improvement in Arabic has important implications for downstream processing for Arabic natural language processing, e.g. speech recognition ([6]; [7]), speech synthesis ([8]), and machine translation [9].

To develop highly accurate speech recognition systems for Arabic, the text or the data has to be diacritized, because it reduces the ambiguity, another reason for the importance of the diacratization data is that, as we know Arabic dialects are essentially spoken varieties, so building an Arabic TTS system would handle the different dialects of the spoken Arabic.
It was found that, the only available corpus that includes detailed phonetic information is the CallHome[1] (CH) Egyptian Colloquial Arabic (ECA) corpus distributed by the Linguistic Data Consortium (LDC). This corpus has been transcribed in both the script form and a so-called romanized form, which is an ASCII representation that includes short vowels and other diacritic information and thus has complete pronunciation information.

Sakhr TTS overcomes the traditional challenges of TTS for Arabic language, including lack of punctuation and diacritisation (vowel placement), with its powerful Automatic Diacritizer. The diacritizer performs language processing, analysis, and word disambiguation. Sakhr has developed a diacritizer engine. This engine can insert the diacritics needed in Arabic texts automatically. It is the main component of Arabic TTS. Without the diacritizer, the output of the TTS engine would be inaccurate and unclear. Since native Arabic speakers write Arabic text without diacritics, the TTS engine should handle the non-diacritized text. The Diacritizer will convert non-diacritized text into a diacritized text, and then the TTS engine converts it to a clear human Arabic voice.
The Arabic language allows one or more pronunciations for most words. It allows users to force the TTS to pronounce a word as they like. i.e. their names, countries and any specials words.

[1] https://catalog.ldc.upenn.edu/LDC97S45

Sakhr TTS supports its own private user dictionary format with an assistant editing tool. It has a custom voice development with extensive experience in creating domain specific custom voices, Sakhr can quickly produce customized TTS voices for clients with specific brand and voice needs. Sakhr TTS is deployed across the Middle East in IVR phone systems, directory assistance, desktop applications, mobile services, and embedded devices and products. Sakhr TTS operates with pre-packaged voices or custom-built voices for clients. Sakhr TTS is available in both desktop and embedded versions (https://www.alibaba.com/product-detail/Sakhr-Text-To-Speech-engine-converting_250414834.html).

Sakhr TTS main feature is that it is supported in both male and female voices with many effects. Volume, pitch and speed rate can be adjusted. It supports major audio formats for both telephone and desktop. In addition, it supports text normalization that enables smart handling of ambiguous text input such as dates, time, currencies, units, abbreviations and exceptions. It has unlimited vocabulary, unlimited text size, raw text, phonetic and prosodic input. It can handle common Arabic errors using Sakhr Corrector.

There is another trial for TTS system for the Arabic language and it uses allophone/diphone concatenation method with two main modules: text/linguistic analyzer and synthesizer core. The system takes a complex text as input; the text includes abbreviations, numbers, dates, times, addresses, e-mails, and so on and the output is corresponding Arabic speech. In this system, the output is available in one male voice only[10]. In the diacratization step, there is a per-processing stage to the complex text where each character and its diacritic must be determined. The system uses a half diacritized lexicon of sample Arabic words (diacritization database) developed using Microsoft Access 2003. But, in order to obtain a fully diacritized Arabic word, the user has to add the missing diacritic of the last character to each half-diacritized word, otherwise the default diacritic "ً" will be added automatically. It is worth mentioning that some English words were not handled in the previous stage. Therefore, the output of this module may include English words from the previous stage in addition to the diacritized Arabic words, in the same sequence used by the user. The system also uses what is called “FIFO” queue to reserve the order [10].

Another Arabic Text-to-Speech (TTS) system was developed at the Human Language Technologies laboratory of IBM Egypt. The system is based on the state of the art IBM trainable concatenative speech synthesizer.

The system is composed of three major modules: a text module, a prosody module, and a back-end module [11]. The text module includes a text normalizer, phonological analyzer, and a prosodic planner. At the text analyzer step, the text is normalized, for example, digits are converted into their words equivalent. Moreover, the Arabic phonological analyzer is the step, where the grapheme is transformed to phoneme, syllabification happens, and syllable stress is assigned. The prosodic planner does some abstract prosodic planning at the text level.

The system takes the decision of relying on a manual diacritization (and not to consider the diacritization module as a part of the IBM Arabic TTS system) is based on the fact that the all existed automatic Arabic diacritization techniques are not mature enough to relay on them. In fact, none of the tested commercial automatic diacritization tools provided the quality required by the current TTS client [11].

The focus in this paper is on investigation the different procedures that could generated diacritized alphabetic characters (numeric nouns) out of numeric values occurring in the Arabic texts. These procedures are developed as a specific module in Alserag diacritization system [12]. Alserag is an Arabic diacritizer that has been developed by Bibliotheca Alexandrina computational linguistics team. Alserag consists of different steps: retrieval of unambiguous lexicon entries, disambiguating between the different stored possible solutions of the words to realize their internal diacritization through the morphological analysis step, the syntactic processing step that is responsible for the case ending detection, which is based on shallow parsing, and finally the morpho-phonological step [12]. By and large, no available diacritization system is able to convert the numbers (digits) into their corresponding numeric nouns yet, which is a very crucial addition to the TTS systems. Thus, our trail and the specialized module we have developed is the contribution we offer to the field.

Section 2 will presents the rules of the numerical nouns in Arabic. Section 3 will present the grammar workflow of Alserag diacritization system that is responsible for generating the diacritized numeric nouns in the Arabic texts. Section 4 presents the numeric values in other available diacritization systems. Section 5 evaluates the output, discusses the results, and is concerned with the benchmarking process. Finally, section 6 concludes the paper.

2 THE LINGUISTIC DESCRIPTION OF THE RULES OF NUMERICAL NOUNS IN ARABIC[1]

This section discusses the problem of converting numerical numbers into numerical nouns so that they couldbepronounced correctly in order to be used in TTS applications. For that purpose, a module has been developed based on rules and dictionary. After converting the numbers into numerical nouns, these numerical nouns have to be diacritized. There are different kinds of numbers. For example, there are ordinal, cardinal, partial, decimal numbers, hours, and so on.

In Arabic, the numerical nouns agree with the preceding or following nouns "تمييز العدد" (tamiez) either in gender only or in both gender and number. The different types of Arabic numeric nounscan be classified as follows: Units, Combined 'Morakkab', Conjuncted 'Maatouf', and "ألفاظ العقود" (AlfazAlokud). Each type will be discussed in details in the following subsections. Each type follows certain rules. The gender of the numeric noun is based on the following noun "تمييز العدد" (tamiez).

a) Units

Numeric nouns of the numbers (1 - 2) always agree with "تمييز العدد" (tamiez) in both gender and number. Tamiez is always preceding the numeric nouns. For example, sentences (1) and (2):

Sentence 1: امرأة واحدة –رجل واحد
Sentence 2: امرأتان اثنتان –رجلان اثنان

Numeric nouns for the numbers from 3 to 10 (3 - 10) disagree with the following noun "تمييز العدد" (tamiez) in gender, which is always 'plural' with a genitive case "مجرور". For example, sentences (3) and (4). Number 8 is written as "ثمان" and treated like the cases of "الاسم المنقوص" (Al-Esm Al-Manqus) whether it is added or not. For example, sentence (5):

Sentence 3: سبع فتيات –سبعة رجال
Sentence 4: عشر سنين –عشرة أعوام
Sentence 5: سافر من النساء ثمانٍ –سافر ثماني نساءٍ

b) Combined 'Morakkab'

Numeric nouns of the numbers (11 -12)consist of two parts that always agree with the following noun "تمييز العدد" (tamiez) in gender. For example, sentences (6) and (7):

Sentence 6: إحدى عشرة امرأة–أحد عشر رجلا
Sentence 7: اثنتا عشرة امرأة –اثنا عشر رجلا

Numeric nouns of the numbers from 13 to 19 (13 - 19) consist of two parts, their first part disagrees with the following noun"تمييز العدد" (tamiez) in gender, while the second part always agrees. The following noun of these numerical nouns is always "singular" with accusative case "منصوب", such as the sentences (8) and (9). However, the numeric noun of number 18 has only one form, which is "ثماني" as in the two sentences in (10):

Sentence 8: سبعة عشر رجلا –سبع عشرة فتاة
Sentence 9: أربعة عشر كتابا –ثلاث عشرة فراشة
Sentence 10: سافر ثماني عشر رجلا –سافر ثماني عشرة امرأة

c) Conjuncted 'Maatouf'

Numeric nouns of the number (21) consist of two parts; the first part agrees with the following noun "تمييز العدد" (tamiez) in gender, while the second part does not change. For example, sentences in (11):

Sentence 11: إحدى وعشرون امرأة –واحد وعشرون رجلا

While, numeric nouns of the numbers from 22 to 99 (22 to 99) consist of two parts; the first disagrees with the following noun in gender and the second part does not change as in sentences in(12). Moreover, the following noun "تمييز العدد" (tamiez) is always singular with accusative case.

[1] http://www.reefnet.gov.sy/education/kafaf/Bohoth/AdadMadoud.htm
http://mawdoo3.com/%D9%82%D9%88%D8%A7%D8%B9%D8%AF_%D8%A7%D9%84%D8%B9%D8%AF%D8%AF_%D9%88%D8%A7%D9%84%D9%85%D8%B9%D8%AF%D9%88%D8%AF

Sentence 12: تسعة وتسعون رجلا –تسع وتسعون فتاة

d) "ألفاظ العقود" (AlfazAlokud)

Numeric nouns of the numbers (20 – 30 – 40……..1000) have their own orthographical form, which is constant regardless of the gender and number of the following noun and they are assigned with case ending according to their position in the sentences. The noun following (AlfazAlokud) "ألفاظ العقود" is always singular assigned with the genitive case. For example, sentences in (13) and (14):

Sentence 13: تسعون بابا –عشرين كتابا
Sentence 14: ثلاثة آلاف كتابٍ –مئة كتابٍ

After describing the Arabic numeric nouns in the previous sub-subsections, there is a type of numerical nouns that should be described which is called the ordinal numeric nouns "العدد الترتيبي", where the numeric nouns (either it consists of only one part or composed of two parts) agree with preceding noun"تمييز العدد" (tamiez) in gender. For example, sentences in (15) and (16):

Sentence 15: المتسابقة العاشرة –المتسابق العاشر
Sentence 16: المتسابقة السابعة عشرة – المتسابق السابع عشر

Definiteness in Arabic numeric nouns has three cases: First, numeric noun can be defined with the prefix "ال" (The)and its "تمييز العدد" (tamiez) is still undefined. Second, cases where both are defined. Third, "تمييز العدد" (tamiez) is defined while the numeric noun is not, for example sentence (17):

Sentence 17: ثلاثة الأطواف –الثلاثة الأطواف –الثلاثة أطواف

3 WORKFLOW OF THE DEVELOPED GRAMMAR

In Alserag system, a grammar module is responsible for converting the numerical values to their diacritized numerical nouns taking into account the Arabic rules listed above in section 2. In addition, a word-form specialized dictionary containing different forms of the Arabic numerical nouns has been developed. For example, the numeric value "5" is converted to the string "خمس" 'five' by the rule in (1) and the string "خمس" is stored in two forms "خُمَس"and "خَمْسَة".

rule (1) : (%y,{BLK|SHEAD})(%x,^@ordinal,"5"):=(%y) (%x,-[5],?[خمس],+units);

The mark "?" in rule (1) means that the form "خمس" will be retrieved from the dictionary. After applying rule (1), the system will find that "خمس" is stored in two forms. These two forms are internally diacritized as shown in figure 1:

```
[خمس] {} "خَمْس"
(LEMMA=خُمَس,BF=خَمْس,LEX=U,POS=CDN,MOR=WFO,LST=WRD,GEN=MCL,NUM=SNG,
PAR=M55,SEM=QTT) <ar,200,200>;

[خمس] {} "خَمْسَة" (LEMMA=خَمْس ,BF=خَمْس ,LEX=U,POS=CDN,
LST=WRD,GEN=FEM,NUM=SNG,PAR=M557,ABN=ABT,SEM=QTT) <ar,200,200>;
```

Figure 1: The stored forms of the numeric noun "خمس" in the dictionary

The developed rules are able to disambiguate between the different forms of the string "خمس" 'five', according to the context and the Arabic rules describe above in section 2.

The system could choose the correct stored form of "خمس" in the context "خمس فتيات" 'Five girls' and "خمسة رجال" 'Five men'; when the noun following the number is feminine the rules could successfully choose the form "خُمَس", but when the noun following the number is masculine, the rules could successfully choose the form "خَمْسَة".

It has to be mentioned that the compound numbers are not stored in the dictionary, because they consist of two parts where the first part is units; numbers form (3-9) as mentioned in section 2 and the second part is the string "عشر" or"عشرة". Both the units and the numerical nouns "عشر" and"عشرة" are stored in Alserag dictionary, but separately. During the conversion of the compound numbers such as "13", the rules convert the units by rule in (2) and then convert the tens into the corresponding numerical noun according to the context; "1" in "13" is converted into "عشر" and assigned with the feature "mrakab" to indicate that it is a compound noun by rule in (3).

rule (2) : (%y,{BLK|SHEAD})(%x,^@ordinal,"3"):=(%y) (%x,-[3],[ثلاثة],+units);

rule (3) : (%x,"1",^@ordinal)(%y,units):=(%y,units,+mrakab,-units)([ ],+blk)(%x,-[1],[عشر],+mrakab);

However, numbers "11" and "12" are treated in a different way; they are stored as compound nouns in the dictionary as follows: "أَحَدَ عَشَر", "إِحْدَى عَشْرَة", "اِثْنَا عَشَر", "اِثْنَتَا عَشْرَة", "اِثْنَيْ عَشَر"and "اِثْنَتَيْ عَشْرَة" as shown in figure 2:

```
أَحَدَ عَشَر"{}[أحد
[عشر"(LEX=U,POS=CDN,MOR=WFO,LST=WRD,GEN=MCL,NUM=SNG,PAR=M557,FRA=Y0,ABN
=ABT,ANI=NANM,SEM=QTT) <ar,200,200>;
إِحْدَى عَشْرَة "{} [أحد
"
[عشر(LEX=U,POS=CDN,MOR=WFO,LST=WRD,GEN=FEM,NUM=SNG,PAR=M557,FRA=Y0,ABN=
ABT,ANI=NANM,SEM=QTT) <ar,200,200>;
```

Figure 2: The stored forms of the numeric noun "أحد عشر" in the dictionary

The numeric nouns of the numbers (20 – 30 – 40……..1000) have two forms; the two forms differ in the case morpheme according to their contexts. For example, 50 has the forms "خمسون" and "خمسين"; however, they are stored in the dictionary as one form which is the nominative form; "خمسون", the system considers this form as the default one in the analysis stage and it will generated in the correct case morpheme according to the contexts at the generation stage.

In the dictionary of Alserag system, all the Arabic numbers are diacratized, but there are not assigned with a case ending, since it depends on the context, for example, both "أربعة" or "أربع" are stored in the dictionary as "أَرْبَعَة" and "أَرْبَع" without case endings.

```
[أربع]{}"أَرْبَع"(LEMMA=أَرْبَع,BF=أَرْبَع,LEX=U,POS=CDN,MOR=WFO,LST=WRD,GEN=MCL,
NUM=SNG,PAR=M557,FRA=Y0,ABN=ABT,ANI=NANM,SEM=QTT) <ar,200,200>;
[أربع]{}"أَرْبَعَة"
(LEMMA=أَرْبَع,BF=أَرْبَع,LEX=U,POS=CDN,MOR=WFO,LST=WRD,GEN=FEM,NUM=SNG,P
AR=M557,FRA=Y0,ABN=ABT,ANI=NANM,SEM=QTT) <ar,200,200>;
```

Figure 3: The stored forms of the string "أربع" in the Alserag dictionary

After converting the numbers into the numeric nouns, the system assigns them their case endings according to their context. For example, in the sequence "جاء 4 أشخاص", "4" will be converted into the string "أربعة", then the diacritized form will be retrieved from the dictionary.The system will find two forms for the string "أربعة",the disambiguation rules are responsible for disambiguating between the two forms in figure 3. The system will select the form"أَرْبَعَة" which is the feminine form, because the following noun is masculine "أشخاص" 'persons', according to the linguistic description listed in section 2. As "أَرْبَعَة" is subject of the sentence, rules will assign the nominative case to it, so it will generated as "أَرْبَعَةُ". Then, the system will assign the genitive case to the noun following the numeric noun "أَرْبَعَةُ", so the word "أشخاص" will be "أَشْخَاصٍ"

For the number "44", the rules that is responsible for converting the units in the number module will convert the first part in the sequence "44". Rule (4) converts "4" into "أربعة" and assign it with the feature "units". After applying rule (4), the rule (5) will convert the second part in the sequence "44"to "أربعون" and assign it the feature "maatof".

rule (1) : (%y,{BLK|SHEAD})(%x,^@ordinal,"4"):=(%y) (%x,-[3],[أربعة],+units);

rule (2) : (%x,"4")(%y,{units|ordinal}):=(%y,units,+maatof,-units)([ ],+blk)(%z,[و])(%x,-[3],[أربعون|],+maatof);

For the number "2,000,000",the system could convert it to its equivalent numeric noun with two rules. Rule in (6) converts "000000" to "مليون" 'million' and assign it with the feature "million".

rule (3) : (%x,^@ordinal,"000000",^million):=(%x,-[000000],[مليون],+million);

Then, rule in (7) will convert the digit "2" in "2,000,000" into "اثنان", and it will be assigned with the feature "units".

rule (4) : (%x,^@ordinal,"2")(%y, million):=(%x,-[2],[اثنان],+units)([ ],+blk)(%y);

For the number “1,000”, the rule in (8) converts “1000” to “ألف”. The converted number is assigned with the feature “allaf”, which means that it is converted to thousands and this feature will be useful in forming rules that will be applied in other cases. The feature “BLK” is a feature that refers to the blank space, and “STAIL” is a feature that refers to the end of the sentence.

rule (5) : (%x,^@ordinal,"1000")({STAIL|BLK}):=(%x,-[1000],[ألف],+allaf)({STAIL|BLK});

For number such as “12,500”, the rule in (9) that converts the sequence “500” to “خمسمائة”.As long as it is not followed by other digits, it will be assigned with the feature “meaat” to indicate that it is converted to hundreds.

rule (6) : (%x,"500")(^DIGIT,%y):=(%x,-[500],[خمسمائة],+meaat)(%y);

After applying rule in (9), the rule in (10) states that if the sequence in “12,500” contains the sequence “12” and is followed by a sequence that has the feature “meaat” by rule in (9), then the string “ألف” which will be retrieved from the dictionary with its internal diacritization will be add. The sequence will be as follows “12 ألف و خمسمائة”

rule (7) : (%x,^@ordinal,"12")(%y,^STAIL,meaat):=(%x)([ ],+blk)(%z,[ألف])([ ],+blk)([و],%w)(%y);

After applying the rule in (10), the rule in (11) will be applied to convert the sequence “12” into "اثنى عشر" and assign the feature “num_generated” to it. The feature “num_generated” is a feature that indicates that the sequence has been fully converted and generated.

rule (8) : (%x,^@ordinal,"12")({STAIL|":"|BLK|N|blk},%y):=(%x,-[12],[اثنى عشر],+num_generated)(%y);

For the number “100”,the rule in (12) will convert the sequence “100” into “مائة” and assign it with the feature “hundred” to indicate that it is converted to hundreds.

rule (9) : (%x,"100")(^DIGIT, %y):=(%x,-[100],[مائة],+hundred)(%y);

Moreover, the system could also diacritize the partial numbers. For example, the partial number “1/2” could be converted to “نصف”‘half’ by the rule (13) and assign it with the feature “par_num” to indicate that it is a partial number. The numericnoun “نصف” will be handled in Alserag diacritization system as a common noun as stated in [12].

rule (10): (%x,"1")(%y,"/")(%z,"2"):=(%x,[نصف],+par_num);

In addition, Alserag system is able to convert and diacritize the numbers that refer to time such as “00:02:00”, the rule in (14) could convert these digits into “دقيقتان” ‘two minutes’.They will be handled as regular nouns,and will be diacritized as such [12]. It will be assigned with the feature “minutes”. In rule (14), the nodes in the left side will be suppressedand the word “دقيقتان” will be generated.

rule (11): (%h,"00")(%x,":")(%m,"02")(%y,":")(%s,"00"):=(%x,[دقيقتان],+minutes);

Finally, Alserag system also converts and diacritizes the decimal numbers such as “4.7”.Rules responsible for converting the units as the rule in (15) could convert the digit “7” in the sequence “4.7” to “سبعة”. Then, the rule in (16) will convert “4” that precedes the dot in the sequence “4.7” and the feature “units” which is assigned to the number “سبعة” will be changed to the feature PTN to indicate that it a decimal number and the sequence “من عشرة” will be added.

rule (12): (%y,{BLK|SHEAD})(%x,^@ordinal,"7"):=(%y) (%x,-[7],[سبعة],+units);

rule (13): (%x,"4")(%y,".")(%z,units):=(%x,[أربعة])([ ],blk)(%w,[و])(%z,+PTN)([ ],blk)(%y,[من عشرة],+mark);

After building the number module,its performance has to be tested. During the evaluation of the module, a specialized corpus has been used. It covers all different types of the Arabic numbers. Table (1) is a sample from the developed tested corpus, the corpus was built in a mechanism that graded in difficulty by starting with simple numbers like units and ending with compound numbers. The corpus also contains set of data especially to test the decimal numbers, time, and partitive ones.

TABLE 1: THE DEVELOPED CORPUS TO TEST THE NUMBER MODULE OF ALSERAG SYSTEM

01:00:00	1/2	100	1
02:00:00	1/3	101	2
00:01:00	2/3	122	3
00:15:00	1/4	233	4
00:00:01	2/4	1000	5
00:00:59	3/4	1001	6
03:50:00	1/5	1100	7
00:50:18	2/5	1144	8
03:50:18	3/5	1255	9
10:00:00	4/5	2000	10
00:30:00	5/6	2366	11
08:05:00	6/7	11000	12
08:15:00	7/8	11467	13
08:30:00	8/9	21588	14
03:50:00	9/10	50699	15
	2 1/2	100001	16
	5 1/2	200027	17
	3 3/4	233003	18
	4.0	1000000	19
	4.7	1000001	20
		1000002	21
		22000002	30
		300000003	32
		1000000000	40
		1000000000000	43
			50
			54
			60
			65
			70
			76
			80
			87
			90
			98

The first results were obtained as shown in table 2 and they were very satisfying concerning the units, tens, hundreds and so on. The percentage of errors was mostly in the times section. After obtaining the results in table (2), they are diacritized as table (3):

TABLE 2: THE RESULT OF THE DATA IN TABLE 1

الساعة الواحدة	نصف	مائة	واحد
الساعة الثانية	ثلث	مائة وواحد	اثنان
دقيقة	ثلثين	مائة اثنان وعشرون	ثلاثة
خمسة عشر دقيقة	ربع	مائتين وثلاثة وثلاثون	أربعة
ثانية	ربعان	ألف	خمسة
تسع وخمسون ثانية	ثلاث أرباع	ألف وواحد	ستة
الساعة الثالثة وخمسون دقيقة	خمس	ألف ومائة	سبعة
خمسون دقيقة وثمانية عشر ثانية	خمسان	ألف مائة أربعة وأربعون	ثمانية
الساعة الثالثة وخمسون دقيقة وثمانية	ثلاث أخماس	ألف مائتين وخمسة وخمسون	تسعة
عشر ثانية	أربع أخماس	ألفان	عشرة
الساعة العاشرة	خمس أسداس	ألفين ثلاثة مائة وستة وستين	إحدى عشر
ثلاثون دقيقة	ست أسباع	أحد عشر ألف	اثنا عشر
الساعة الثامنة وخمس دقائق	سبع أثمان	أحد عشر ألف أربعمائة وسبعة وستون	ثلاثة عشر
الساعة الثامنة وخمس عشرة دقيقة	ثمن أتساع	واحد وعشرون ألف خمسمائة ثمانية وثمانون	أربعة عشر
الساعة الثامنة وثلاثون دقيقة	تسع أعشار	خمسون ألف ستمائة تسعة وتسعون	خمسة عشر
الساعة الثالثة وخمسون دقيقة	اثنان ونصف	مائة ألف وواحد	ستة عشر
	خمسة ونصف	مائتين ألف وسبعة وعشرون	سبعة عشر
	ثلاثة وثلاث أرباع	مائتين ثلاثة وثلاثون ألف وثلاثة	ثمانية عشر
	أربعة من عشرة	مليون	تسعة عشر
	أربعة وسبعة من عشرة	مليون وواحد	عشرون
		مليون واثنان	واحد
		اثنان وعشرون مليون واثنان	وعشرون
		ثلاثمائة مليون وثلاثة	ثلاثون
		بليون	اثنان وثلاثون
		تريليون	أربعون
			ثلاثة وأربعون
			خمسون
			أربعة
			وخمسون
			ستون
			خمسة وستون
			سبعون
			ستة وسبعون
			ثمانون
			سبعة وثمانون
			تسعون
			ثمانية
			وتسعون

TABLE 3: THE DIACRITIZED CORPUS OF THE DATA IN TABLE 2

اَلسَّاعَة الْوَاحِدَة	نِصْف	مِائَة	وَاحِد
اَلسَّاعَة الثَّانِيَة	ثُلُث	مِائَة وَوَاحِد	اِثْنَان
دَقِيقَة	ثُلُثَيْن	مِائَة اِثْنَان وَعِشْرُون	ثَلَاثَة
خَمْسَة عَشَر دَقِيقَة	رُبُع	مِائَتَيْن وَثَلَاثَة وَثَلَاثُون	أَرْبَعَة
ثَانِيَة	رُبْعَان	أَلْف	خَمْسَة
تِسْع وَخَمْسُون ثَانِيَة	ثَلَاث أَرْبَاع	أَلْف وَوَاحِد	سِتَّة
اَلسَّاعَة الثَّالِثَة وَخَمْسُون دَقِيقَة	خُمُس	أَلْف وَمِائَة	سَبْعَة
خَمْسُون دَقِيقَة وَثَمَانِيَة عَشَر ثَانِيَة	خُمْسَان	أَلْف مِائَة أَرْبَعَة وَأَرْبَعُون	ثَمَانِيَة
اَلسَّاعَة الثَّالِثَة وَخَمْسُون دَقِيقَة وَثَمَانِيَة	ثَلَاث أَخْمَاس	أَلْف مِائَتَيْن وَخَمْسَة وَخَمْسُون	تِسْعَة
عَشَر ثَانِيَة	أَرْبَع أَخْمَاس	أَلْفَان	عَشَرَة
اَلسَّاعَة اَلْعَاشِرَة	خُمْس أَسْدَاس	أَلْفَيْن ثَلَاثَة مِائَة وَسِتَّة وَسِتِّين	إِحْدَى عَشَر
ثَلَاثُون دَقِيقَة	سُتّ أَسْبَاع	أَحَد عَشَر أَلْف	اِثْنَا عَشَر
اَلسَّاعَة الثَّامِنَة وَخَمْس دَقَائِق	سُبْع أَثْمَان	أَحَد عَشَر أَلْف أَرْبَعْمِائَة وَسَبْعَة وَسَبْعُون	ثَلَاثَة عَشَر
اَلسَّاعَة الثَّامِنَة وَخَمْس عَشَرَة دَقِيقَة	ثُمْن أَتْسَاع	وَاحِد وَعِشْرُون أَلْف خَمْسْمِائَة ثَمَانِيَة وَثَمَانُون	أَرْبَعَة عَشَر
اَلسَّاعَة الثَّامِنَة وَثَلَاثُون دَقِيقَة	تِسْع أَعْشَار	خَمْسُون أَلْف سِتْمِائَة تِسْعَة وَتِسْعُون	خَمْسَة عَشَر
اَلسَّاعَة الثَّالِثَة وَخَمْسُون دَقِيقَة	اِثْنَان وَنِصْف	مِائَة أَلْف وَوَاحِد	سِتَّة عَشَر
	خَمْسَة وَنِصْف	مِائَتَيْن أَلْف وَسَبْعَة وَعِشْرُون	سَبْعَة عَشَر
	ثَلَاثَة وَثَلَاث أَرْبَاع	مِائَتَيْن ثَلَاثَة وَثَلَاثُون أَلْف وَثَلَاثَة	ثَمَانِيَة عَشَر
	أَرْبَعَة مِنْ عَشَرَة	مِلْيُون	تِسْعَة عَشَر
	أَرْبَعَة وَسَبْعَة مِنْ عَشَرَة	مِلْيُون وَوَاحِد	عِشْرُون
		مِلْيُون وَاثْنَان	وَاحِد
		اِثْنَان وَعِشْرُون مِلْيُون وَاثْنَان	وَعِشْرُون
		ثَلَاثْمِائَة مِلْيُون وَثَلَاثَة	ثَلَاثُون
		بِلْيُون	اِثْنَان وَثَلَاثُون
		تِرِيلْيُون	أَرْبَعُون
			ثَلَاثَة وَأَرْبَعُون
			خَمْسُون
			أَرْبَعَة
			وَخَمْسُون
			سِتُّون
			خَمْسَة وَسِتُّون
			سَبْعُون
			سِتَّة وَسَبْعُون
			ثَمَانُون
			سَبْعَة وَثَمَانُون
			تِسْعُون
			ثَمَانِيَة
			وَتِسْعُون

4 NUMERIC NOUNS OF OTHER AVAILABLE DIACRITIZATION SYSTEMS

In Arabic, several available diacritization systems help the reader to articulate the text correctly. The most known systems are the following: Harakat[1], Mishkal[2], Farasa[3]. Harakat is a system of automatic translation. It has the advantages of all three existing approaches syntax, statistics and semantics. Mishkal is a system for diacratizing Arabic texts automatically in order to be used in reading, teaching, and resolving the ambiguity; it uses a linguistic approach based on stages of morphological, grammatical and semantic analysis to select the appropriate form from several possible cases. Farasa is a national research institute; it focuses on the future needs of its stakeholders by developing cutting-edge applied computing research, helping to identify specific problems and generating tested and proven solutions. Farasa vision is to be a global leader of computing research in identified areas that will bring positive impact to the lives of citizens and society.

Benchmarking the three systems concerning the generation of diacratized alphabetic characters out of numerical values shows that Harakat, Mishkal and Farasa were not able to change the numerical values to alphabetic diacratized characters as shown in figures 4, 5 and 6.

However, Alserag results show that the system was able to convert the numerical to alphabetic characters and diacritize them as shown in figure7. The input used during the benchmarking is shown in sentences number 18, 19, 20, 21 and 22.

Sentence 18: 12 مجلة
Sentence 19: 4 صحف
Sentence 20: 1000 كتاب
Sentence 21: 22 تلميذة
Sentence 22: 66 بنت

Figure 4: The output of Harakat.

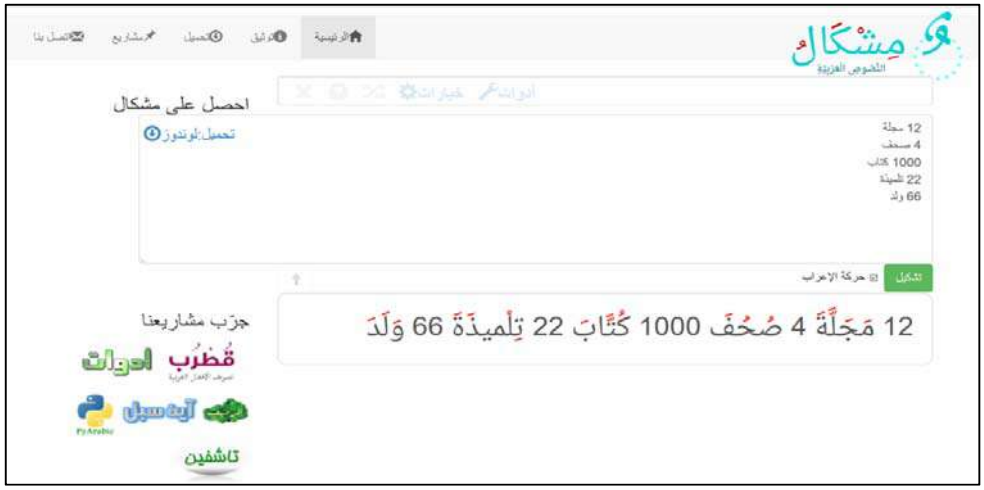


Figure 5: The output of Meshkal

[1] https://harakat.ae/
[2] http://tahadz.com/mishkal
[3] http://qatsdemo.cloudapp.net/farasa/demo.html

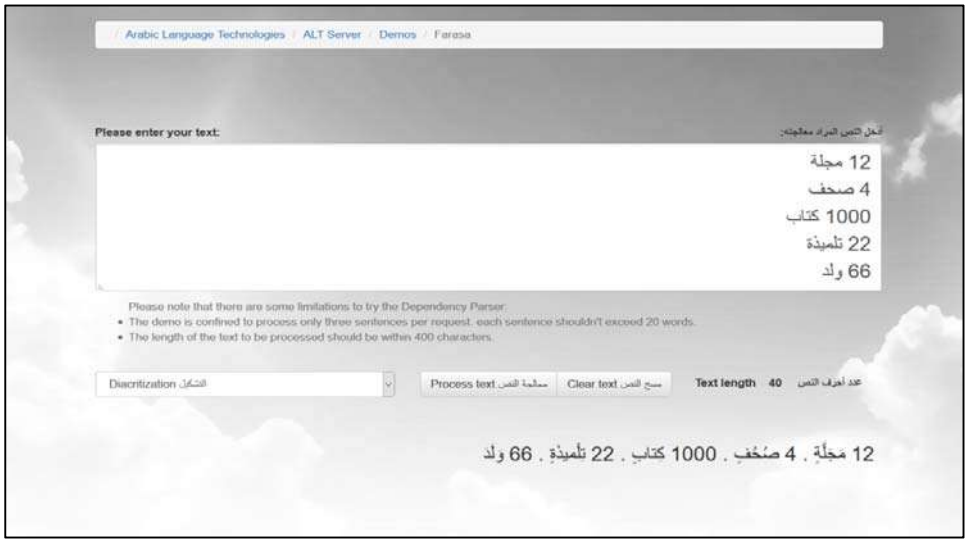


Figure 6: The output of Farasa.

Figure 7: The output of Alserag.

5 RESULTS AND EVALUATION

This section discusses the process of evaluating and testing the number module built in Alserag Arabic diacratization system using testing data from Arabic Language Technologies Society (ALTEC). The testing included about 3355 sentences from ALTEC data. A sample of the tested data is shown in figure 8.Those sentences in figure 8 were tested and benchmarked with the ALTEC reference corpus and the output is shown in figure 9. According to the results obtained, Alserag system scored less than 10% error rate.


" -ليست لنا أية علاقة بالقرار 425وهو لا يخصنا كثيرة."
" -هزيمة حزيران 1967ليست مجرد لكسة عسكرية، إنها هزيمة أوضاع عربية وهزيمة أنظمة عسكرية وهزيمة بداية معينة للحركة الوطنية العربية."
1968 أحرز المركز الأول في بطولة سوريا للأشبال في الخطى والجمباز عام 1968م.
68 -م (وصل إلى العرش لأنه كان ابن كلوديوس بالتبني.
-أبرز نجوم سوريا في دورة المتوسط في المغرب عام 1983.
-أدرك أبو زرد أن رياضة كمال وبناء الأجسام في منطقة الشرق الأوسط لا زالت بعيدة جدا عن المستوى العالمي آنذاك بالمقارنة ما بين المحترفين في منطقة الشرق الأوسط والغرب)أوربا وأميركا (لذا عزم أبو زرد على السفر إلى الغرب لاكتشاف وتعلم أسرار وكمالات رياضة كمال الأجسام وهذا ما حصل عندما تمت دعوته للمشاركة بأحد البطولات في إيطاليا فور حصوله على المركز الأول في بطولة الجمهورية العربية السورية 1987.
-أعلن الفلكيون أن 99 – 90من كتلة الكون مفقودة أو غير مرئية.
-أنشأ مدرسة كيفورك مردكيان لتعليم كرة القدم للصغار عام 1985.
-أول مشاركة رسمية لأبي زرد كانت في عام 1987حيث شارك في بطولتي الجمهورية العربية السورية وأحرز المركز الأول وبطولة فلسطين العامة وأحرز أيضا المركز الأول على جميع الفئات.
-استلم فريق رجال حطين نهاية عام 2005.
-استلم منتخب ناشئي سوريا عام 2001وشارك معهم بتصفيات كأس آسيا واستمر معهم حتى منتخب شباب سوريا وتأهل بهم لكأس العالم 2005في هولندا وقدم هذا المنتخب مستوى مشرفا جدا وتأهل المنتخب للدور الثاني بعد فوز على إيطاليا وتعادل مع كلدا وخسارة من كولومبيا وخرج من البطولة بعد خسارة مباراة مشرفة مع البرازيل بهدف وحيد جاء من ضربة جزاء ظالمة أهداها حكم المباراة للمنتخب البرازيلي ولعب المنتخب السوري مباراة لن تنسى.
-افتتح أبو زرد الفرع الثاني للنادي الأولمبي عام 2004في منطقة الحدوي في دمشق -سوريا.
البداية كانت لا تزيد على الثمانين ألف جنيه مصري عام 1920، لتصل قيمة رأس مال مجموعة شركاته عند اندلاع الحرب العالمية الثانية إلى خمسة ملايين جنيه، ولم يقتصر تأثير البنك على مجموعة شركاته فقط وإنما لعب دورا رئيسيا في تشكيل السياسة المالية للحكومة خلال عقدي العشرينيات والثلاثينيات من القرن العشرين كما كان مؤثرا في مجال تنمية الاقتصاد المصري حيث امتدت أنشطته خارج القطر المصري إلى دول الوطن العربي.
-بداية عام 1964، غادر جورج حبش إلى القاهرة، لمقابلة الرئيس عبد الناصر، وتكونت بينهما علاقات صداقة ومودة.
-بعد اعتزاله عاد إلى ناديه الأم حطين كمدرب من عام 1985حتى 1990وصعد به إلى الدرجة الأولى.
-بعد الاستقالة، تم الاتفاق أن يغادر جورج حبش إلى عمان)سقف السيل (لافتتاح عيادة 1952، وبعد فترة يلحق به وديع حداد.
-بعد مرور سنتين على سفر أبي زرد إلى إيطاليا شارك في أول بطولة خارج الشرق الأوسط على صعيده الشخصي وكانت بطولة روما 1989أحرز فيها المركز الثاني، وفي عام 1991شارك أبو زرد للمرة الثانية في بطولة وسط إيطاليا وأحرز المركز الأول.
-تصفيات أولمبياد لوس أنجلوس 1983.
-ثورة -39- 36في فلسطين -خلفيات وتفاصيل وتحليل.
-جائزة ابن رشد للفكر الحر- أكتوبر /تشرين الأول 2008
-حاز على الكرة الذهبية)هداف العرب (كأفضل لاعب عربي في استفتاء مجلة الصقر القطرية عام 1978.

Figure 8:Sentences with numbers "Digits".


" -لَيْسَتْ لَنا أَيَّةُ عَلاقَةٍ بِالقَرارِ [أَرْبَعُمِائَةٍ وَخَمْسَةٍ وَعِشْرِينَ] وَهُوَ لا يَخُصُّنا كَثِيرًا."
" -هَزِيمَةُ حُزَيْرانَ [أَلْفٍ وَتِسْعُمِائَةٍ وَسِتَّةٍ وَسِتِّينَ] لَيْسَتْ مُجَرَّدَ لَكْسَةٍ عَسْكَرِيَّةٍ، إِنَّها هَزِيمَةُ أَوْضاعٍ عَرَبِيَّةٍ وَهَزِيمَةُ أَنْظِمَةٍ عَسْكَرِيَّةٍ وَهَزِيمَةُ بِنْيَةٍ مُعَيَّنَةٍ لِلْحَرَكَةِ الْوَطَنِيَّةِ الْعَرَبِيَّةِ."
-[أَلْفٍ وَتِسْعُمِائَةٍ وَثَمانِيَةٍ وَسِتِّينَ] – أَحْرَزَ الْمَرْكَزَ الأَوَّلَ فِي بُطولَةِ سُورِيَا لِلأَشْبالِ فِي الْخُطَى وَالْجُمْبازِ عامَ [أَلْفٍ وَتِسْعُمِائَةٍ وَثَمانِيَةٍ وَسِتِّينَ]م.
-[ثَمانِيَةٍ وَسِتِّينَ] م (وَصَلَ إِلَى الْعَرْشِ لأَنَّهُ كانَ ابْنَ كَلُودِيُوسَ بِالتَّبَنِّي.
-أَبْرَزُ نُجُومِ سُورِيَا فِي دَوْرَةِ الْمُتَوَسِّطِ فِي الْمَغْرِبِ عامَ [أَلْفٍ وَتِسْعُمِائَةٍ وَثَلاثَةٍ وَثَمانِينَ].
-أَدْرَكَ أَبُو زَرْدٍ أَنَّ رِياضَةَ كَمالِ وَبِناءِ الأَجْسامِ فِي مِنْطَقَةِ الشَّرْقِ الأَوْسَطِ لا زالَتْ بَعِيدَةً جِدًّا عَنِ الْمُسْتَوَى الْعالَمِيِّ آنَذاكَ بِالْمُقارَنَةِ ما بَيْنَ الْمُحْتَرِفِينَ فِي مِنْطَقَةِ الشَّرْقِ الأَوْسَطِ وَالْغَرْبِ)أُورُبَّا وَأَمِيرِكا (لِذا عَزَمَ أَبُو زَرْدٍ عَلَى السَّفَرِ إِلَى الْغَرْبِ لاكْتِشافِ وَتَعَلُّمِ أَسْرارِ وَكَمالاتِ رِياضَةِ كَمالِ الأَجْسامِ وَهَذا ما حَصَلَ عِنْدَما تَمَّتْ دَعْوَتُهُ لِلْمُشارَكَةِ بِأَحَدِ الْبُطُولاتِ فِي إِيطالْيا فَوْرَ حُصُولِهِ عَلَى الْمَرْكَزِ الأَوَّلِ فِي بُطُولَةِ الْجُمْهُورِيَّةِ الْعَرَبِيَّةِ السُّورِيَّةِ [أَلْفٍ وَتِسْعُمِائَةٍ وَسَبْعَةٍ وَثَمانِينَ].
-أَعْلَنَ الْفَلَكِيُّونَ أَنَّ [تِسْعِينَ] – [تِسْعَةً وَتِسْعِينَ] مِنْ كُتْلَةِ الْكَوْنِ مَفْقُودَةٌ أَوْ غَيْرُ مَرْئِيَّةٍ.
-أَنْشَأَ مَدْرَسَةَ كِيفُورْكْ مَرْدِكْيانْ لِتَعْلِيمِ كُرَةِ الْقَدَمِ لِلصِّغارِ عامَ [أَلْفٍ وَتِسْعُمِائَةٍ وَخَمْسَةٍ وَثَمانِينَ].
-أَوَّلُ مُشارَكَةٍ رَسْمِيَّةٍ لأَبِي زَرْدٍ كانَتْ فِي عامِ [أَلْفٍ وَتِسْعُمِائَةٍ وَسَبْعَةٍ وَثَمانِينَ] حَيْثُ شارَكَ فِي بُطُولَتَيِ الْجُمْهُورِيَّةِ الْعَرَبِيَّةِ السُّورِيَّةِ وَأَحْرَزَ الْمَرْكَزَ الأَوَّلَ وَبُطُولَةَ فِلَسْطِينَ الْعَاشِرَةَ وَأَحْرَزَ أَيْضًا الْمَرْكَزَ الأَوَّلَ عَلَى جَمِيعِ الْفِئاتِ.
-اسْتَلَمَ فَرِيقَ رِجالِ حِطِّينَ نِهايَةَ عامِ [أَلْفَيْنِ وَخَمْسَةٍ].
-اسْتَلَمَ مُنْتَخَبَ ناشِئِي سُورِيَا عامَ [أَلْفَيْنِ وَواحِدٍ] وَشارَكَ مَعَهُمْ بِتَصْفِياتِ كَأْسِ آسْيا وَاسْتَمَرَّ مَعَهُمْ حَتَّى مُنْتَخَبِ شَبابِ سُورِيَا وَتَأَهَّلَ بِهِمْ لِكَأْسِ الْعالَمِ [أَلْفَيْنِ وَخَمْسَةٍ] فِي هُولَنْدا وَقَدَّمَ هَذا الْمُنْتَخَبُ مُسْتَوًى مُشَرِّفًا جِدًّا وَتَأَهَّلَ الْمُنْتَخَبُ لِلدَّوْرِ الثّانِي بَعْدَ فَوْزٍ عَلَى إِيطالْيا وَتَعادُلٍ مَعَ كَلُدا وَخَسارَةٍ مِنْ كُولُومْبِيا وَخَرَجَ مِنَ الْبُطُولَةِ بَعْدَ خَسارَةٍ مُبارَاةٍ مُشَرِّفَةٍ مَعَ الْبَرازِيلِ بِهَدَفٍ وَحِيدٍ جاءَ مِنْ ضَرْبَةِ جَزاءٍ ظالِمَةٍ أَهْداها حَكَمُ الْمُبارَاةِ لِلْمُنْتَخَبِ الْبَرازِيلِيِّ وَلَعِبَ الْمُنْتَخَبُ السُّورِيُّ مُبارَاةً لَنْ تُنْسَى.
-افْتَتَحَ أَبُو زَرْدٍ الْفَرْعَ الثّانِيَ لِلنّادِي الأُولَمْبِيِّ عامَ [أَلْفَيْنِ وَأَرْبَعَةٍ] فِي مِنْطَقَةِ الْحَدُوِيِّ فِي دِمَشْقَ -سُورِيَا.
الْبِدايَةُ كانَتْ لا تَزِيدُ عَلَى الثَّمانِينَ أَلْفَ جُنَيْهٍ مِصْرِيٍّ عامَ [أَلْفٍ وَتِسْعُمِائَةٍ وَعِشْرِينَ] ، لِتَصِلَ قِيمَةُ رَأْسِ مالِ مَجْمُوعَةِ شَرِكاتِهِ عِنْدَ انْدِلاعِ الْحَرْبِ الْعالَمِيَّةِ الثّانِيَةِ إِلَى خَمْسَةِ مَلايِينِ جُنَيْهٍ، وَلَمْ يَقْتَصِرْ تَأْثِيرُ الْبَنْكِ عَلَى مَجْمُوعَةِ شَرِكاتِهِ فَقَطْ وَإِنَّما لَعِبَ دَوْرًا رَئِيسِيًّا فِي تَشْكِيلِ السِّياسَةِ الْمالِيَّةِ لِلْحُكُومَةِ خِلالَ عَقْدَيِ الْعِشْرِينِيّاتِ وَالثَّلاثِينِيّاتِ مِنَ الْقَرْنِ الْعِشْرِينَ كَما كانَ مُؤَثِّرًا فِي مَجالِ تَنْمِيَةِ الاقْتِصادِ الْمِصْرِيِّ حَيْثُ امْتَدَّتْ أَنْشِطَتُهُ خارِجَ الْقُطْرِ الْمِصْرِيِّ إِلَى دُوَلِ الْوَطَنِ الْعَرَبِيِّ.
-بِدايَةُ عامِ [أَلْفٍ وَتِسْعُمِائَةٍ وَأَرْبَعَةٍ وَسِتِّينَ] ، غادَرَ جُورْجْ حَبَشْ إِلَى الْقاهِرَةِ، لِمُقابَلَةِ الرَّئِيسِ عَبْدِ النّاصِرِ، وَتَكَوَّنَتْ بَيْنَهُما عَلاقاتُ صَداقَةٍ وَمَوَدَّةٍ.
-بَعْدَ اعْتِزالِهِ عادَ إِلَى نادِيهِ الأُمِّ حِطِّينَ كَمُدَرِّبٍ مِنْ عامِ [أَلْفٍ وَتِسْعُمِائَةٍ وَثَمانِينَ] حَتَّى [أَلْفٍ وَتِسْعُمِائَةٍ وَتِسْعِينَ] وَصَعَدَ بِهِ إِلَى الدَّرَجَةِ الأُولَى.
-بَعْدَ الاسْتِقالَةِ، تَمَّ الاتِّفاقُ أَنْ يُغادِرَ جُورْجْ حَبَشْ إِلَى عَمّانَ)سَقْفِ السَّيْلِ (لافْتِتاحِ عِيادَةٍ [أَلْفٍ وَتِسْعُمِائَةٍ وَاثْنَيْنِ وَخَمْسِينَ] ، وَبَعْدَ فَتْرَةٍ يَلْحَقُ بِهِ وَدِيعُ حَدّادٌ.
-بَعْدَ مُرُورِ سَنَتَيْنِ عَلَى سَفَرِ أَبِي زَرْدٍ إِلَى إِيطالْيا شارَكَ فِي أَوَّلِ بُطُولَةٍ خارِجَ الشَّرْقِ الأَوْسَطِ عَلَى صَعِيدِهِ الشَّخْصِيِّ وَكانَتْ بُطُولَةَ رُومَا [أَلْفٍ وَتِسْعُمِائَةٍ وَتِسْعَةٍ وَثَمانِينَ] أَحْرَزَ فِيها الْمَرْكَزَ الثّانِي، وَفِي عامِ [أَلْفٍ وَتِسْعُمِائَةٍ وَواحِدٍ وَتِسْعِينَ] شارَكَ أَبُو زَرْدٍ لِلْمَرَّةِ الثّانِيَةِ فِي بُطُولَةِ وَسْطِ إِيطالْيا وَأَحْرَزَ الْمَرْكَزَ الأَوَّلَ.
-تَصْفِياتُ أُولَمْبِيادِ لُوسْ أَنْجِلُوسَ [أَلْفٍ وَثَلاثَةٍ وَثَمانِينَ]
-ثَوْرَةُ [سِتَّةٍ وَثَلاثِينَ] - [تِسْعَةٍ وَثَلاثِينَ] فِي فِلَسْطِينَ -خَلْفِيّاتٌ وَتَفاصِيلُ وَتَحْلِيلٌ.
-جائِزَةُ ابْنِ رُشْدٍ لِلْفِكْرِ الْحُرِّ- أُكْتُوبَرَ /تِشْرِينَ الأَوَّلِ [أَلْفَيْنِ وَثَمانِيَةٍ].
-حازَ عَلَى الْكُرَةِ الذَّهَبِيَّةِ)هَدّافِ الْعَرَبِ (كَأَفْضَلِ لاعِبٍ عَرَبِيٍّ فِي اسْتِفْتاءِ مَجَلَّةِ الصَّقْرِ الْقَطَرِيَّةِ عامَ [أَلْفٍ وَتِسْعُمِائَةٍ وَثَمانِيَةٍ وَسَبْعِينَ].

Figure 9: The diacritized sentences with the diacritized numeric nouns (Alserag output).

6 CONCLUSION

This paper has addressed the issue of converting the numbers into their equivalent numeric nouns and diacritizing them, after reviewing the importance of the issue. Moreover, all the previous trails were presented. The paper presents a module that is capable of converting the numeric values into their numeric nouns, taken into consideration, the case morpheme and case ending of the numeric nouns, whichis considered as our contribution to the automatic diacritization field and to the text to speech applications. Alserag system is developed based on the rule-based approach. In addition, the paper presents other available systems and how they handle the numeric nouns; however, none of the other available systems that were mentioned could handle the numeric nouns. The diacritization system is tested using the ALTEC data and the results of the system were evaluated against the reference. The system scored less than 10% error rate in the generation of the diacritized numeric nouns.

REFERENCES

[1] I. Rebai and Y. BenAyed, "Text-to-speech synthesis system with Arabic diacritic recognition system", Computer Speech & Language, 34(1), 43-60, 2015.

[2] K. Darwish, H. Mubarak and A. Abdelali "Arabic Diacritization: Stats, Rules, and Hacks." Proceedings of the Third Arabic Natural Language Processing Workshop, 2017.

[3] S. Ananthakrishnan, S. Narayanan, and S. Bangalore "Automatic diacritization of Arabic transcripts for automatic speech recognition" In Proceedings of the 4th International Conference on Natural Language Processing (pp. 47-54), 2005.

[4] D. Vergyri and K. Kirchhoff "Automatic diacritization of Arabic for acoustic modeling in speech recognition" In Proceedings of the workshop on computational approaches to Arabic script-based languages (pp. 66-73). Association for Computational Linguistics, 2004.

[5] I. Rebai and Y. BenAyed "Arabic text to speech synthesis based on neural networks for MFCC estimation" In Computer and Information Technology (WCCIT), 2013 World Congress on (pp. 1-5). IEEE, 2013.

[6] F. Biadsy, N. Habash, and J. Hirschberg "Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules" In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 397–405, Boulder, Colorado, 2009.

[7] M. Elshafei, H. Al-Muhtaseb, and M. Al-Ghamdi "Techniques for high quality Arabic speech synthesis" Information sciences. 140(3):255–267, 2002.

[8] M. Diab, M. Ghoneim, and N. Habash "Arabic Diacritization in the Context of Statistical Machine Translation" In Proceedings of Machine Translation Summit (MT-Summit) , Copenhagen, Denmark, 2007.

[9] R. Zbib, S. Matsoukas, R. Schwartz, and J. Makhoul "Decision trees for lexical smoothing in statistical machine translation." In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, pages 428–437, Uppsala, Sweden, July. Association for Computational Linguistics, 2010.

[10] M. Hamad, M. Hussain "Arabic Text-To-Speech Synthesizer." Research and Development (SCOReD), IEEE Student Conference on, 19-20 Dec. 2011.

[11] A. Youssef, O. Emam "An arabic TTS system based on the ibm trainable speech synthesizer." Le traitement automatique de l'arabe, JEP–TALN 2004.

[12] S. Alansary, "Alserag: An Automatic Diacritization System for Arabic". In International Conference on Advanced Intelligent Systems and Informatics (pp. 182-192). Springer International Publishing. (2016, October).

BIOGRAPHY

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

TRANSLATED ABSTRACT

التشكيل الآلي للنصوص العربية من أجل تطبيقات تحويل النص المكتوب إلى منطوق

سامح الأنصاري

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
sameh.alansary@bibalex.org

ملخص—**التشكيل الآلي للنصوص العربية له أهمية كبيرة على تطبيقات المعالجة الآلية للغة العربية. تشكيل النصوص مفيد جدا في تطوير تطبيقات تحويل النص المكتوب إلى كلام منطوق (TTS)، لأنه يقلل اللبس اللغوي. أصبحت تطبيقات تحويل النص المكتوب إلى كلام منطوق محور العديد من الدراسات في وقتنا الحالي. في هذه الورقة، نقدم تقنيات للتشكيل الآلي النصوص العربية ونتائج التقييم لهذه التقنيات على مجموعة من البيانات المتاحة، يتم تناول قضية توليد وتشكيل أسماء العدد بمختلف أشكالها باستخدام نظام السراج للتشكيل الآلي. استطاع النظام تحقيق أقل من 10% نسبة أخطاء في توليد وتشكيل أسماء العدد.**

DiaVator: A Tool for Evaluating Arabic Diacritization Systems

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
sameh.alansary@bibalex.org

***Abstract*— The evaluation process of any system is very important and must be subject to certain criteria. This paper discusses the problems of lacking a standard evaluation tool for evaluating the diacritized Arabic texts. It reconsiders the evaluation criteria and tries to tackle the issues of calculating the evaluation measures. The paper reviews some online automatic diacritization systems and their outputs. It presents an attempt to build an automatic evaluation tool to evaluate the quality of the diacritization systems outputs against any reference. It also discusses the results of evaluating some available diacritization systems outputs using the proposed tool and it sheds light on tool functionalities and the capabilities of handling different linguistic issues.**

1 INTRODUCTION

The process of evaluating any system is important in order to know the quality and effectiveness of using this system. Evaluating the quality of automatically diacritized Arabic systems is done through measuring the extent of similarity between the output of the diacritization system and the reference. Almost all earlier attempts for automatic diacritization have evaluated their diacritization systems and achieved various results according to their data and their diacritization approaches. For example, Gal (2002), tested on the Quran text, and achieved 14% word error rate (WER) [1]. Vergyri and Kirchhoff (2004) used acoustic features in conjunction with morphological and contextual constrains to train a diacritizer. They evaluated their automatic diacritization system on two corpora, namely FBIS and LDC Call Home ECA, and reported a 9% diacritics error rate (DER) without case ending, and 28% DER with case endings [2]. Nelken and Shieber (2005) used a cascade of probabilistic finite state transducers trained on the LDCs Arabic treebank news stories (Part 2). They achieved an accuracy of 7.33% and 23.61% WER without and with case ending respectively [3]. Zitouni et al. (2006) trained a maximum entropy model for sequence classification to restore diacritics for each character in a word. For training, they used the LDCs Arabic Tree-bank (Part 3, version 1.0) diacritized corpus. The maxEnt system achieved 5.5% DER and 18% WER on words without case ending [4]. Habash and Rambow (2007) presented "MADA-D" a system that combines a tagger and a lexeme language model. The system showed that the morphological tagger along with a 3-gram language model were able to achieve the best performance of 5.5% and 14.9% WER respectively for diacritized words without and with case ending [5]. Later work by Rashwan et al. (2009) (2015) used deep learning to improve diacritization accuracy and they reported a WER of 3.0% without case ending and 9.9% WER for guessing case ending [6][7].

They all follow the same the evaluation measures both at the word level (WER) and at the character level (DER). However, there is no criteria for calculating DER and WER. Moreover, none of them have mentioned how they calculated DER and WER, which would lead to the lack of standardization of the evaluation results that would cause unfair comparison between different diacritization systems. Moreover, there is not an evaluation tool that different diacritization systems could use to evaluate their diacritized texts against their references. Therefore, there is an urgent need to developing an evaluation system in order to achieve standardization, which will greatly benefit the field.

This paper presents an automatic evaluation tool (DiaVator) that has been built to use in evaluating the quality of any automatic Arabic diacritization system. Section 2 shows some available diacritization systems and their outputs. Section 3 discusses some general criteria to be considered in building an evaluator for evaluating the output of different diacritization systems. Section 4 presents the evaluation tool in details, how it works, the functionality of the tool, and the challenges of building the tool. Section 5 discuses experimenting the tool and discusses the results of evaluation. Section 6 includes the conclusion and future work.

2 ARABIC DIACRITIZATION SYSTEMS

A. Alserag Automatic Arabic Diacritization System[1]

Alserag is an automatic Arabic diacritization system developed under the umbrella of Bibliotheca Alexandrina. It is based on the UNL technology. It can diacritize any Arabic text at different levels of details. It is a rule-based diacritization system and has two phases. The first phase deals with analyzing the input text lexically, morphologically, and syntactically using a highly detailed set of disambiguation rules. The second phase deals with generating a diacritized

[1] Will be released soon

text by applying a set of case-sensitive rules responsible for generating a well formed diacritized text. The system was tested on a large data and the results ware promising [8]. Figure 1 shows the interface.

Figure 1. Alserag Diacritization System

B. Mishkal[2]

This project comes in a great vacuum, to provide an open source project for the diacritization. This project is called "Mishkal" which means "kaleidoscope"; a visual tool for making decorative colored shapes. The program is accompanied by other tools such as, morphological analysis, convert text to word list, generate various forms of name by adding appendices such as conjunctions, definite articles and pronouns, convert numbers to words, etc. Figure 2 shows the interface.

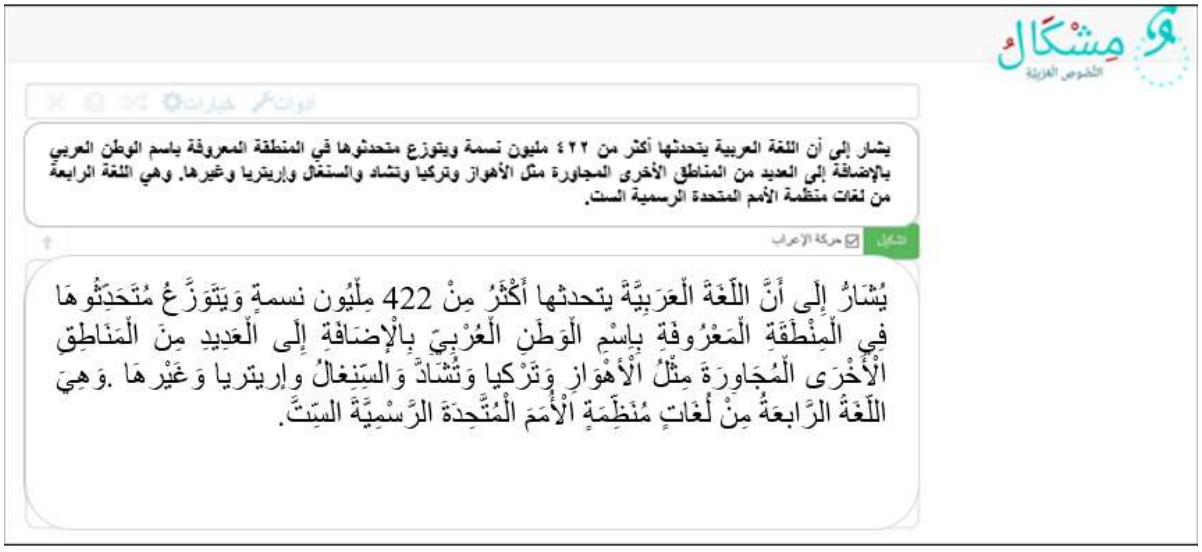


Figure 2. Mishkal Diacritization System

C. Harakat[3]

[2]https://tahadz.com/mishkal/
[3]https://harakat.ae/ar

Harakat is a system for diacritizing Arabic texts; it was developed under "Multillect" project, which is a unique idea for automatic translation technology started in 2007 in Dubai, United Arab Emirates. The website did not mention any other information about the diacritization project, but the service is available online.

Figure 3. Harakat Diacritization System

D. FARASA[4]

Farasa (means "insight" in Arabic), is a fast and accurate text processing toolkit for Arabic text. Farasa consists of a segmentation/tokenization module, POS tagger, Arabic text Diacritizer, and Dependency Parser. The core component of Farasa is the segmentation/tokenization module which is based on SVM-rank. The linear kernels used in the SVM use a variety of features and lexicons to rank possible segmentations of a word. The features include likelihoods of stems, prefixes, suffixes, and their combinations; presence in lexicons containing valid stems or named entities; and underlying stem templates. Figure 4shows the interface.

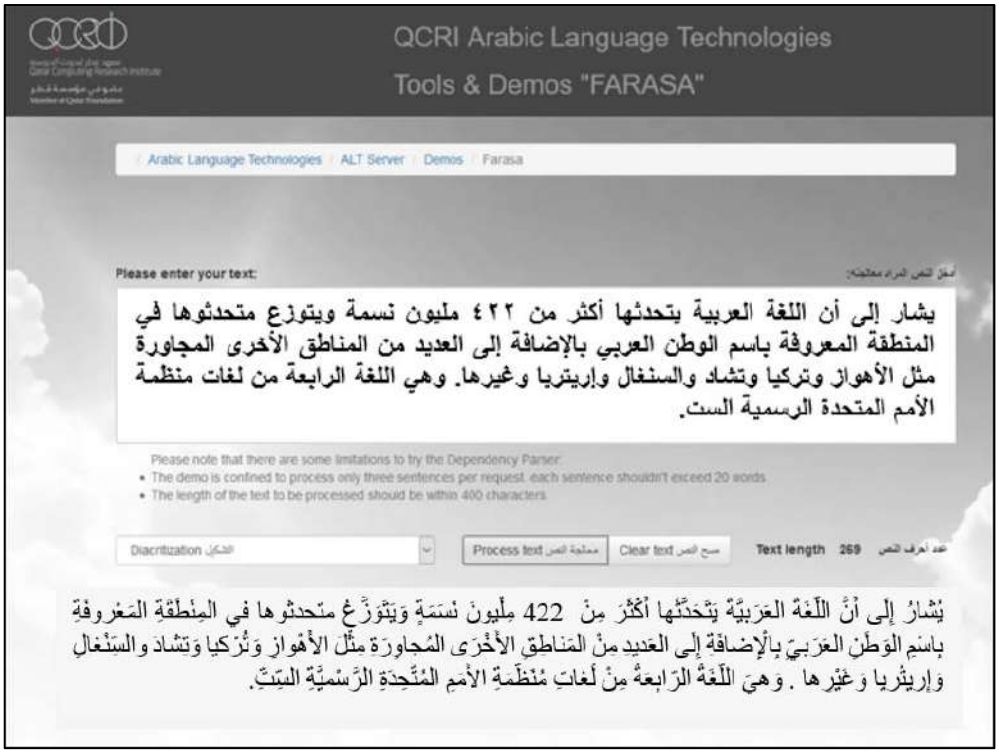


Figure 4. Farasa Diacritization System

[4]http://qatsdemo.cloudapp.net/farasa/demo.html

The output of some of these online diacritization systems may have some problems such as hidden characters, missed words during the diacritization, failure to diacritized some sentences, different document structure, different order of sentences, etc. which may cause problem when evaluating them. These problems should be solved and the two documents (reference and diacritized output) should be handled.

3 EVALUATION CRITERIA FOR THE OUTPUT OF ARABIC DIACRITIZATION SYSTEMS

According to the literature, there are two standard calculations; WER and DER; however, there can be some variations in the process of applying them, which may affect the precision of benchmarking and the comparison between the various diacritization systems. Generally, diacritization error rate (DER) which represents the proportion of characters with incorrectly restored diacritics. Word error rate (WER) which represents the percentage of incorrectly diacritized white-space delimited words: in order to be counted as incorrect, at least one letter in the word must have a diacritization error. Sometimes, the WER is divided into two types, word error rate internally (WERI) and case ending. Morphological WER is the percentage of words that have at least one morphological diacritic error, and syntactic WER is the percentage of words that have one syntactic error.

If we have a closer look at some issues, we will find that not setting standards for the way in which the error rate is calculated will cause problems. We have to be able to obtain fair evaluation results that are comparable between different diacritization systems. For example, when counting the diacritic error rate (DER) in any word that has shadda "ّ" combined with another diacritic fatha "َ", kasra "ِ", or dama "ُ" would that be counted as one error or two. Like the word "كَتَّبَ", if the reference is "كَتَّبَ" and the output is "كَتبَ", would this be considered as one diacritic or two diacritics and would it be counted as one diacritic error or two errors in the word?

Another issue is the determination of case ending diacritics in the words with attached pronouns. Such as "كتبهم", if the reference is "كُتُبُهُمْ" and the output is "كُتُبُهُمُ" is the case ending of this word right or wrong? Actually, this word is morphologically analyzed into two words; "كتب" + "هم", since "هم" is an attached pronoun, it is attached to the word "كتب" to compose the orthographic form "كتبهم" which may cause a confusion during the evaluation process for determining the case ending. Orthographically, the sokon "ْ" on the final "م" is the final diacritic, but grammatically the diacritic mark on "ب" is the case ending. So, we should have clear criteria to take this decision.

Another issue that needs a decision during the evaluation is the extra diacritics in the output over the reference. Should these extra diacritics counted as of the error rate or as an advantage of the diacritization system? For example, sometimes the reference does not diacritize the characters followed by "alef mad" such as "بابٌ", because according to some diacritization approaches, this is predictable, but the output is diacritized as "بَابٌ". Also, the final diacritic on the past active verbs with feminine singular third person subject, such as "كتبت", the reference is "كَتَبَت" and the output is "كَتَبَتْ".

Moreover, we have to consider dealing with the syntactic errors resulting from the morphological errors. In other words, if the word has an internal diacritic error that causes a case ending error. For example, if the word "طَالِبٌ" diacritized as "طَالَبَ" will it be counted once as DER only and will not be counted as a case ending error. This method is adopted by Aya S. Metwally. et al (2015), they said "Syntactic WER: the percentage of words that have one syntactic error, but don't have any morphological error. This means that if a word has both syntactic error and morphological error, it will be counted in the morphological error only." [9]. While Alansary (2015) adopted a different evaluation measure as counting the same error twice; as DER and as case ending or syntactic error.

Furthermore, the modifiers case ending that is depending on the main predicate have to be considered, for example, the word "الطالب" in the sentence like "كُتُبُ اَلطَّالِبِ جَدِيدَةٌ" is genitive, because it is "modaf elih" to the word "كُتُبُ" which is a nominativetopic. But if the word "كُتُبُ" is diacritized wrongly as "كَتَبَ", it will lead to wrong diacritization of the word "الطالب". Should the case ending of the word "الطالب" be counted as an error or not?

Finally, the things to be evaluated and the things that are excluded from the evaluation process must be agreed upon. For example, non-Arabic character, numerical symbols, symbols, and punctuations should be counted as a percentage of the total number of words to be evaluated or not.

So, if we do not consider these criteria in the evaluation tool to compare the different diacritization systems, we will not be able to know which is better. Hence, the need to build a tool that takes into account all the mentioned criteria for evaluating different diacritization systems has emerged.

4 DIAVATOR SYSTEM

There are three important questions should be answered when building any tool: 1) Is the tool effective? Does it operate quickly, smoothly and with minimal waste? 2) Is the tool easy to use? Are all of the tool's users able to use the system easily and effectively? Moreover, can anyone understand and use the system with minimal training? 3) Is the tool appropriate?

DiaVator is not just an evaluator; It could be used both by specialists and non-specialists, since it classifies the errors linguistically, which specialists could use in enhancing their diacritization systems. This section will discuss the methodology of building the evaluator, and how it works. There will be a detailed explanation for the interface and the use of each button, starting from uploading the two files up to obtaining the evaluation results. In addition, the functions

and outputs that the user can obtain from the evaluator will be explained, as well as the challenges that have been overcame in order to build a linguistically based evaluator and not just a counting tool.

The evaluation criteria of DiaVator is done using mainly two metrics; diacritization error rate (DER) which is the proportion of characters with incorrectly restored diacritics. Word error rate (WER) which is the percentage of incorrectly diacritized white-space delimited words: in order to be counted as incorrect, at least one letter in the word must have a diacritization error. These two metrics were calculated as: (1) all words are counted excluding numbers and punctuators, (2) each letter in a word is a potential host for a set of diacritics, and (3) all diacritics on a single letter are counted as a single binary (True or False) choice. Moreover, the target letter that is not diacritized is taken into consideration, as the output is compared to the reference. In addition to calculating DER and WER, DiaVator calculates internal diacritics error rate, case ending error rate and word error rate internally.

A. DiaVator (Methodology)

The evaluation tool is a desktop application that can be installed on any computer or laptop. The tool is designed to evaluate any diacritized text against any reference. The philosophy of building the evaluator is to work in a multi-dimensional array. Once the user uploads the two files; the reference and the diacritized text, the tool checks if there are any hidden characters and deletes them, then it normalizes the lines of the two files to ensure that the structure of the two documents is the same and the number of lines are equal.

B. System Interface

The evaluation process passes through four phases and each phase contains a number of functions. Phase one is the preprocessing phase where the user can upload the two files then the evaluator eliminates the non-Arabic words. This process is a preparation of the two files and some statistical information about the two documents such as the number of lines in each document, the total number of words after the elimination process with the number of the eliminated words, and the total number of valid Arabic words to be evaluated. This can be seen in figure 5.

Figure 5. The screen shot of loading button

Phase two is the process of aligning the two documents by listing the valid Arabic words of the reference and the diacritized text to ensure that they are equal, this is a kind of document normalization. During this phase, the user can show the matched two lists of words as well as the matched two lists of sentences and the errors list if the aligning process failed. This can be seen in figure 6.

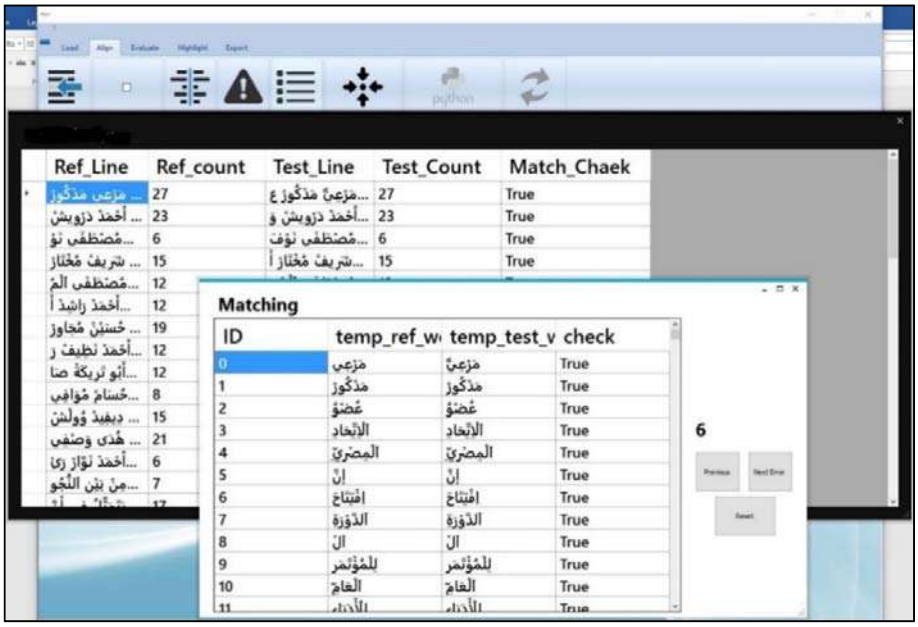


Figure 6. The screen shot of the two documents after the alignment process

The third and main phase is the evaluation process itself. In this phase, the user will have more than one option to evaluate the diacritized text. The evaluator allows evaluating the document as a whole or evaluating it as separated sentences. In addition to different types of evaluating the documents, the evaluator also take into account different reference types such as fully and partially diacritized reference. This tab also contains the calculation button, which calculates the evaluation statistics. Finally, the evaluator takes into consideration that the calculation results are detailed as possible. The evaluator provides not only the WER and DER, but also a detailed statistics about the internal diacritics that is the morphological structure of the words and the case ending diacritics that is the syntax. In addition to the statistics about which diacritics are missing and which one are wrong. The evaluation options and calculations are shown in figure 7.

Figure 7. The evaluation button options

The fourth and last phase of the evaluation is the phase of building the tables of errors and extracting different information in different forms, this will be discussed in the next subsection (C).

C. Functions and outputs

After conducting the evaluation, it will be useful for users, especially specialists, to obtain information not only about the number of errors, but also about the types of errors. DiaVator provides the users with various information about the types of errors and their linguistic classification, because DiaVator is linguistically based evaluator as we have mentioned before. Figure 8 shows that the evaluator can export the errors in an access database, the classification of the errors, the distinct list of errors with their frequency of occurrence, the list of correct sentences, or the list of wrong sentences. Figure 9 presents the exported errors in an access database, classified on three levels of details. First level classifies the errors into main two types, internal and case ending. The second level of classification is according to the source of error (dictionary, or grammar). The third level is a more linguistically detailed level, which classifies the grammar errors into errors due to disambiguation and errors due to transformation.

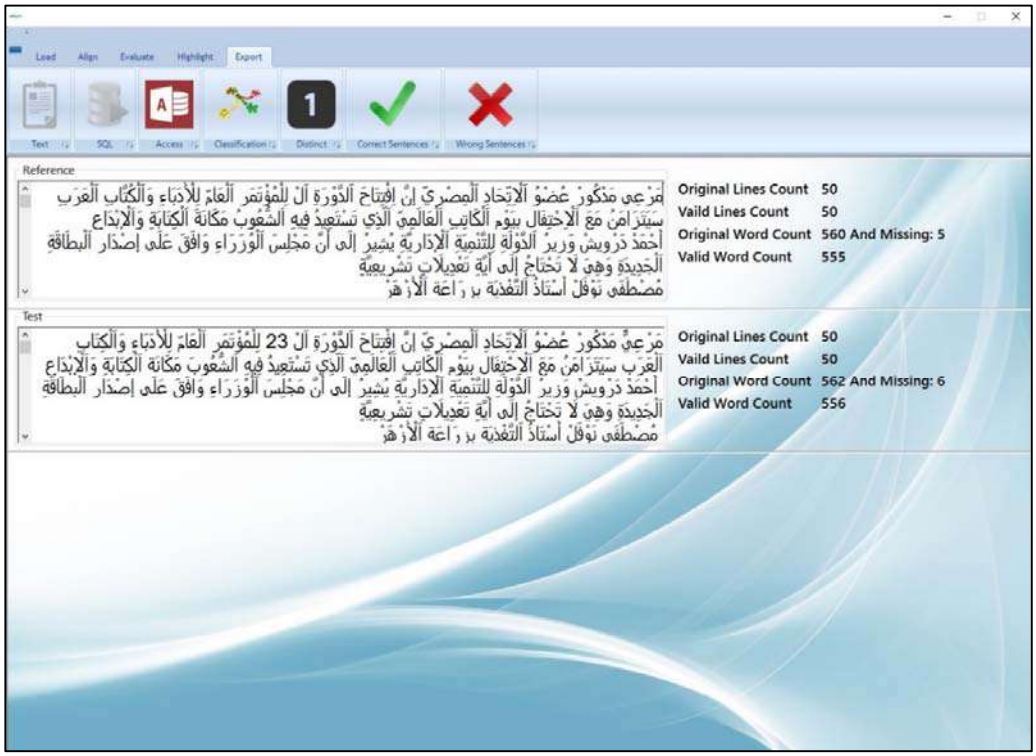


Figure 8. The export button

Input	Output	Reference	Error	Intern	Case_Enc	Calssification	freque
15 من يوم أمس الجمعة بقصف الغابة حيث كان يوجد ،	15 منْ { يَوْم } أَمسْ الْجُمْعَةْ بِقَصْفِ الْغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	15 من { يَوْم } أَمس الْجُمْعَة بِقَصْفِ الغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	يَوْم		Wrong Answered	Grammar problem – Transformation|False|False	2
15 من يوم أمس الجمعة بقصف الغابة حيث كان يوجد ،	15 منْ يَوْم { أَمسْ } الْجُمْعَةْ بِقَصْفِ الْغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	15 من يَوْم { أَمس } الْجُمْعَة بِقَصْفِ الغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	أَمسْ	Wrong Answered	Wrong Answered	Grammar problem – Disambiguation|False|False	106
15 من يوم أمس الجمعة بقصف الغابة حيث كان يوجد ،	15 منْ يَوْم أَمسْ { الْجُمْعَةْ } بِقَصْفِ الْغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	15 من يَوْم أَمس { الْجُمْعَة } بِقَصْفِ الغَابَةِ حَيْثُ كَانْ يُوجَدْ ،	الْجُمْعَةْ		Wrong Answered	Grammar problem – Transformation|True|False	63
بحسب معلوماتنا ،	{ بِحَسْبِ } مَعْلُومَاتِنَا ،	{ بِحَسْبِ } مَعْلُومَاتِنَا ،	بِحَسْبِ	Wrong Answered	Not Answered	Grammar problem – Disambiguation|False|False	42
بحسب معلوماتنا ،	بِحَسْبِ { مَعْلُومَاتِنَا } ،	بِحَسْبِ { مَعْلُومَاتِنَا } ،	مَعْلُومَاتِنَا	Not Answered		Grammar problem - disambiguation	1
لصوص (وهي التسمية التي يطلقها الروس في إشارة إلى المقاتلين الشيشان) .	{ لُصُوص } (وَهِيَ التَّسْمِيَةُ الَّتِي يُطْلِقُهَا الرُّوسُ فِي إِشَارَةٍ إِلَى الْمُقَاتِلِينَ الشِّيشَانِ) .	{ لُصُوصٌ } (وَهِيَ التَّسْمِيَةُ الَّتِي يُطْلِقُهَا الرُّوسُ فِي إِشَارَةٍ إِلَى الْمُقَاتِلِينَ الشِّيشَانِ) .	لُصُوص		Not Answered	Grammar problem – Transformation|False|False	1
لصوص (وهي التسمية التي يطلقها الروس في إشارة إلى المقاتلين الشيشان) .	لُصُوص (وَهِيَ التَّسْمِيَةُ الَّتِي { يُطْلِقُهَا } الرُّوسُ فِي إِشَارَةٍ إِلَى الْمُقَاتِلِينَ الشِّيشَانِ) .	لُصُوصٌ (وَهِيَ التَّسْمِيَةُ الَّتِي { يُطْلِقُهَا } الرُّوسُ فِي إِشَارَةٍ إِلَى الْمُقَاتِلِينَ الشِّيشَانِ) .	يُطْلِقُهَا	Wrong Answered		Grammar problem – Disambiguation|True|False	1

Figure 9. The exported database of errors

D. Challenges

The aim of building this tool is not only to build an evaluator, but also to build an intelligent evaluator with a linguistic value and not only a counting tool. Therefore, there were some linguistic challenges that have been considered and addressed to obtain the best results from the evaluation process. One of these challenges is to provide the evaluator with a simple morphological analyzer and a dictionary to handle some issues such as dealing with the attached pronouns. In example (1), the word "إنتاجها" is wrongly diacritized in the output, but what is the class of this error? Is it an internal or case ending error? During the development of DiaVator this problem has been faced, so it was important to support the evaluator with a morphological analyzer that can analyze the words like "إنتاجها" and decide that the wrong diacritic "ـَ" on the letter "جَ" is a case ending error not an internal, and the diacritization on "هَا"is not the case of the word.

(1) Reference: "إِنَّها مُسْتَعِدَّةٌ لِرَفْعِ {**إِنْتاجِهَا**} مِن النَّفْطِ بِمُعَدَّل 500 أَلْفِ بَرْمِيلٍ فِي اليَوْمِ"

Output: "إِنَّهَا مُسْتَعِدَّةٌ لَرَفَعَ {**إِنْتَاجَهَا**} مِنْ اَلنَّفْطِ بِمُعَدَّل 500 أَلْف بَرْمِيلٍ فِي اَلْيَوْمِ"

More complicated example of the same issue is deciding that the final "هـ" in words like "شِبْهِ، وَجْهِ، تِجَاهَ، جُنَيْهٍ" are different from the words that end with "هـ" as an attached pronoun, such as "رَصِيدُهُ، بِلادُهُ، لَوْحاتُهُ، وُصُولِهِ". The evaluator can distinguish between these two types of words during the evaluation process in order to calculate the errors correctly.

Other linguistic issue that has been considered during building the evaluator is the problem of classifying the errors as disambiguation problem, which means that the word is wrongly diacritized due to wrong selection, although the word is stored in the dictionary. In example (2), the word "كتاب" is internally diacritized in two different ways in the reference, and the output. The output has selected the wrong form, which may be due to one of two reasons. First reason, the word does not exist in the dictionary with this diacritization form, so the error will be classified as a dictionary problem. The second reason is that the two diacritization forms are stored in the dictionary, but the grammar rules failed to distinguish between them in this context, so this is a grammar; disambiguation problem.

(2) Reference: "وَتَتَشَكَّلُ الحُكُومَةُ الجَدِيدَةُ مِن 17 وَزِيرًا وَاِثْنَيْنِ مِن {**كُتَّابِ**} دَوْلَةٍ لَدَى الرِّئاسَةِ"

Output: "وَتَتَشَكَّلُ اَلْحُكُومَةُ اَلْجَدِيدَةُ مِنْ 17 وَزِيرًا وَاِثْنَيْنِ مِنْ {**كِتَابِ**} دَوْلَةٍ لَدَى اَلرِّئَاسَةَ"

In addition to the linguistic challenges, there are other types of challenges related to different diacritization approaches of the references such as undiacritizing the characters followed by "alef mad", for example, "كِتابٌ", and undiacritized definite articles. Different positions of nunation (Tanween Fatha "ـً") on the final character or pre-final character. Undiacritiez the final character when it should take skoun "ـْ" such as "كَتَبَت".

5 TESTING DIAVATOR FUNCTIONALITY

The evaluator has been tested by evaluating three different samples from different sources against three different references. The first sample is selected from the international Corpus of Arabic (ICA) [10]. The second sample is selected from LDC [11]. The third sample is selected from ALTEC data.

The first sample from ICA has been diacritized automatically using Alserag diacritization system. Figure 10 shows the diacritized output of Alserag system and figure 11 shows its reference.

مَرْعِيٌّ مَذْكُورْ عُضْوُ اَلْاِتِّحَادِ اَلْمِصْرِيِّ: إِنَّ اِفْتِتَاحَ اَلدَّوْرَةِ اَلْ23 لِلْمُؤْتَمَرِ اَلْعَامِّ لِلْأُدَبَاءِ وَاَلْكِتَّابِ اَلْعَرَب سَيَتَزَامَنُ مَعَ اَلْاِحْتِفَالِ بِيَوْمِ اَلْكَاتِبِ اَلْعَالَمِيِّ اَلَّذِي تَسْتَعِيدُ فِيهِ اَلشُّعُوبَ مَكَانَةَ اَلْكِتَابَةِ وَاَلْإِبْدَاعِ

أَحْمَدْ دَرْوِيشْ وَزِيرُ اَلدَّوْلَةِ لِلتَّنْمِيَةِ اَلْإِدَارِيَّةِ يُشِيرُ إِلَى أَنَّ مَجْلِسَ اَلْوُزَرَاءِ وَافَقَ عَلَى إِصْدَارِ اَلْبِطَاقَةِ اَلْجَدِيدَةِ وَهِيَ لَا تَحْتَاجُ إِلَى أَيَّةِ تَعْدِيلَاتٍ تَشْرِيعِيَّةٍ

مُصْطَفَى نَوْفَلْ أُسْتَاذُ اَلتَّغْذِيَةِ بِزِرَاعَةِ اَلْأَزْهَرْ

شَرِيفْ مُخْتَارْ أُسْتَاذُ اَلْقَلْبِ بِقَصْرْ اَلْعَيْنِيْ اَلَّذِي قَرَّرَ أَنَّهَا بِحَاجَةٍ إِلَى جِهَازٍ لِتَنْظِيمِ ضَرَبَاتِ اَلْقَلْبِ

مُصْطَفَى اَلْمُنَيِّرِي طَبِيبُ اَلْفَرِيقِ يَخْضَعُ لِجِرَاحَةٍ فِي يَدِهِ يَوْمَ اَلسَّبْتْ بِأَحَدِ اَلْمُسْتَشْفَيَاتِ

أَحْمَدْ رَاشِدْ أُسْتَاذُ أَمْرَاضِ اَلنِّسَاءِ وَاَلتَّوْلِيدِ بِطِبِّ عَيْنْ شَمْسْ و رَئِيسُ جَمْعِيَّةِ اَلْمِينُوبُوزْ

حُسَيْنْ مُجَاوِرْ رَئِيسُ اِتِّحَادِ اَلْعُمَّالِ وَرَئِيسِ اَللَّجْنَةِ اَلْمُشْرِفَةِ عَلَى اَلْاِنْتِخَابَاتِ قَالَ: إِنَّ نَتَائِجَ اَلْاِنْتِخَابَاتِ اَلَّتِي سَتُجْرَى اَلْيَوْمَ سَتُعْلَنُ غَدًا

أَحْمَدْ نَظِيفْ رَئِيسُ مَجْلِسِ اَلْوُزَرَاءِ وَعَدَد مِنْ اَلْوُزَرَاءِ وَاَلْقِيَادَاتِ اَلشَّعْبِيَّةِ وَاَلتَّنْفِيذِيَّةِ بِاَلْمُحَافَظَةِ

أَبُو تَرِيكَةْ صَاحِبُ هَدَفَ اَلْفَوْزُ اَلْقَاتِلُ عَلَى اَلصَّفَاقِسِيِّ اَلتُّونِسِيِّ فِي عُقْرِ دَارِهِ

حُسَامْ مُوَافِي أُسْتَاذُ طِبِّ اَلْحَالَاتِ اَلْحَرِجَةِ جَامِعَةُ اَلْقَاهِرَةْ

دِيفِيدْ وُولْشْ مُسَاعِدْ وَزِيرَةُ اَلْخَارِجِيَّةِ اَلْأَمْرِيكِيَّةِ: نَنْتَظِرُ اَللَّحْظَةَ اَلْمُنَاسِبَةَ لَحَلَّ اَلْقَضِيَّةَ اَلْفِلَسْطِينِيَّةَ وَلَا نَعُدُ بِشَيْءٍ

Figure 10. Automatically diacritized text from ICA

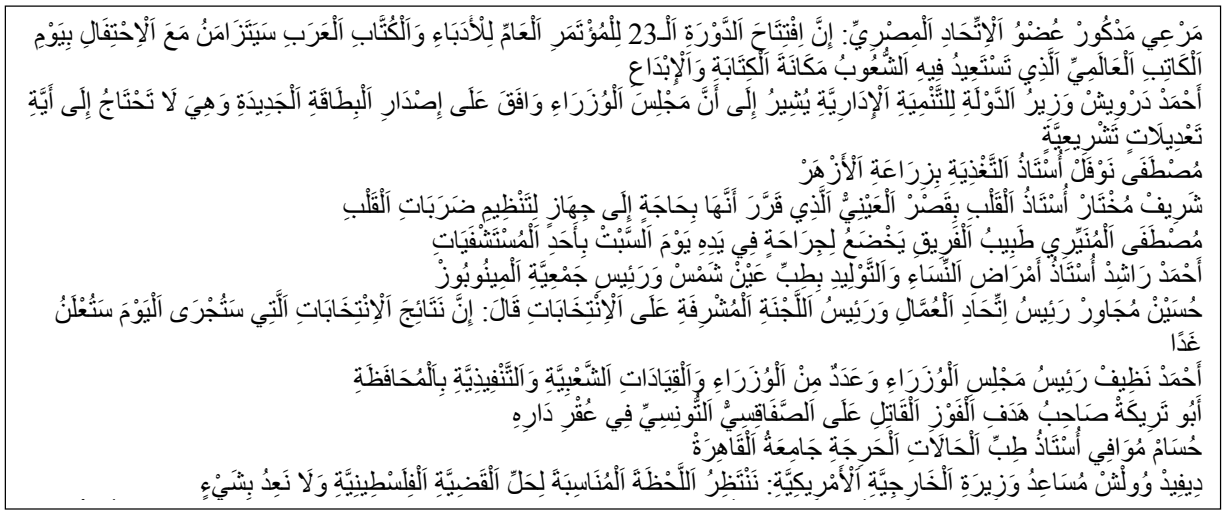

مَرْعِي مَذْكُورْ عُضْوُ اَلْإِتِّحَادِ اَلْمِصْرِيِّ: إِنَّ اِفْتِتَاحَ اَلدَّوْرَةِ اَلـ23 لِلْمُؤْتَمَرِ اَلْعَامِّ لِلْأَدَبَاءِ وَاَلْكُتَّابِ اَلْعَرَبِ سَيَتَزَامَنُ مَعَ اَلْإِحْتِفَالِ بِيَوْمِ اَلْكَاتِبِ اَلْعَالَمِيِّ اَلَّذِي تَسْتَعِيدُ فِيهِ اَلشُّعُوبُ مَكَانَةَ اَلْكِتَابَةِ وَاَلْإِبْدَاعِ
أَحْمَدْ دَرْوِيشْ وَزِيرُ اَلدَّوْلَةِ لِلتَّنْمِيَةِ اَلْإِدَارِيَّةِ يُشِيرُ إِلَى أَنَّ مَجْلِسَ اَلْوُزَرَاءِ وَافَقَ عَلَى إِصْدَارِ اَلْبِطَاقَةِ اَلْجَدِيدَةِ وَهِيَ لَا تَحْتَاجُ إِلَى أَيَّةِ تَعْدِيلَاتٍ تَشْرِيعِيَّةٍ
مُصْطَفَى نَوْفَلْ أُسْتَاذُ اَلتَّغْذِيَةِ بِزِرَاعَةِ اَلْأَزْهَرْ
شَرِيفْ مُخْتَارْ أُسْتَاذُ اَلْقَلْبِ بِقَصْرْ اَلْعَيْنِيْ اَلَّذِي قَرَّرَ أَنَّهَا بِحَاجَةٍ إِلَى جِهَازٍ لِتَنْظِيمِ ضَرَبَاتِ اَلْقَلْبِ
مُصْطَفَى اَلْمُنَيِّرِي طَبِيبُ اَلْفَرِيقِ يَخْضَعُ لِجِرَاحَةٍ فِي يَدِهِ يَوْمَ اَلسَّبْتْ بِأَحَدِ اَلْمُسْتَشْفَيَاتِ
أَحْمَدْ رَاشِدْ أُسْتَاذُ أَمْرَاضِ اَلنِّسَاءِ وَاَلتَّوْلِيدِ بِطِبِّ عَيْنْ شَمْسْ وَرَئِيسِ جَمْعِيَّةِ اَلْمِينُوبُوزْ
حُسَيْنْ مُجَاوِرْ رَئِيسُ اِتِّحَادِ اَلْعُمَّالِ وَرَئِيسُ اَللَّجْنَةِ اَلْمُشْرِفَةِ عَلَى اَلْإِنْتِخَابَاتِ قَالَ: إِنَّ نَتَائِجَ اَلْإِنْتِخَابَاتِ اَلَّتِي سَتُجْرَى اَلْيَوْمَ سَتُعْلَنُ غَدًا
أَحْمَدْ نَظِيفْ رَئِيسُ مَجْلِسِ اَلْوُزَرَاءِ وَعَدَدٌ مِنْ اَلْوُزَرَاءِ وَاَلْقِيَادَاتِ اَلشَّعْبِيَّةِ وَاَلتَّنْفِيذِيَّةِ بِاَلْمُحَافَظَةِ
أَبُو تَرِيكَةْ صَاحِبُ هَدَفِ اَلْفَوْزِ اَلْقَاتِلِ عَلَى اَلصَّفَاقِسِيِّ اَلتُّونِسِيِّ فِي عُقْرِ دَارِهِ
حُسَامْ مُوَافِي أُسْتَاذُ طِبِّ اَلْحَالَاتِ اَلْحَرِجَةِ جَامِعَةُ اَلْقَاهِرَةْ
دِيفِيدْ وُولْشْ مُسَاعِدُ وَزِيرَةِ اَلْخَارِجِيَّةِ اَلْأَمْرِيكِيَّةِ: نَنْتَظِرُ اَللَّحْظَةَ اَلْمُنَاسِبَةَ لِحَلِّ اَلْقَضِيَّةِ اَلْفِلَسْطِينِيَّةِ وَلَا نَعِدُ بِشَيْءٍ

Figure 11. Reference from ICA

Table 1 shows the evaluation results and calculations, which highlights the detailed statistics about the diacritization accuracy. Our evaluator calculates the internal diacritic errors, which is the number of internal characters that are wrongly stored, and this is classified into two categories, the number of characters that are not diacritized at all and the number of characters that are diacritized but wrongly. As well as it calculates the case ending diacritic errors which is also classified into two categories; missing case ending refers to the characters that do not have final case, and wrong case ending refers to the characters that have wrong case. In addition to the general calculations; DER, WER, and DERI

TABLE 1. ICA SAMPLE EVALUATION RESULTS

Internal Diac. Errors	43/2191	%1.962574
Wrong Internal	29/2191	%1.323596
Missing Internal	14/2191	%0.6389776
Case Ending Diac. Errors	83/545	%15.22936
Wrong Case Ending	52/545	%9.541285
Missing Case Ending	31/545	%5.688073
Diac Error Rate (DER)	126/2736	%4.605263
Word Error Rate (WER)	94	%17.24771
Word Error Rate Internal (WERI)	27/545	%4.954128

Table 2 shows the statistics about the classification of the errors according to the source for the evaluated sample. The number of disambiguation errors is 48, the number of transformation errors is 44 and the number of dictionary errors is one in the tested sample.

TABLE 2. STATISTICS OF CLASSIFICATION OF THE IN ICA SAMPLE

Source	No. of Errors
Dictionary	1
Morphological rules	48
Syntactic rules	44

The evaluator has succeeded in classifying the errors correctly using the diacritized dictionary and the morphological analysis module, which is illustrated in table 3 as the word "كتاب" is stored in the diacritization dictionary with two diacritization forms "كُتَّاب" and "كِتَاب", but the diacritizer failed to choose the right one. Therefore, the evaluator classified the problem as a disambiguation problem.

Table 3. ERROR CLASSIFICATION SAMPLE FROM ICA

Input	Output	Reference	Error	Internal	Case ending	classification
مرعي مدكور عضو الاتحاد المصري : إن افتتاح الدورة الـ 23 للمؤتمر العام للأدباء والكتاب العرب سيتزامن مع الاحتفال بيوم الكاتب العالمي الذي تستعيد فيه الشعوب مكانة الكتابة والإبداع	مَرْعِيٌّ مَذْكُورْ عُضْوُ اَلْإِتِّحَادِ اَلْمِصْرِيِّ : إِنَّ اِفْتِتَاحَ اَلدَّوْرَةِ اَلْ 23 لِلْمُؤْتَمَرِ اَلْعَامِّ لِلْأُدَبَاءِ **{وَاَلْكُتَّابِ}** اَلْعَرَب سَيَتَزَامَنُ مَعَ اَلْإِحْتِفَالِ بِيَوْمِ اَلْكَاتِبِ اَلْعَالَمِيِّ اَلَّذِي تَسْتَعِيدفِيهِ اَلشُّعُوبَ مَكَانَةَ اَلْكِتَابَةِ وَاَلْإِبْدَاع	مَرْعِي مَذْكُورْ عُضْوُ اَلْإِتِّحَادِ اَلْمِصْرِيِّ : إِنَّ اِفْتِتَاحَ اَلدَّوْرَةِ اَلْ 23 لِلْمُؤْتَمَرِ اَلْعَامِّ لِلْأُدَبَاءِ **{وَاَلْكُتَّابِ}** اَلْعَرَبِ سَيَتَزَامَنُ مَعَ اَلْإِحْتِفَالِ بِيَوْمِ اَلْكَاتِبِ اَلْعَالَمِيِّ اَلَّذِي تَسْتَعِيدُ فِيهِ اَلشُّعُوبُ مَكَانَةَ اَلْكِتَابَةِ وَاَلْإِبْدَاع	وَاَلْكِتَابِ	Wrong Answered		Grammar problem – Disambiguation

The second sample from LDC has also been diacritized automatically using Alserag diacritization system. Figure 13 shows the diacritized output of Alserag system and figure 14 shows its reference. However, it has been noticed that some words in the reference are not fully diacritized.

لُونْغْ بِيتْشْ(اَلْوِلَايَاتُ اَلْمُتَّحِدَةُ) 7-15(إِفْ بِـ)-
عفوا، حدث خطأ أثناء التشكيل
وَغَادَرَ كِنْتْ(45 عَامًا) مَسَاءُ اَلْأَرْبِعَاءْ اَلْمَدِينَةَ مُتَوَجِّهًا إِلَى وِلَايَةِ أُوهَايُو(شَمَالُ شَرْقٍ) بَعْدَ أَنْ اِسْتَقَلَّ أَحَدَ بَاصَّاتِ شَرِكَةِ غِرِيهَاوْنْدْ اَلشَّهِيرَةِ اَلَّتِي تَجُوبُ كُلَّ اَلْوِلَايَاتِ اَلْأَمِيرْكِيَّةِ.
وَبَدَا سِتِيفِنْ كِنْتْ نَحِيلًا جِدًا،
إِلَّا أَنَّهُ اِغْتَسَلَ وَحَلَقَ ذَقْنَهُ لِلْمَرَّةِ اَلْأُولَى مُنْذُ فَتْرَةٍ لَا بُدَّ أَنْ تَكُونَ طَوِيلَةً.
وَبِمَا أَنَّ اَلْمُنَاسَبَةَ تَسْتَحِقُّ اَلْعَنَاءَ اِشْتَرَى سِرْوَالًا أَزْرَقَّ وَحِذَاءً جَدِيدَيْنِ وَمِعْطَفًا خَفِيفًا بِهَا حَيَاتَهُ اَلْجَدِيدَةَ.
وَلَمْ يَكُنْ مِنْ اَلسَّهْلِ عَلَيْهِ مُوَاجَهَةَ كَامِيرَاتِ اَلتِّلِْفِزْيُونِ وَعَدَسَاتِ اَلْمُصَوِّرِينَ وَهُوَ يَصْعَدُ اَلْبَاصّ.
وَقَالَ بِصَوْتٍ خَافِتٍ يَكَادُ لَا يَسْمَعُ " " اَلْأَمْرُ يُخِيفُنِي بَعْضُ اَلشَّيْءِ(...
) إِنَّهَا مُفَاجَأَةٌ بِالْفِعْلِ.
لَنْ أَكُونَ مُتَشَرِّدًا بَعْدَ اَلْيَوْمِ " ".
وَبَعْدَ أَنْ وَدَّعَ أَصْحَابَهُ وَعَنَاصِرِ مِنْ اَلشُّرْطَةِ وَأَبْنَاءِ اَلْحَيِّ اَلَّذِي كَانَ يَبِيتُ فِي شَوَارِعِهِ فِي اَلْعَرَاءِ صَعِدَ إِلَى اَلْبَاصِّ لِيَتْرُكَ لُونْغْ بِيتْشْ اَلْوَاقِعَةَ عَلَى بُعْدِ ثَلَاثِينَ كِلَمِ جَنُوبَ لُوسْ أَنْجِلِيسْ بَعْدَ أَنْ أَمْضَى فِيهَا نَحْوَ عُشْرَيْنِ عَامًا.
وَرَوَى اَلشُّرْطِيُّ فِي لُونْغْ بِيتْشْ دِيفِيدْ مَارِنْدَرْ أَنَّهَا اَلْمَرَّةُ اَلثَّانِيَةُ خِلَالَ يَوْمَيْنِ اَلَّتِي يُحَاوِلُ فِيهَا سِتِيفِنْ كِنْتْ اِسْتِقْلَالَ اَلْبَاصِّ لِلْحُصُولِ عَلَى اَلْمِيرَاثِ اَلْمَوْعُودِ.

Figure 12. Automatically diacritized text from LDC

لُونْغ بِيتش (الولاياتُ المُتَّحِدَةُ) 7-15 (إف ب) -
كُلُّ شَيْءٍ تَغَيَّرَ فِي حَياةِ المُتَشَرِّدِ ستِيفِن كِنْت عِنْدَما عَثَرَت عَلَيهِ شَقِيقَتُهُ بَعْدَ عَناءٍ طَوِيلٍ لِتُبْلِغَهُ بِأَنَّهُ وَرِثَ 300 أَلْفَ دُولارٍ وَبِأَنَّهُ باتَ قادِرًا عَلَى وَضْعِ حَدٍّ لِعِشْرِينَ سَنَةً مِن حَياةِ التَّشَرُّدِ فِي شَوارِعِ مَدِينَةِ لُونْغ بِيتش فِي وِلايَةِ كالِيفُورْنِيا.
وَغادَرَ كِنْت (45 عامًا) مَساءَ الأَرْبِعاءِ المَدِينَةَ مُتَوَجِّهًا إِلَى وِلايَةِ أُوهايُو (شَمالَ شَرْقٍ) بَعْدَ أَن اِسْتَقَلَّ أَحَدَ باصاتِ شَرِكَةِ غريهاوند الشَّهِيرَةِ الَّتِي تَجُوبُ كُلَّ الوِلاياتِ الأَمِيرْكِيَّةِ .
وَبَدا ستِيفِن كِنْت نَحِيلًا جِدًّا،
إِلّا أَنَّهُ اِغْتَسَلَ وَحَلَقَ ذَقْنَهُ لِلمَرَّةِ الأُولَى مُنْذُ فَتْرَةٍ لا بُدَّ أَن تَكُونَ طَوِيلَةً.
وَبِما أَنَّ المُناسَبَةَ تَسْتَحِقُّ العَناءَ اِشْتَرَى سِرْوالًا أَزْرَقَ وَحِذاءً جَدِيدَيْنِ وَمِعْطَفًا خَفِيفًا بِها حَياتَهُ الجَدِيدَةَ.
وَلَم يَكُنْ مِن السَّهْلِ عَلَيهِ مُواجَهَةُ كامِيراتِ التِّلْفِزْيُونِ وَعَدَساتِ المُصَوِّرِينَ وَهُوَ يَصْعَدُ الباصَ.
وَقالَ بِصَوْتٍ خافِتٍ يَكادُ لا يُسْمَعُ "" الأَمْرُ يُخِيفُنِي بَعْضَ الشَّيْءِ (...
) إِنَّها مُفاجَأَةٌ بِالفِعْلِ.
لَن أَكُونَ مُتَشَرِّدًا بَعْدَ اليَوْمِ "".
وَبَعْدَ أَن وَدَّعَ أَصْحابَهُ وَعَناصِرَ مِن الشُّرْطَةِ وَأَبْناءِ الحَيِّ الَّذِي كانَ يَبِيتُ فِي شَوارِعِهِ فِي العَراءِ صَعِدَ إِلَى الباصِ لِيَتْرُكَ لُونْغ بِيتش الواقِعَةَ عَلَى بُعْدِ ثَلاثِينَ كلم جَنُوبَ لُوس أَنْجِلِيس بَعْدَ أَن أَمْضَى فِيها نَحْوَ عِشْرِينَ عامًا.
وَرَوَى الشُّرْطِيُّ فِي لُونْغ بِيتش دِيفِيد مارندر أَنَّها المَرَّةُ الثّانِيَةُ خِلالَ يَوْمَيْنِ الَّتِي يُحاوِلُ فِيها ستِيفِن كِنْت اِسْتِقْلالَ الباصِ لِلْحُصُولِ عَلَى المِيراثِ المَوْعُودِ.

Figure 13. Reference from LDC

Table 4shows the same results classification as discussed in table 1. There are two extra highlighted calculations; over reference internal diacritics, which refers to the extra internal diacritic marks that have been retrieved and are not found in the reference, and over reference case ending diacritics, which are the extra case ending that have been retrieved and are not found in the reference. These two additional information are calculated according to the nature of the reference, which showed that not all characters are diacritized, which means that the reference is partially diacritized.

TABLE 4. LDC SAMPLE EVALUATION RESULTS

Internal Diac. Errors	13 ~ 9/907	%1.433 ~ %0.992
Wrong Internal	10/907	%1.102536
Missing Internal	3/907	%0.3307607
Case Ending Diac. Errors	44/339	%12.97935
Wrong Case Ending	23/339	%6.784661
Missing Case Ending	21/339	%6.19469
Diac Error Rate (DER)	57/1246	%4.574639
Word Error Rate (WER)	49	%14.45428
Word Error Rate Internal (WERI)	11/339	%3.244838
Over Reference Internal Diac.	231/907	%25.46858
Over Reference Case Ending Diac.	53/339	%15.63422

Table 5 shows the statistics about the classification of the errors of the evaluated sample according to the source. The number of disambiguation errors is 15, the number of transformation errors is 31, and the number of dictionary errors is 3 in the tested sample.

TABLE 5. STATISTICS OFCLASSIFICATION OF THE ERRORS IN LDC SAMPLE

Source	No. of Errors
Dictionary	3
Morphological rules	15
Syntactic rules	31

The evaluator has succeeded in classifying the errors correctly using the diacritized dictionary and the morphological analysis module, which is illustrated in table 6. The word "إِدْمَانِهِ"is classified as a transformation error, because the evaluator recognized that "ن" letter should have the case ending diacritic, and that "ه" is an attached pronoun.

TABLE 6.CLASSIFICATION OF THE ERRORS SAMPLE FROM LDC

Input	Output	Reference	Error	Internal	Case ending	classification
وقال في المرة الأولى كانت رائحته نتنة إلى حد دفع سائق الباص إلى منعه من الصعود في إشارة إلى رفضه الاغتسال وإدمانه على الكحول .	وَقَالَ فِي اَلْمَرَّةِ اَلْأُولَى كَانَتْ رَائِحَتُهُ نَتِنَةً إِلَى حَدٍّ دَفَعَ سَائِقَ اَلْبَاصِّ إِلَى مَنْعِهِ مِنْ اَلصُّعُودِ فِي إِشَارَةٍ إِلَى رَفْضِهِ اَلْإِغْتِسَالَ {وَإِدْمَانِهِ} عَلَى اَلْكُحُولِ .	وَقالَ فِي المَرَّةِ الأُولَى كانَت رائِحَتُهُ نَتِنَةً إِلَى حَدٍّ دَفَعَ سائِقَ الباصِ إِلَى مَنْعِهِ مِن الصُّعُودِ فِي إِشارَةٍ إِلَى رَفْضِهِ الإِغْتِسالَ {وَإِدْمانِهِ} عَلَى الكُحُولِ .	وَإِدْمَانِهِ		Not Answered	Grammar problem – Transformation

The third sample from ALTEC data has been diacritized automatically using Alserag diacritization system as well. Figure 15 shows the diacritized output of Alserag system and figure 16 shows its reference. This reference also have some words that are not fully diacritized.

مَكَانَتُهُ اَلْعِلْمِيَّةُ
رَوَّى أَبُو نَعِيمٍ فِي اَلْحِلْيَةِ أَقْوَالًا لِكَثِيرٍ مِمَّنْ عَاصَرُوهُ تُشِيدُ بِمَكَانَتِهِ اَلْعِلْمِيَّةِ وَزُهْدِهِ فِي اَلدُّنْيَا وَمِنْ ذَلِكَ مَا قَالَهُ عَمْرُو بْنُ دِينَارٍ وَهُوَ أَحَدُ عُلَمَاءِ اَلتَّابِعِينَ:(مَا رَأَيْتُ أَحَدًا أَعْلَمَ بِالْفَتْوَى مِنْ جَابِرْ بْنُ زَيْدَ). وَكَانَ إِيَاسُ بْنُ مُعَاوِيَةٍ وَهُوَ قَاضِي اَلْبَصْرَةِ فِي عَهْدِ عُمَرَ بِنْ عَبْدْ اَلْعَزِيزْ يَقُولُ:(أَدْرَكْتُ أَهْلَ اَلْبَصْرَةِ وَمُفْتِيهِمْ جَابِرْ بْنُّ زَيْدَ). أَمَّا عَبْدُ اَللَّهْ بِنْ عَبَّاسْ فَكَانَ يَقُولُ:(لَوْ أَنْ أَهَلَّ اَلْبَصْرَةَ نَزَلُوا عِنْدَ قَوْلِ جَابِرْ بْنُّ زَيْدَ لِأَوْسَعِهِمْ عِلْمًا عَمَّا فِي كِتَابِ اَللَّهْ). كَمَا وَصَفَهُ(اِبْن عُمَرْ)(إِنَّهُ مِنْ فُقَهَاءِ اَلْبَصْرَةِ اَلْبَارِزِينَ) بَيْنَمَا قَالَ عَنْهُ قَتَادَةْ:(إِنَّهُ عَالِمُ اَلْعَرَب). وَيَصِفُهُ أَبُو نَعِيمٍ الأصبهي بُقُولُهُ:(كَانَ لِلْعِلْمِ عَيْنًا مُعِينًا. وَرُكْنًا مكينًا. وَكَانَ إِلَى اَلْحَقِّ آبِيًا. وَمِنْ اَلْخِلَقِ هَارِبًا). كَمَا ذَكَرَهُ(اِبْن اَلْقِيَمِ) فِي أَعْلَامِ اَلْمُوَقِّعِينَ بَعْدَمَا ذَكَرَ اَلْمفتين مِنْ اَلصَّحَابَةِ ذَكَرَ اَلْمفتين مِنْ اَلتَّابِعِينَ. فَابْتَدَأَ بِالْمَدِينَةِ وَفُقَهَائِهَا. وَثَنَّى بِمَكَّةَ اَلْمُكَرَّمَةْ وَفُقَهَائِهَا. ثُمَّ ثَلَّثَ بِالْبَصْرَةِ وَذَكَرَ مِنْ فُقَهَائِهَا اَلْمفتين جَابِرْ بْنّ زَيْدَ. وَلِذَلِكَ يُعْتَبَرُ جَابِرْ بْنُّ زَيْدَ مِنْ أَبْرَزَ عُلَمَاءِ اَلْبَصْرَةِ فِي عَصْرِهِ وَأَجْمَعَ عُلَمَاءِ اَلْحَدِيثِ عَلَى عَدَالَتِهِ وَضَبْطَهُ. فَقَدْ رَوَّى عَنْهُ اَلْبُخَارِيُّ وَمُسْلِم وَأَبُو دَاوُدْ وَاَلتِّرْمِذِيُّ وَاَلنِّسَائِيّ وَمَجْمُوعَة مِنْ اَلْمُفَسِّرِينَ. وَوَرَدَتْ إِشَارَاتٍ بِمَكَانَتِهِ اَلْعِلْمِيَّةِ عِنْدَ اَلسُّيُوطِيُّ وَابْنِ حَجَرٍ وَقَالَ عَنْهُ اِبْنِ تيمية بِأَنَّهُ أَعْلَمَ اَلنَّاسَ فِي زَمَانِهِ. وَنَظَرًا لِهَذِهِ اَلْمَكَانَةِ اَلْعِلْمِيَّةِ لِجَابِرْ بْنّ زَيْدَ فَلَمْ يَسْتَطِعْ أَحَدٌ أَن يقدح فِيهِ إِلَّا أَنَّ بَعْضَ اَلْمُؤَرِّخِينَ أَنْكَرُوا عَلَاقَتَهُ بِالإباضية وَاسْتَنَدُوا عَلَى رِوَايَاتٍ ضَعِيفَةٍ أَوْ مَكْذُوبَةٍ تَقُولُ بِأَنَّهُ تَبَرَّأَ مِنْ اَلْإباضية قَبْلَ مَوْتِهِ. وَاسْتَنَدَ كُلُّ متحامل عَلَى اَلْإباضية عَلَى هَذِهِ اَلرِّوَايَاتِ لِيَبْعُدَ اَلْإباضية عَنْ جَابِرْ بْنّ زُيَّدَ. وَمِنْهُمْ مِنْ قَالٍ بِأَنَّ جَابِرْ بِنْ زَيْدْ اَلْمُحَدِّثْ وَالتَّابِعِيُّ اَلْمَشْهُور غَيْرَ جَابِرْ بْنّ زَيْدَ شَيْخُ اَلْإباضية.
وَقَدْ قَامَ اَلدُّكْتُورُ عَوَّضَ خليفات فِي كِتَابِهِ " نَشْأَةُ اَلْحَرَكَةِ اَلْإباضية " بِالرَّدِّ عَلَى هَذِهِ اَلشُّبْهَاتِ وَتَحْلِيلِهَا وَاِنْتَهَى إِلَى اَلْقَوْلِ:(بَعْدَ هَذَا اَلْعَرْضِ وَالتَّحْلِيلِ يَبْدُو أَنَّ إِنْكَارَ جَابِرْ لِعَلَاقَتِهِ بِالإباضية كَمَا تُوردُهَا بَعْضُ اَلْمَصَادِرِ اَلسُّنِّيَّةِ إِنَّمَا اِخْتَرَعَتْ مِنْ بَعْضِ رُوَاةِ اَلسُّنَّةِ اَلَّذِينَ يَرَوْنَ جابرًا شِيخًا جليلًا وَمُحَدِّثًا ثِقَةٍ. وَبِالتَّالِي فَيَجِبُ عَدَمُ إِلْصَاقِ تُهْمَةِ اَلْإباضية بِهِ حَتَّى يَعْتَبِرَ مَجْرُوحًا. وَخَاصَّةً أَن نقدة اَلْحَدِيثُ قَدْ رَفَضُوا رِوَايَات " أَصْحَاب اَلْبَدْعِ "). ثُمَّ قَالَ يَتَّضِحُ مِمَّا سَبَقَ أَنَّ جَابِرْ بِنْ زَيْدْ كَانَ وَثِيقَ اَلصِّلَة بِالْحَرَكَةِ اَلْإباضية مُنْذُ وَقْتٍ مُبَكِّرٍ. وَكَانَ لِهِ دَوْرٌ كَبِيرٌ فِي تَنْظِيمِ اَلْحَرَكَةِ وَتَطَوُّرِهَا.

Figure 14. Automatically diacritized text from ALTEC Data

مَكَانَتُهُ اَلْعِلْمِيَّةُ
رَوَى أَبُو نُعَيْمٍ فِي اَلْحِلْيَةِ أَقْوَالًا لِكَثِيرٍ مِمَّنْ عَاصَرُوهُ تُشِيدُ بِمَكَانَتِهِ اَلْعِلْمِيَّةِ وَزُهْدِهِ فِي اَلدُّنْيَا وَمِنْ ذَلِكَ مَا قَالَهُ عَمْرُو بْنُ دِينَارٍ وَهُوَ أَحَدُ عُلَمَاءِ اَلتَّابِعِينَ: (مَا رَأَيْتُ أَحَدًا أَعْلَمَ بِالْفَتْوَى مِنْ جَابِرِ بْنِ زَيْدٍ)، وَكَانَ إِيَاسُ بْنُ مُعَاوِيَةَ وَهُوَ قَاضِي اَلْبَصْرَةِ فِي عَهْدِ عُمَرَ بْنِ عَبْدِ اَلْعَزِيزِ يَقُولُ: (أَدْرَكْتُ أَهْلَ اَلْبَصْرَةِ وَمُفْتِيَهُمْ جَابِرَ بْنَ زَيْدٍ). أَمَّا عَبْدُ اَللَّهِ بْنُ عَبَّاسٍ فَكَانَ يَقُولُ: (لَوْ أَنَّ أَهْلَ اَلْبَصْرَةِ نَزَلُوا عِنْدَ قَوْلِ جَابِرِ بْنِ زَيْدٍ لَأَوْسَعَهُمْ عِلْمًا عَمَّا فِي كِتَابِ اَللَّهِ)، كَمَا وَصَفَهُ (اِبْنُ عُمَرَ) (إِنَّهُ مِنْ فُقَهَاءِ اَلْبَصْرَةِ اَلْبَارِزِينَ) بَيْنَمَا قَالَ عَنْهُ قَتَادَةُ: (إِنَّهُ عَالِمُ اَلْعَرَبِ). وَيَصِفُهُ أَبُو نُعَيْمٍ اَلْأَصْبَهَانِيُّ بِقَوْلِهِ: (كَانَ لِلْعِلْمِ عَيْنًا مَعِينًا، وَرُكْنًا مَكِينًا، وَكَانَ إِلَى اَلْحَقِّ آبِيًا، وَمِنْ اَلْخَلْقِ هَارِبًا). كَمَا ذَكَرَهُ (اِبْنُ اَلْقَيِّمِ) فِي أَعْلَامِ اَلْمُوَقِّعِينَ بَعْدَمَا ذَكَرَ اَلْمُفْتِينَ مِنْ اَلصَّحَابَةِ ذَكَرَ اَلْمُفْتِينَ مِنْ اَلتَّابِعِينَ، فَابْتَدَأَ بِالْمَدِينَةِ وَفُقَهَائِهَا، وَثَنَّى بِمَكَّةَ اَلْمُكَرَّمَةِ وَفُقَهَائِهَا، ثُمَّ ثَلَّثَ بِالْبَصْرَةِ وَذَكَرَ مِنْ فُقَهَائِهَا اَلْمُفْتِينَ جَابِرَ بْنَ زَيْدٍ. وَلِذَلِكَ يُعْتَبَرُ جَابِرُ بْنُ زَيْدٍ مِنْ أَبْرَزِ عُلَمَاءِ اَلْبَصْرَةِ فِي عَصْرِهِ وَأَجْمَعَ عُلَمَاءُ اَلْحَدِيثِ عَلَى عَدَالَتِهِ وَضَبْطِهِ. فَقَدْ رَوَى عَنْهُ اَلْبُخَارِيُّ وَمُسْلِمٌ وَأَبُو دَاوُدَ وَاَلتِّرْمِذِيُّ وَاَلنِّسَائِيُّ وَمَجْمُوعَةٌ مِنْ اَلْمُفَسِّرِينَ. وَوَرَدَتْ إِشَارَاتٌ بِمَكَانَتِهِ اَلْعِلْمِيَّةِ عِنْدَ اَلسُّيُوطِيِّ وَابْنِ حَجَرٍ وَقَالَ عَنْهُ اِبْنُ تَيْمِيَةَ بِأَنَّهُ أَعْلَمُ اَلنَّاسِ فِي زَمَانِهِ. وَنَظَرًا لِهَذِهِ اَلْمَكَانَةِ اَلْعِلْمِيَّةِ لِجَابِرِ بْنِ زَيْدٍ فَلَمْ يَسْتَطِعْ أَحَدٌ أَنْ يَقْدَحَ فِيهِ إِلَّا أَنَّ بَعْضَ اَلْمُؤَرِّخِينَ أَنْكَرُوا عَلَاقَتَهُ بِالْإِبَاضِيَّةِ وَاسْتَنَدُوا عَلَى رِوَايَاتٍ ضَعِيفَةٍ أَوْ مَكْذُوبَةٍ تَقُولُ بِأَنَّهُ تَبَرَّأَ مِنْ اَلْإِبَاضِيَّةِ قَبْلَ مَوْتِهِ، وَاسْتَنَدَ كُلُّ مُتَحَامِلٍ عَلَى اَلْإِبَاضِيَّةِ عَلَى هَذِهِ اَلرِّوَايَاتِ لِيُبْعِدَ اَلْإِبَاضِيَّةَ عَنْ جَابِرِ بْنِ زَيْدٍ. وَمِنْهُمْ مَنْ قَالَ بِأَنَّ جَابِرَ بْنَ زَيْدٍ اَلْمُحَدِّثَ وَالتَّابِعِيَّ اَلْمَشْهُورَ غَيْرُ جَابِرِ بْنِ زَيْدٍ شَيْخِ اَلْإِبَاضِيَّةِ.
وَقَدْ قَامَ اَلدُّكْتُورُ عَوَضُ خَلِيفَاتٍ فِي كِتَابِهِ " نَشْأَةُ اَلْحَرَكَةِ اَلْإِبَاضِيَّةِ" بِالرَّدِّ عَلَى هَذِهِ اَلشُّبَهَاتِ وَتَحْلِيلِهَا وَانْتَهَى إِلَى اَلْقَوْلِ: (بَعْدَ هَذَا اَلْعَرْضِ وَالتَّحْلِيلِ يَبْدُو أَنَّ إِنْكَارَ جَابِرٍ لِعَلَاقَتِهِ بِالْإِبَاضِيَّةِ كَمَا تُوردُهَا بَعْضُ اَلْمَصَادِرِ اَلسُّنِّيَّةِ إِنَّمَا اُخْتُرِعَتْ مِنْ بَعْضِ رُوَاةِ اَلسُّنَّةِ اَلَّذِينَ يَرَوْنَ جَابِرًا شَيْخًا جَلِيلًا وَمُحَدِّثًا ثِقَةً، وَبِالتَّالِي فَيَجِبُ عَدَمُ إِلْصَاقِ تُهْمَةِ اَلْإِبَاضِيَّةِ بِهِ حَتَّى يُعْتَبَرَ مَجْرُوحًا، وَخَاصَّةً أَنَّ نَقَدَةَ اَلْحَدِيثِ قَدْ رَفَضُوا رِوَايَاتِ "أَصْحَابِ اَلْبِدَعِ")، ثُمَّ قَالَ يَتَّضِحُ مِمَّا سَبَقَ أَنَّ جَابِرَ بْنَ زَيْدٍ كَانَ وَثِيقَ اَلصِّلَةِ بِالْحَرَكَةِ اَلْإِبَاضِيَّةِ مُنْذُ وَقْتٍ مُبَكِّرٍ، وَكَانَ لَهُ دَوْرٌ كَبِيرٌ فِي تَنْظِيمِ اَلْحَرَكَةِ وَتَطَوُّرِهَا.

Figure 15. Reference from ALTEC Data

Table 7 shows the same results classification as discussed in table 4, but these results show that the over reference case ending diacritics is 0 (zero) which means that every word in the reference has case ending. Nonetheless,it is not the same for the internal characters and not all characters are fully diacritized internally.

TABLE 7. ALTEC DATA SAMPLE EVALUATION RESULTS

Internal Diac. Errors	152 ~ 147/1149	%13.22889 ~ %12.79373
Wrong Internal	69/1149	%6.005222
Missing Internal	83/1149	%7.223673
Case Ending Diac. Errors	124/347	%25.73487
Wrong Case Ending	79/347	%13.76657
Missing Case Ending	45/347	%12.9683
Diac Error Rate (DER)	276/1496	%18.4492
Word Error Rate (WER)	143	%30.21037
Word Error Rate Internal (WERI)	71/347	%20.4611
Over reference character diacritics	20/1149	%1.740644
Over reference case ending diacritics	0/347	%0

Table 8 shows the statistics about the classification of the errors of the evaluated sample according to the source. The number of disambiguation errors is 43, the number of transformation errors is 52, and the number of dictionary errors is 47 in the tested sample.

TABLE 8. STATISTICS OF CLASSIFICATION OF THE ERRORS IN ALTEC DATA SAMPLE

Source	No. of Errors
Dictionary	47
Morphological rules	43
Syntactic rules	52

The evaluator succeeded in classifying the problem in the word "تطورها" as shown in table 9. It is classified as a morphological problem,although it has a case ending problem,which is usually a syntactic problem.The evaluator was able to detect the reason correctly, because this word has wrong diacritization on the morphological level. Thus, the evaluator was succsessful in predicting that the case ending problem was the result of the morphological problem.

Table 9. ERROR CLASSIFICATION SAMPLE FROM ALTEC Data

Input	Output	Reference	Error	Internal	Case ending	Classification
وكان له دور كبير في تنظيم الحركة وتطورها.	وَكَانَ لِهِ دَوْرٌ كَبِيرٌ فِي تَنْظِيمِ ٱلْحَرَكَةِ {وَتُطَوِّرُهَا}.	وَكَانَ لَهُ دَوْرٌ كَبِيرٌ فِي تَنْظِيمِ ٱلْحَرَكَةِ {وَتَطَوُّرُهَا}.	وَتُطَوِّرُهَا	Wrong Answered	Wrong Answered	Grammar problem – Disambiguation

Another function of the evaluator is to provide users with a distinct list of errors with frequency of occurrence of each error, which highlights the most frequent error, in order to improve the results quickly and noticeably. This can help specialist users to discover the weakness of their diacritization system and deal with the most frequent issues.This is shown in table 10 which is a sample output of the distinct list of errors of a diacritized text from ICA.

TABLE 10. DISTINCT ERRORS WITH FREQUENCIES

Reference	Error	Internal	Case ending	Frequency
لَهُ	لِهِ	Wrong Answer	Wrong Answered	125
أَمْسِ	أَمَسَّ	Wrong Answer	Wrong Answered	69
لَهُمْ	لَهِمْ	Wrong Answered		41
غَيْرَ	غَيْرِ		Not Answered	33
فِي	فِي	Not Answered		27
عَلَى	عَلَّى	Wrong Answer		26
بِلَا	بَلَا	Wrong Answered		26
أَنَّنِي	أَننِي		Not Answered	23
آخَرَ	آخَرِ		Wrong Answered	22
بَيْنَ	بَيَّنَ	Wrong Answered		21
حِينَ	حِينَ		Wrong Answered	21

Moreover, the evaluator can evaluate sentence by sentence, not only the diacritized text as a whole, and give statistics about the accuracy of each sentence separatly. Table 11 shows the evaluation results by sentence for a sample diacritized text from ICA. The evaluator can calculate all evaluation measures for each sentence and give a detailed statistics about the correctness of each diacritized sentence separately. It gives information about the total number of

characters and total number of words of each sentence. It is also able to highlight the correct sentences with zero errors, sentences with internal errors only, and sentences with case ending errors only.

TABLE 11. EVALUATION RESULTS BY SENTENCE

Sentence	Internal Errors	Wrong Internal	Missing Internal	Case Ending Errors	Wrong Case Ending	Missing Case Ending	DER	WER	WERI
مَرْعِيٌّ مَذْكُورْ عُضْوُ اَلْإِتِّحَادِ اَلْمِصْرِيِّ: إِنَّ اِفْتِتَاحَ اَلدَّوْرَةِ اَلْ23 لِلْمُؤْتَمَرِ اَلْعَامِّ لِلْأُدَبَاءِ وَاَلْكِتَّابِ اَلْعَرَب سَيَتَزَامَنُ مَعَ اَلْاِحْتِفَالِ بِيَوْمِ اَلْكَاتِبِ اَلْعَالَمِيِّ اَلَّذِي تَسْتَعِيدُ فِيهِ اَلشُّعُوبَ مَكَانَةَ اَلْكِتَابَةِ وَاَلْإِبْدَاع	2/118 = %1.69	2/118 = %1.69	0/118 = %0	4/27 = %14.81	2/27 = %7.40	2/27 = %7.40	6/145 = %4.13	5/27 = %18.51	1/27 = %3.70
مَحْمَدْ دَرْوِيشْ وَزِيرُ اَلدَّوْلَةِ لِلتَّنْمِيَةِ اَلْإِدَارِيَّةِ يُشِيرُ إِلَى أَنَّ مَجْلِسَ اَلْوُزَرَاءِ وَافَقَ عَلَى إِصْدَارِ اَلْبِطَاقَةِ اَلْجَدِيدَةِ وَهِيَ لَا تَحْتَاجُ إِلَى أَيَّةِ تَعْدِيلَاتٍ تَشْرِيعِيَّةٌ	0/87 = %0	0/87 = %0	0/87 = %0	1/23 = %4.34	1/23 = %4.34	0/23 = %0	1/110 = %0.90	1/23 = %4.34	0/23 = %0
مُصْطَفَى نَوْفَلْ أُسْتَاذُ اَلتَّغْذِيَةِ بِزِرَاعَةِ اَلْأَزْهَرْ	0/27 = %0	0/27 = %0	0/27 = %0	0/6 = %0	0/6 = %0	0/6 = %0	0/33 = %0	0/6 = %0	0/6 = %0
شَرِيفْ مُخْتَارْ أُسْتَاذُ اَلْقَلْبِ بِقَصْرْ اَلْعَيْنِيِّ اَلَّذِي قَرَّرَ أَنَّهَا بِحَاجَةٍ إِلَى جِهَازٍ لِتَنْظِيمِ ضَرَبَاتِ اَلْقَلْبِ	0/53 = %0	0/53 = %0	0/53 = %0	0/15 = %0	0/15 = %0	0/15 = %0	0/68 = %0	0/15 = %0	0/15 = %0
مُصْطَفَى اَلْمُنَيَّرِي طَبِيبُ اَلْفَرِيقِ يَخْضَعُ لِجِرَاحَةٍ فِي يَدِهِ يَوْمَ اَلسَّبْتْ بِأَحَدِ اَلْمُسْتَشْفَيَاتِ	0/47 = %0	0/47 = %0	0/47 = %0	0/12 = %0	0/12 = %0	0/12 = %0	0/59 = %0	0/12 = %0	0/12 = %0
أَحْمَدْ رَاشِدْ أُسْتَاذُ أَمْرَاضِ اَلنِّسَاءِ وَاَلتَّوْلِيدِ بِطِبِّ عَيْنْ شَمْسْ و رَئِيسُ جَمْعِيَّةِ اَلْمِينُوبُوزْ	1/48 = %2.08	0/48 = %0	1/48 = %2.08	1/12 = %8.33	1/12 = %8.33	0/12 = %0	2/60 = %3.33	1/12 = %8.33	1/12 = %8.33

Another test that has been done to illustrate the evaluator capability to evaluate any diacritized output of any diacritization system in order to held a benchmarking between different diacritization systems. A sample data from ICA has been diacritized automatically using the two systems, Alserag and Farasa, and the output has been evaluated against the reference. Figure 16 shows the input sample, figure 17 shows the output of Alserag system, and figure 18 shows the output of Farasa system. Figure 18 shows that although Farasa usesdifferent format for the output than that of the input, the evaluator can handle this issue and provide the results as shown in table 12.

وإن ظلت المسألة على ما هي عليه فإن العراق في طريقه الحتمي إلى التقسيم
سواء أخذ بالنظام الفيدرالي أو أبقى على النظام الحالي
وإذا كان هؤلاء الذين يعيشون في الماضي ويهتفون للعراق يحبونه بالفعل
وأزعم أنهم حينئذ سيهتدون إلى الحل الفيدرالي للعراق
اتفاق مبارك ونزار باييف على الضرورة العاجلة لإطلاق عملية السلام ودفع العلاقات الثنائية
دورة دبي الدولية الثامنة عشرة لكرة السلة
حكومة لبنان الوضع مختلف تماما حكومة قوية وقادرة وكل قراراتها بالإجماع وبتأييد وطني
مندوب إيران لدى الوكالة الدولية للطاقة الذرية
خروج أبي فلوس وأبي منشار
تكريم فاروق العقدة وكمال إسماعيل في المهرجان الثقافي لمحافظة الدقهلية
متحف شرم الشيخ القومي
تصريحات إيهود أولمرت رئيس الوزراء الإسرائيلي حول عقد لقاء بينهما
رسالة تونس محمد الخولي
مسجد أبو النصر شتا بدسوق
ميلاد نجيب محفوظ في معرض بالهناجر

Figure 16. Sample input from ICA

وَإِنْ ظَلَّتْ اَلْمَسْأَلَةُ عَلَى مَا هِيَ عَلَيْهِ فَإِنَّ اَلْعِرَاقْ فِي طَرِيقِهِ اَلْحَتْمِيِّ إِلَى اَلتَّقْسِيمِ
سَوَاءٌ أَخَذَ بِاَلنِّظَامِ اَلْفِيدِرَالِيِّ أَوْ أَبْقَى عَلَى اَلنِّظَامِ اَلْحَالِيِّ
وَإِذَا كَانَ هَؤُلَاءِ اَلَّذِينَ يَعِيشُونَ فِي اَلْمَاضِي وَيَهْتِفُونَ لِلْعِرَاقْ يُحِبُّونَهُ بِاَلْفِعْلِ
وَأَزْعُمُ أَنَّهُمْ حِينَئِذٍ سَيَهْتَدُونَ إِلَى اَلْحَلِّ اَلْفِيدِرَالِيِّ لِلْعِرَاقْ
اِتِّفَاقٌ مُبَارَكٌ وَنِزَارْ بَاييفْ عَلَى اَلضَّرُورَةِ اَلْعَاجِلَةِ لِإِطْلَاقِ عَمَلِيَّةِ اَلسَّلَامِ وَدَفْعَ اَلْعَلَاقَاتِ اَلثُّنَائِيَّةَ
دَوْرَةُ دُبَيِّ اَلدَّوْلِيَّةِ اَلثَّامِنَةَ عَشْرَةَ لِكُرَةِ اَلسَّلَّةِ
حُكُومَةُ لُبْنَانْ اَلْوَضْعُ مُخْتَلِفٌ تَمَامًا حُكُومَة قَوِيَّة وَقَادِرَة وَكُلّ قَرَارَاتِهَا بِاَلْإِجْمَاعِ وَبِتَأْيِيدٍ وَطَنِيّ
مَنْدُوبُ إِيرَانْ لَدَى اَلْوَكَالَةَ اَلدَّوْلِيَّةَ لِلطَّاقَةِ اَلذَّرِّيَّةِ
خُرُوجُ أَبِي فُلُوسْ وَأَبِي مِنْشَارْ
تَكْرِيمُ فَارُوقْ اَلْعُقْدَةُ وَكَمَالْ إِسْمَاعِيلْ فِي اَلْمِهْرَجَانِ اَلثَّقَافِيِّ لِمُحَافَظَةِ اَلدَّقَهْلِيَّةْ
مَتْحَفُ شَرْمْ اَلشَّيْخْ اَلْقَوْمِيُّ
تَصْرِيحَاتُ إِيهُودْ أُولْمِرْتْ رَئِيسُ اَلْوُزَرَاءِ اَلْإِسْرَائِيلِيِّ حَوْلَ عَقْدِ لِقَاءٍ بَيْنَهُمَا
رِسَالَةُ تُونِسْ مُحَمَّدْ اَلْخُولِيُّ
مَسْجِدُ أَبُو اَلنَّصْرْ شِتَا بِدُسُوقْ
مِيلَادُ نَجِيبْ مَحْفُوظْ فِي مَعْرِضٍ بِاَلْهَنَاجِرِ

Figure 17. Output of Alserag System

وَإِنْ ظَلَّتْ المَسْأَلَةُ عَلَى ما هيَ عَلَيْهِ فَإِنَّ العِراقَ في طَرِيقِهِ الحَتْمِيِّ إِلَى التَّقْسِيمِ . سَواءٌ أَخَذَ بِالنِّظامِ الفِيدْرالِيِّ أَوْ أَبْقَى عَلَى النِّظامِ الحالِيِّ . وَإِذا كانَ هَؤُلاءِ اَلَّذِينَ يَعيشونَ في الماضي وَيَهْتِفونَ لِلْعِراقِ يُحِبّونَهُ بِالْفِعْلِ . وَأَزْعُمَ أَنَّهُمْ حِينَئِذٍ سيهتدونْ إِلَى الحَلِّ الفِيدْرالِيِّ لِلْعِراقِ . اتِّفاقُ مُبارَكُ وَنِزارٌ باييفْ عَلَى الضَّرورَةِ العاجِلَةِ لِإِطْلاقِ عَمَلِيَّةِ السَّلامِ وَدَفْعِ العَلاقاتِ الثُّنائِيَّةِ . دَوْرَةُ دُبَيِّ الدَّوْلِيَّةِ الثّامِنَةَ عَشْرَةَ لِكُرَةِ السَّلَّةِ . حُكومَةُ لُبْنانَ الوَضْعِ مُخْتَلِفٌ تَمامًا حُكومَةً قَوِيَّةً وَقادِرَةٍ وَكُلُّ قَراراتِها بِالْإِجْماعِ وَبِتَأْيِيدٍ وَطَنِيٍّ . مَنْدوبُ إيرانُ لَدَى الوِكالَةِ الدَّوْلِيَّةِ لِلطّاقَةِ الذَّرِّيَّةِ . خُروجُ أَبي فُلوسٌ وَأَبي مِنْشارٍ . تَكْريمُ فاروقُ العُقْدَةِ وَكَمالُ إِسْماعيلَ في المَهْرَجانِ الثَّقافِيِّ لِمُحافَظَةِ الدَّقَهْلِيَّةِ . مَتْحَفُ شَرْمِ الشَّيْخِ القَوْمِيِّ. تَصْريحاتُ إيهودْ أولْمِرْتْ رَئيسِ الوُزَراءِ الإِسْرائيلِيِّ حَوْلَ عَقْدِ لِقاءِ بَيْنَهُما . رِسالَةُ تونِسَ مُحَمَّدِ الخولي . مَسْجِدُ أَبو النَّصْرِ شَتا بِدُسوقٍ . ميلادُ نَجيبُ مَحْفوظٍ في مَعْرِضٍ بِالْهَناجِرِ.

Figure 18. Output of Farasa System

TABLE 12. BENCHMARKING RESULT

Benchmarking		
Evaluation Criteria	**Alserag Evaluation**	**Farasa Evaluation**
Internal Diac. Errors	1/510 = 0.19%	179/510 = 35.09%
Wrong Internal	1/510 = 0.19%	10/510 = 1.96%
Missing Internal	0/510 = 0%	169/510 = 33.13%
Case Ending Diac. Errors	16/126 = 12.69%	40/126 = 31.74%
Wrong Case Ending	13/126 = 10.31%	38/126 = 30.15%
Missing Case Ending	3/126 = 2.38%	2/126 = 1.58%
Diac Error Rate (DER)	17/636 = 2.67%	219/636 = 34.43%
Word Error Rate (WER)	16 = 12.69%	103 = 81.74%
Word Error Rate Internal (WERI)	1/126 = %0.79	97/126 = 76.98%

6 CONCLUSION AND FUTURE WORK

The paper presented an evaluation tool (DiaVator) that can be used to evaluate any diacritized Arabic text against any reference. DiaVatoris suitable to be used by specialists and non-specialists and it can help specialists in enhancing their dacritization systems.DiaVator is a promising evaluation system; it is planned to be more developed and enhanced concering some issues. Itwill be provided with statistical analysis to be able to draw the normal distribution curve automatically. The tool will be supplied with syntactic rules as it was provided with morphological rules so that it can classify the problems according to their syntactic function in the sentence, not just their form. It is also planned to develop the information obtained from the evaluation results such as the relationship between the length of the sentence and the number of errors. In addition to some modifications in the interface.

REFERENCES

[1] Y. Gal, *"An hmm approach to vowel restoration in Arabic and Hebrew*". In Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages, pages 1–7. Association for Computational Linguistics, 2002.

[2] D. Vergyri, and K. Kirchhoff, "*Automatic diacritization of Arabic for acoustic modeling in speech recognition*". In Proceedings of the workshop on computational approaches to Arabic script-based languages, COLING'04, pages 66-73, Geneva, Switzerland. Association for Computational Linguistics, 2004.

[3] R. Nelken and S. M Shieber, "*Arabic diacritization using weighted finite-state transducers*". In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 79–86. Association for Computational Linguistics, 2005.

[4] I. Zitouni, J. S. Sorensen, and R. Sarikaya, "*Maximum entropy based restoration of Arabic diacritics*". In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 577–584. Association for Computational Linguistics, 2006.

[5] N. Habash and O. Rambow, "*Arabic diacritization through full morphological tagging*". In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, pages 53–56. Association for Computational Linguistics, 2007.

[6] M. Rashwan, M. Al-Badrashiny, M. Attia, and S. Abdou, "*A hybrid system for automatic Arabic diacritization*". In The 2nd International Conference on Arabic Language Resources and Tools, pages 54–60, 2009.

[7] M. Rashwan, A. Al Sallab, M. Raafat, and A. Rafea, "*Deep learning framework with confused sub-set resolution architecture for automatic arabic diacritization*". In IEEE Transactions on Audio, Speech, and Language Processing, pages 505–516, 2015.

[8] S. Alansary, "Alserag: *An Automatic Diacritization System for Arabic*". Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016, AISI 2016, pp 182-192, 2016.

[9] A., S.Metwally, M., A.Rashwan, and F., A.Atiya,"*A Multi-Layered Approach for Arabic Text Diacritization*". In proceeding of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 389-393). IEEE, (2016).

[10] S. Alansary,"MASAR: *A Morphologically Annotated gold Standard Arabic Resource*". In 16th International Conference on Language Engineering. The Egyptian Society of Language Engineering (ESOLE), 2016b.

[11] M.Diab,N. Habash, O.Rambow, and R. Roth. LDC Arabic treebanks and associated corpora: Data divisions manual. arXiv preprint arXiv:1309.5652, (2013).

BIOGRAPHY

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

TRANSLATED ABSTRACT

ديافاتور: أداة لتقييم أنظمة التشكيل الآلي للنصوص العربية

سامح الأنصاري

مكتبة الإسكندرية، الشاطبي، الإسكندرية، مصر
قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر
sameh.alansary@bibalex.org

ملخص—إن عملية تقييم أي نظام مهمة جدا ويجب أن تخضع لمعايير معينة .تناقش هذه الورقة مشاكل الافتقار إلى أداة تقييم قياسية لتقييم النصوص العربية المشكولة. وتعيد النظر في معايير التقييم وتحاول معالجة المسائل المتعلقة بطريقة حساب معايير التقييم .وتستعرض الورقة بعض أنظمة التشكيل الآلي على الإنترنت ومخرجاتها .وتقدم محاولة لبناء أداة للتقييم الآلي تستخدم لتقييم جودة مخرجات أنظمة التشكيل الآلي مقارنة بالنص المرجع .كماتناقش نتائج تقييم بعض مخرجات أنظمة التشكيل باستخدام الأداة المقترحة وتسلط الضوء على وظائف الأداة وقدراتها على التعامل مع مختلف المسائل اللغوية.

How to Store a Syntactically Annotated Corpus in a Database?

Israa Elhosiny*[1], SamehAl-ansary*[2]

**Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt*

[1]israa_elsayed@yahoo.com

[2]s.alansary@alexu.edu.eg

***Abstract*—there is an increasing demand for linguistic databases, as it permits studying many linguistic aspects of text necessary for building natural language processing applications. The current research focusses on one type of databases; the syntactic database. It proposes a method and design for building an Arabic syntactic dependency database. A sample Arabic corpus has been compiled; from various genres and with different Arabic structures, analysedand stored in a relational database. Extra linguistic information has been added for the database enriching purpose. The designed database enables efficient querying, quantitative analysis and statistical modelling. Some statistical operations have been achieved using R language, to prove the efficiency of the database.**

***Keywords:* Linguistic Database, Syntactic Annotation, quantitative analysis, Arabic Natural Language Processing, R language**

1 INTRODUCTION

The term 'database' refers to collections of electronic records of linguistic data. As a simple example, this might be a file or set of files of sentences. Furthermore, a database records data in a DECLARATIVE form, i.e. independent of the particular procedures needed to interpret, display or modify it. This means that the creators and maintainers of databases have avoided storage forms like the files of word-processing packages, which rely on extensive programs for display, etc. Similarly, something like a parser or generator of sentences is not a database, even if it could be argued to contain the same information. Even if that were so, these forms are by definition abstract and require either a special procedural interpretation, in which case they fail to be declarative, or, in the case of logic grammars, they require sophisticated inference if they are to be used as characterizations of data [1]. The grammars used in parsers and generators are hypotheses about data, and thus serve a completely different purpose from databases, even if ideal grammars would characterize the same information. But in fact, the information is never the same, the data is not exactly as even the best hypotheses predict [2].General-purpose database management systems are based on some formal, general model for organizing data. By far the most common type of database in use today is the so-called relational database. All the well-known DBMSs are relational databases, including Oracle, MySQL, Postgres, FileMaker Pro and Microsoft Access. The simplest type of data model is to have a single table, or "file". Each row corresponds to some object (e.g., a language) being described and each column represents a property ("attribute"), such as name, location, or Basic Word Order. A relational database consists of several tables ("relations") of this sort, linked to each other in complex ways. In an object-oriented database, data are modeled as "objects" of various types. Objects share or inherit properties according to their type; e.g., a database about word classes could let objects of the type transitive verb inherit properties of the type verb [3]. Moreover, many projects have been tried in this field and created reliable linguistic databases as in Survey of English usage (1959), Brown Corpus (1976), Computer Corpus Pilot Project (1981), TOSCA (1991), and Penn Treebank (1993)[3].

The primary goal for the creation of the database is to produce a usable research tool for the academic community. Determining syntactic relationships, though, not only require judgment, which is necessarily subjective, but also depend on one's theory of grammar. To think that such a project can be accomplished "without" a theory would be like saying that exegesis can happen without an explicit methodology or that interpretation can exist in a vacuum, without a hermeneutical theory [4].

The study proposes a method for designing an Arabic syntactic dependency database based on syntactically analyzed corpus. Additionally, linguistic and logical aspects have been considered to build the database. Ambiguous structures and subordinates have been considered in dependency tree storage. Extra linguistic information has been added to the database for enriching. Moreover, querying logic and statistical modeling were considered and planned in the database design.

This paper is divided into three sections; section 2 exhibits the bases of corpus compilation and syntactic analysis, database design and building, and adding extra linguistic information. Section 3 presents the verbs syntax-semantic classification and extraction; section 4 discusses the quantitative linguistic analysis and statistical modeling. Finally, section 5 concludes the paper.

2 THE ARABIC SYNTACTIC DATABASE

In order to build the Arabic syntactic database, three different steps have been achieved. The following subsections present three steps in details. The first subsection discusses how the sample corpus collected, and analyzed. The second presents the proposed method and structure of the database. Finally, the third subsection, exhibits the extra linguistic information which can enhance the database from the searching and statistical modeling point of view.

A. Corpus compilation and analysis

The corpus is composed of Arabic verbal sentences represent the selected Arabic verbs in [5]. The corpus is collected from the Egyptian newspaper; Al-Ahram 1999 as it is considered as being representative of modern standard Arabic. The pages of Al-Ahram are collected on the Arabicorpus website; ArabiCorpus1 allows the researcher to search in large, untagged Arabic corpora. 'Untagged' means that the words in the corpora have not been assigned to a particular part of speech. ArabiCorpus is divided into five main categories or genres: Newspapers, Modern Literature, Nonfiction, Egyptian Colloquial, and Pre-modern. User can search any text individually by using the Advanced Search mode. You can even search all of the texts at the same time. It allows search in combined, individual or all texts. The total number of words of the whole ArabiCorpus is: 173,600,000.

In order to collect an appropriate size of data for linguistic analysis, the size of the corpus to be analyzed has to be precisely estimated: it should not be too small, because it would raise the risk of not containing enough data. On the other hand, the corpus should not be too big either, since the time needed for analysis has to be also taken into account when planning corpus building. Other sentences types have been added to the corpus to address all Arabic verb types in the syntactic database; nominal sentences were added.8 words long to contain all verbs arguments with their modifiers. The corpus is 300 sentences.

Corpus has been syntactically analyzed using grammar modules in the Analysis UNL Engine; IAN [6].The grammar has several modules such as; morphological, syntactic modules. The morphological module is responsible for removing the linguistic obstacles between words to make them ready to be linked in the following module which is the syntactic module. These obstacles are blank spaces, punctuations, the accusative suffix "ا", and definite articles. For the example in (1), the morphological module will produce the output in figure (1):

(1) يصلي العرب في القدس عاصمة دولة فلسطين الحرة

```
#L(يصلي:01,العرب:18)
#L(العرب:18,في:06)
#L(في:06,القدس:08)
#L(القدس:08,عاصمة:10)
#L(عاصمة:10,دولة:12)
#L(دولة:12,فلسطين:14)
#L(فلسطين:14,الحرة:19)
```

Figure 1: The output of the morphological module

```
[S:18]
   {org}
      يصلي العرب في القدس عاصمة دولة فلسطين الحرة
   {/org}
   {unl}
      link(يصلي:01,في:06)
      sbj(يصلي:01,العرب:18)
      gen(في:06,القدس:08)
      adj(فلسطين:14,الحرة:19)
      app(عاصمة:10,القدس:08)
      poss(دولة:12,فلسطين:14)
      poss(عاصمة:10,دولة:12)
   {/unl}
[/S]
```

Figure 2: The syntactic dependency graph for " يصلي العرب في القدس عاصمة دولة فلسطين الحرة"

The syntactic module depends on the morphological module; it uses the output of the morphological analysis as its input. It is responsible for linking the words of the sentence with syntactic dependency relations. The set of syntactic relations (syntactic tags) that are used in the grammar are those used in the Quranic Arabic Dependency Treebank [7]as shown in table (1). The reason for choosing this set of relations is that it is well-equipped to provide the technical means for describing any syntactic behavior properly. The final output dependency syntactic graph is shown in figure (2).

[1] http://arabicorpus.byu.edu/

TABLE I

THE DIFFERENT TYPES OF RELATIONS

Relation	Arabic Name	Dependency Relation	Example
adj	صفة	Adjective	فلسطين الحرة
poss	مضاف إليه	Possessive construction	دولة فلسطين
app	بدل	Apposition	القدس عاصمة
spec	تمييز	Specification	ثلاثون جنيها
subj	فاعل	Subject of a verb	أكل الولد
obj	مفعول به	Object of a verb	**أكل الولد تفاحة**
subjx	اسم كان	Subject of a special verb or particle	**كان الولد** نشيطا
predx	خبر كان	Predicate of a special verb or particle	**كان** الولد **نشيطا**
gen	جار ومجرور	Preposition phrase	في الحديقة
link	متعلق	PP attachment	**الولد في** الحديقة
Conj	معطوف	Coordinating conjunction	**الولد والبنت**
Circ	حال	Circumstantial accusative	يأكل دائما
emph	توكيد	Emphasis	قد أعلن
Sub	صلة	Subordinate clause	**المشكلة التي** يعاني منها الشعب

B. Database design

The primary technological objectives of the study is to design and implement a linguistic database that can ease the storage, maintenance and retrieval of natural language data. According to the databases survey, the research considered the main three requirements; usability, extensibility, portability and simplicity [8]:

- Usability: to facilitate the application of the methodology
- Suitability: to meet the specific necessities of storing and maintaining natural language data.
- Extensibility: to enable and encourage users of the database to add linguistic data and annotations according to their needs without changes to the underlying data model.
- Portability and simplicity: to make its results available on several different hard- and software platforms easy to use.
- The design of database is based on the format of the Syntactic output; the UNL graph structure, as in figure (2). The UNL graph structure follows the SGML format. In the first line contains the sentence number which mark the beginning of the new sentence to be analyzed ; i.e. [s:18] and [/s] marks the end of the sentence. The second and fourth to mark the row sentence in between "يصلي العرب في القدس عاصمة دولة فلسطين" in the third line by "{org}" and "{/org}". The fifth line contains the tag which marks the beginning of the analyzed sentence in the UNL format by "{unl}", while "{/unl}" marks the end of analysis. The dependency based syntactic analysis follows the format "rel (node1:id, node2:id)"; the "rel" expresses the name of the binary syntactic relation which links the nodes between "(" and ")". The relation nodes are separated using ",". An automatic ID is assigned to each relation node separated by ":". The syntactic analysis output appears as structured UNL document. A "VBA" code has been written to parse automatically the syntactically analyzed document, to tansform the syntactic dependency trees to a relational structured data; access relational database.

The created relational database following the E-R Diagram in figure (3). The database contained five data tables designed and structured to serve the querying for data retrieval, exploration and modeling. The first for storing the sentences and their information, such as sentence index "ID" , length, number of relations and the ambigiouty status. The second is the table for relations; which stores the relations of each sentence, their word's nodes, and their nodes' IDs.

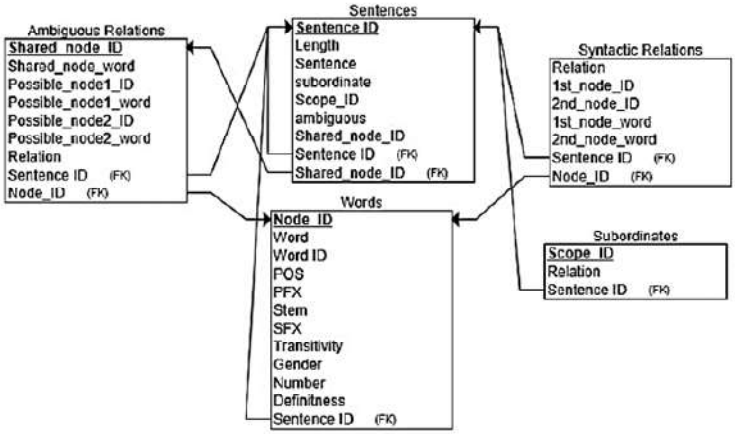


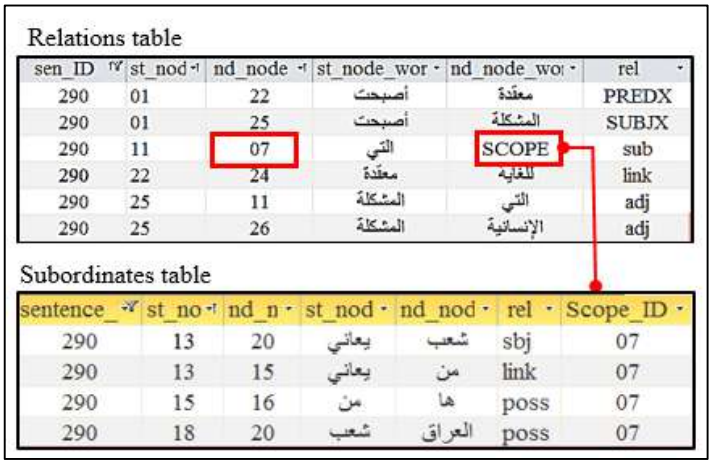
Figure 3: E-R Diagram for the Arabic Syntactic Database

Handeling subordinates: The output syntactic analysis in the complex cases, the subordinate clause, appears inside a scope (:07) as in figure (4) for the sentence "أصبحت المشكلة التي يعاني منها شعب العراق معقدة للغاية" ' the problem which is suffered by Iraq people became extremely complex '. The clause introduced by the relative pronoun "التي" which' is considered an adjectival clause for the head noun"المشكلة" 'the problem'.A scope is a group of relations between nodes that work as a single syntactic entity in a syntactically analyzed sentence. The phrase "يعاني منها شعب العراق", represented as a hyper-node (i.e., as a sub-graph) as indicated below. The nodes inside the scope; are marked in the rectangle part for the subordinate clause in figure (4).Figure (5) shows how such structures were stored in the database.

```
[S:61]
   {org}
      أصبحت هذه المشكلة الإنسانية التي يعاني منها شعب العراق معقدة للغاية
   {/org}
   {unl}
      SUBJX(أصبحت:01,المشكلة:25)
      PREDX(أصبحت:01,معقدة:22)
      adj(المشكلة:25,التي:11)
      sub(التي:11,:07)
      sbj:07(يعاني:13,شعب:18)
      poss:07(شعب:18,العراق:20)
      gen:07(من:15,ها:16)
      link:07(يعاني:13,من:15)
      link(معقدة:22,للغاية:24)
      app(هذه:03,المشكلة:25)
      adj(المشكلة:25,الإنسانية:26)
   {/unl}
[/S]
```

Figure 4: The output for subordinate clause

The scope is considered a second node; 07, in the relations table in figure (5). The node word is considered empty in the syntactic representation in figure (4), which left a null value in the field of the second node word. The word "SCOPE" has been generated automatically in the transformation (from syntactic output file to database process). The subordinate table in figure (5) stores the nodes inside scope and their relations for the organization and retrival purposes.



Relations table

sen_ID	st_nod	nd_node	st_node_wor	nd_node_wo	rel
290	01	22	أصبحت	معقدة	PREDX
290	01	25	أصبحت	المشكلة	SUBJX
290	11	07	التي	SCOPE	sub
290	22	24	معقدة	للغاية	link
290	25	11	المشكلة	التي	adj
290	25	26	المشكلة	الإنسانية	adj

Subordinates table

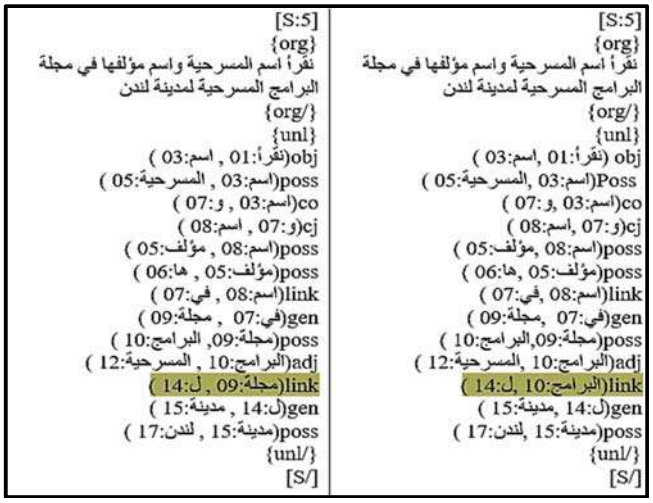
sentence_	st_no	nd_n	st_nod	nd_nod	rel	Scope_ID
290	13	20	يعاني	شعب	sbj	07
290	13	15	يعاني	من	link	07
290	15	16	من	ها	poss	07
290	18	20	شعب	العراق	poss	07

Figure 5: Subordinates in the Database

Handeling ambiguous structures: a sentence may be interpreted in more than one way due to ambiguous sentence structure. Incorrect attachment of prepositional phrases often constitutes the largest single source of errors in current parsing systems. Correct attachment of PPs is necessary to construct a parse tree, which will support the proper interpretation in the sentence.

(2) نقرأ اسم المسرحية واسم مؤلفها في مجلة البرامج المسرحية لمدينة لندن

For the sentence in (2), there are two possible interpertations for this sentence; the frist is that: "**لمدينة لندن**" 'for London city' is modifying the word "مجلة" 'magazine' to be interperted as "the magazine of London city". The second possiblity is, "**لمدينة لندن**" is modifying "البرامج" 'the programs' to be interperted as "the theater programs of London city" . Since there are two possible interpertations, there are two syntactic representation output for the input sentence as in figure (6). Both representations are the same except the highlighted relations with their nodes.



[S:5]
{org}
نقرأ اسم المسرحية واسم مؤلفها في مجلة البرامج المسرحية لمدينة لندن
{org/}
{unl}
obj (نقرأ:01, اسم:03)
Poss(اسم:03, المسرحية:05)
co(اسم:03, و:07)
cj(و:07, اسم:08)
poss(اسم:08, مؤلف:05)
poss(مؤلف:05, ها:06)
link(اسم:08, في:07)
gen(في:07, مجلة:09)
poss(مجلة:09,البرامج:10)
adj(البرامج:10, المسرحية:12)
link(البرامج:10, ل:14)
gen(ل:14, مدينة:15)
poss(مدينة:15, لندن:17)
{unl/}
[S/]

[S:5]
{org}
نقرأ اسم المسرحية واسم مؤلفها في مجلة البرامج المسرحية لمدينة لندن
{org/}
{unl}
obj(نقرأ:01, اسم:03)
poss(اسم:03, المسرحية:05)
co(اسم:03, و:07)
cj(و:07, اسم:08)
poss(اسم:08, مؤلف:05)
poss(مؤلف:05, ها:06)
link(اسم:08, في:07)
gen(في:07, مجلة:09)
poss(مجلة:09, البرامج:10)
adj(البرامج:10, المسرحية:12)
link(مجلة:09, ل:14)
gen(ل:14, مدينة:15)
poss(مدينة:15, لندن:17)
{unl/}
[S/]

Figure 6: The syntactic representations of an ambiguous structure

The 'link' dependency relation is the expressive relation for the PP attatchment. The frist node in the 'link' relation represents to which node the preposition is related; the PP attached to. For the example in hand, the preposition "ل" 'of' with the ID "14", may be relataed \ attached to the word "مجلة"; 09 or "البرامج"; 10, as two different interpertations higlighted in figure (6). Therfore, for orgnizational, economical and retrival reasons, such information for ambiguity is stored in the table of ambiguous structures. The sentence which contained ambigous structure is marked in three tables as in figure (7). In the sentences table, the 'Ambg' field is checked for the ambiguous sentence. In the relations table, all of the sentence 5 relation is stored, except the preposition "ل" and its linked word possiblities. The three nodes; the preposition "ل", the frist possiblity word "البرامج" and the second possiblity word "مجلة" and their node indexes are stored in the ambiguty table as in figure (7), in addition to the'link'relation and sentence ID.

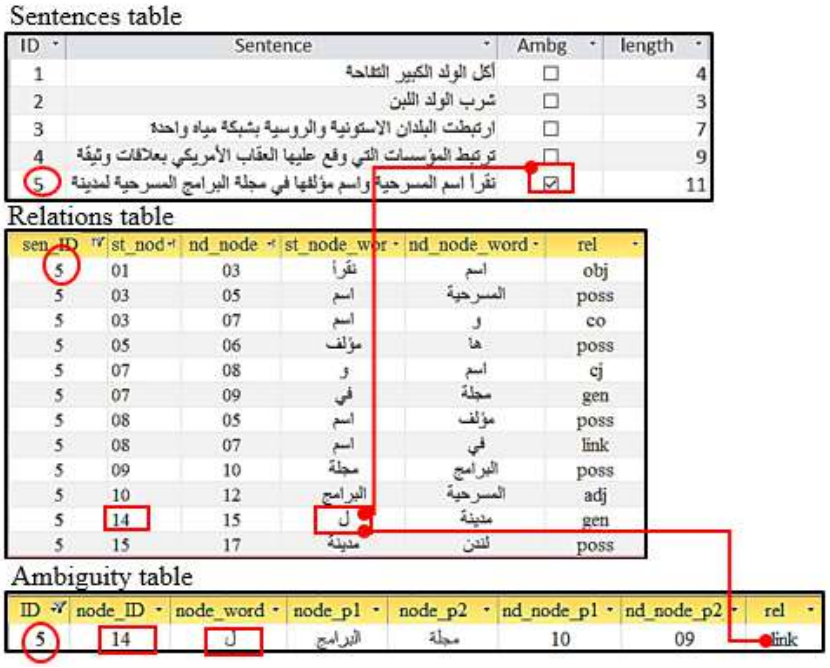


Figure 7: Ambiguous structures in the Database

The fifth table in the database stores the words of each relation occurred in the corpus, and each words' sentence are linked by sentence ID.Moreover, another process achieved on the sentences' words; morphological analysis and disambiguation to assign extra linguistic information. An automatic morphological analysis proceeded the task, by generating multiple solutions to each word as in string like "أكل", it may be a verb, a noun as shown in figure (8).

أكل	VER	_Pref-0	أكل_VER _suf-0
أكل	NOU	_Pref-0	أكل_NOU _suf-0
أكل	QUA	أ_Pef	كل_QUA _suf-0

Figure 8: The possible solutions for the input "أكل"

Therefore, POS has been disabiguated manually and other linguistic information have been added based on mapping and filtration from the UNL Arabic dictionary. Transivity, tense, voice , person, gender and Number attributes have been assigned to verbs. Moreover, gender, number and animacy attributes have been assigned to nouns in the Words' table as in figure(9).

Words | Query3

Sentence_ID	word	node_II	wordD	POS	pr	sfx	word_tag	GEN	NUM	ANI	DEC
168	التاجر	05	تاجر	NOU	DET_PREF_ال	_suf-0	NOU_تاجر	MCL	SNG		
88	التاريخ	14	تاريخ	NOU	DET_PREF_ال	_suf-0	NOU_تاريخ	MCL	SNG	NANM	
13	التجاري	24	تجاري	ADJ	DET_PREF_ال	_suf-0	ADJ_تجاري	MCL	SNG		PST
64	التجاري	24	تجاري	ADJ	DET_PREF_ال	_suf-0	ADJ_تجاري	MCL	SNG		PST
267	التجدد	29	تجدد	NOU	DET_PREF_ال	_suf-0	NOU_تجدد	MCL	SNG		
265	التجربة	31	تجربة	NOU	DET_PREF_ال	_suf-0	NOU_تجربة	FEM	SNG	NANM	
266	التجربة	28	تجربة	NOU	DET_PREF_ال	_suf-0	NOU_تجربة	FEM	SNG	NANM	
268	التجربة	35	تجربة	NOU	DET_PREF_ال	_suf-0	NOU_تجربة	FEM	SNG	NANM	
189	التحليل	08	تحليل	NOU	DET_PREF_ال	_suf-0	NOU_تحليل	MCL	SNG	NANM	
97	التصدي	27	تصدي	NOU	DET_PREF_ال	_suf-0	NOU_تصدي	MCL	SNG		
277	التصدي	27	تصدي	NOU	DET_PREF_ال	_suf-0	NOU_تصدي	MCL	SNG		
58	التصنيع	22	تصنيع	NOU	DET_PREF_ال	_suf-0	NOU_تصنيع	MCL	SNG	NANM	

Figure 9: Linguistic attributes in the words table

3 THE QUANTITATIVE LINGUISTIC ANALYSIS

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing [9]. The R language is widely used among statisticians and data miners for developing statistical software[10] and data analysis[11]. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C, C++, Java, .NET or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its lexical scoping rules. Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages.

In this study, R is used to prove the database structure quality. Three database parameters are detrmined to test the database structure. The first is; to which extent the linguist can query to extract the linguistic information he asked for. The second, is the statistcal analysis can be achieved on both before and after corpus analysis. The third parameter is , how database permits statistical language modelling for machine learning. The following subsections will explain and show with examples, how the the parameters tested.

A. Linguistic information retrieval

In order to retrieve and query linguistic information exists in the database, two packages in R were very helpful to achieve the retrival porocess; “RODBC” and “sqldf” libraries. “RODBC” library implements ODBC database connectivity. To access the database tables, “sqlFetch” command is written to read some or all of a table from an ODBC database into a data frame as in (3). The command is used 5 times to read the 5 tables of the database as in (3).

(3)
```
sentences_tb<- sqlFetch(ch, "Sentences")
relations_tb<- sqlFetch(ch, "relations")
words_tb<- sqlFetch(ch, "words")
Ambigous_tb<- sqlFetch(ch, "Ambiguous")
subordinates_tb<- sqlFetch(ch, "subordinates")
```

The other package used is “sqldf” library to provide an easy way to perform SQL selects on R data frames. The linguist may ask the database to exhibit the sentences which have an object dependency relation and doesn’t have subject as in figure (10).

```
sqldf("SELECT distinct sentences_tb.sentence , relations_tb.st_node_word, relations_tb.nd_node_word, relations_tb.rel
      FROM sentences_tb, relations_tb
      WHERE
            relations_tb.sen_ID=Sentences_tb.ID
            And relations_tb.rel = 'obj'
            And sentences_tb.all_rel not like '%sbj_%'");
```

Figure 10: Query to exhibit sentences without subject and contain object

The Answer for the query in figure (11) shows that there are 18 sentences fit the query requirements. Moreover, sentences appear with different lengths and nodes, which held the object relations ‘obj’. We can note that sentence (3) its object is a scope node; subordinate clause. In addition, we can see to which extent the object and its verb are far from each other as in sentence (9). Some other questions, the linguist may ask to the database like: a) Which POS tags can hold a predicate relation “pred”. b) Which sentences have subject relation; “sbj” relation with specific verbs such as, verb “ارتبط” or the verb “أكل” .

	Sentence	st_node_word	nd_node_word	rel
1	نقرأ اسم المسرحية واسم مؤلفها في مجلة البرامج المسرحية	نقرأ	اسم	obj
2	نقرأ في الجريدة بعد ذلك أسماء عشرات من الجماعات والأفراد	نقرأ	أسماء	obj
3	...في القرن الثامن عشر كان الفرنسيون يقولون إن لافونتين هو آخر	يقولون	Scope	obj
4	كسرت الحصالة	كسرت	الحصالة	obj
5	كسر الزجاج الخلفي	كسر	الزجاج	obj
6	سيكسر عظامكم	سيكسر	عظام	obj
7	قرأت هذا الكتاب	الكتاب	قرات	obj
8	قرأت الكتاب	الكتاب	قرات	obj
9	رأيت في الطريق الواسع الطويل السيارة الجميلة	رأى	السيارة	obj
10	رأى في طريقه الواسع الطويل السيارة الجميلة	رأى	السيارة	obj
11	يتنفس هواء	يتنفس	هواء	obj

Figure 11: The output for the query in figure (10)

B. Quantitative linguistic Exploration

Data exploration is the first process in the analytical treatment of data. In sufficient attention is often given to preliminary investigations of data, and researchers often jump straight into the formal statistical analysis phase of analysis of regression, analysis of variance etc. without making sure that data have been entered correctly. Furthermore, following a series of preliminary procedures the researcher should be able to identify definite patterns in the data, gain insight into the variability contained within the data, detect any strange observations that need following up and decide how to proceed with formal analysis.

In the current research, R language used to report the status of the syntactically analyzed data. For example, most and least frequently occurring words, relations, pos tags. In addition to, types and tokens counting and ratio. In addition, maximum and minimum number of relations per sentence and finally graphical data exploration which can represent the statistical counts in an easy way.

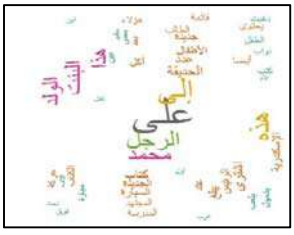
The NLP – R library "tm" used to process the text. The main structure for managing documents in "tm" is a so-called "Corpus", representing a collection of text documents. A corpus is an abstract concept, and there can exist several implementations in parallel. The default implementation is the so-called "VCorpus" (short for Volatile Corpus) which realizes semantics as known from most R objects: corpora are R objects held fully in memory. Another implementation is the "PCorpus" which implements a Permanent Corpus semantics, i.e., the documents are physically stored outside of R(e.g., in a database), corresponding R objects are basically only pointers to external structures, and changes to the underlying corpus are reacted to all R objects associated with it. Compared to the volatile corpus the corpus encapsulated by a permanent corpus object is not destroyed if the corresponding R object is released. The "corpus" method used in the second line of code in figure (12a) to proceed on the sentence column in the sentences table.

```
# Data Exploration Quires

MyData <- sentences_tb$Sentence
docs <- Corpus(VectorSource(MyData))
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head (d)
```

(a)

```
66  min(sentences_tb$length)
67  max(sentences_tb$length)
68
69
67:25  the most frequent words

Console
> view(sentences_tb)
> min(sentences_tb$length)
[1] 2
> max(sentences_tb$length)
[1] 20
```

(b)



(c)

```
> findAssocs(dtm, terms = "بلغ", corlimit = 0.3)
$بلغ
  100  الأرض  4300  مليار  لدول  فائض  الميزان  المجموعة  التجاري
 0.57   0.57   0.57   0.57   0.57   0.57   0.57   0.57   0.40
متسابقا  سكان  دولار  جنيها  نهي  المصنع  المتسابقين  الغربة  الصناعية
 0.40   0.40   0.40   0.40   0.40   0.40   0.40   0.40   0.40
```

(d)

Figure 12: R- libraries implementations

A common approach in text mining is to create a term-document matrix from a corpus. In the "tm" package, the classes Term Document Matrix and Document Term Matrix(depending on whether terms desired as rows and documents as columns, or vice versa) employ sparse matrices for corpora. Inspecting a term-document matrix displays a sample, whereas matrix () yields the full matrix in dense format (which can be very memory consuming for large matrices).The third line code in (12a) creates the Term Document Matrix of the corpus; stored in the variable (dam).Term Document Matrixes very useful for calculating to which extent words or element of matrix in general, related to each other. For example, the association of specific element to the rest of elements of the matrix as in figure (12d). The association of the word "يبلغ"with the surround words are calculated given the document term matrix and correlation limit.

A word cloud shows the frequency of words in a document by varying the size of words in a visualization. Word clouds are great for a quick, qualitative view of your open-ended survey responses, collection of tweets, or website content. There are many word cloud creating websites; the most popular one being wordle.net box. Once the matrix is obtained (dtm), word cloud can be created. The primary arguments for the *wordcloud()* function are the list of words and the list of frequencies in the same order as the word list. There are a number of other arguments you can be used to customize the appearance of the word cloud; the maximum and minimum size using the *scale* argument. a minimum frequency can be set to limit the words to only those that appear x times. In addition, the colors can be set, as in the words clouds above in figure (12c).

C. Towards syntactic modelling

Language Modeling (LM) is a central task to Natural Language Processing and Language Understanding. Models which can accurately place distributions over sentences not only encode complexities of language such as grammatical structure, but also extract a fair amount of information about the knowledge that a corpus may contain. Indeed, models that are able to assign a low probability to sentences that are grammatically correct but unlikely may help other tasks in fundamental language understanding like question answering, machine translation, or text summarization.

LMs have played a crucial role in traditional NLP tasks such as speech recognition [12] [13], machine translation, or text summarization [14] [15]. Often (although not always), training better language models improves the underlying metrics of the downstream task (such as word error rate for speech recognition, or BLEU score for translation), which makes the task of training better LMs valuable by itself. Further, when trained on vast amounts of data, language models compactly extract knowledge encoded in the training data. For example, when trained on movie subtitles [16], these language models are able to generate basic answers to questions about object colors, facts about people, etc. Lastly, recently proposed sequence-to-sequence models employ conditional language models [17] as their key component to solve diverse tasks like machine translation [18] [19] [20] or video generation [21]. Deep Learning and Recurrent Neural Networks (RNNs) have fueled language modeling research in the past years as it allowed researchers to explore many tasks for which the strong conditional independence assumptions are unrealistic. Despite the fact that simpler models, such as N- grams, only use a short history of previous words to predict the next word, they are still a key component to high quality, low perplexity LMs. Indeed, most recent work on large scale LM has shown that RNNs are great in combination with N-grams, as they may have different strengths that complement N-gram models, but worse when considered in isolation [22] [23] . We believe that, despite much work being devoted to small data sets like the Penn Tree Bank (PTB) [24], research on larger tasks is very relevant as over fitting is not the main limitation in current language modeling, but is the main characteristic of the PTB task. Results on larger corpora usually show better what matters as many ideas work well on small data sets but fail to improve on larger data sets. Further, given current hardware trends and vast amounts of text available on the Web, it is much more straightforward to tackle large scale modeling than it used to be. Thus, we hope that our work will help and motivate researchers to work on traditional LM beyond PTB – for this purpose; we will open-source our models and training recipes.

N-grams are essential in any task in which we have to identify words in noisy, ambiguous input. In speech recognition, for example, the input speech sounds are very confusable and many words sound extremely similar. [25]Gives an intuition from handwriting recognition for how probabilities of word sequences can help. In the movie Take the Money and Run, Woody Allen tries to rob a bank with a sloppily written hold-up note that the teller incorrectly reads as "I have a gub". Any speech and language processing system could avoid making this mistake by using the knowledge that the sequence "I have a gun" is far more probable than the non-word "I have a gub" or even "I have a gull"

Language models are very useful in a broad range of applications, the most obvious perhaps being speech recognition and machine translation. In many applications, it is very useful to have a good "prior" distribution p ($x_1 \ldots x_n$) over which sentences are or are not probable in a language. For example, in speech recognition the language model is combined with an acoustic model that models the pronunciation of different words: one way to think about it is that the acoustic model generates a large number of candidate sentences, together with probabilities; the language model is then used to reorder these possibilities based on how likely they are to be a sentence in the language.

Considering the database design discussed in (section 2 - B), language model cannot be built on the table form for 2 reasons. The first is the **nodes IDs** are arbitrary and cannot reflect the structure of related nodes inside the same sentence to enable machine learning. For example, the same structure may occur in two sentences; for one of them, a verb node

with the ID (x) related to two noun nodes with IDs (y and z). And in the other sentence, a verb node with the ID (a) related to two noun nodes with IDs (b and c). Although both sentences have the same structure, they will be considered as two because of nodes IDs. This issue can be solved using a numbering system that limits the numbers of IDs; replacing "01" and "0X" with "1" and "2". The proposed numbering system is resetting from 1 when the new sentence began as in figure (13). The first column represents the automatic assigned ID, the second represents the word, the third for the tag, the fourth represents the sentence ID, and the fifth represents the new numbered tag based on the sentence ID.

File Edit Format View Help

01	أكل	VER	1	VER:1
03	الولد	NOU	1	NOU:2
05	التفاحة	NOU	1	NOU:3
07	الكبير	ADJ	1	ADJ:4
0A	شرب	VER	2	VER:1
0C	اللبن	NOU	2	NOU:3
0E	الولد	NOU	2	NOU:2
01	ارتبطت	VER	3	VER:1
03	البلدان	NOU	3	NOU:2
05	الاستونية	ADJ	3	ADJ:3
07	و	COO	3	COO:4
08	الروسية	ADJ	3	ADJ:5
09	ب	PRE	3	PRE:6
11	شبكة	NOU	3	NOU:7
13	مياه	NOU	3	NOU:8
15	واحدة	ADJ	3	ADJ:9
01	ترتبط	VER	4	VER:1
03	المؤسسات	NOU	4	NOU:2
05	التي	RPR	4	RPR:3
07	وقع	VER	4	VER:4
09	علي	PRE	4	PRE:5

Figure 13: The proposed Numbering System

Using the new numbered tag in figure (13) above, the sequences according to its occurrence in each sentence can be represented. Each element in the sequence represented by 3 elements; the relation, the first node and the second node. Separators are determined as follows to output the file (considered as input file for language modeling) in figure (15) below:

- The relation is followed by "_",
- Nodes are separated by "-",
- Sequences are separated by blank space,
- Sentences are separated by the tag [/S].

sbj_VER:1-NOU:2 link_ADJ:4-NOU:2 obj_VER:1-NOU:3 [/S] obj_VER:1-NOU:3 sbj_VER:1-NOU:2 [/S] sbj_VER:1-NOU:2 adj_NOU:2-ADJ:3 co_ADJ:3-COO:4 cj_COO:4-ADJ:5 link_VER:1-PRE:6 gen_PRE:6-NOU:7 poss_NOU:7-NOU:8 adj_NOU:8-ADJ:9 [/S] sbj_VER:1-NOU:2 link_NOU:2-RPR:3 link_VER:4-PRE:5 sub_RPR:3-SCOPE gen_PRE:5-PRON:6 sbj_VER:4-NOU:7 adj_NOU:7-ADJ:8 link_VER:1-PRE:9 gen_PRE:9-NOU:10 adj_NOU:10-ADJ:11 [/S] obj_VER:1-NOU:2 poss_NOU:2-NOU:3 poss_NOU:8-NOU:4 poss_NOU:4-PRON:5 co_NOU:2-COO:6 link_NOU:8-PRE:7 cj_COO:6-NOU:8 gen_PRE:7-NOU:9 poss_NOU:9-NOU:10 adj_NOU:10-NOU:11 link_NOU:10-PRE:12 gen_PRE:12-NOU:13 poss_NOU:13-PPN:14 [/S] sbj_VER:1-NOU:2 link_VER:1-PRE:3 gen_PRE:3-NOU:4 poss_NOU:4-NOU:5 adj_NOU:5-ADJ:6 link_VER:1-PRE:7 gen_PRE:7-NOU:8 poss_NOU:8-NOU:9 link_NOU:9-PRE:10 gen_PRE:10-NOU:11 poss_NOU:11-NOU:12 [/S] sbj_VER:1-NOU:2 link_VER:1-PRE:3 gen_PRE:3-NOU:4 poss_NOU:4-NOU:5 adj_NOU:4-ADJ:6 link_VER:1-PRE:7 gen_PRE:7-NOU:8 poss_NOU:8-NOU:9 link_NOU:9-PRE:10 gen_PRE:10-NOU:11 poss_NOU:11-NOU:12 [/S] sbj_VER:1-NOU:2 link_VER:1-PRE:3 gen_PRE:3-NOU:4 link_NOU:4-PRE:5 gen_PRE:5-NOU:7 link_NOU:7-PRE:8 gen_PRE:8-NOU:9 poss_NOU:9-NOU:10 poss_NOU:10-NOU:11 adj_NOU:11-ADJ:12 [/S] poss_NOU:11-NOU:1 gen_PRE:12-NOU:2 adj_NOU:10-ADJ:4 adj_NOU:6-ADJ:5 sbj_VER:3-NOU:6 gen_NOU:9-NOU:7 link_VER:3-PRE:8 link_NOU:11-NOU:9 gen_PRE:8-NOU:10 poss_NOU:2-NOU:11 link_NOU:10-PRE:12 [/S] sbj_VER:1-NOU:2 obj_VER:1-NOU:4 poss_NOU:3-NOU:4 poss_NOU:4-NOU:5 poss_NOU:5-PRON:6 adj_NOU:5-ADJ:7 adj_NOU:5-ADJ:8 [/S] sbj_VER:1-PPN:2 co_PPN:2-COO:3 link_VER:1-NOU:4 poss_NOU:4-NOU:5 link_VER:1-PRE:6 gen_PRE:6-PPN:7 link_VER:1-PRE:8 link_NOU:13-PRE:9 gen_PRE:9-PRON:10 cj_COO:3-NOU:11 poss_NOU:5-NOU:12 poss_PRE:8-NOU:13 [/S] link_VER:1-ADV:2 link_VER:1-PRE:3 gen_PRE:3-NOU:4 adj_NOU:4-ADJ:5 adj_NOU:4-ADJ:6 sbj_VER:1-NOU:7 [/S] link_NOU:2-PRE:3 gen_PRE:3-NOU:4 obj_VER:1-NOU:7 spec_NUM-NOU:7 spec_NOU:6-NOU:7 poss_NOU:2-NOU:8 poss_NOU:4-NOU:9 adj_NOU:8-ADJ:10 sbj_VER:1-NOU:11 [/S] sbj_VER:1-NOU:2 link_VER:1-PRE:3 gen_PRE:3-NOU:4 poss_NOU:4-NOU:5 adj_NOU:5-ADJ:6 [/S] obj_VER:1-NOU:3 poss_NOU:2-NOU:3 sub_PTC:4-SCOPE obj_VER:5-NOU:6 sbj_VER:1-NOU:7 poss_NOU:3-NOU:8 poss_NOU:6-NOU:9 sbj_VER:5-PRON:10 [/S] sbj_VER:1-NOU:2 link_VER:1-PRE:3 gen_PRE:3-PRE:4 link_PRE:4-PRE:5 gen_PRE:5-NOU:6 link_NOU:7-NOU:9 poss_NOU:9-NOU:10 poss_NOU:10-NOU:11 adj_NOU:11-NOU:12 [/S] adj_NOU:2-ADJ:1 poss_NOU:5-NOU:2 link_VER:4-ADV:3 gen_PRE:6-NOU:5 link_VER:4-PRE:6 [/S] adj_NOU:12-ADJ:2 gen_PRE:7-NOU:3 app_NOU:8-NOU:4 link_ADJ:2-PRE:5 gen_PRE:5-NOU:6 link_ADJ:2-PRE:7 sbj_VER:1-NOU:8 adj_NOU:12-ADJ:9 poss_NOU:6-NOU:10 adj_NOU:6-ADJ:11 obj_VER:1-NOU:12 [/S] adj_NOU:2-ADJ:1 gen_PRE:7-NOU:2 gen_PRE:4-DEM:3 link_VER:6-PRE:4 app_DEM:3-NOU:5 link_VER:6-PRE:7 [/S] poss_NOU:10-NOU:1 adj_NOU:1-ADJ:2 sbj_VER:8-NOU:3 poss_NOU:3-NOU:4 link_VER:8-ADV:5 link_VER:8-PRE:6 link_PPN:9-PRE:7 gen_PRE:6-PPN:9 gen_PRE:7-NOU:10 [/S] cj_COO:3-NOU:1 poss_NOU:11-NOU:2 co_NOU:7-COO:3 gen_PRE:8-DEM:4 gen_PRE:1
-NOU:5 link_NOU:2-PRE:6 gen_PRE:6-NOU:7 link_VER:9-PRE:8 link_VER:9-PRE:10 obj_VER:9-NOU:11 [/S] gen_PRE:14-VER:1 poss_NOU:17-NOU:3 co_NOU:3-COO:4
poss_NOU:18-NOU:5 link_VER:1-PRE:6 poss_NOU:10-NOU:7 adj_NOU:3-ADJ:8 poss_NOU:11-PRON:9 sbj_VER:12-NOU:10 gen_PRE:13-NOU:11 link_NOU:10-PRE:13 link_VER:12-PRE:14 link_NOU:16-PER:15 cj_COO:4-NOU:16 gen_PRE:6-NOU:17 gen_PER:15-NOU:18 app_NOU:11-PPN:19 [/S] obj_VER:4-SCOPE PREDX_VER:1-SCOPE PREDX_PTC:10-SCOPE poss_NOU:9-PRON:2 gen_PRE:12-NOU:3 adj_NOU:3-ADJ:6 SUBJX_VER:1-NOU:7 SUBJX_PTC:10-PPN:8 poss_NOU:11-NOU:9 pred_PRON:5-NOU:11 link_VER:1-PRE:12 [/S] gen_PRE:7-NOU:2 link_VER:1-PER:3 poss_NOU:2-NOU:4 gen_PER:3-NOU:5 adj_NOU:5-ADJ:6 link_ADJ:6-PRE:7 sbj_VER:1-PPN:8 [/S] poss_NOU:11-NOU:1 gen_PRE:12-NOU:2 adj_NOU:10-ADJ:4 adj_NOU:6-ADJ:5 sbj_VER:3-NOU:6 gen_NOU:9-NOU:7 link_VER:3-PRE:8 link_NOU:11-NOU:9 gen_PRE:8-NOU:10 poss_NOU:2-NOU:11 link_NOU:10-PRE:12 [/S] poss_NOU:1-NOU:2 poss_NOU:2-ADJ:3 gen_PRE:4-NOU:5 poss_NOU:5-NOU:6 poss_NOU:6-NOU:7 obj_VER:9-NOU:10 link_NOU:15-PRE:12 poss_NOU:7-NOU:13 app_DEM:8-NOU:14 gen_PRE:12-NOU:16 sbj_VER:11-NOU:17 [/S] poss_VER:2-PRE:3 gen_PRE:4-NOU:5 adj_NOU:5-ADJ:6 sbj_VER:1-NOU:7 obj_VER:1-RPR:8 poss_PRE:3-NOU:9 adj_NOU:9-ADJ:10 [/S] link_NOU:7-PRE:2 link_NOU:11-PRE:4 gen_PRE:4-NOU:5 poss_NOU:5-NOU:6 gen_PRE:2-NOU:8 poss_NOU:3-NOU:8 poss_NOU:6-NOU:9 adj_NOU:6-ADJ:10 sbj_VER:1-NOU:12 [/S] link_NOU:7-PRE:2
gen_PRE:2-NOU:3 poss_NOU:3-NOU:5 adj_NOU:5-ADJ:6 sbj_VER:1-NOU:8 [/S] link_VER:1-NOU:2 co_NOU:8-COO:5 cj_COO:5-NOU:6 gen_NOU:2-NOU:7 gen_PRE:3-

Figure 15: The input file for language modeling

The basic and necessary model for the NLP tasks is the N-gram model. The R library (ngram) is used for generating the bi-gram model for the relations – tag file in figure (15). Frequencies and Probability distribution for each bi-gram can be calculated as in figure (17)

1,"gen_PRE:5-NOU:6 link_NOU:7-NOU:9"
2,"link_NOU:1-PER:3 gen_PER:3-NOU:4"
3,"sbj_VER:1-NOU:2 link_VER:1-ADV:3"
4,"gen_PRE:2-NOU:3 poss_NOU:3-SCOPE"
5,"sub_PTC:4-SCOPE obj_VER:5-NOU:6"
6,"gen_PRE:3-NOU:4 poss_NOU:4-NOU:5"
7,"link_VER:1-ADV:2 link_VER:1-NOU:3"
8,"gen_PRE:2-NOU:4 spec_NUM-NOU:4"
9,"link_VER:1-PER:3 poss_NOU:2-NOU:4"
10,"spec_NUM-NOU:9 obj_VER:2-NOU:10"
11,"adj_NOU:3-ADJ:8 poss_NOU:11-PRON:9"
12,"gen_PRE:5-NOU:6 link_NOU:6-PRE:7"
13,"poss_NOU:4-NOU:12 adj_PTL:9-ADJ:13"
14,"link_VER:1-PRE:5 gen_PRE:5-NOU:6"
15,"link_VER:5-ADV:6 [/S]"
16,"gen_PRE:4-NOU:5 poss_NOU:5-NOU:6"
17,"gen_PRE:8-PPN:11 adj_NOU:7-ADJ:12"
18,"neg_NOU:1-VER:2 [/S]"
19,"sbj_VER:1-NOU:12 [/S]"
20,"link_VER:1-PRE:6 gen_PRE:6-NOU:7"
21,"link_VER:9-ADJ:10 link_VER:9-PRE:11"
22,"app_DEM:2-NOU:5 sbj_VER:1-NOU:6"
23,"poss_NOU:5-NOU:6 poss_NOU:3-NOU:7"

Figure 16: The bigram output

ngrams	freq	prop
[/S] sbj_VER:1-NOU:2	43	0.024487472
[/S] link_VER:1-PRE:2	30	0.017084282
link_VER:1-PRE:2 gen_PRE:2-NOU:3	21	0.011958998
[/S] obj_VER:1-NOU:2	17	0.009681093
[/S] link_NOU:2-PRE:3	13	0.007403189
[/S] sbj_VER:1-PPN:2	12	0.006833713
link_VER:1-PRE:3 gen_PRE:3-NOU:4	12	0.006833713
gen_PRE:2-NOU:3 poss_NOU:3-NOU:4	10	0.005694761
adj_NOU:2-ADJ:3 [/S]	9	0.005125285
sbj_VER:1-NOU:2 poss_NOU:2-NOU:3	9	0.005125285
obj_VER:1-NOU:3 [/S]	9	0.005125285

Figure 17: Probabilities and frequencies of bigrams

A correlation coefficient is a number that quantifies a type of correlation and dependence, meaning statistical relationships between two or more values in fundamental statistics. In this study, for finding the relationships between two syntactic relations and their nodes, correlation coefficient can be sufficient in prediction with the syntactic relation.

The cor( ) function can be used to produce correlations and the cov( ) function to produces covariances.

A simplified format is cor(x, use=, method=) where:

x	Matrix or data frame
use	Specifies the handling of missing data. Options are **all.obs** (assumes no missing data - missing data will produce an error), **complete.obs** (listwise deletion), and **pairwise.complete.obs** (pair wise deletion)
method	Specifies the type of correlation. Options are **pearson**, **spearman** or **kendall**.

The bi-gram output has been used to output the matrix as an input for the correlation function to output the correlation matrix in figure (18).

	adj_NOU:1-ADJ:2	adj_NOU:1-ADJ:3	adj_NOU:2-ADJ:3	adj_NOU:2-NOU:5	adj_PPN:2-ADJ:3	app_DEM:1-NOU:3	app_DEM:2-NOU:3
Scope	-0.06420289	0.14512399	0.46539418	-0.04514817	0.21671122	-0.07877819	-0.10277763
adj_NOU:2-ADJ:3	-0.04228469	0.24698235	1.00000000	-0.02973505	-0.02973505	-0.05188413	-0.06769041
sbj_VER:1-NOU:3	-0.03896186	-0.03885408	0.26124486	-0.02739840	-0.02739840	-0.04780695	-0.06237114
gen_PRE:2-NOU:3	-0.03888814	-0.03878056	-0.07399678	-0.02734656	-0.02734656	-0.04771649	-0.06225313
app_DEM:2-NOU:3	-0.03557390	-0.03547549	-0.06769041	-0.02501595	-0.02501595	-0.04364986	1.00000000
link_VER:1-PRE:3	-0.03155874	-0.03147144	0.33671068	-0.02219245	0.47713761	-0.03872318	-0.05052003
obj_VER:1-NOU:3	-0.03155874	-0.03147144	0.14798113	-0.02219245	0.49378194	-0.03872318	-0.05052003
obj_VER:1-NOU:4	-0.03155874	-0.03147144	0.14369182	0.47713761	0.47713761	-0.03872318	0.18692412
poss_NOU:2-NOU:3	-0.03155874	-0.03147144	0.35601257	-0.02219245	-0.02219245	-0.03872318	-0.05052003

Figure 18: The correlation matrix for all syntactic relations

4 CONCLUSION

Considering the importance of databases in data organization, the researcher suggests a database design for the syntactic dependency trees which enable syntactic data search for linguists to enhance their models. The database is a usable research tool for the academic community; the extensibility condition considered in design. Moreover, relational database was the best way for syntactic relations storage and retrieval given other information about relation nodes. R language proved its efficiency in data statistical operations; it is a very promising language programming for building NLP tools. Furthermore, R is very suitable for building language models such as n-gram model which is the core of the statistical NLP tasks. Using n-gram model, probability distribution, frequencies and correlations calculation were

possible to be conducted on the syntactic relations. All such calculations can be used for building the offline phase in the supervised machine learning application, as the analyzed data is considered as a training data.

REFERENCES

[1] Baranowski, Krzysztof J., et al. "Kleine Untersuchungen zur Sprache des Alten Testaments und seiner Umwelt."
[2] Nerbonne, John A., ed. *Linguistic databases*. Stanford, CA: CSLI Publications, 1998.
[3] Van Halteren H. *Excursions into syntactic databases*. Rodopi; 1997.
[4] Elhosiny, I. and Alansary, S. "Syntax-Semantics Classification of Arabic verbs for Semantic Annotation", 2014.
[5] S. Alansary, M. Nagi, N. Adly, "IAN: A tool for Natural Language Analysis", *in Proceeding of 12th Conference on Language Engineering*, Cairo, Egypt, 2012.
[6] K. Dukes and T. Buckwalter (2010). A Dependency Treebank of the Quran using Traditional Arabic Grammar. In Proceedings of the 7th International Conference on Informatics and Systems (INFOS). Cairo, Egypt.
[7] Van Halteren, Hans. *Excursions into syntactic databases*. Vol. 21. Rodopi, 1997.
[8] *Hornik, Kurt .*"R FAQ"*. The Comprehensive R Archive Network. 2.1 What is R?. Retrieved 2015-12-06.*
[9] R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2016.
[10] *Vance, Ashlee* "Data Analysts Captivated by R's Power"*.* New York Times*. Retrieved 2009-04-28. R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca.2009...*
[11] Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. 2010.
[12] Arisoy, Ebru, et al. "Deep neural network language models." *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012.
[13] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).
[14] Filippova, Katja, et al. "Sentence Compression by Deletion with LSTMs." EMNLP. 2015.
[15] Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).
[16] Mikolov, Tomas, and Geoffrey Zweig. "Context dependent recurrent neural network language model." SLT 12 (2012): 234-239.
[17] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
[18] Cho, Kyunghyun, et al. "On the properties of neural machine translation: Encoder-decoder approaches." arXiv preprint arXiv:1409.1259 (2014).
[19] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modeling sentences." arXiv preprint arXiv:1404.2188 (2014).
[20] Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations using LSTMs." ICML. 2015.
[21] Mikolov, Tomáš, et al. "Extensions of recurrent neural network language model." Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.
[22] Mikolov, Tomáš. "Statistical language models based on neural networks." Presentation at Google, Mountain View, 2nd April (2012).
[23] Chelba, Ciprian, et al. "One billion word benchmark for measuring progress in statistical language modeling." arXiv preprint arXiv:1312.3005 (2013).
[24] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19.2 (1993): 313-330.
[25] Russell, Stuart, Peter Norvig, and Artificial Intelligence. "A modern approach." *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25 (1995): 27.

BIOGRAPHIES

Israa Elhosiny

Principal Grammar Developer in the Arabic Computational Linguistics Center Bibliotheca Alexandrina.
PhD. Student in the Department of Phonetics and Linguistics. Faculty of Arts, Alexandria University (2016). She obtained her MA, Department from Phonetics and Linguistics 2015. She obtained her BA, Department from Phonetics and Linguistics 2004. She obtained the UNL certificates; CLEA250, CLEA750, CUP250, and CUP500. She attended the X UNL School organized by the UNDL foundation and Bibliotheca Alexandrina (7-11 October 2012).

She has an experience in Python programming, and obtained the New Horizon certificates in Java programming (OOP, JDBC and JSF) (2017). She is also Principal Grammar Developer of Arabic

Computational Linguistics Center in Bibliotheca Alexandrina. She is working in the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now. She has an experience in Arabic language analysis and generation and text tokenization. She participated in building grammars using UNL for library information system (LIS) and Knowledge Extraction sYStem (Keys). She has an experience in automatic Arabic diacritization. She participated in building grammars, and developing a diacritization evaluation tool. She is a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

Dr. Sameh Alansary

Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.
He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

كيف تُخزن مدونة مُحللة نحويا في قاعدة بيانات؟

إسراء الحسيني[1] ,سامح الأنصاري[2]

قسم الصوتيات واللسانيات، كلية الآداب جامعة الإسكندرية

[1]israa_elsayed@yahoo.com

[2]s.alansary@alexu.edu.eg

ملخص – نظرا لزيادة الاحتياج لقواعد البيانات اللغوية؛ حيث أنها تتيح دراسة العديد من الظواهر اللغوية الضرورية لبناء تطبيقات المعالجة الآلية، يركز البحث على نوع واحد من قواعد البيانات وهو قواعد البيانات النحوية. فإن الدراسة وضعت منهجية وتصميم لبناء قاعدة بيانات نحوية إعتمادية. فقد جُمعت عينة لغوية من مجالات متعددة تمثل تراكيب نحوية مختلفة بهدف التعرض لكيفية تخزين مثل هذه التراكيب في قاعدة البيانات. وقد أُمدت قاعدة البيانات بمعلومات لغوية إضافية بهدف إثراءها. وقد برهنت المنهجية على مدى فاعلية تخزين التراكيب النحوية عن طريق قياس إمكانية الاستعلام اللغوي بداخل مدونة محللة نحويا، بالإضافة إلى استخراج النمذجة الإحصائية للمركبات النحوية عن طريق لغة الآر (R language).

Predicting Diacritics for Arabic Unknown Words

Amany Fashwan[1], Sameh Alansary[2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

[1]amany.fashwan@alexu.edu.org

[2]s.alansary@alexu.edu.eg

***Abstract*—This paper presents a methodology for predicting the diacritics of Out Of Vocabulary (OOV) for Automatic Diacritization of Modern Standard Arabic (MSA) texts. An Arabic annotated corpus of 550,000 words is used; the International Corpus of Arabic (ICA). In addition, a testing data of 52,000 words from Arabic Tree Bank (ATB) have been handled and used. For predicting the OOV words, a prior step is taken; pre-processing the stems of the training data. A list of unique 4307 diacritized patterns with their templates and frequencies have been used. At this point, the morphological Word Error Rate (WER) is 4.56% while the morphological Diacritic Error Rate (DER) is 1.88% and the syntactic WER is 9.36%. The best WER is 14.78% compared to the best published results, of [1]; 11.68%, [13]; 12.90%and [10]; 13.60%.**

Keywords: **Automatic Diacritization · Out Of Vocabulary · Arabic Natural Language Processing· Internal Diacritics · Case Ending Diacritics**

1 INTRODUCTION

Modern Standard Arabic (MAS) is currently the sixth most widely spoken language in the world with estimated 422 million native speakers. It is usually written without diacritics that make it difficult for performing Arabic text processing. In addition, this often leads to considerable ambiguity since several words that have different diacritic patterns may appear identical in a diacritic-less setting. In fact, a text without diacritics may bring difficulties for Arabic readers. It is also problematic for Arabic processing applications where the lack of diacritics adds another layer of ambiguity when processing the input data [17].

Diacritics restoration is the problem of inserting diacritics into a text where they are missing. Predicting the correct diacritization of the Arabic words elaborates the meaning of the words and leads to better understanding of the text, which in turn is much useful in several real life applications such as Information Retrieval (IR), Machine Translation (MT), Text-to-speech (TTS), Part-Of-Speech (POS) tagging and others.

For full diacritization of an Arabic word, two basic components are needed: 1) Morphology-dependent that selects the best internal diacritized form of the same spelling; e.g. the word "علم" "Elm" has different diacritized forms; "عِلْم" "Eilom" "science", "عَلَم" "Ealam" "flag","عَلَّمَ" "Eal~ama" "taught" and "عَلِمَ" "Ealima" "knew".2) Syntax-dependent that detects the best syntactic case of the word within a given sentence; i.e. its role in the parsing tree of that sentence. For example; دَرَسْتُ عِلْمَ الرِّيَاضِيَّاتِ "darasotu Eiloma Alr~iyADiy~Ati" "I studied Mathematics" implies the syntactic diacritic of the target word - which is an "object" in the parsing tree - is "Fatha", while يُفِيدُ عِلْمُ الرِّيَاضِيَّاتِ جَمِيعَ الْعُلُومِ "yufiydu Eilomu Alr~iyADiy~Ati jamiyEa AloEuluwmi" "Mathematics benefits all sciences" implies the syntactic diacritic of the target word which is a "subject" in the parsing tree - is "Damma" [14].

OOV words are unknown words that appear in the testing process of the diacritization system but not in the diacritizer vocabulary. Most automatic diacritization systems can only diacritize words that belong to a fixed finite vocabulary. When encountering an OOV word, the diacritizer will not diacritize it. In addition, OOV words also affect the diacritizer performance of their surrounding words. Furthermore, OOV words are usually important content words, such as names, locations, etc., which incorporate crucial information for understanding the diacritized text.

Due to its importance in giving the correct more accurate diacritization results and consequently gives the correct understanding of Arabic texts, we present a methodology for predicting the diacritics of OOV words. Section 2 reviews some related work to automatic diacritization systems. Section 3 describes the used data sets. Section 4 reviews the methodology of our system in general and details the built OOV Arabic diacritization module. Section 5 discusses the output results. Finally, section 6 concludes the paper.

2 RELATED WORK

Diacritic restoration has been receiving increasing attention and has been the focus of several studies. Different methods such as rule-based, example-based, hierarchical, morphological and contextual-based as well as methods with Hidden Markov Models (HMM) and weighted finite state machines have been applied for the diacritization of Arabic text. Among these trials, that are most prominent, [1-2], [5-18] and [20].

In addition, some software companies have developed commercial products for the automatic diacritization of Arabic; Sakhr Arabic Automatic Diacritizer [21], Xerox's Arabic Morphological Processor [22] and RDI's Automatic Arabic Phonetic Transcript or (Diacritizer/Vowelizer) [23]. Moreover, there are also other free online available systems; Meshkal Arabic Diacritizer [24], Harakat Arabic Diacritizer [25], Al-Du'aly [26], Farasa [27] and Google Tashkeel which is no longer working where the tool is not available now. To our knowledge, none of the previous systems makes use of syntax with the exception of [2], [6] and [19] who have integrated syntactic rules.

3 DATA SETS

The used **"Training Data Set"** in the current module is about 450,000 words that were selected from a Morphologically Annotated Gold Standard Arabic Resource (MASAR) for MSA [7]. The texts were selected from different sources; Newspapers, Net Articles and Books. Moreover, these selected texts covered more than one genre. Each word is tagged with features, namely, Lemma, Gloss, prefixes, Stem, Tag, suffixes, Gender, Number, Definiteness, Root, Stem Pattern, Case Ending, Name Entity and finally Vocalization.

In MASAR, some words are manually analyzed, because the used analyzer; Buckwalter Arabic Morphological Analyzer (BAMA 0.2) [4], may provide many solutions, but none of them is right, or it may be unable to provide any solutions for the input word. Consequently, solutions enhancement is needed in these cases. The solutions are analyzed manually as if they are analyzed by BAMA and they are added in BAMA's dictionaries so that they would be analyzed correctly the next time these words are used. This helps in making sure that the training data set contains all the vocabulary. Consequently, the out of vocabulary (OOV) problem will not form a problem while working in the used training data sets, but it may appear in handling new texts.

In order to have an objective evaluation of the system, the same testing data (LDC's Arabic Treebank) that was used in the other systems was used to compare the results. The **"Testing Data Set"** is a part of Arabic Tree Bank part 3 (ATB3) form "An-Nahar" Lebanese News Agency that was annotated depending on BAMA. It consists of 91 articles (about 52.000 words) covering the period from October 15, 2002 to December 15, 2002 [20].

Preprocessing the ATB testing data is a very important step that allows the researcher to represent the data and its features in a useful way that everything becomes ready for calculating any statistics or system parameters.

Every file in this data represents one journal article. The article is represented in the form of sequences of Arabic morphological analysis and part-of-speech tagging or sequences of parsing trees of the vocalized words. The first step in preprocessing the data is to generate the original sentences, from the Arabic morphological analysis and part-of-speech tagging. This is accomplished through the following sub-steps:

1. Arabic morphological analysis and part-of-speech tagging input, and extracting the selected solution for each word with its vocalized form POS tag.
2. Generating the non-vocalized form of each word by removing all the diacritics from the corresponding vocalized words.

By the end of these steps, the data is stored in a database table, so that they can be used in creating both the raw and diacritized text of each article. After finishing the preprocessing of the testing data, the researcher finds that:

- 3.16% of the words do not have selected solutions in the ATB testing data. This is due to some wrongly joined words; there are no solutions to be disambiguated from BAMA or none of the found BAMA's solutions is available for the context. Some of these words are found with their corresponding solution in ICA data sets. Other words have been handled by editing them as "والاحضارة", "ابوجودة" and "لجنةمستقلة". Finally, the remaining words have been dealt with as **OOV**. To evaluate the OOV words, they are reviewed according to their contexts and assigned the suitable diacritics by the researcher just to be used in the testing process.

- It has been noticed that the letter before 'ا', 'A' that is written within the word is never diacritized in BAMA since it depends on that 'ا' 'A' is a prolongation letter 'حرف مدّ' which leads to have words that have no morphological diacritics at all such the word 'قال' 'qAl' 'said'. This problem has not been fixed in ATB, although it has been fixed in the training data sets. To overcome this problem that leads to inconsistency while evaluating the system with other state-of-the-art systems, the rules that are responsible for restoring these missing diacritics in morpho-phonological modules have been disabled.

4 ARABIC DIACRITIZATION SYSTEM

In this system, the diacritization problem will be handled through two levels; morphological processing level (for detecting the internal diacritics)and syntactic processing level (for detecting the case ending diacritics).The morphological processing level depends on four layers. The first three layers are similar to BASMA's (Alansary, 2015). The first layer is direct matching between the words that have only one morphological analysis and their diacritized forms, the second is disambiguating the input by depending on contextual morphological rules, and the third is disambiguating the input statistically by using machine learning techniques. However, the three layers in this algorithm are applied sequentially for the whole input, unlike BASMA's system that applies the layers word by word. In each of these layers, a step towards the morphological diacritization of the input text is performed as figure 1 shows. Moreover, this algorithm makes use of the relations between the words and their contexts, whether the preceding or the succeeding words, but BASMA depends only on the morphological disambiguation of the preceding words. In addition to these three layers, another layer is used; the Out Of Vocabulary (OOV) layer. The adopted syntactic algorithm is a rule based approach that detects the main constituents of the morphological analysis output and applies the suitable syntactic rules to detect the case ending.

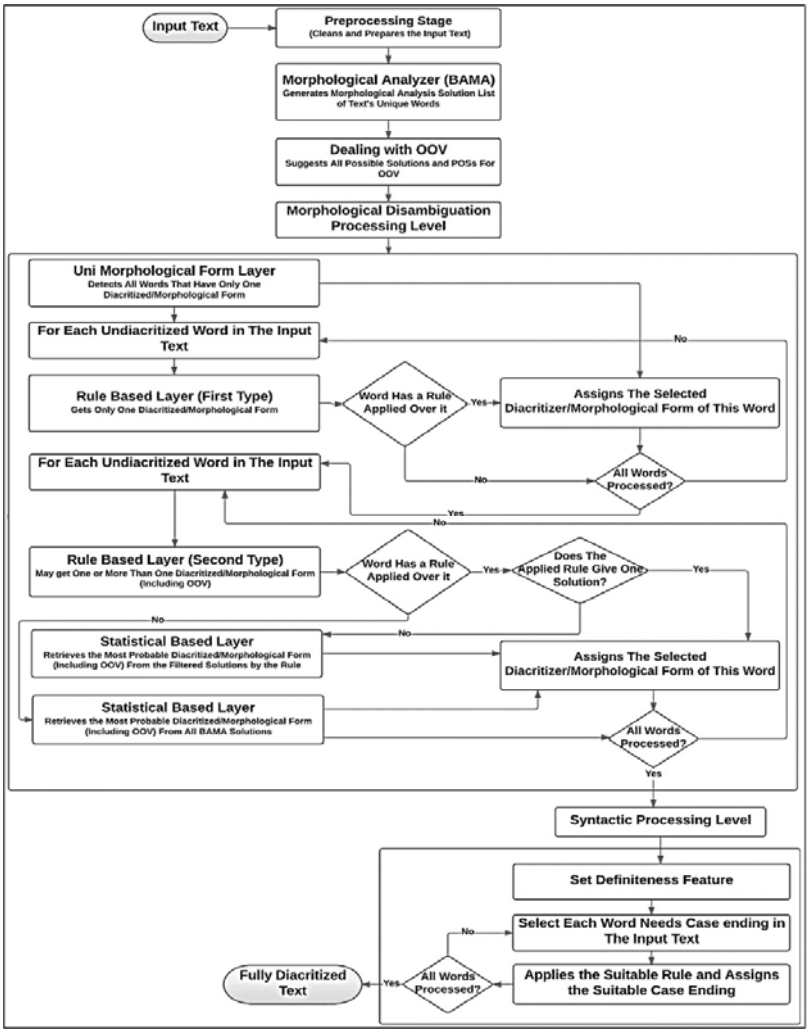


Figure 1: General Design of SHAKKIL System

3.1 OOV Module

As mentioned before, there are no out of vocabulary (OOV) words in the training data set, but the system encounter OOV when handling new texts. Handling OOV words means getting the most probable morphological diacritization of the OOV words and predicting their POS tags.

For predicting the OOV words, a prior step is taken; preprocessing the stems of the training data. The stems of the training data are used to get a list of unique 4307 diacritized patterns with their templates and frequencies. The patterns

are prepared by converting the consonants in the stem to placeholder while keeping the vowels, hamazat (" "أ" ,"<" "إ" ,"ئ" ,"&" "ؤ" ,">"" "}" … etc.) and weak letters 'واي" "حروف العلة" "wAy". In addition, POS of patterns are taken into consideration as figure 2 shows.

The POS helps, in some cases, in limiting the scope of the search of the matched pattern, where, for example, if the OOV word has been detected as having 'ال' 'Al' at the beginning of it, this means the system should search for the detected pattern in the patterns of nouns or adjectives.

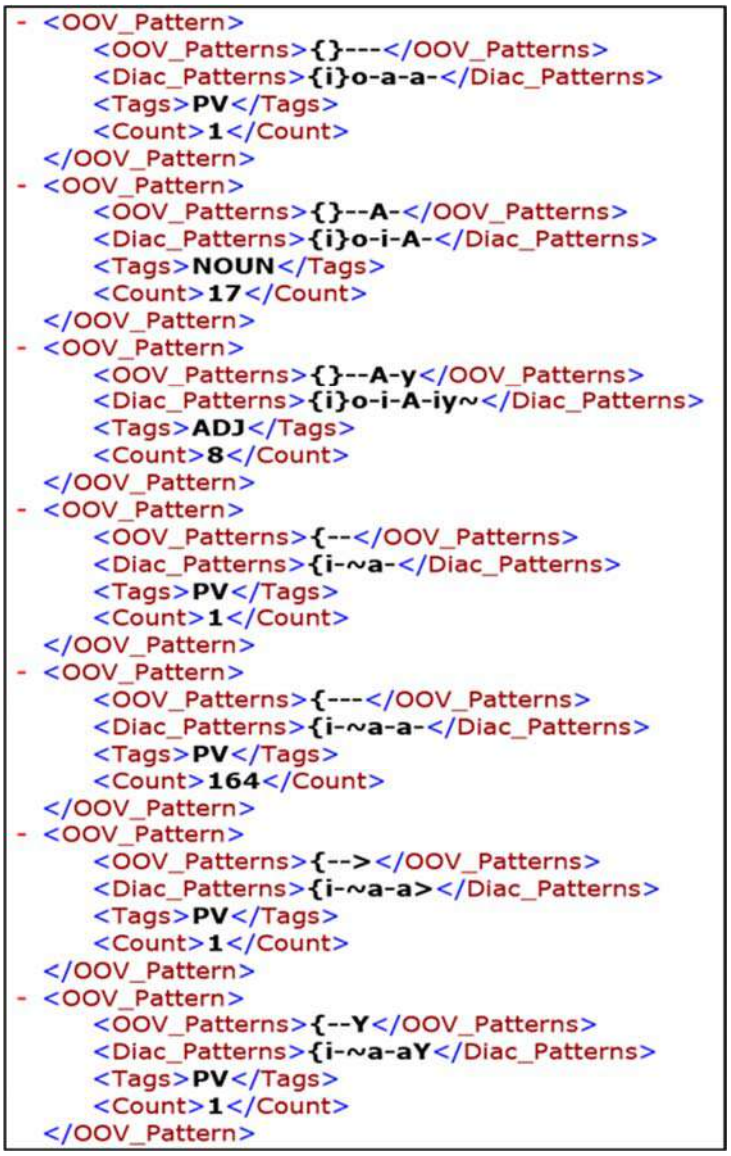


Figure 2: Pattern List with their Diacritized Patterns and Tags

While detecting the input text analysis solutions, each word is checked by the system to determine whether it has analyses solutions from BAMA or it is OOV. When the word form is checked as OOV, the system switches to the OOV module. In this module, the system tries to get all word's possible morphological constituents (a combination of prefixes, stem and suffixes). Then, it uses the list of detected stems and gets their counterpart diacritized patterns. The selected pattern is used to retrieve the suitable diacritic for the stem. Moreover, the system chooses the POS tag of the diacritized pattern and assign it to the diacritized stem where each selected solution is added to text's solutions:

While working in the morphological disambiguation processing level, if the OOV word has more than one matched POS tag, the system detects the best one depending morphological processing level. Figure 3 shows an example for some OOV words highlighted by red square:

Figure 3: An Example for OOV Words

After detecting the suitable diacritized stem, tags, the system concatenates the prefixes and suffixes with the selected pattern to get the full morphological diacritized form depending on the morpho-phonological rules. After all these steps, it can be said that the system analyzes the input text morphologically. In order to achieve the morphological diacritized form, some morpho-phonological rules are applied.

5 Results and Evaluation

A blind copy of the testing data set is used to evaluate the system versus the gold annotated data. Two error rates are calculated: diacritic error rate (DER) which indicates how many letters have been incorrectly restored with their diacritics, and word error rate (WER) which indicates how many words have at least one diacritic error. In the testing process, 51.63% of the words are diacritized in the first layer, 5.56% of the words are diacritized by rule-based layer only, 8.26% of the words are diacritized by both rule-based and statistical-based layers, 32.99% of the words are diacritized by statistical-based layer only, and finally 1.56% of the word is diacritized in OOV layer.

In OOVlayer, the system could predict the words with WER of 11.2% and DER of 6.7%. Table 1 summarizes the results of the current proposed system in comparison with other systems.

Table 1:
Summary of the Comparison between the State-Of-The-Art Systems.

System	Total WER		Ignoring Last	
	WER	DER	WER	DER
Zitouni (2006)	17.30%	5.10%	7.90%	2.50%
Habash, Rambow & Roth (2009)	13.60%	NA	5.20%	NA
Rashwan (2015)	12.90%	NA	NA	NA
Abandah (2015)	11.68%	NA	3.54%	1.28%
Metwally, Rashwan & Atiya (2016)	13.70%	NA	NA	NA
Chennoufi & Mazouri (2016)	NA	NA	1.86%	0.71%
Current System	**14.78%**	**4.11%**	**4.81%**	**1.93%**

The comparison indicates that [1], [13] and [10] outperform the current system's results. However, the results are still close to [11].

6 CONCLUSION

In this work, we depend on Arabic morphological rules as well as different machine learning techniques for detecting the morphological diacritics (internal diacritics). In addition, we adopted a rule based syntactic algorithm that detects the main constituents of the morphological analysis output and applies the suitable syntactic rules to detect the case ending. Moreover, we have dealt with OOV words to get the most probable morphological diacritization and predict their POS tags depending on unique 4307 diacritized patterns with their templates and frequencies. Evaluation of the proposed system is made in comparison with other best state of the art systems. The best WER of the morphological diacritization achieved by the system is 4.81% and the best syntactic diacritization achieved is 9.97% compared to the best-published results. Since this work is in progress, these results are expected to be enhanced by extracting more Arabic linguistic rules (morphological and syntactic), adding more semantic features, using different machine learning techniques for morphological and syntactic processing levels and implementing the improvements by working on larger amounts of data. For enhancing the OOV results, more patterns with more features need to be handled.

ACKNOWLEDGMENT

Thanking Bibliotheca Alexandrina for their permission to use a sample of their morphologically analyzed Arabic Corpus.

REFERENCES

[1] Abandah, G. A., Graves, A., Al-Shagoor, B., Arabiyat, A., Jamour, F., & Al-Taee, M. (2015). Automatic diacritization of Arabic text using recurrent neural networks. *International Journal on Document Analysis and Recognition (IJDAR), 18(2)*, 183-197.

[2] Alansary, S. (2016a). Alserag: A Rule Based Approach for Arabic Text Diacrtization System. *In proceedings of 2nd International Conference on Adavanced Intelligent Systems and Informatics (AISI2016).* Cairo, Egypt.

[3] Alansary, S. (2016b). MASAR: A Morphologically Annotated Gold Standard Arabic Resource., (p. In proceedings of the 16th International Conference on Language Engineering). Cairo, Egypt, 7-8 December.

[4] Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0.* Linguistic Data Consortium, University of Pennsylvania, 2004. LDC Catalog No.: LDC2004L02.

[5] Chennoufi, A., & Mazroui, A. (2016a). Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization. *International Journal of Speech Technology, 19: 269*. doi:10.1007/s10772-015-9313-5

[6] Chennoufi, A., & Mazroui, A. (2016b). Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences. *Journal of King Saud University-Computer and Information Sciences.*

[7] Diab, M.; Ghoneim, M.; Habash, N. (2007, September). Arabic diacritization in the context of statistical machine translation. *In proceeding of MT-Summit.* Copenhagen, Denmark.

[8] Habash, N., & Rambow, O. (2005, June). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 573-580). Ann Arbor: Association for Computational Linguistics.

[9] Habash, N., & Rambow, O. (2007, April). Arabic Diacritization through Full Morphological Tagging. *In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics.Companion Volume. Short Papers* , pp. 53-56. Rochester, NY: Association for Computational Linguistics.

[10] Habash, N., Rambow, O., & Roth, R. (2009, April). 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)*, (pp. 102-109). Cairo, Egypt.

[11] Metwally A., S., Rashwan M., A., & Atiya F., A. (2016). A Multi-Layered Approach for Arabic Text Diacritization. *In proceeding of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* (pp. 389-393). IEEE.

[12] Rashwan, M. A., Al Sallab, A. A., Raafat, H. M., & Rafea, A. (2014). Automatic Arabic diacritics restoration based on deep nets. *In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, (pp. 65–72). Doha, Qatar. October 25.

[13] Rashwan, M. A., Al Sallab, A. A., Raafat, H. M., & Rafea, A. (2015). Deep Learning Framework with Confused Sub-Set Resolution Architecture for Automatic Arabic Diacritization. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 23(3)*, (pp. 505-516).

[14] Rashwan, M., Al-Badrashiny, M., Attia, M., & Abdou, S. (2009). A Hybrid System for Automatic Arabic Diacritization. *In the 2nd International Conference on Arabic Language Resources and Tools.* Cairo, Egypt.

[15] Rashwan, M., Al-Badrashiny, M., Attia, M., Abdou, S. M., & Rafea, A. (2011). A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Un-factorized Textual Features. 166-175. Audio, Speech, and Language Processing, IEEE Transactions on, 19 (1).

[16] Roth, R., Rambow, O., Habash, N., Diab, M., & Rudin, C. (2008, June). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 117-120). Columbus, Ohio, USA: Association for Computational Linguistics.

[17] Shaalan, K., Abo Bakr, H. M., & Ziedan, I. (2009, March). A Hybrid Approach for Building Arabic Diacritizer. *In Proceedings of the 9th EACL Workshop on Computational Approaches to Semitic Languages* (pp. 27-35). Association for Computational Linguistics.

[18] Shaalan, K., Abo Bakr, H. M., & Ziedan, I. A. (2008). A Statistical Method for Adding Case Ending Diacritics for Arabic Text. *In proceedings os Language Engineering Conference*, (pp. 225-234). Cairo, Egypt. 17-18 December.

[19] Shahrour, A., Khalifa, S., & Habash, N. (2015). Improving Arabic Diacritization through Syntactic Analysis. *In proceedings of Empirical Methods in Natural Language Processing Conference (EMNLP)* (pp. 1309–1315). Association for Computational Linguistics.

[20] Zitouni, I., Sorensen, J. S., & Sarikaya, R. (2006, July). Maximum Entropy Based Restoration of Arabic Diacritics. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of Association for Computational Linguistics* (pp. 577-584). Association for Computational Linguistics.

[21] http://aramedia.com/nlp2.htm [Acc. 12-2-2015].

[22] http://aramedia.com/diacritizer.htm [Acc. 12-2-2015].

[23] http://www.rdi-eg.com/technologies/arabic nlp.htm [Acc. 12-2-2015].

[24] http://tahadz.com/mishkal [Acc. 4-4-2015].

[25] http://harakat.ae/ [Acc. 4-4-2015].

[26] http://faraheedy.mukhtar.me/du2alee/tashkeel [Acc. 20-8-2016]

[27] http://qatsdemo.cloudapp.net/farasa/ [Acc. 28-12-2016]

BIOGRAPHIES

Amany Fashwan: *Head of International Corpus of Arabic Unit, Arabic Computational Linguistic Center, Bibliotheca Alexandrina, Alexandria, Egypt.*

She graduated from Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University (2005). She got her MSA thesis in 'Automatic Diacritization of Modern Standard Arabic Texts: A corpus based approach' with excellent degree (2016). She participated in building the International Corpus of Arabic. She participated in developing the BibAlex Standard Arabic Morphological Analyzer (BASMA). She participated with a team in building a tool for morphological analysis and generation of Arabic roots with excellent degree (field study). Her main areas of interest are building Arabic corpora, corpus based studies, Arabic morphology, Arabic syntax, Arabic semantics, Machine Learning Techniques and Language Modeling. She has experienced in morphological analysis and extracting and implementing Arabic linguistic rules depending on morphologically analyzed Arabic words.

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

التنبؤ بعلامات التشكيل للكلمات غير المعروفة في العربية

Amany Fashwan[1], Sameh Alansary[2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt

[1]amany.fashwan@alexu.edu.org

[2]s.alansary@alexu.edu.eg

ملخص—تقدم هذه الورقة البحثية نظاما لتحديد العلامة الإعرابية للكلمات غير المعروفة في اللغة العربية الفصحى،ومن أجل بناء هذا النظام تم الاعتماد على عينة لغوية محللة تحليلا صرفيا وتحتوي على العلامة الإعرابية للكلمات تتكون من حوالي 500 ألف كلمة تم اختيارها من المدونة اللغوية العربية العالمية. بالإضافة إلى عينة لغوية لاختبار هذا النظام مكونة من 52000 ألف كلمة. وقد تم تقييم النتائج الخاصة بالنظام بالاعتماد على قياسين أساسيين للتقييم هما معدل الخطأ على مستوى الكلمة ومعدل الخطأ هلى مستوى الحرف واللذان سجلا في مرحلة التشكيل الآلي لبينة الكلمة إلى الآن 4.56% و1.88% على التوالي، كما تم تسجيل 9.36% على مستوى تشكيل الحالة الإعرابية.وقد تم مقارنة نتائج النظام مع نتائج أكثر أنظمة التشكيل الآلي انتشارا بالاعتماد على نفس عينة الاختبار اللغوية المستخدمة في هذه الأنظمة من أجل موضوعية النتائج. وقد سجل النظام معدل خطأ14.78%مقارنة بـ[1] الذي سجل أقل معدل خطأ 11,68% [13]12.90% و[10]اللذين سجلا معدل خطأ 13,60%.

Quantitative Analysis of Egyptian Aphorisms Using R Language

Eman M. Yousri El-Gamal*[1], Sameh Alansary**[2]

*Ph.D researcher, Dept. of Phonetics & linguistics, Faculty of Arts, University of Alexandria. Alexandria, Egypt.
e.yousri@alexu.edu.eg[1]

**Professor of computational linguistics, Dept. of Phonetics & Linguistics, Faculty of Arts, University of Alexandria. Alexandria, Egypt.
s.alansary@alexu.edu.ed[2]

***Abstract*— Quantitative linguistics is the comparative study of the frequency and distribution of words and syntactic structures in different texts. The aim of qualitative analysis is a complete, detailed description of the linguistic data. Quantitative analysis is made to assign frequencies to the linguistic features which are identified in the data, and rare phenomena receives (or should receive) the same amount of attention as more frequent phenomena. Thus, the aim of the presented study is to make a compared quantitative linguistic analysis of the most common Egyptian aphorisms in both languages Arabic and English by using the statistical system of R language. The comparison is done between one corpus of Egyptian aphorisms in Arabic language and another corpus of the Egyptian aphorisms in English language. All the data of the present study is analysed morphologically and semantically. Then the output of the analysed corpora is use as an input for doing the quantitative linguistic analysis by using R language statistical analysis; which make us able to query about the most frequent linguistic features of the Egyptian aphorisms database in both languages Arabic and English. Further, a comparison is done between the Egyptian aphorisms in Arabic language and the Egyptian aphorisms in English language, moreover; many more statistical analyses will be investigated from the present database of the Egyptian aphorisms.**

Keywords: **Quantitative linguistics, statistical linguistics, corpus linguistics, NLP, R language.**

1 INTRODUCTION

In quantitative linguistic analysis the features were classified, counted, and even constructed more complex statistical models in an attempt to explain what is observed. Quantitative findings can be generalized to a larger population, and direct comparisons can be made between two corpora, so long as valid sampling and significance techniques have been used. Thus, quantitative analysis allows us to discover which phenomena are likely to be genuine reflections of the behavior of a language or variety, and which are merely chance occurrences. The more basic task of just looking at a single language variety allows one to get a precise picture of the frequency and rarity of particular phenomena, and thus their relative normality or abnormality.

Hurford, Heasley & Smith (2007) state that, in a corpus; the description of words of any language is a central part of its components. Thus, a good corpus should typically gives (at least) three kinds of information about words, such as; grammatical information which includes syntactical or morphological information about its part of speech (e.g. noun, verb) and inflections (e.g. for plural number or past tense or gender), and semantic information about the word's animacy.

The aim of the presented study is to make a comparison quantitative linguistic analysis between two corpora of Egyptian aphorisms; one corpus is in Arabic language and the other is in English language, by using the statistical system of R language. The analysis of these corpora is done on two levels of linguistic analyses (morphological analysis and semantic analysis):

- First: is the morphological analysis, whereas; the part of speech tagging of each word and their description and inflections are analyzed. The tagging process is done by using the online Stanford Parser and the description and inflection of part of speech is done manually based

on the part-of-speech tags which used in the Penn Treebank Project. As well as, the gender of each word in the two corpora, that done manually too.

- Second: is the semantic analysis, in which, the animacy (animate and inanimate) of each word in the two corpora where analyzed manually.

After that, the output of the two analyzed corpora will be used as an input for doing the quantitative linguistic analysis by using R language statistical analysis; which makes us able to query about the frequency of any word included in each corpus of the Egyptian aphorisms corpora and their part of speech tagging. Moreover, the numbers of words of each sentence of the two corpora are measured, that enable us to compare between the two corpora of Egyptian aphorisms. That analyses enable us to make a comparative study between the number of tokens of each Egyptian aphorisms corpus, and many more statistical analyses can be investigated of the present database of the Egyptian aphorisms. Another aim can be achieved throughout the present study is to evaluate the online Stanford Parser by evaluating the time taken for marking the part of speech tagging of different Egyptian aphorisms with different lengths (different number of tokens) and different languages too, by making a relation between the length of the aphorism and the time taken for tagging its words by using R language too.

2 METHOD

The adopted method for accomplishing the quantitative linguistics analysis of the two corpora (in Arabic language & in English language) of Egyptian aphorisms will be clarified in details in this section.

A. Collecting Data

In the present study, the linguistic data which collected for the two corpora are sentences from the Egyptian aphorisms. One corpus is in Arabic language and composed of 50 sentences of Egyptian aphorisms. And, the other corpus is composed of 50 English sentences of Egyptian aphorisms too. The two corpora are collected from the online database of Egyptian aphorisms.

B. Analyzing Data

After collecting the two corpora of Egyptian aphorisms in Arabic language and in English language; analyzing the two corpora is done through to steps:

- First: automatically; by using the online Stanford Parser, for counting the number of tokens of each aphorism of the two corpora of Egyptian aphorisms. In addition, for marking the part of speech tagging (which refer to a syntactic function) of each word of the included two corpora, as well as, measuring the taken time for tagging the words of each aphorism.
- Second: manually; by adding the descriptions and inflections of each part of speech tagging with the aid of the list of the parts of speech encoded in the annotation system of the Penn Treebank Project (which includes the parts of speech with their corresponding abbreviations "tags" and some additional information). Also, analyzing the animacy (animate / inanimate) and the gender (masculine / feminine) of each annotated word of the two corpora of the Egyptian aphorisms.

C. Manipulating Data

Manipulating the analyzed data of the two corpora of the Egyptian aphorisms includes converting and tabulating all the analyzed data into excel sheet (which will be saved as CSV file) to be accepted and read by R language as CSV file. The present study includes four CSV files named as a1, a2, e1 & e2. Files named a1 & e1 (see table 1 & table 3) includes the querying Egyptian aphorisms sentences in Arabic language and in English language, respectively, with their ID, Tokens, Time, and Language. Consequently, the total number of the collected Egyptian aphorisms of the two corpora is 100 sentences (50 in Arabic language & 50 in English language).

Whereas, files named a2 & e2 (see table 2 & table 4) includes the Query sentence ID and the list of words of each sentence with their word ID, respectively. Files named a2 & e2 also includes Tag, Description, Animacy, and Gender for each word that included in the present study. Therefore, the total number of words of the two corpora is 1007 words (512 in Arabic language & 495 in English language).

TABLE 1: SAMPLE OF EXCEL SHEET FILE THAT NAMED A1; INCLUDES THE QUERYING EGYPTIAN APHORISMS SENTENCES IN ARABIC LANGUAGE WITH THEIR ID, TOKENS, TIME, AND LANGUAGE.

ID	Tokens	Time	Query
Arabic 1	11	0.183	لا تكفى مطالعة كل شئ بل ينبغى فهم كل ما تقرأ
Arabic 2	5	0.026	الصبر عند القتال هو الشجاعة
Arabic 3	14	0.702	لا تقس الأمور بعقلك، فاللعقل حد ينتهى إليه كما للبصر حد ينتهى إليه
Arabic 4	10	0.140	بالمال لا تعرف نفسك و بدون المال لا يعرفك أحد
Arabic 5	9	0.193	العاقل يغذى صحته بماله أما الأحمق فيغذى ماله بصحته
Arabic 6	7	0.057	كن كالنحلة تأكل طيبا و تضع طيبا
Arabic 7	11	0.177	يستطيع المال أن يشترى الطعام الشهى و لكنه لا يفتح الشهية
Arabic 8	4	0.022	نسيان الإساءة انتقام رقيق
Arabic 9	14	0.224	من لم يصبر على العلم فى بداية حياته ، صبر على الذل حتى مماته
Arabic 10	13	0.297	البخيل من يعيش فى الدنيا عيش الفقراء و يحاسب فى الآخرة محاسبة الأغنياء
Arabic 11	9	0.085	الصحة تاج على رءوس الأصحاء لا يراه إلا المرضى
Arabic 12	6	0.037	الصبر عند شهوة الطعام يسمى قناعة
Arabic 13	13	0.210	كن شجاعا تكن صادقا، فإذا اعتدت الصدق مشت الفضائل كلها فى ركابك
Arabic 14	10	0.120	إذا نطق السفيه فلا تجبه , فخير من إجابته السكوت
Arabic 15	10	0.161	إذا دعتك قدرتك على ظلم الناس فتذكر قدرة الله عليك
Arabic 16	7	0.090	الجمال هو طهارة القلب و نقاء الضمير
Arabic 17	6	0.056	الجاهل سكران لا يفيق إلا بالمعرفة
Arabic 18	8	0.142	لكل امرئ فى ماله شريكان الوارث والحوادث
Arabic 19	13	0.145	البخيل من يعيش فى الدنيا عيش الفقراء و يحاسب فى الآخرة محاسبة الأغنياء
Arabic 20	14	0.302	لا تر كل ما تراه عيناك ، و لا تسمع كل ما تسمعه أذناك

TABLE 2: SAMPLE OF EXCEL SHEET FILE THAT NAMED A2; INCLUDES THE QUERY SENTENCE ID AND THE LIST OF WORDS OF EACH SENTENCE WITH THEIR WORD ID WITH THEIR TAG, DESCRIPTION, ANIMACY, AND GENDER OF ARABIC EGYPTIAN APHORISMS.

Languag	QueryID	WordID	Word	Tag	Description	Inflection	Animacy	Gender
Arabic	1	1	لا	RP	Particle	null	null	null
Arabic	1	6	بل	CC	Coordinating conjunction	null	null	null
Arabic	1	10	ما	WP	Wh-pronoun	null	null	null
Arabic	3	1	لا	RP	Particle	null	null	null
Arabic	3	5	,	PUNC	Punctuation	null	null	null
Arabic	3	10	كما	CC	Coordinating conjunction	null	null	null
Arabic	4	2	لا	RP	Particle	null	null	null
Arabic	4	6	بدون	IN	Preposition or subordinating conjunction	null	null	null
Arabic	4	8	لا	RP	Particle	null	null	null
Arabic	5	5	اما	RP	Particle	null	null	null
Arabic	6	5	و	CC	Coordinating conjunction	null	null	null
Arabic	7	3	ان	IN	Preposition or subordinating conjunction	null	null	null
Arabic	7	7	و	CC	Coordinating conjunction	null	null	null
Arabic	7	9	لا	RP	Particle	null	null	null
Arabic	9	1	من	WP	Wh-pronoun	null	null	null
Arabic	9	2	لم	RP	Particle	null	null	null
Arabic	9	4	على	IN	Preposition or subordinating conjunction	null	null	null
Arabic	9	6	فى	IN	Preposition or subordinating conjunction	null	null	null
Arabic	9	9	,	PUNC	Punctuation	null	null	null

TABLE 3: SAMPLE OF EXCEL SHEET FILE THAT NAMED E1; INCLUDES THE QUERYING EGYPTIAN APHORISMS SENTENCES IN ENGLISH LANGUAGE WITH THEIR ID, TOKENS, TIME, AND LANGUAGE.

ID	Tokens	Time	Query
English 51	13	0.054	**Empty not your soul to everybody and do not diminish thereby your importance.**
English 52	9	0.025	**The nut doesn't reveal the tree it contains.**
English 53	13	0.116	**The seed cannot sprout upwards without simultaneously sending roots into the ground.**
English 54	12	0.036	**A house has the character of the man who lives in it.**
English 55	17	0.081	**A pupil may show you by his own efforts how much he deserves to learn from you.**
English 56	10	0.029	**Social good is what brings peace to family and society.**
English 57	5	0.007	**Knowledge is not necessarily wisdom.**
English 58	19	0.094	**Each truth you learn will be, for you, as new as if it had never been written.**
English 59	13	0.043	**Listen to your convictions, even if they seem absurd to your reason.**
English 60	10	0.026	**Man, know yourself and you shalt know the gods**
English 61	9	0.025	**People bring about their own undoing through their tongues.**
English 62	13	0.051	**If you search for the laws of harmony, you will find knowledge.**
English 63	12	0.046	**Experience will show you, a Master can only point the way.**
English 64	8	0.012	**Love is one thing, knowledge is another.**
English 65	14	0.062	**Organization is impossible unless those who know the laws of harmony lay the foundation.**
English 66	8	0.013	**The only thing that is humiliating is helplessness.**
English 67	21	0.156	**The first thing necessary in teaching is a master, the second is a pupil capable of carrying on the tradition.**
English 68	10	0.018	**For every joy there is a price to be paid.**
English 69	11	0.048	**The best and shortest road towards knowledge of truth is Nature.**
English 70	8	0.019	**Leave him in error who loves his error.**
English 71	4	0.007	**Understanding develops by degrees.**

TABLE 4: SAMPLE OF EXCEL SHEET FILE THAT NAMED E2; INCLUDES THE QUERY SENTENCE ID AND THE LIST OF WORDS OF EACH SENTENCE WITH THEIR WORD ID WITH THEIR TAG, DESCRIPTION, ANIMACY, AND GENDER OF ENGLISH EGYPTIAN APHORISMS.

Language	QueryID	WordID	Word	Tag	Description	Inflection	Animacy	Gender
English	51	1	Empty	VB	Verb	base form	null	null
English	51	2	not	RB	Adverb	null	null	null
English	51	3	your	PRP$	Possessive pronoun	null	animate	masculine
English	51	4	soul	NN	Noun	singular or mass	animate	feminine
English	51	5	to	TO	to	null	null	null
English	51	6	everybody	NN	Noun	singular or mass	animate	null
English	51	7	and	CC	Coordinating conjunction	null	null	null
English	51	8	do	VBP	Verb	non-3rd person singular present	animate	masculine
English	51	9	not	RB	Adverb	null	null	null
English	51	10	diminish	VB	Verb	base form	null	null
English	51	11	thereby	RB	Adverb	null	null	null
English	51	12	your	PRP$	Possessive pronoun	null	animate	masculine
English	51	13	importance	NN	Noun	singular or mass	animate	masculine
English	52	1	The	DT	Determiner	null	null	null
English	52	2	nut	NN	Noun	singular or mass	animate	masculine
English	52	3	does	VBZ	Verb	3rd person singular present	null	null
English	52	4	n't	RB	Adverb	null	null	null
English	52	5	reveal	VB	Verb	base form	null	null
English	52	6	the	DT	Determiner	null	null	null
English	52	7	tree	NN	Noun	singular or mass	animate	feminine
English	52	8	it	PRP	Personal pronoun	null	inanimate	masculine
English	52	9	contains	VBZ	Verb	3rd person singular present	null	null

D. Querying The Data

This section includes running the manipulated data and making the statistical measurements, to find different relations between both corpora (Arabic & English) by using R language to investigate the quantitative linguistic (lexical) characteristics of Egyptian aphorisms in Arabic and English languages.
To make our data readable by R system, the corpora should be loaded as csv files; this step is done by using the following codes:

First, for reading the Arabic Egyptian aphorisms corpus, the following codes are used:

```
a = read.csv ("C:/Users/eman/Desktop/a1.csv", header = TRUE, sep =';')
aa = read.csv ("C:/Users/eman/Desktop/a2.csv",header = TRUE, sep =';')
```

Second, for reading the English Egyptian aphorisms corpus, the following codes are used:

```
e = read.csv ("C:/Users/eman/Desktop/e1.csv", header = TRUE, sep =';')
ee =  read.csv ("C:/Users/eman/Desktop/e2.csv", header = TRUE, sep =';')
```

To select which of Words' Description is a "Verb" in the Arabic corpus or in English corpus, respectively; the following codes are used:

```
aa[aa$Description=="Verb", ]
ee[ee$Description=="Verb", ]
```

To select which of Words' Tag is a "NN" in the Arabic corpus or in English corpus, respectively; the following codes are used:

```
ee[ee$Tag=="NN", ]
aa[aa$Tag=="NN", ]
```

To select which of Words' Tag is a "DTNN" and occurred as a first word in the sentences in the Arabic corpus or in English corpus, respectively; the following code is used (Note that English corpus did not has DTNN tagging, though we select only the sentences which start with a DT):

```
aa[aa$Tag=="DTNN" & aa$WordID== 1 , ]
ee[ee$Tag=="DT" & ee$WordID== 1 , ]
```

For more constrictions, in both corpora, the following codes are used:

```
ee[ee$Tag=="NN" & ee$WordID== 1 & ee$Animacy=="animate"  , ]
aa[aa$Tag=="NN" & aa$WordID== 1 & aa$Animacy=="animate"  , ]
```

For ordering the sentences of both corpora (Arabic and English, separately) according to their number of tokens, the following codes are used:

```
a[order(a$Tokens), ]
e[order(e$Tokens), ]
```

For ordering the Words of both corpora (Arabic and English, separately) alphabetically, the following codes are used:

```
aa[order(aa$Word), ]
```

```
ee[order(ee$Word), ]
```

For adding a column (column called "Length" and counts the number of letters or characters for each word) in the data frame of both corpora of Egyptian aphorisms, the following code is used:

```
aa$Length= nchar(as.character(aa$Word))
```

For calculating the number of characters or letters for each sentence of both corpora, the following code is used:

```
a$Length= nchar(as.character(a$Query))
```

Then:

```
a[1:3, c("Length", "Query")]
```

To tabulate the word and its tag, the following code is used:

```
xtabs(~Word+ Tag, data= aa)
```

Also, other relations can be tabulated, using the following codes:

```
xtabs(~Gender+ Tag, data= aa)
xtabs(~Gender+ Animacy, data= aa)
xtabs(~Description+ Tag, data= aa)
xtabs(~Inflection+Tag, data= aa)
xtabs(~Inflection + Tag + Animacy, data= aa)
xtabs(~Gender + Animacy, data= aa)
```

For chai square test, for both corpora, the following codes are used:

```
chisq.test(a$Token)
chisq.test(e$Token)
```

For calculating the frequency of words in both corpora (Arabic & English) of Egyptian aphorisms, the following steps are done by the following codes, using R language:

```
temple.freq.list=table(ee$Word)
table(ee$Word)
```

```
> temple.freq.list=table(ee$Word)
> table(ee$Word)

           ,            a            A        about       absurd
          13            8            2            1            1
      active          all        alone           an          and
           1            1            1            1            8
     another    architect          are       arises           as
           1            1            2            1            3
       based           be          bed         been      beliefs
           1            6            1            1            1
        best       better        blind        bliss         both
           2            1            2            1            1
       bread    brightest        bring       brings       buried
           1            1            1            1            1
          by          can     capable    carrying    character
           3            3            1            1            1
       comes communication     contains  convictions      corrupt
           1            1            1            1            1
     courage          defy     degrees     deserves     develops
           1            1            1            1            1
    diminish     discover        ditch           do         does
           1            1            1            4            2
        done       double     doubting         Each        early
           1            1            1            1            1
```

Then, to organize the frequency of words decreasingly (the most frequent followed by the least ones), by R language; the following codes are used:

```
temple.sorted.freq.list=sort(temple.freq.list, decreasing=TRUE)
```

temple.sorted.freq.list

```
> temple.sorted.freq.list=sort(temple.freq.list, decreasing=TRUE)
> temple.sorted.freq.list
```

the	is	you	of	,	to
25	22	17	15	13	12
The	a	and	not	who	be
10	8	8	7	7	6
in	it	do	his	If	know
6	6	4	4	4	4
knowledge	thing	will	your	as	by
4	4	4	4	3	3
can	He	learn	makes	man	no
3	3	3	3	3	3
on	only	own	that	A	are
3	3	3	3	2	2
best	blind	does	error	Every	for
2	2	2	2	2	2
from	good	half	harmony	has	him
2	2	2	2	2	2
if	into	laughs	laws	one	out
2	2	2	2	2	2
possession	pupil	show	their	truth	way
2	2	2	2	2	2
weep	wise	without	yourself	about	absurd
2	2	2	2	1	1

To sort the list of word frequency in different way by R language (in that sort the word is separated from its frequency of occurrence by "\t" followed by the frequency number), the following codes are used:

```
temple.sorted.table=paste(names(temple.sorted.freq.list),
temple.sorted.freq.list, sep="\t")
temple.sorted.table
```

```
> temple.sorted.table=paste(names(temple.sorted.freq.list), temple.sorted.freq.list, sep="\t")
> temple.sorted.table
  [1] "the\t25"           "is\t22"             "you\t17"          "of\t15"
  [5] ",\t13"             "to\t12"             "The\t10"          "a\t8"
  [9] "and\t8"            "not\t7"             "who\t7"           "be\t6"
 [13] "in\t6"             "it\t6"              "do\t4"            "his\t4"
 [17] "If\t4"             "know\t4"            "knowledge\t4"     "thing\t4"
 [21] "will\t4"           "your\t4"            "as\t3"            "by\t3"
 [25] "can\t3"            "He\t3"              "learn\t3"         "makes\t3"
 [29] "man\t3"            "no\t3"              "on\t3"            "only\t3"
 [33] "own\t3"            "that\t3"            "A\t2"             "are\t2"
 [37] "best\t2"           "blind\t2"           "does\t2"          "error\t2"
 [41] "Every\t2"          "for\t2"             "from\t2"          "good\t2"
 [45] "half\t2"           "harmony\t2"         "has\t2"           "him\t2"
 [49] "if\t2"             "into\t2"            "laughs\t2"        "laws\t2"
 [53] "one\t2"            "out\t2"             "possession\t2"    "pupil\t2"
 [57] "show\t2"           "their\t2"           "truth\t2"         "way\t2"
 [61] "weep\t2"           "wise\t2"            "without\t2"       "yourself\t2"
 [65] "about\t1"          "absurd\t1"          "active\t1"        "all\t1"
 [69] "alone\t1"          "an\t1"              "another\t1"       "architect\t1"
 [73] "arises\t1"         "based\t1"           "bed\t1"           "been\t1"
 [77] "beliefs\t1"        "better\t1"          "bliss\t1"         "both\t1"
 [81] "bread\t1"          "brightest\t1"       "bring\t1"         "brings\t1"
 [85] "buried\t1"         "capable\t1"         "carrying\t1"      "character\t1"
 [89] "comes\t1"          "communication\t1"   "contains\t1"      "convictions\t1"
```

To sort the list of word frequency in different way by R language (in that sort the word is separated from its frequency of occurrence by "-" followed by the frequency number), the following codes are used:

```
temple.sorted.table=paste(names(temple.sorted.freq.list),
temple.sorted.freq.list, sep="-")
temple.sorted.table
```

```
> temple.sorted.table=paste(names(temple.sorted.freq.list), temple.sorted.freq.list, sep="-")
> temple.sorted.table
  [1] "ء-٨"       "في-١١"      "ان-١٥"       "و-١٨"       "لا-٢٠"       "من-٢٧"
  [7] "على-٥"      "اذا-٥"       "هو-٦"       "كل-٦"       "اليه-٦"        "ما-٨"
 [13] "ليس-٣"       "كن-٣"      "عند-٣"    "الناس-٣"    "الدنيا-٣"     "الخلق-٣"
 [19] "الاخرة-٢"   "الاحمق-٢"      "الا-٢"     "اكثر-٢"      "احد-٢"     "يفتح-٣"
 [25] "السفيه-٢"    "السخاء-٢"     "الذي-٢"   "البخيل-٢"  "الانسان-٢" "الاغنياء-٢"
 [31] "الله-٢"   "الفقير-٢"  "الفقراء-٢"   "الظلام-٢"    "الطعام-٢"     "الصبر-٢"
 [37] "حد-٢"    "تعطيني-٢"       "تحب-٢"       "بل-٢"       "الى-٢"     "المال-٢"
 [43] "فلا-٢"       "عيش-٢"      "طيبا-٢"      "شمعة-٢"       "رجل-٢"       "خير-٢"
 [49] "لك-٢"       "كما-٢"    "كثيرا-٢"      "قبل-٢"       "في-٢"       "فهو-٢"
 [55] "يحتاج-٢"     "يحاسب-٢"    "محاسبة-٢"      "ماله-٢"        "له-٢"        "لم-٢"
 [61] "اجابته-١"      "يوجد-٢"     "ينتهي-٢"      "يغلق-٢"      "يعيش-٢"    "يستطيع-٢"
 [67] "اذنيه-١"      "اذنب-١"     "اذناك-١"     "اذا انهم-١"     "اختفى-١"     "احترس-١"
 [73] "اعتدت-١"      "اصبحت-١"     "اشترى-١"     "ازدادت-١"      "ازداد-١"       "اردت-١"
 [79] "الاساءة-١"     "الارض-١"   "الاخرين-١"    "اقدامك-١"   "افواههم-١"      "اعطيت-١"
 [85] "الجروح-١"     "الجاهل-١"     "الباب-١"     "الامور-١"      "الالم-١"   "الاصحاء-١"
 [91] "الذى-١"       "الذل-١"   "الحوادث-١"      "الحق-١"  "الحسنات-١"     "الجمال-١"
 [97] "السيئات-١"      "السوط-١"      "السن-١"     "السماء-١"     "السكوت-١"     "السفح-١"
[103] "الصدق-١"     "الصحة-١"    "الشهية-١"     "الشهى-١"    "الشجرة-١"   "الشجاعة-١"
[109] "العدل-١"    "العاقل-١"    "العاقر-١"    "الطريق-١"    "الضمير-١"    "الصعود-١"
[115] "القلب-١"    "القتال-١"     "القاع-١"     "الفقر-١"  "الفضائل-١"      "العلم-١"
[121] "المرضى-١"      "المرء-١"     "الكسل-١"     "الكثير-١"      "القمة-١"    "القليل-١"
[127] "الوقوف-١"     "الوارث-١"     "النار-١"     "الناجح-١"      "المهم-١"   "المضروب-١"
[133] "انتقاد-١"        "انت-١"        "انا-١"       "امرئ-١"        "اما-١"      "اليها-١"
[139] "انهم-١"       "انما-١"       "انك-١"      "انظر-١"     "انسان-١"    "انتقام-١"
[145] "بحجر-١"        "بان-١"  "بالمعرفة-١"    "بالمال-١"       "باع-١"       "اولا-١"
[151] "بعقلك-١"     "بصديق-١"     "بصحته-١"      "بدون-١"     "بداية-١"        "بدا-١"
[157] "به-١"    "بماله-١"     "بليغا-١"     "بلغت-١"       "بلا-١"      "بقى-١"
```

For saving the output list of words frequencies resulted by R language in txt, the following codes are used:

```
cat("Word\tFREQ", temple.sorted.table, file=choose.files(), sep="\n")
```

All the previous steps for calculating the frequency of words by R language are done for both corpora of Egyptian aphorisms (Arabic corpus and English corpus). Also all the previous steps are done for calculating the tagging frequency in both corpora, for making a comparison of the lexical characteristics between the both corpora of Egyptian aphorisms, in order to investigate the quantitative linguistic characteristics of Egyptian aphorisms in Arabic and English languages. The following table showing us a sample of the excel sheet of the final word frequency list of both corpora of Egyptian aphorisms after saving the list by R language.

3 RESULTS (QUERY OUTPUT OF R)

This section includes the results of the querying system by using R language, In order to extract the quantitative linguistic characteristics of Arabic and English corpora of Egyptian aphorisms. See Table 5 & Table 6

Egyptian Aphorisms			
English corpus		Arabic Corpus	
Word	Frequency	Word	Frequency
the	25	من	27
is	22	لا	20
you	17	و	18
of	15	ان	15
,	13	في	11
to	12	,	8
The	10	ما	8
a	8	اليه	6
and	8	كل	6
not	7	هو	6
who	7	اذا	5
be	6	على	5
in	6	الخلق	3
it	6	الدنيا	3
do	4	الناس	3
his	4	عند	3
if	4	كن	3
know	4	ليس	3
knowledge	4	يفتح	3
thing	4	احد	2
will	4	اكتر	2

TABLE 5: SAMPLE OF EXCEL SHEET FILE TO SHOW US THE MOST FREQUENT WORDS THAT ACCUERED IN BOTH LANGUAGES ARABIC AND ENGLISH APHORISMS.

TABLE 6: SAMPLE OF EXCEL SHEET FILE TO SHOW US THE MOST FREQUENT TAGGS THAT ACCUERED IN BOTH LANGUAGES ARABIC AND ENGLISH APHORISMS.

Egyptian Aphorisms			
English corpus		Arabic Corpus	
Tagging	Frequency	Tagging	Frequency
NN	85	NN	89
DT	56	DTNN	80
IN	52	VBP	74
VBZ	43	IN	59
JJ	34	NNP	48
PRP	33	RP	27
RB	27	VBD	25
VB	23	CC	22
NNS	18	WP	20
VBP	16	JJ	15
PUNC	13	DTJJ	13
NNP	12	PRP	8
TO	12	PUNC	8
MD	11	NOUN_QUANT	6
PRP$	11	VBN	4
CC	8	DTNNS	3
VBN	8	JJR	3
WP	8	NNS	2
JJS	4	PRON	2
VBG	4	ADJ_NUM	1
JJR	3	DT	1

The results section is divided into two subsections:

A. Statistical measurements by R; which contains the statistical measurements that is done by R statistics.

B. Visualizing the output by R: which contains the visualization of the output of the R results by R graphics. Both subsections are all done by R language; in order to compare between the Arabic corpus and the English corpus of Egyptian aphorisms.

A. *Statistical measurements by R*

Table 7 showing us the statistical measurements of the Egyptian aphorisms corpora in Arabic & English languages; the measurements includes a comparison between the Arabic corpus and the English corpus of the Egyptian aphorisms. All of the following measurements are done by using R language statistics.

TABLE 7
THE OUTPUT OF QUERY ON THE EGYPTIAN APHORISMS IN ARABIC CORPUS AND ENGLISH CORPUS USING R

Querying Objects		**Egyptian Aphorisms**	
		Arabic Corpus	**English Corpus**
1) Sum of tokens		512	495
2) Mean number of tokens		10.24	9.9
3) Sum time		11.428	1.763
4) Mean time		0.22856	0.03526
5) Mean length of words		3.794922	4.179798
6) Median length of words		4	4
7) Range length of words		1 to 8	1 to 14
8) Mean length of aphorisms sentences		48.18	50.66
9) Median length of aphorisms sentences		45	45.5
10) Range length of aphorisms sentences		21 to 107	17 to 111
11) Minimum number of tokens		4	4
12) Maximum number of tokens		22	21
13) Chi-squared test of aphorism length (df = 49, p-value < 2.2e-16)		X-squared = 367.82	X-squared = 444.48
14) Chi-squared test of number of tokens (df = 49)		X-squared = 75.695, p-value = 0.00853	X-squared = 80.051, p-value = 0.003364
15) Animacy of words frequencies	Animate	254	184
	Inanimate	117	52
	Null	141	259
16) Gender of words frequencies	Masculine	292	179
	Feminine	72	51
	Null	148	265
17) Tagging frequencies	NN	89	85
	IN	59	52
	CC	22	8
	NNP	48	12
	VBN	4	8
	WRB	1	3
	NNS	2	18
	PRP	8	33
	VBP	74	16
	RP	27	1
	WP	20	8

B. *Visualizing the output by R*

Figure 1 and Figure 2 showing us the bar plots and the histograms of the querying objects to visualize the output of the R results for both corpora (English and Arabic) of the Egyptian aphorism. Note that; on the left hand occurred the results of the English corpus and on the right hand the results of the Arabic corpus.
Listed below; the querying object followed by the used codes to create those figures for both corpora (English and Arabic) of Egyptian aphorisms.

For visualizing the word length in English corpus and in Arabic corpus, the following codes are used by R language; respectively: (see Figure 1)

```
barplot(xtabs(~ee$Length), xlab= "word length in English corpus", col= "grey")
barplot(xtabs(~aa$Length), xlab= "word length in Arabic corpus", col= "grey")
```

For visualizing the number of tokens of each sentence of the query in English corpus and in Arabic corpus, the following codes are used by R language, respectively: (see Figure 1)

```
barplot(xtabs(~e$Tokens), xlab= "English corpus tokens numbers", col= "grey")
barplot(xtabs(~a$Tokens), xlab= "Arabic corpus tokens numbers", col= "grey")
```

For visualizing the query length in English corpus and in Arabic corpus, the following codes are used by R language, respectively: (see Figure 1)

```
barplot(xtabs(~e$Length), xlab= "English query length", col= "grey")
barplot(xtabs(~a$Length), xlab= "Arabic query length", col= "grey")
```

For visualizing the Animacy of each word in the query in English corpus and in Arabic corpus, the following codes are used by R language, respectively: (see Figure 2)

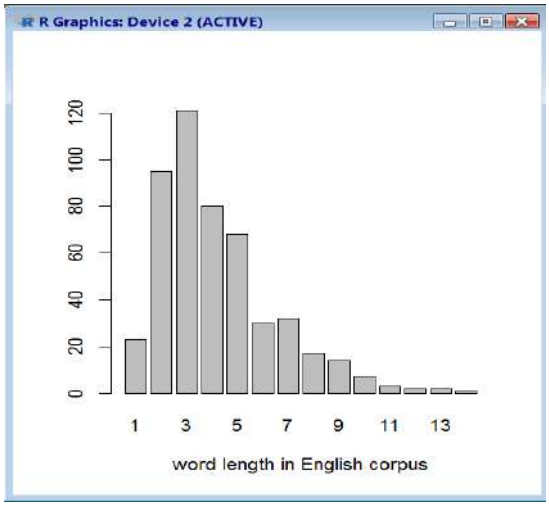
```
barplot(xtabs(~ee$Animacy), xlab= "Animacy in English corpus", col= "grey")
barplot(xtabs(~aa$Animacy), xlab= "Animacy in Arabic corpus", col= "grey")
```

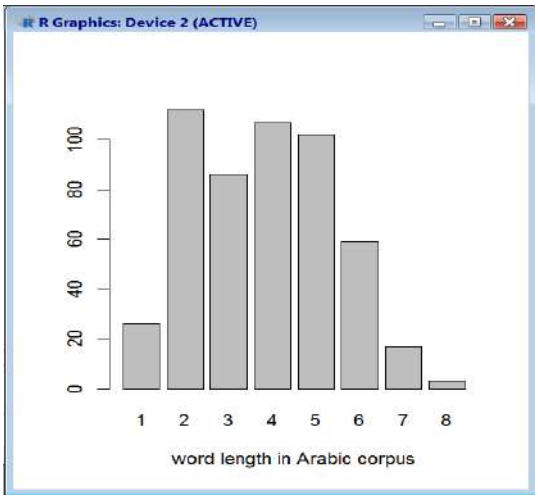
For visualizing the Gender of each word of the query in English corpus and in Arabic corpus, the following codes are used by R language, respectively: (see Figure 2)

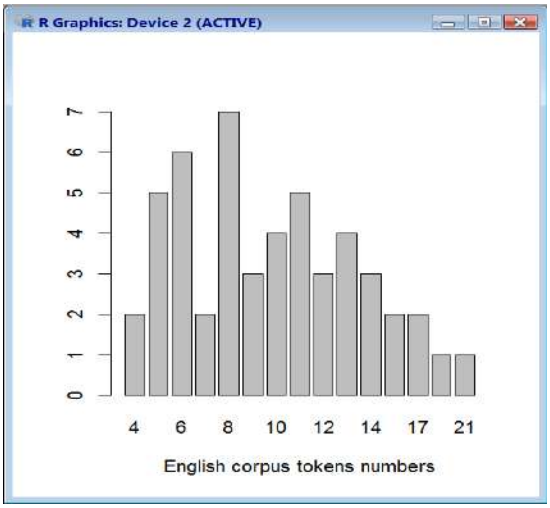
```
barplot(xtabs(~ee$Gender), xlab= "Gender in English corpus", col= "grey")
barplot(xtabs(~aa$Gender), xlab= "Gender in Arabic corpus", col= "grey")
```

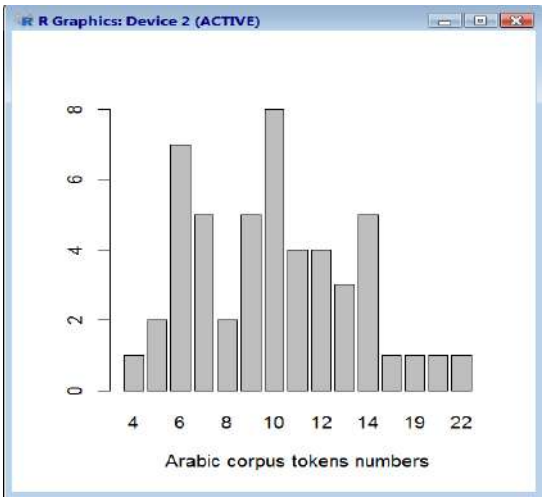
For visualizing tokens numbers in English corpus and in Arabic corpus, the following codes are used by R language, respectively: (see Figure 2)

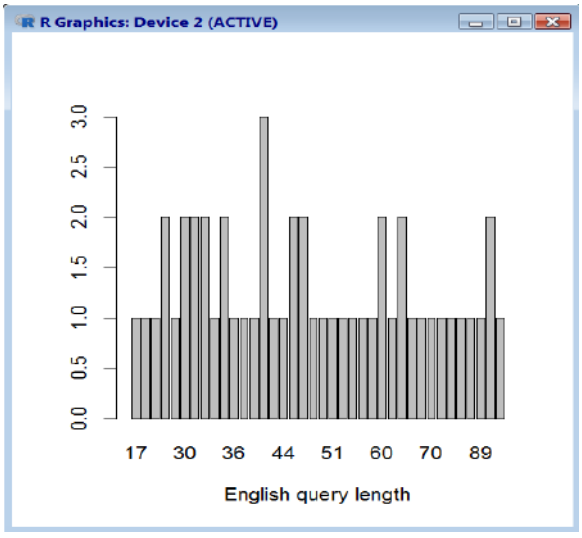
```
truehist(e$Tokens, col="lightblue", xlab="English tokens numbers")
truehist(a$Tokens, col="lightblue", xlab="Arabic tokens numbers")
```











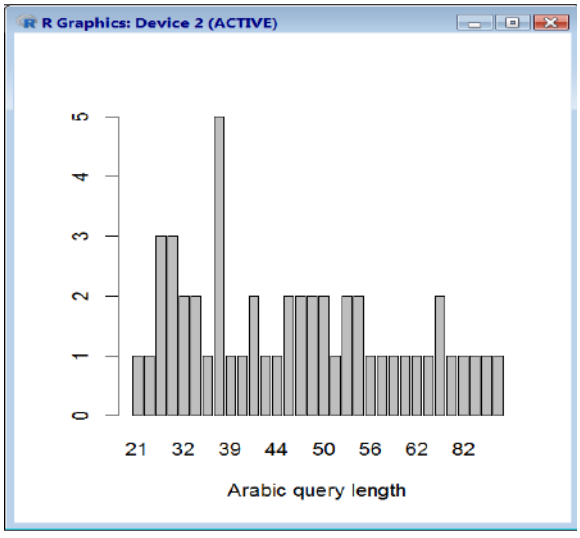
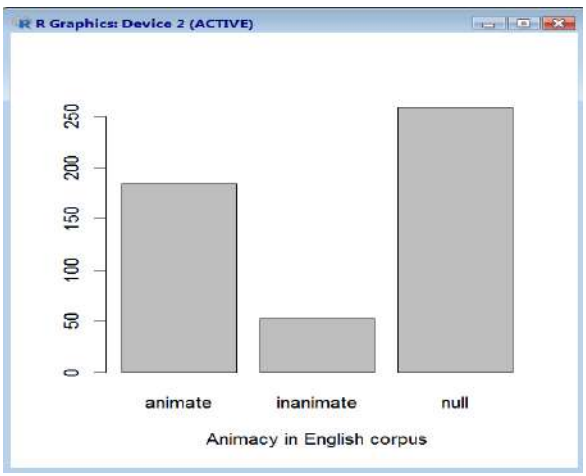
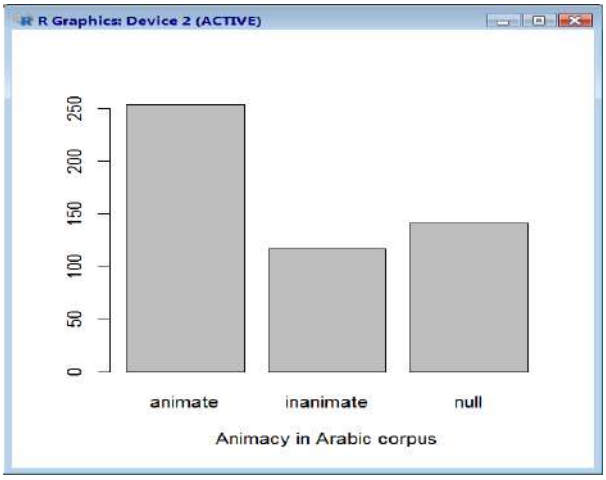
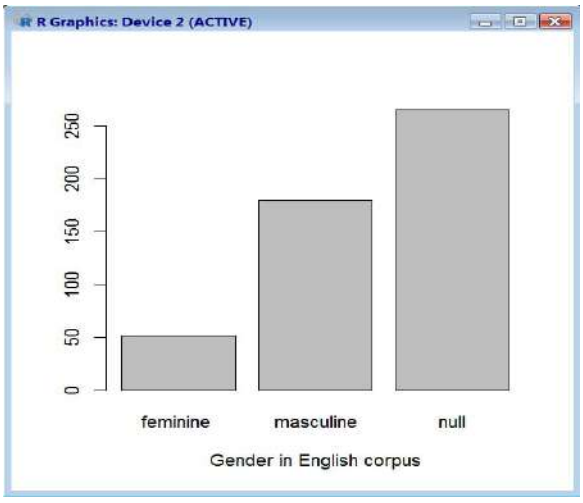
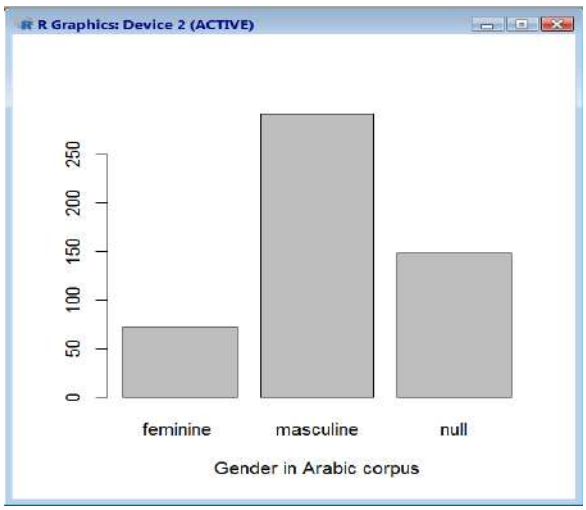


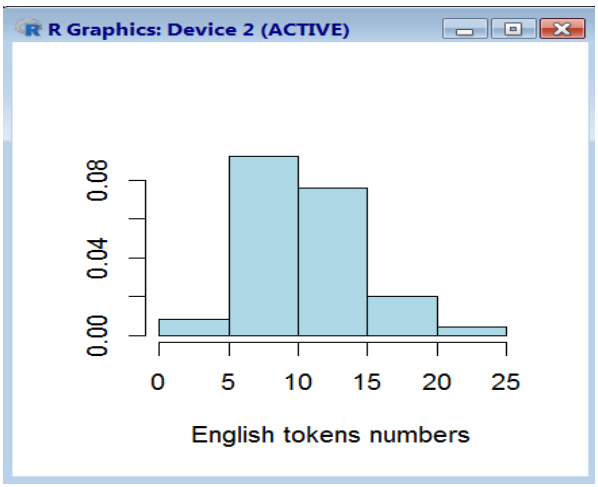
Figure 1: showing a comparison of the query results of word length, number of tokens and query length (sentence length) by R output for both corpora (Arabic & English) of English aphorisms. English corpus results are on the left hand and Arabic corpus results are on the right hand.











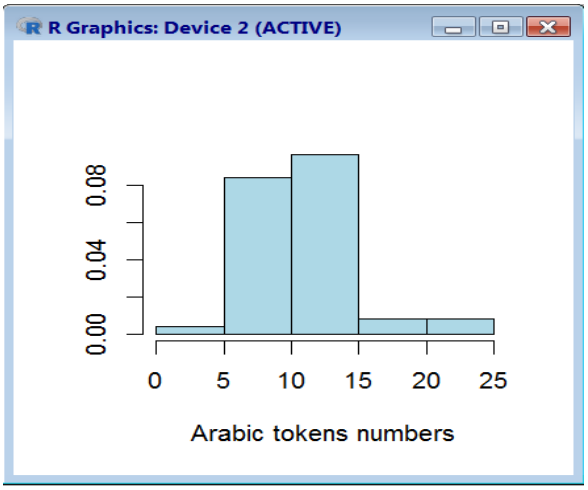


Figure 2: showing a comparison of the query results of gender, animacy and tokens numbers by R output for both corpora (Arabic & English) of English aphorisms. English corpus results are on the left hand and Arabic corpus results are on the right hand.

4 CONCLUSIONS

The present study is designed to make a quantitative analysis of the most common observed linguistic features which are identified in the data of the Egyptian aphorisms. This aim is reached by making a comparison between two corpora of Egyptian aphorisms in two languages; Arabic and English by using the statistical system of R language. After analyzing and querying the data of the two corpora in both languages; the results indicate the following:

- The mean length of an Egyptian Aphorism in the Arabic corpus (48.18 per letter) is lesser than its counterpart in English corpus (50.66 per letter), which means that the Egyptian aphorism in English language is longer than its counterpart in Arabic language.
- The number of tokens that used in expressing an Egyptian aphorism in Arabic language is more than the number of tokens which used in expressing its English counterpart. Also, the mean number of tokens of an Arabic Egyptian aphorism (mean number of tokens = 10.24) is greater than in an English Egyptian aphorism (mean number of tokens = 9.9). See table 5
- Minimum and maximum numbers of tokens for both corpora are quite the same, whereas the minimum number of tokens is 4 tokens for booth corpora, and the maximum number of tokens is 22 for Arabic corpus & 21 for English corpus. Although, the English corpus contains larger number of sentences that have less number of tokens (5 to 10 tokens per sentence); whereas, the Arabic corpus contains larger number of sentences that have larger number of tokens (10 to 15 tokens per sentence). See figure 2.
- The mean of words length in Arabic corpus (3.794922) is lesser than in English corpus (4.179798), wherein, the range of Arabic words in Arabic corpus is from 1 to 8 words, and in English corpus is from 1 to 14. Although the median length of words is the same in both corpora (which calculate 4 letters in both corpora).
- According to the animacy of words; the Arabic corpus has a larger set of animate words (254) than English corpus (184), and also in inanimate words (Arabic corpus has 117 inanimate words and English corpus has only 52 inanimate words). Whereas English corpus has huge number of null words (259) than in Arabic corpus (141). See table 7 and figure 2.
- In respect to gender; Arabic corpus has a huge number of masculine words (292) than in English corpus (179). And also, Arabic corpus has a larger number of feminine words (72) than English corpus (51). Whereas, English corpus has a larger set of null words (265) than Arabic corpus (148). See table 7 and figure 2.
- According to the most general frequent words in both corpora of Egyptian aphorisms, in Arabic corpus, the words (ان, و, لا, من ; respectively from right to left) are the most frequent words. Whereas, in English corpus the words (the, is, you, of; respectively from left to right) are the most frequent words.
- Regarding the tag set of Egyptian aphorisms of both corpora, NN (Noun) are the most frequent tag for both corpora followed by DT (Determiner), followed by Verbs and Prepositions.

All the preceding results indicate that there is a quite difference between the Arabic corpus and the English corpus of Egyptian aphorisms; especially in the length of the sentence of the aphorism and the number of tokens of a given aphorism.

BIOGRAPHY

Eman El-Gamal: *Ph.D. researcher, Dept. of Phonetics and linguistics, Faculty of Arts, Alexandria University.*

She graduated from the department of phonetics and linguistics in 2009. Her Graduation Project is entitled as "Forensic Phonetics – Voice Print Analysis"; she got an excellent degree for that research. Master Degree of Arts in Phonetic Science, Department of Phonetics and Linguistics in June 2015. The title of the thesis is "Speaker Identification Based on Temporal Parameters". It was published at The Fifteenth Conference of Language Engineering that held on A'ain Shams University at Cairo in December 2015. Since January 2016, I'm a PHD researcher in Phonetics and Linguistics Department, Faculty of Arts, Alexandria University.

Dr. Sameh Alansary: Professor *of Computational Linguistic, Phonetics and linguistics, Faculty of Arts, Alexandria University.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now. Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

REFERENCES

[1] Pawlowski, (1999). The quantitative approach in cultural anthropology: Application of linguistic corpora in the analysis of basic colour terms. Journal of Quantitative Linguistics 6(3), 222-234.

[2] Santorini, "Part-of-Speech Tagging Guidelines for the Penn Treebank Project". The University of Pennsylvania, 1990. https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

[3] G. Herdan, "The Relation between The Dictionary Distribution And The Occurrence Distribution Of Word Length And Its Importance For The Study Of Quantitative Linguistics" 1958. Biometrika, Volume 45, Issue 1-2, 1 June 1958, Pages 222–228, https://doi.org/10.1093/biomet/45.1-2.222

[4] G. Wimmer, G. Altmann, (19996). The theory of word length: some results and generalizations. Glottometrika 15, 112-133

[5] H. F. Wolcott, "Transforming Qualitative Data: Description, Analysis, and Interpretation" https://books.google.com.eg/books?hl=en&lr=&id=BMqxX_TaWNEC&oi=fnd&pg=PA1&dq=Analyzing+Linguistic+Data+biography&ots=1Cn0gQJMuy&sig=6JcY5-JaDLrVKCQyPpke5DFzD-I&redir_esc=y#v=onepage&q&f=false

[6] I.-I. Popescu, G. Altmann, (2007a). On the diversity of word frequencies and language typology. Göttinger Beiträge zur Sprachwissenschaft 14, 81-91.

[7] I.-I. Popescu, G. Altmann, (2008a). Hapax legomena and language typology.Journal of Quantitative Linguistics 15(4), 370-378.

[8] I.-I. Popescu, J. Mačutek, G. Altmann, Aspects of word frequencies. 2009, IV + 198 pp. http://library2.nipne.ro/sites/default/files/iovitzubook2-aspects_of_word_frequencies-july_2009.pdf

[9] J. R. Hurford, B. Heasley, and M. B. Smith , "Semantics: A Course book". Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK, second edition, 2007.

[10] K. Johnson, "Quantitative Methods In Linguistics", First published 2008 by Blackwell Publishing Ltd 1 2008 Library of Congress Cataloging-in-Publication Data Johnson, Keith, 1958. https://books.google.com.eg/books?hl=en&lr=&id=uF1oLql0-0wC&oi=fnd&pg=PT8&dq=quantitative+linguistics+references&ots=Uuihk_puG9&sig=jHmip934pDLETTB972WL7UGHn4&redir_esc=y#v=onepage&q=quantitative%20linguistics%20references&f=false

[11] M. A. Montemurro, "Physica A: Statistical Mechanics and its Applications", Volume 300, Issues 3–4, 15 November 2001, Pages 567-578. http://www.sciencedirect.com/science/article/pii/S0378437101003557

[12] M. Stubbs, "Text And Corpus Analysis: Computer-Assisted Studies Of Language And Culture". This is chapter 1 of Text and Corpus Analysis which was published by Blackwell in 1996. © Copyright Michael Stubbs 1996. https://www.uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-1996-text-corpus-ch-1.pdf

[13] R. H. Baayen "Analyzing Linguistic Data: A Practical Introduction to Statistics using R", Cambridge University Press, baayen@mpi.nl. www.sfs.uni-tuebingen.de/~hbaayen/publications/baayenCUPstats. pdf , 2008.

[14] U. Strauss, F. Fan, G. Altmann, Problems in quantitative linguistics 1. 2008, VIII +134 pp.

Online Resources:

[1] http://anabeebsuz.hooxs.com/t162-topic

[2] http://johnvictoranderson.org/?p=115

[3] http://nlp.stanford.edu:8080/parser/

[4] http://www.almstba.net/t17244.html

[5] http://www.goldenproverbs.com/tp_egyptian.html

[6] http://www.sal.tohoku.ac.jp/ling/corpus3/3qual.htm

[7] https://cran.r-project.org/package=languageR

[8] https://en.oxforddictionaries.com/definition/quantitative_linguistics

[9] https://explorable.com/quantitative-research-design

[10] https://www.egyptabout.com/2012/11/10-popular-ancient-egyptian-sayings.html

[11] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

TRANSLATED ABSTRACT

دراسة تحليلية مقارنة للحكم المصرية المترجمة للإنجليزية باستخدام لغة الآر

علم اللغة الكمي هو دراسة مقارنة لأعداد وتوزيع الكلمات والتراكيب اللغوية في النصوص المختلفة. هدف هذه الدراسة هو وصف تحليلي كامل ومفصل للمدخلات اللغوية من الحكم المصرية باللغة العربية وبين ما يقابلها باللغة الإنجليزية, وذلك لإكتشاف بعض الظواهر اللغوية المختلفة التي يمكن استخراجها من الحكم بين اللغتين وذلك باستخدام لغة التحليل الإحصائي الآر التحليلية. المدخلات اللغوية من الحكم المصرية باللغة العربية وبين ما يقابلها باللغة الإنجليزية تم تحليليها على المستوى المورفولوجي والدلالي. تشير النتائج إلى أن طول الجملة التركيبية في الحكم المصرية باللغة الإنجليزية أكثر طولا من ما يقابلها باللغة العربية من حيث عدد الكلمات ومتوسط طول الكلمة. وبالنسبة لأكثر الكلمات حدوثا في اللغة العربية هي (من, لا, و, إن) وفي اللغة الإنجليزية (the, is, you, of). وأما بالنسبة أنواع الكلمات (tag set) فكان الأكثر حدوثا هم الاسماء (Noun,NN) ويليها أدوات التعريف (Determiner, DT) ويليها الأفعال (Verbs) وحروف الجر (Prepositions) وذلك بالنسبة للحكم في اللغتين العربية في الإنجليزية على حد سواء.

Recent Advances in Speech and Speaker Recognition Using Deep Learning Techniques

Mohammed Affifi

Microsoft Advanced Technology Lab Cairo, Egypt

mafify@microsoft.com

Abstract: **Deep learning techniques have resulted in impressive performance improvements for large vocabulary continuous speech recognition (CSR) compared to Gaussian mixture models that dominated the field for almost two decades. Recently researchers from different laboratories claim to achieve human parity for conversational telephony speech in English. During the past several years there have been a lot of innovations for deep learning architectures and training methods for deep learning-based CSR. Architectures include, feed forward networks, convolutional networks, recurrent networks and many variants and combinations of these models. We will review some of the most prominent architectures and show how some of these are combined to achieve human parity on some challenging tasks. While, deep learning methods show robustness to varying conditions, adaptation is also a key in improving performance for varying acoustic conditions. We quickly review some of the adaptation techniques used for deep learning. Interestingly, in addition to their applications to acoustic modeling, deep learning techniques show a lot of promise for language modeling. We briefly overview the application of recurrent networks, in particular LSTM, for improving CSR performance. While deep learning techniques are mostly language agnostic, we quickly review recent advances in Arabic CSR especially for the MGB challenge. Some interesting results have been obtained by different techniques including improved architectures, data augmentation, adaptation and LSTM language models just to name a few.**

Speech and speaker recognition have been mainly addressed by seemingly different techniques e.g. i-vectors for speaker recognition. Interestingly, with deep learning both fields can use very similar architectures. We briefly review recent text-dependent and text-independent speaker recognition results using deep learning techniques. We also briefly address how both architectures could be used to improve the performance of each individual system.

Arabic Dialect Identification from Dialects Motifs using UBM-GMM

Mohsen Moftah[*1], Waleed Fakhr[*2], Salwa El Ramly[*3]

[*1]*Electronics &Communications Engineering Department, Faculty of Engineering, Ain Shams University*
Cairo, Egypt

[*2] *College of Computing, Arab Academy for Science and Technology*
Ahmed Ismail street, Heliopolis, Cairo, Egypt

[1]mohsen.moftah@barmagyat.com

[2]waleedf@aast.edu

[3]salwahelramly@gmail.com

***Abstract*— Over the last decade, Arabic Dialect Identification DID has attracted the attention of many researchers. The first step in DID/LID is analyzing and extracting the characteristics of the target dialect/language. In this paper we introduce a new technique for extracting the features and characteristics of different Arabic dialects direct from the speech signal by discovering the repeated sequences that characterize each dialect. These repeated sequences are called motifs. We adopted an extremely fast parameter-free Self Join based motif discovery algorithm called STOMP to overcome the computation time barrier. Using motif discovery directly from the speech audio signal eliminates the intermediate step of having an ASR or Phone analyzer to extract dialect/language characteristics. Our approach is based on extracting MFCC features from discovered motifs. For classification we applied the widely used techniques GMM-UBM approach. We applied our new approach on the two most common Arabic dialects; the Egyptian (EGY) and Levantine (LEV). The data set was downloaded from Qatar-Computing-Research- Institute domain for free. The test data is a high quality audio from Aljazeera channel covering the period of July 2104 until January 2015.**

Keywords: motif discovery, dialect identification, language identification.

1 INTRODUCTION

The main Arabic dialects can be classified as: Egyptian, Gulf, Levantine, and North Africa. Automatic Dialect Identification (DID) is a special case of the more general task which is Automatic Language Identification (LID). LID became a mature technology and has various applications [1]. An Arabic DID system is required to automatically identify the dialect of the input speech; this is a challenging task since there are no solid boundaries between different Arabic dialects. As mentioned above, DID is a special case of LID, therefore, we can apply the same techniques used in LID to establish an Arabic DID system. Most LID systems, and therefore DID systems, operate in two phases, a training phase and recognition phase. In training phase, the system is trained using examples of every target dialect. This training data can be as simple as the digitized speech utterances mapped to the corresponding spoken language. More sophisticated system may require more data such as phonetic transcription in a form of sequence of symbols of the spoken sounds, and an orthographic transcription of the spoken words. From the training speech, fundamental characteristics of each language are analyzed to produce language-dependent models. The second phase, recognition, makes use of the language-dependent models produced in the training phase to identify new unknown utterances [2]. Based on the type of dialect features extraction and modeling, DID approaches can be divided into two main classes, a high level lexical and phonetic features approach such as Phone Recognition followed by Language Modeling (PRLM) and Parallel Phone Recognition followed by Language Modeling (PPRLM), and low level acoustic features concerned with spectral characteristics of speech such as Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) as acoustic front end and Gaussian Mixture Model (GMM), Universal Background Model-Gaussian Mixture Model (UBM-GMM) as acoustic backend [3] [4]. In this paper we are presenting a new approach based on the acoustic features of the spoken dialect. This approach is based on first discovering the repeated sequences/patterns i.e. motifs, of the speech signal directly, and then extract the MFCC features of the motifs. To examine the new approach we selected the well known UBM-GMM method for modeling and classification. The reset of this paper is organized as follows: section 2 will present a brief description of the most popular DID/LID approaches; section 3 will discuss the motif discovery approach. Section 4 will be dedicated to explain our proposed approach; section 5 will show the experiments results, while the last section 6 will be a conclusion and future work.

2 DID/LID APPROACHES

A. High Level Lexical and Phonetic Features Approach

1) PRLM Approach

In Phone Recognition followed by Language Modeling (PRLM) approach, a phone recognizer is used to tokenize the training dataset of the target dialects to produce phone sequences. The phone sequences are used to train a statistical language model to generate phonotactic language model for the dialects in question. These phonotactic language models are used to compute the dialect likelihood for the unknown utterances [5][6][7].

2) PPRLM Approach

In Parallel Phone Recognition followed by Language Modeling (PPRLM), phonotactic statistics of a language are extracted using multiple phone recognizers. Every phone recognizer is trained on different languages to capture acoustic characteristics of each language. The recognizers are combined to form a parallel recognizer PPR to characterize the spoken language [4].

B. Acoustic Approach

The implementation of the acoustic approach is comprised of two phases, a feature extraction phase, followed by a classification phase [8] [3][9]. The most popular features used in this phase are:

1) Mel Frequency Cepstral Coefficients (MFCC): Frequency domain features characterized by their robustness and reliability to variations of speakers and recording conditions.

2) Shifted Delta Cepstral coefficients (SDC): SDC is a stack of delta spectra computed across multiple speech frames.

3) Relative Spectra Filtering (RASTA): Filtering of cepstral trajectories is used to remove slowly varying, linear channel effects from raw feature vectors.

The second phase in acoustic based approach is the classification phase. The following are the most popular classifiers applied in DID/LID. These classifiers are used successfully in speaker recognition:

1) *GMM-UBM*: GMM is extensively used in speaker recognition. In GMM-UBM approach the first step is to create a Universal Background Model (UBM) by training the GMM with a large amount of data using iterative Expectation Maximization (EM) algorithm to maximize the likelihood of the GMM. To create a speaker specific model, GMM parameters; the mixture weight, mean vector, and covariance matrix are adapted to specific speaker using Maximum a Posteriori (MAP) scheme. During the adaptation process, parameters for the Gaussian mixtures which bear a high probabilistic resemblance to the language specific training data will tend towards the parameters of that training data whereas the parameters of the Gaussian mixtures bearing little resemblance to the language specific data will remain fairly close to their original UBM values[4].
2) *GMM-SVM:* Support Vector Machines (SVM) became as popular as GMM, it uses a linear kernel in a supervector space for rapid computation of language distance. The kernel computes the distance between two supervectors one represents the GMM model and the other represents the target language [4].
3) *i-vector*: Dehak [10] developed a new classifier by finding a low dimensional subspace from the GMM super-vector space based on Joint Factor Analysis (JFA) as feature extractor. The low dimensional subspace is called total variability space since it includes both speaker and channel variations. The dimensionality of the low-dimensional space is reduced using Linear Discriminant Analysis (LDA). The vectors in the low-dimensional space are called i-vectors, which are of small size compared with those in GMM super-vector to reduce execution time while keeping the recognition rate acceptable.

3 MOTIF DISCOVERY

Motif discovery has been applied in many applications such as summarizing and visualizing massive time series databases, in addition to various data mining tasks, including the discovery of association rules. Figure 1, shows an example of motifs discovered in a time series [11].

One common approach of Motif discovery applies similarity search approach which depends on similarity threshold, a value that is difficult to determine [12]. Another approach called All-Pairs-Similarity-Search, Similarity Join, or Self Join approach. A brief explanation of this approach is introduced in the following paragraphs showing how to apply it on speech signals.

In speech, a speech audio signal can be easily considered a time series. As will be explained, a time series is defined as a sequence of real-valued numbers, in digital audio these valued numbers are the audio sample values. A motif in a speech time series can represent repeated words or sub words. The following is background on motif discovery in speech and a brief explanation of the self-join algorithm used as the base of our approach in Arabic DID [13].

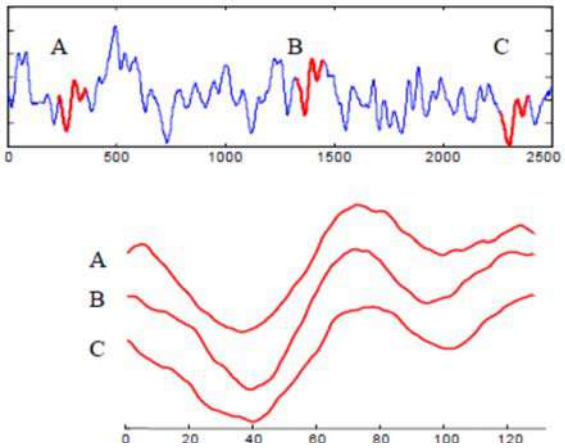


Figure 1: An astronomical time series (above) contains 3 near identical subsequences. A "zoom-in" (below) reveals just how similar to each other the 3 subsequences are.

Definition 1: A time series T is a sequence of real-valued numbers t_i: $T = t_1, t_2, \ldots, t_n$ where n is the length of T.
A local region of time series is called a *subsequence*:

Definition 2: A *subsequence* $T_{i,m}$ of a time series T is a continuous subset of the values from T of length m starting from position i. Formally, $T_{i,m} = t_i, t_{i+1}, \ldots, t_{i+m-1}$, where $1 \leq i \leq n-m+1$.

If we compute the distance of a subsequence to *all* subsequences in the same time series; we come up with a *distance profile*:

Definition 3: A *distance profile* D_i of time series T is a vector of the Euclidean distances between a given query subsequence $T_{i,m}$ and each subsequence in time series T. Formally, $D_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,n-m+1}]$, where $d_{i,j}$ $(1 \leq i, j \leq n-m+1)$ is the distance between $T_{i,m}$ and $T_{j,m}$ where the distance is measured by Euclidean distance between z-normalized subsequences. Equation 1 shows how to calculate distance between two z-normalized subsequences. A z-normalized subsequence has a mean value of zero and standard deviation value of one [14].

$$d_{i,j} = \sqrt{2m - \left(\frac{QT_{i,j} - m\mu_i m\mu_j}{m\sigma_i\sigma_j}\right)} \qquad (1)$$

where m is the subsequence length, μ_i is the mean of $T_{i,m}$, μ_j is the mean of $T_{j,\mathrm{m}}$, σ_i is the standard deviation of $T_{i,m}$, and σ_j is the standard deviation of $T_{j,\mathrm{m}}$, $QT_{i,j}$ is the dot product of $T_{i,m}$ and $T_{j,\mathrm{m}}$.
The mean can be calculated by [14]

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (2)$$

and the standard deviation can be calculated by [14]

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 \; - \; \mu^2 \qquad (3)$$

We use a vector called matrix profile to represent the distances between all subsequences and their nearest neighbors:

Definition 4: A *matrix profile* P of time series T is a vector of the Euclidean distances between each subsequence $T_{i,m}$ and its nearest neighbor (closest match) in time series T.
Formally, $P = [\min(D_1), \min(D_2), \ldots, \min(D_{n-m+1})]$, where D_i $(1 \leq i \leq n-m+1)$ is the distance profile D_i of time series T Figure 2 .

The i^{th} element in the matrix profile P tells us the Euclidean distance from subsequence $T_{i,m}$ to its nearest neighbor in time series T. However, it does not tell us *where* that neighbor is located. This information is recorded in a companion data structure called the *matrix profile index*.

Definition 5: A *matrix profile index* I of time series T is a vector of integers: $I=[I_1, I_2, \ldots I_{n-m+1}]$, where $I_i=j$ if $d_{i,j} = \min(D_i)$.

We can use the *matrix profile P* and the *distance profile D* to extend the notion of motifs to sets of subsequences that are very similar to each other [15].

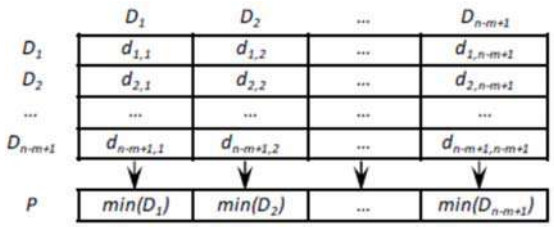


Figure 2: An illustration of the relationship between the distance profile, the matrix profile and the full distance matrix.

Definition 6: The *Range motif* with range r is the maximal set of subsequences that have the property that the maximum distance between them is less than $2r$. More formally S is a *range motif* with range r iff $\forall T_x, T_y \in S$, $\text{dist}(T_x, T_y) \leq 2r$ and $\forall T_d \in D\text{-}S$ $\text{dist}(T_d, T_y) > 2r$.

Figure 3 shows a flowchart of the motif discovery algorithm.

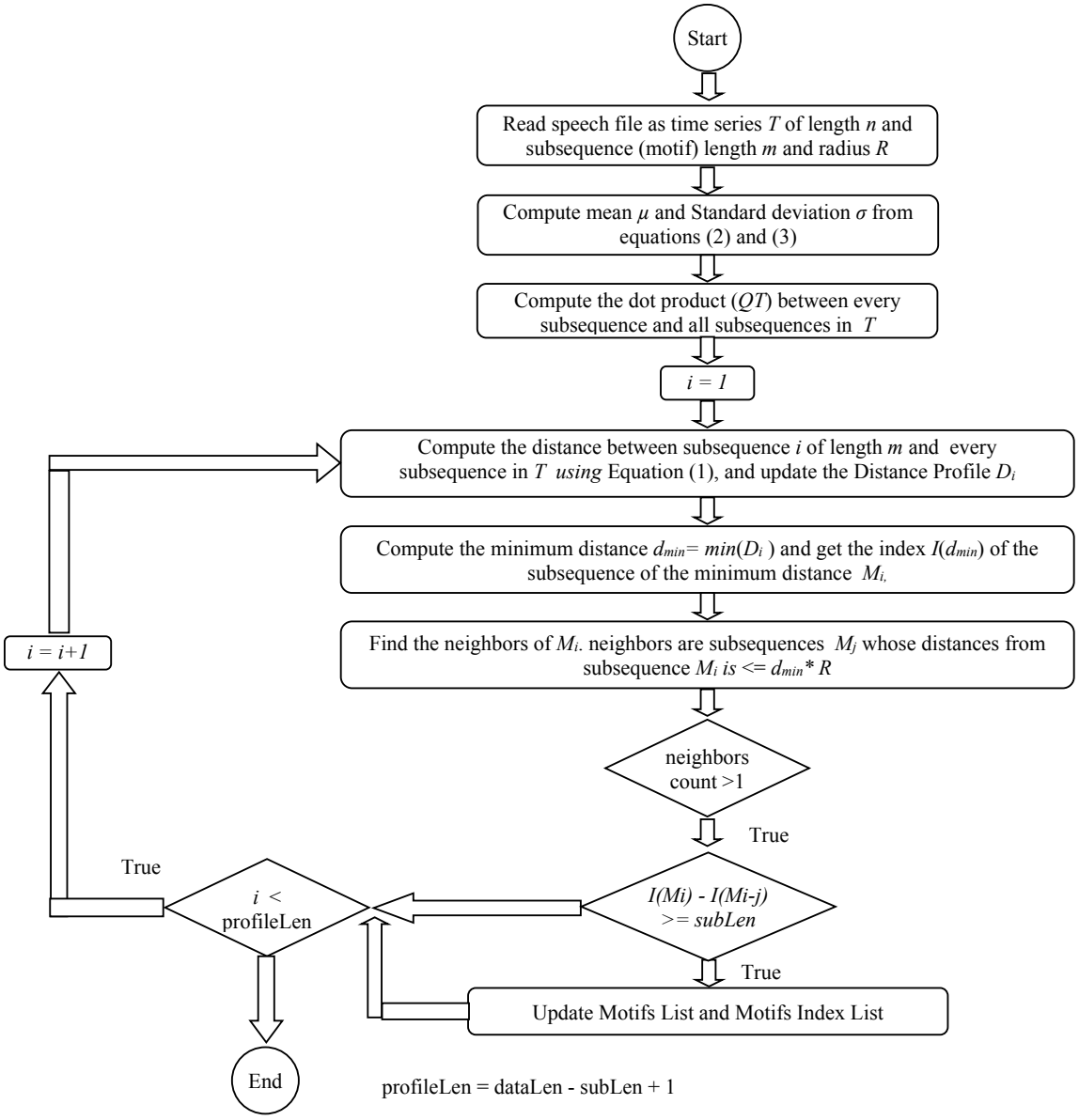


Figure 3: Motif Discovery Algorithm

4 PROPOSED APPROACH

In DID and other speech processing tasks, the traditional approach always starts by extracting speech features from complete utterances, in our new proposed approach we first extract motifs from speech utterances. Motif discovery was originally applied to time series; a time series is defined as an ordered set of real-valued variables. Since digitized audio/speech is a set of samples with real values, speech and audio signals can be considered as time series. Discovery of frequently occurring patterns (motifs) is very important in knowledge discovery [11]. If we apply this concept to dialects or languages, motif discovery can be used to characterize a dialect/language by extracting the frequently repeated patterns; these patterns may be parts of words, words, or even a group of words depending on the length of the motif. In our approach, the acoustic features are extracted from motifs rather than from the complete utterance.

The realization of the proposed approach comprises many steps. The training corpus was downloaded from Qatar-Computing-Research-Institute domain for free; we selected to apply our new approach in the most common Arabic dialects; the Egyptian (EGY) and the Levantine (LEV). It consists of audio files at sampling rate 16k. Table I and Table II show the statistics of the data used in experiments.

TABLE I

INPUT DATASET

Dialect	Train			Test		
	Hours	Utterances	motifs	Hours	Utterances	motifs
EGY	9.7	3053	4917	0.7	132	253
LEV	8.7	3040	4814	0.8	240	435

TABLE III

INPUT DATA USAGE DISTRIBUTION IN HOURS, UTTERANCES, AND MOTIFS

Dialect	training			testing
	motif count	UBM-GMM		motif count
		60% UBM	40% enrollment	
EGY	4917	2950	1967	100
LEV	4814	2888	1926	100

Tools used were Microsoft MSR Identity Toolbox v1.0: A MATLAB Toolbox, that consists of functions needed for training and scoring for UBM-GMM system and VOICEBOX: Speech Processing Toolbox for MATLAB provides a rich library for handling speech and audio signals such as computing the MFCC coefficients. The implementation consists of the following phases:

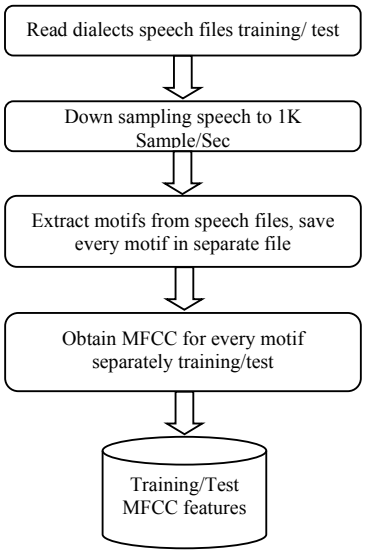


Figure 4: Preprocessing Phase Steps

A. Preprocessing phase

Figure 4 shows the preprocessing phase for the implementation of UBM-GMM. In step one, speech files training/test of every dialect are read. In step two, speech utterances are down sampled to 1K sample/Sec, the purpose of this operation is to represent the utterance with the outer envelope of the signal which is more representative for motif extraction, in addition to

reduce processing time knowing that the time complexity of the motif extraction algorithm is $O(n^2)$ where n is the length of the speech signal. Motifs of length 1second are extracted from speech files using STOMP algorithm [13] as explained in section 3. Every motif was saved in a separate audio file. In step three, MFCC coefficients are calculated for every motif and saved in a separate file.

B. UBM-GMM Implementation

At this stage we are dealing with MFCC features of discovered motifs, therefore the data used will be in terms of number of motifs rather than audio duration or number of utterances. Training MFCC Features created in the reprocessing phase are divided into two parts, 60% for development, and 40% for enrollment.

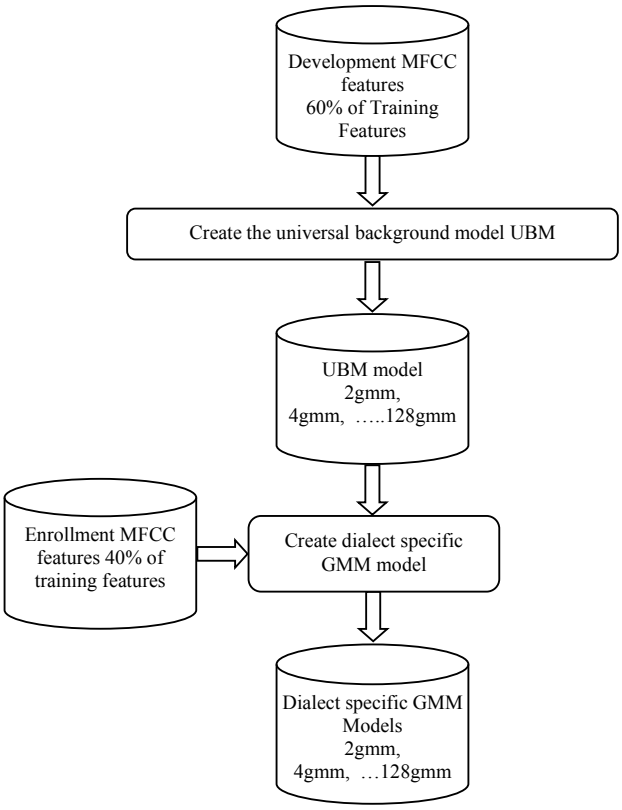


Figure 5: UBM-GMM Training Phase

Training Phase: Figure 5, shows the steps for Training Phase of the UBM-GMM approach using MSR toolbox. In Training Phase, the development part (60% of the training data) is used to create the Universal background Model UBM, for our experiments, we created UBM models for 2gmm, 4gmm, …, 128gmm. The UBM created along with the Enrollment MFCC features (40% of the training data) are used to create dialect specific models using *maximum a posteriori* estimation MAP for 2gmm, 4gmm, …,128gmm models.

Test Phase: Figure 6 shows the steps of the Test Phase. The resulting dialect specific models are tested using the test data to compare results. The scores are computed as the log-likelihood ratio between the given dialects models and the UBM given the test observations.

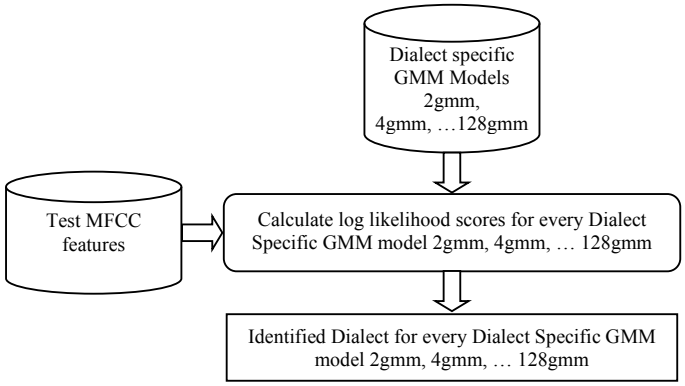


Figure 6: UBM-GMM Test Phase

TABLE III

RESULTS SUMMARY FOR UBM-GMM

classifier		Correct Classification %		Accuracy %
		EGY	LEV	
UBM-GMM	**2gmm**	**62**	**65**	**63.5**
	4gmm	60	61	60.5
	8gmm	64	60	62
	16gmm	58	60	59
	32gmm	51	57	54
	64gmm	52	52	52
	128gmm	47	51	49

5 EXPERIMENTS RESULTS

We started our experiments with 32gmm and increased the number of gmms up to 128gmm; we noticed that the accuracy decreases with the increase of the number of Gaussian Mixtures gmm, so we decided to check the effect of reducing the number of gmms, and we carried out the experiments for 16gmm, 8gmm, 4gmm, and 2gmm, the accuracy reached a maximum of 63.5% at 2gmm. Table III shows correct classification results using GMM-UBM implementation with different number of Gaussian Mixtures.

6 CONCLUSION AND FUTURE WORK

The proposed approach is a good candidate to be a new technique in Dialect Identification field. In our experiments we intentionally used simple implementation for the sake of proof of concept. Only MFCC was used as features and one classifier UM-GMM was used. The results are encouraging to do more sophisticated experiments like using delta and shifted delta Coefficients, in addition to using other classification approaches such as Support Vector Machine (SVM), i-vector, and neural networks.

REFERENCES

[1] Santhi.S , Raja Sekar, "An Automatic Language Identification Using Audio Features", *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Special Issue 1, pp 358-364, January 2013

[2] Marc A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE *Transactions On Speech And Audio Processing,* VOL. 4, NO. 1, January 1996.

[3] Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell5, Steve Renals, "Automatic Dialect Detection in Arabic Broadcast Speech" in *INTERSPEECH,* pp 2943-2938, San Francisco, USA, September 8–12, 2016,

[4] Fadi Biadsy, Julia Hirschberg, and Nizar Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling" in *Workshop on Computational Approaches to Semitic Languages*, pp 53–61, Athens, Greece, 31 March, 2009.

[5] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu, "Language Identification A Tutorial", *IEEE Circuits And Systems Magazine* Second Quarter 2011, pp 82-108 vol?, no.?

[6] Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, "Prosody-based Spoken Algerian Arabic Dialect Identification" in *International Conference on Natural Language and Speech Processing*, Algiers, Algeria 2015.

[7] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models", in *Proc. Interspeech*, 2011, pp. 141--144.

[8] Kshirod Sarmah and Utpal Bhattacharjee, "GMM based Language Identification using MFCC and SDC Features", *International Journal of Computer Applications (0975 – 8887)* Volume 85 – No 5, pp 36-42, January 2014

[9] Rania R Ziedan · Michael Nasief, · Abdulwahab K. Alsammak · Mona F. M. Mursi · Adel S. Elmaghraby , "A Unified Approach for Arabic Language Dialect Detection". *29th International Conference on Computer Applications in Industry and Engineering (CAINE 2016)*, pp 165-170, Denver, USA, September 26-28, 2016

[10] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Br¨ummer, Pierre Ouellet, and Pierre Dumouchel. "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification". *In Interspeech*, volume 9, pages 1559– 1562, 2009.Place?, date?

[11] Jessica Lin, Eamonn Keogh, Stefano Lonardi, Pranav Patel, "Finding Motifs in Time Series" *Proceedings of the Second Workshop on Temporal Data Mining at the 8th SIGKDD, pp 53-68*, Edmonton, Alberta, Canada — July 23 - 26, 2002

[12] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh, "Matrix Profile I: All Pairs Similarity Joins for Time Series:A Unifying View that Includes Motifs, Discords and Shapelets". *IEEE International Conference on Data Mining IEEE ICDM 2016*, Pp 1317-1326, Barcelona, Spain 2016.

[13]- Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh, "Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins', *IEEE International Conference on Data Mining IEEE ICDM 2016*, pp 739-748, Barcelona, Spain 2016.

[14] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, Eamonn Keogh, "Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping" *Conference: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp 262-270, Beijing, China — August 12 - 16, 2012

[15] Abdullah Mueen , Eamonn Keogh , Qiang Zhu , Sydney Cash, "Exact Discovery of Time Series Motifs", *Conference: Proceedings of the SIAM International Conference on Data Mining, SDM 2009*, pp 473-484 Sparks, Nevada, USA, April 30 - May 2, 2009

الخلاصة - على مدى العقد الماضي استرعى التعرف الآلي للهجات العربية انتباه كثير من الباحثين. و يعتبر البحث عن خصائص و مميزات اللهجة خطوة رئيسة في اي نظام للتعرف على اللهجات المختلفة. في هذا البحث نقدم تقنية جديدة للتعرف على اللهجات من أشارة الصوت مباشرة تقوم على إستخراج الانماط المتكررة للهجات موضوع البحث. و قد استخدمنا في هذا البحث خوارزم فائق السرعة يدعى STOMP حتى نستطيع اختصار وقت التشغيل, كما يتميز هذا الخوارزم بأنه لا يحتاج الى إعدادات مسبقة. أن التعرف على الانماط المتكررة مباشرة من اشارة الصوت يغني عن وجود خطوة وسيطة مثل التعرف الآلي على الكلام ASR او التحليل الآلي لاصوات الكلام Phone Analyzer. في عملية التصنيف classification فقد استخدمنا أحد أشهر الاساليب وهو GMM-UBM. اما بالنسبة لبيانات التشغيل فقد استخدمنا بيانات معهد قطر لبحوث الحوسبة و هي عبارة عن تسجيلات عالية الجودة من قناة الجزيرة تغطي الفترة من يوليو 2014 الى يناير 2015 و تشمل اللهجات العربية الرئيسة و هي المصرية و الشامية و الخليجية و شمال إفريقيا بالإضافة إلى اللغة العربية القياسية. لقد أخترنا اللهجتين الاكثر انتشارا و هما اللهجتين المصرية و الشامية لإختبار التقنية الجديدة.

BIOGRAPHY

Mohsen Moftah holds a M.Sc. degree in Communications and Electronics Engineering, now perusing PhD in the same discipline. In the academic side, his research interest is Arabic language engineering, and participated in local as well as international conferences covering this field. He also has papers published in those conferences. In the industrial side, ha has more than 30 years of experience in the IT arena. His experience covers Application Development, Technical Support, and Operations Management, Projects Delivery involving multi-party in addition to deep exposure to many technologies such as ERP, Hospital Management Systems, And Content Management Systems. And other technologies such as wired/wireless networking, Access Control and Time Attendance using RFID technology. In addition to security applications, using Video Analytics based surveillance.

Dr. Fakhr finished his Ph.D. at the University of Waterloo, Canada, 1993, in the field of neural networks and machine learning; he then joined the speech research lab at NORTEL, Montreal, Canada, for 5 years where he was a researcher investigating and implementing different speech processing, speech recognition, language modeling, and statistical error analysis techniques and has 2 patents with NORTEL. Since 1999 he has been a professor with the Arab academy for science and technology (Cairo, Egypt) with 3 years sabbatical at the University of Bahrain. He has been doing research in the areas of Multimedia processing, Arabic Language Processing, Printed and handwritten character recognition, Statistical machine translation, Language modeling, Neural networks and Sparse coding.

Prof. Salwa Elramly, BSc. Degree 1967, MSc. Degree 1972 from Faculty of Engineering, Ain Shams University, Egypt & PhD degree 1976 from Nancy University, France. She is now professor Emeritus with the Electronics and Communications Engineering Department, Faculty of Engineering, Ain Shams University; where she was the Head of the Department (2004-2006). Her research field of interest is Wireless Communication Systems and Signal Processing, Language Engineering, Coding, Encryption, and Radars. She is a Senior Member of IEEE and Signal Processing Chapter chair in Egypt. She was awarded Ain Shams Award of Appreciation in Engineering Sciences (2010), Award of Excellence from the Society of Communications Engineers (2009) & Award of Excellence from the Egyptian Society of Language Engineering.

Effect of Reducing the Number of Linear Predictive Coefficients on the Voice Quality of the CELP Vocoder using the Arabic Words

Nayra Abd Elhalim*[1], Noha Korany*[2], Onsy Abdel Alim*[3]

**Electrical Department, Faculty of Engineering and Alexandria University, Egypt*

[1]eng_nayra158@outlook.com

[2]nokorany@hotmail.com

[3]onsy2066@hotmail.com

***Abstract*—The aim of all speech coding is to reach the best quality with the least bandwidth. At bit rates above 4 kbps, speech-specific hybrid coders based on code excited linear prediction (CELP) can produce good quality speech. But at bit rate less than 4 kbps becomes very difficult. CELP is the quite efficient closed loop analysis-by-synthesis method.**

In this paper, (CELP) is implemented using Matlab. The paper aims to find the minimum number of linear predictive coefficients that yields to good codec quality and hence the bit rate is reduced. Arabic words are used. The quality of the codec is influenced by the language used. More compression with language or accents other than English leads to less quality. This coding technique is analysed on the basis of subjective and objective tests. Intelligibility, Mean Opinion Score (MOS) and Signal-to-Noise Ratio (SNR) are employed.

***Keywords*: Reducing number of linear predictive coefficients, CELP Implementation, Voice Quality using Arabic words, analysis-by-synthesis**

1 INTRODUCTION

Speech coding is an important aspect of modern telecommunications. Speech coding is the process of digitally representing a speech signal. The primary objective of speech coding is to represent the speech signal with the fewest number of bits, while maintaining a sufficient level of quality of the retrieved or synthesized speech with reasonable computational complexity. To achieve high quality speech at a low bit rate, coding algorithms apply sophisticated methods to reduce the redundancies, that is, to remove the irrelevant information from the speech signal. Naturalness and intelligibility of the produced sounds are important and desired criteria. The speech quality can be determined by listening tests which compute the mean opinion of the listeners. The bit rate of the encoder is the number of bits per second the encoder needs to transmit [1].

Linear Predictive Coding begins in the 1970s with the development of the first linear predictive coding algorithms. Linear predictive coding reduces it to2400 bits/second but there is a noticeable loss of quality. Extensions of Linear predictive coding such as Code Excited Linear Predictive (CELP) algorithms and Vector Selectable Excited Linear Predictive (VSELP) algorithms were developed in the mid1980s .This principle leads to acceptable speech quality in the rate range 4.8-16 kbps [2].

Various applications employ CELP algorithms. GSM uses 13 kbps speech data rate using CELP technique. CDMA uses various CELP codec at rates 8.55kbps, 9.6kbps, 13.3kbps. VOIP uses CELP algorithm. All these applications need a reduced bit rate with good quality. Also the effect of compression on the speech quality of Arabic words should be considered [3].

Most of these codec have been built for 7 languages not including the Arabic or its accents. It is known that the MOS of the English recorded speech is very slightly higher for most cases than either Arabic or Cairo accented Arabic ([4]-[6]). Also the quantization distortion is not uniform across languages and it influences codec codebook performance and overall quality of decoded speech, when the coders are used for heavily accented English or other languages significant performance degradation is noted. The compression with language or accents other than English is inversely proportional with the speech quality. The influence of the Arabic language on the performance of the codec has to be investigated. So, the paper employs Arabic words to find the minimum number of linear predictive coefficients that yield to good codec quality and hence the bit rate is reduced without significant degradation of the voice quality.

In this paper CELP codec of 9.6kbps is implemented. The speech quality is investigated for different number of LPC to find the minimum number yields to good quality using Arabic words. The paper is organized as follows next Section

describes CELP algorithm. Section (3) simulates the CELP using Matlab. Section (4) presents subjective and objective tests. Section (5) discusses the results. Section (6) concludes the paper.

2 CODE EXCITED LINEAR PREDICTIVE ANALYSIS

This section explains the algorithm using the flow chart.

A. CELP Encoder

Figure 1 shows the CELP encoder flow chart.

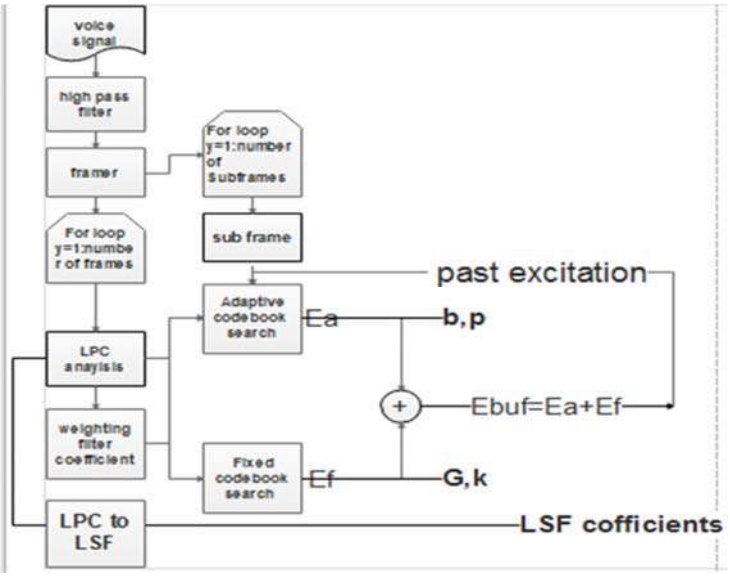


Figure 1: CELP encoder flow chart

The operation of CELP coders is described as follows:
The input voice to the handset's microphone is filtered by high pass filter to remove the DC component and sampled at a rate of 8000 samples per second then signal is segmented into frames and sub-frames. The duration of the frame is usually around 20ms to 30ms, while for the sub-frame it is in the range of 5 to 7.5ms [7]. Then speech frame is modeled by a linear prediction model [8]. The z-transfer function of the linear prediction filter is given by equation 1 [9], where p is the filter order and a_k is a set of linear predictive coefficients (LPC).

$$\mathbf{H(z)} = \frac{\mathbf{1}}{\mathbf{1} + \sum_{\mathbf{k=1}}^{\mathbf{p}} \mathbf{a_k z^{-k}}} \quad (1)$$

For stability and efficiency, LP filter coefficients are transformed into LSF then they are vector- quantized for transmission. The weighting filter coefficients are calculated from LPC. The z-transfer function of the perceptual weighting filter is given by equation 2 [10], where α is a parameter in the range $0 < \alpha < 1$ that is used to control the noise spectrum weighting. In practice, listening tests prove that α=0.85 is the best choice [11].

$$\mathbf{w(z)} = \frac{\mathbf{A(z)}}{\mathbf{A(z/\alpha)}} = \frac{\mathbf{1-\sum_{k=1}^{p} a_k z^{-k}}}{\mathbf{1-\sum_{k=1}^{p} a_k \alpha^k z^{-k}}} \quad (2)$$

CELP is based on the technique analysis by synthesis ([13], [14]) because the speech is encoded and then decoded the speech at the encoder to find the parameters that minimize the energy of the error signal. Adaptive and fixed codebook search are used to estimate the gain G, and the index of the fixed codebook k, the gain b and the pitch period P in samples of the adaptive codebook. P lies between Pmin=16 and Pmax=160 in samples.

To simplify the optimization process, the minimization of the energy of error is performed in two steps. First b and P are determined to minimize the error energy. Figure 2 shows us the minimization process for adaptive codebook search to estimate b, P. Second G, k are estimated Figure 3 shows us the minimization process for fixed codebook search to estimate G, k.

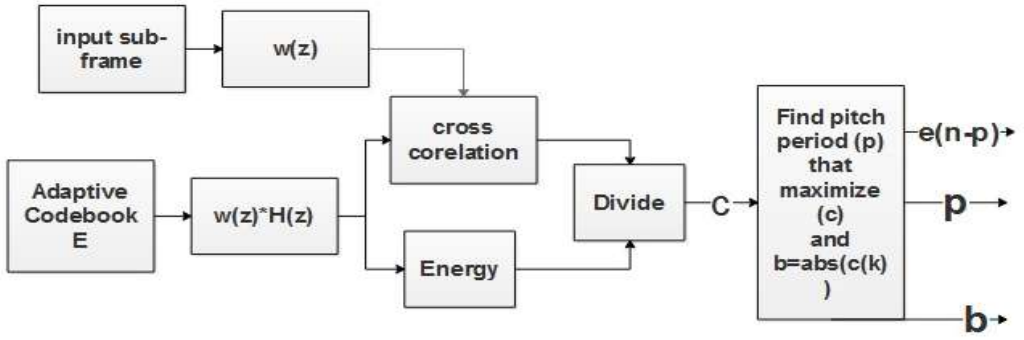


Figure 2: Adaptive codebook search

As shown in figure 2, the pitch prediction signal E is calculated from past excitation frame $Ebuf$ with initial zeros. The past excitation is updated each frame. The calculated pitch prediction signal E in adaptive codebook is convolved with W(z)*H(z). The convolution is calculated for each of pitch period. Each pitch period convolution is then correlated with the weighted filter input sub-frame. The optimum pitch period (P) maximizes C with positive gain (b) ([14] – [19]). The adaptive excitation e(n-P) is calculated from past excitation as shown in equation 3.

$$e(n-p)=b* Ebuf(n-p) \quad (3)$$

The fixed codebook search estimate G, k , figure 3 show us the minimization process for fixed codebook.

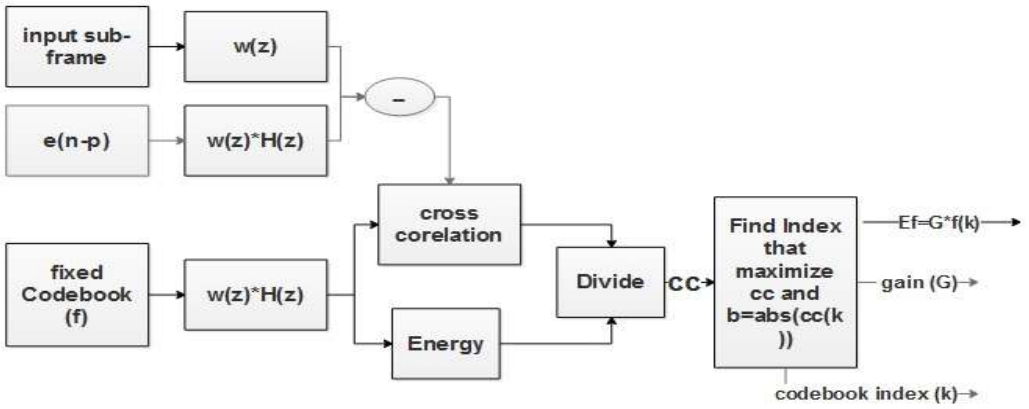


Figure 3: Fixed codebook search

As shown in figure 3, the fixed codebook (f) is Gaussian random noise codes. Each Gaussian code is convolved with W(z)*H(z). The adaptive excitation e(n-P) is convolved with W(z)*H(z), then subtracted from the weighted filter sub-frame. The resulted difference is correlated with each Gaussian code convolution. The optimum codeword maximizes CC with positive gain is chosen ([14] – [19]).

Finally, as shown in figure 1, the excitation ($Ebuf$) is calculated in the encoder and is saved as past excitation for the next frame. The excitation $Ebuf$(n) is produced by summing the contributions from an adaptive codebook and fixed codebook so that $Ebuf$=Ea+Ef where Ea=b* $Ebuf$(n-p) and the fixed codebook Ef=G*f(k) ([14] – [19]).

B. CELP Decoder

The decoder receives five parameters to extract the voice. The five parameters are LPC, the gain G and the index k of the fixed codebook, the gain b and the pitch period P of the adaptive codebook. Figure 4 shows the CELP decoder procedures. The excitation e(n) is produced by summing the contributions from an adaptive codebook and fixed codebook so that e(n)=e_a+e_f where $e_a = G * f(k)$ and $e_f = b * Ebuf2(n-P)$ where $Ebuf2$ is the past excitation saved in decoder and updated each frame and the excitation *e(n)* is filtered by the LP filter 1/*A*(z) and synthesized speech is reconstructed ([14] – [19]).

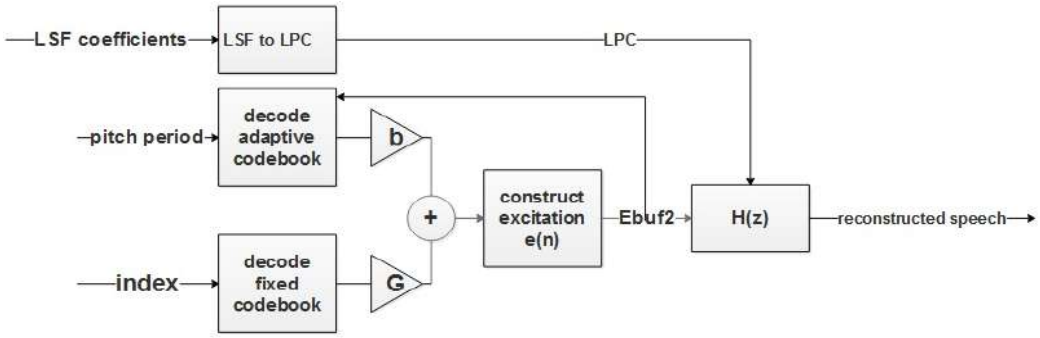


Figure 4: CELP Decoder flow chart

3 SIMULATION

This section simulates CELP codec using MATLAB ([20], [21]). The bit rate is calculated for various number of LPC, different size of fixed codebook, various frame length of the speech segment. The pitch period (P) is searched between 16 to 160 samples, 8 bits are used to represent it. The gain of the fixed codebook (G) is coded by 7 bits per sub-frame. The gain of the adaptive codebook (b) is coded by 8 bits per sub-frame. a 6-bit uniform quantizer is used for each LPC (a). The number of bits used to transmit the index of fixed codebook K=$\log_2(fixed\ codebook\ s\ ize)$.

Table 1 shows total number of bits per frame, each frame is divided into four sub-frame (m=4). 10 LPC and 1024 fixed codebook size are used. From table I total bits per each frame equal 192 bits.

Frame rate = (8000 samples/second) / (160 samples/second) = 50 Frames/second then the bit rate = (192 bits/frame) * (50 frames/second) = 9600 bits/second**.**

Table I
TOTAL BITS PER FRAME

No. of bits	Parameters
8*4	P*m
8*4	b*m
7*4	G*m
10*6	filter order*a
10*4	K*m
192 bits	Total bits per frame

The bit rate using various number of LPC are given in Table II at frame length equal 160 and fixed codebook size equal 1024. The sampling frequency 8000 Hz is used.

Table II
TOTAL BIT RATES for DIFFERENT NUMBER of LPC

Number of LPC	Bits per sec	Bits per frame
10	9600	192
8	9000	180
5	8100	162

Table II concludes thatusing 8 LPC save 12 bit per frame than using 10 LPC and using 5 LPC save 30 bit per frame than using 10 LPC.

4 SUBJECTIVE &OBJECTIVE TESTS

This section presents the Mean opinion square and signal to noise ratio tests.

A. Mean Opinion Square

Mean opinion square (MOS) is a test for measuring the acceptability or quality of speech over a Communication system. Table III shows MOS scale range. The scale allows the listener to judge the overall quality of a communication system [22].

Table III
MOS SCALE RANGE

MOS scale	Speech quality
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

B. Signal to Noise Ratio

Signal-to-noise ratio (SNR) is one of the most common objective measures for evaluating the performance of a compression algorithm. Equation 4 defines the SNR [23], Where $s(n)$ is the original speech data while $\hat{s}(n)$ is the coded speech data.

$$\text{SNR} = 10\ \log_{10}\frac{\sum s^2(n)}{\sum(s(n)-\hat{s}(n))^2} \qquad (4)$$

C. Methodologies

10 Arabic words are used. Each word contains 3 Arabic phonemes altered by a male speaker. The number of subjects is 30. Two tests are conducted .Tables IV and V show the words used within test 1 and test 2 respectively. The subjects are asked to write each word as they hear, and estimate the corresponding MOS scale range. The MOS results are averaged over 30 subjects for each word. We calculate SNR for various parameters.

TABLE IV
TEST 1 WORDS

Word	تاب	باب	سام	غاب	جاب	غاب	سوق	عيب

TABLE V
TEST 2 WORDS

Word	صام	باب	سام	سوق	عيب	خاب	طاب	غاب

5 RESULTS & DISCUSSIONS

In this section the effect of varying the number of LPC of CELP coder is discussed by means of subjective and objective tests.

A. MOS Test

Figure 5 and figure 6show the MOS scale for different number of LPC. It is concluded that using LPC value equal 8 and LPC value equal 10, MOS are slightly different whereas five LPC value yields to a significant decrease of the MOS scale. Also figure 7 shows percentage number of subjects that heard correctly 3 Characters out of 3 (100% intelligibility) 93% of the subjects heard correctly at 10 LPC, whereas the percentage is significantly reduced at 5 LPC. This means that there is a significant decrease in intelligibility using LPC value 5.

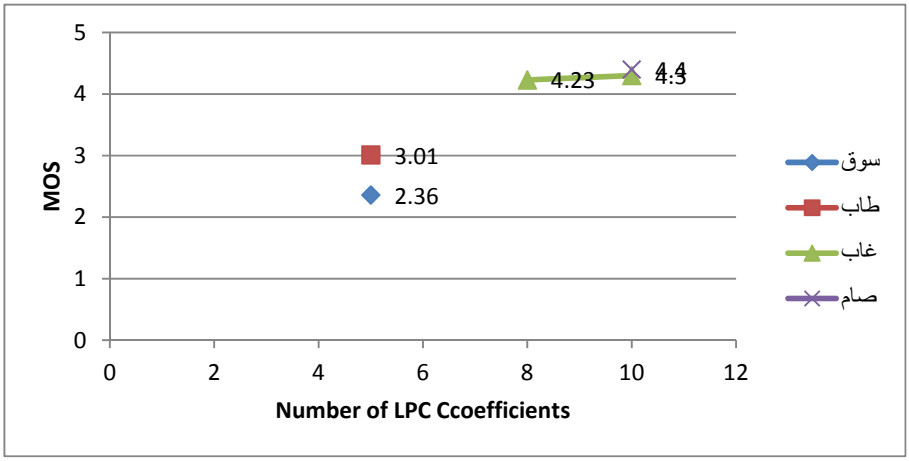


Figure 5 : Number of lpc coefficients with mos scale

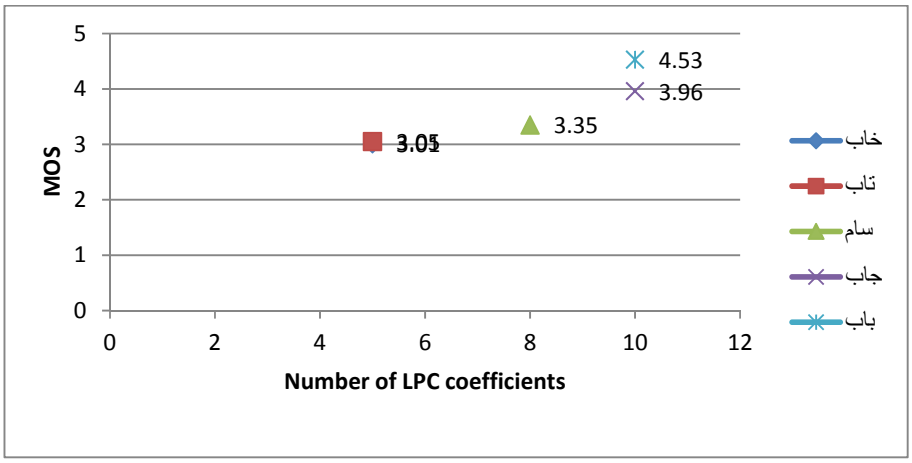


Figure 6 : Number of lpc coefficients with mos scale

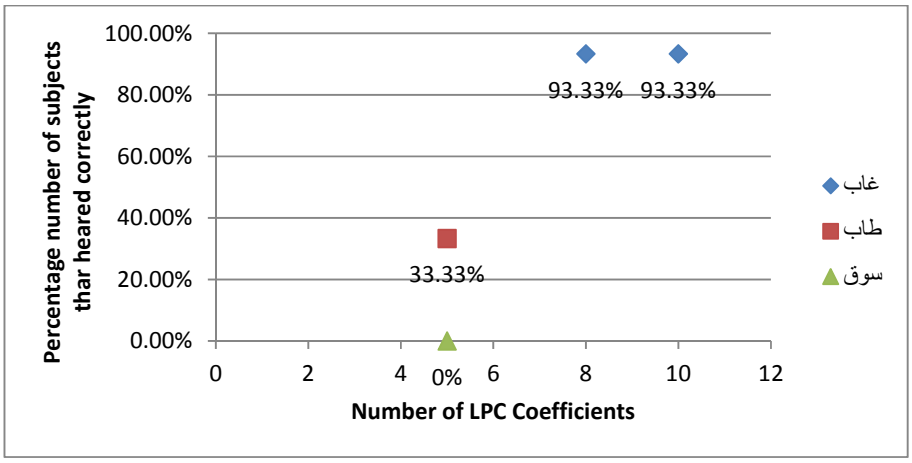


Figure 7: Number of lpc coefficients with percentage number of subjects that heard the word correctly(3/3)

Figure 8 shows the MOS scale for different numbers of LPC. It is concluded that using 10 LPC the MOS is high whereas five LPC yield to a significant decrease of the MOS scale. Exception is shown in figure 8 where the word (عيب) gives MOS at LPC value 5 & 8 respectively which means that the word (عيب) records high MOS even LPC is decreased . The same exception is cleared in figure 9 where the number of subjects that heard correctly at LPC value 5 & 8 is 100%. The reason is that word (عيب) has higher signal strength. The ratio of the signal power of the word عيب to that of the word غاب is calculated and it is found to be equal 1.72.

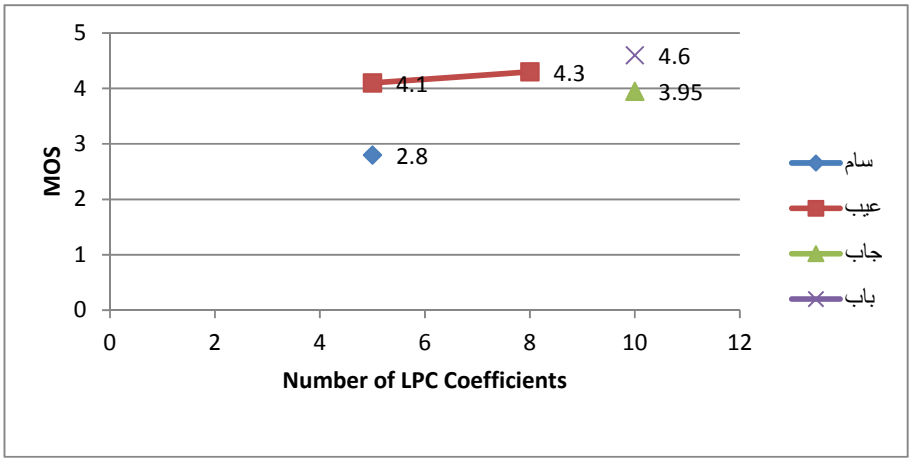


Figure 8: Number of lpc coefficients with mos scale

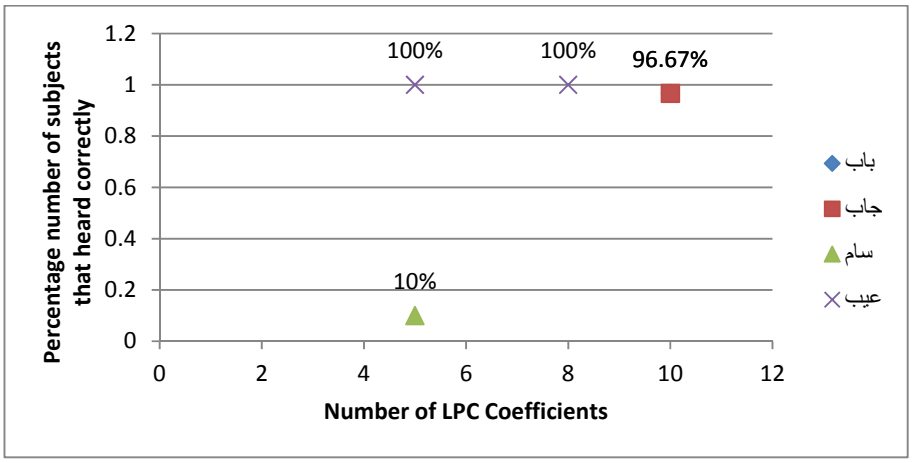


Figure 9: number of lpc coefficients with percentage number of subjects that heard the word correctly(3/3)

Figure 10 shows the number of subjects that heard correctly for various number of LPC coefficients. It is concluded that using 10 LPC the number of subjects that heard correctly is 96.66% whereas five LPC yields to a significant decrease in number of subjects that heard correctly. Exception is shown in figure 10, the word (سام) is poorly perceived although eight LPC are used. The reason is that the character (م) has weak signal strength so listeners cannot recognize all the 3 character, they heard only first 2 character. Figure 11and 12 show the plot for words (سام) and (خاب) respectively. The ratio of the signal power of the word خاب to that of the word سام is calculated and it is found to be equal 2.1.

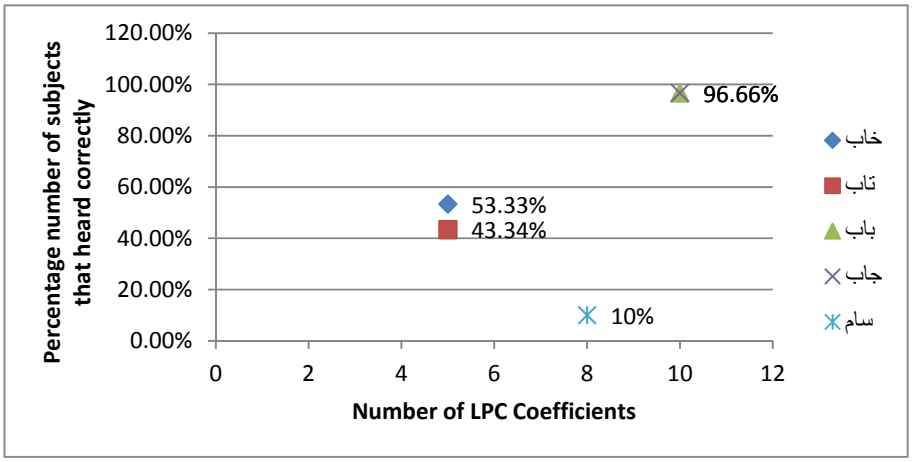


Figure 10: Number of lpc coefficients with percentage number of subjects that heard the word correctly(3/3)

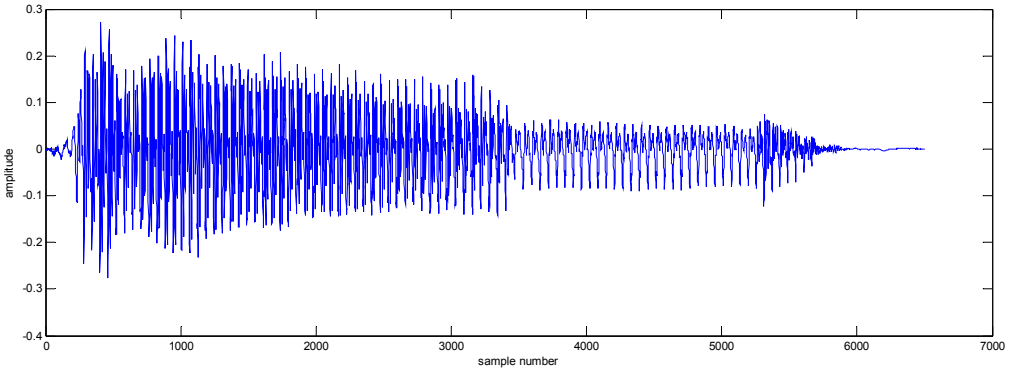


Figure 11: Plot for the word (سام)

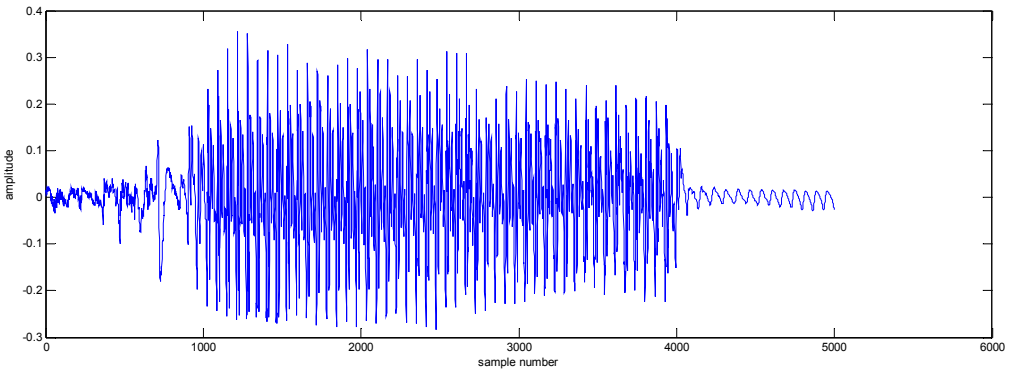


Figure 12: Plot for the word (خاب)

B. SNR Test

SNR is calculated for frame length=160 and fixed codebook size =1024. Table VI shows SNR for different number of LPC.

TABLE VI
DIFFERENT NUMBER of LPC with SNR

Number of LPC	**SNR (dB)**
8	11.348
5	10.990

Table VI concludes that decreasing the number of LPC SNR decreases.

6 CONCLUSION

This paper implements the CELP coder for Arabic words with different number of LPC coefficients. The paper investigates their effect on the coder quality by means of subjective and objective tests. The bit rate is calculated for each case. It is concluded that using 5 LPC yields to low speech quality. Employing LPC either eight or ten improves the speech quality. The paper specifies 8 LPC as they provide good speech quality. It is found that employing the parameters that are listed in table VII lowered bit rate of the CELP coder from 9.6 kbps to 9 kbps. The MOS varies from 4.2 to 3.5 for bit rate of 9.6 and 9 kbps respectively. Future work on Arabic words must be considerable to all researches also the variation in the fixed codebook size and frame length has to be considered.

TABLE VII
PROPOSED CELP PARAMETERS for ARABIC WORDS

Value	**Parameter name**
160 (20ms)	Frame length (N)
40 (5ms)	Sub-frame length
8	Order of LP filter
0.85	Constant parameter for perceptual weighted filter (c)
[16,160]	Estimate of number of samples in the pitch period
(40,1024)	Size of fixed codebook

REFERENCES

[1]BishnuS.Atal, "The History of Linear Prediction," *IEEE signal processing magazine*, vol. 23, no. 2, pp.154-161, march 2006.
[2]Wai C. Chu, speech coding algorithms foundation and evolution of standardized coders, *John Wiley & Sons*, Hoboken, New Jersey, 2003.
[3]S.K Jagtap,M.S Mulye, M.D Uplane , "Speech Coding Techniques, " *4th International Conference on Advances in Computing, Communication and Control*, Elsevier, available online on Science Direct, pp. 253 – 263, India, 2015.
[4] Mohamad Itani, Sarunas Paulikas, "Influence of Languages on CELP Codecs Performance," *Information Technology and Control, 2008, Vol. 37, No. 2*, Lithuania, January 2008.
[5] Michael Nasief , "Performance Evaluation of Speech CODECs against the Change in the Spoken Language and Accent, " *30th NATIONAL RADIO SCIENCE CONFERENCE,Egypt,2013.*

[6]Mansour Alsulaiman, Ghulam Muhammad, Zulfiqar Ali, "Comparison of voice features for Arabic speech recognition," *Sixth IEEE International Conference on Digital Information Management, ICDIM 2011*, pp.92-95, Melbourne, Australia, September, 2011.
[7]Ankita Anand, Richa Bhatia , "Performance evaluation of band-limited LPC vocoder and band-limited RELP vocoder in adaptive feedback cancellation, " *International Conference on Advances in Computing, Communications and Informatics,* Aug. 2015 India.
[8] N. R. Chong-White and R. V. Cox, "An intelligibility enhancement for the mixed excitation linear prediction speech coder," *IEEE Signal Processing Letters*, vol. 10, no. 9, pp. 263 – 266, September 2003.
[9]Mohit Narayanbhai Raja, Priyanka Richhpal Jangid, Sanjay M. Gulhane ,"Linear Predictive Coding," *International Journal of Engineering Sciences & Research Technology*, pp.373-379, India, April-2015.
[10]Atsushi Murashima, Masahiro Serizawa, Kazunori Ozawa , "A post-processing technique to improve coding quality of CELP under background noise, " *IEEE Workshop Speech coding Proceedings*, p.102-104, 2000.
[11] Awais M. Kamboh, Krispian C. Lawrence, Aditya M. Thomas, Philip I. Tsai, "Design of A CELP Coder and Analysis of Various Quantization Techniques, " EECS 651 Project, University of Michigan Ann Arbor 2005
[12]C. Li and V. Cuperman, "Analysis-by-synthesis multimode harmonic speech coding at 4 kb/s, " in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, vol. 3, pp. 1367–1370, 2000.
[13]L.Hanzo, F.C.A. Somerville, J.P. Woodard, Department of Electronics and Computer Science, University of Southampton, UK, voice and audio compression for wireless communications, second edition, *John Wiley & sons*, 2007.
[14] A. McCree, J.Stachurski, T. Unno, E. Ertan, E. Paksoy, V. Viswanathan, A. Heikkinen,A. Ramo, S. Himanen, P. Blocher and Dressler, "A 4 kb/s hybrid MELP/CELP speech coding candidate for ITU standardization, " *Proc. IEEE Int. Conf. Acoustic, Speech Signal Processing* ,vol. 1 pp. 629–632 , 2002.
[15]RhutujaJage, SavithaUpadhya, "Implementation of CELP and MELP speech coding techniques," *international journal of advanced research in computer and communication engineering*, vol.5, issue 6, pp.702-708, India, June 2016.
[16]Lani Rachel Mathew, Ancy S. Anselam, Sakuntala S. Pillai, "Performance Comparison of Linear Prediction Based Vocoder in Linux Platform, " *International Journal of Engineering Trends and Technology*, vol.10, Issue 11, pp.554-558, April2014, India.
[17]Neel Kamal Saha , Rajendra N Sarkar , Dr. Miftahur Rahman, "Comparison Of Musical Pitch Analysis Between LPC And CELP , " *International Journal of Advances In Engineering Sciences*,vol.1, Issue 1, pp.35-39, North South University ,Jan, 2011.
[18]Lila Madour "A Low-Delay Code Excited Linear Prediction Speech Coder at 8 k b it/s, "Master degree to Department of Electrical Engineering McGill University, Montreal, Canada ,March, 1994.
[19] MinalMulye, SonalJagtap, "Overview of Code Excited Linear Predictive Coder, " India, Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, accessed July 2015.
[20]*P. Kabal,* "ITU-T G.723.1 Speech coder- MATLAB Implementation, " McGill University, Technical Report, 3-August 2011.
[21]Ian Mcloughlin, "Applied Speech and Audio Processing with Matlab Examples, "Cambridge University, New York, 2009.
[22]NasirSaleem, Usman Khan, Imad Ali, "implementation of low complexity CELP coder and performance evaluation in terms of speech quality," *International Journal of Computer Applications*, vol.54, no.9, pp.12-16, Pakistan, September 2012.
[23]Lingfen Sun, "Speech Quality Prediction for Voice over Internet Protocol Networks," Doctoral dissertation,University of Plymouth, Jan. 2004.

BIOGRAPHY

Nayra A. ESSA is a graduated of Faculty of Engineering, Electrical Department, Alexandria University, 2011. Graduation Project is Matlab Implementation for LTE Advanced (4G). She Studies a Master Degree in Code Excited Linear Predictive, at Faculty of Engineering, Alexandria University.
Email: eng_nayra158@outlook.com

Noha O. Korany is Currently Professor at the Department of Electrical Engineering (Communications and Electronics), University of Alexandria, Egypt. She Received her B. Sc. Eng. in 1992, M. Sc. in 1995 from Alexandria University, and her Ph.D. from Alexandria University, Egypt and Fellowship Ruhr-Universitaet Bochum, Germany in 2000. She was Member of the Scientific Staff at the Institute of Communication-Acoustics, Ruhr-Universitaet Bochum, Germany from 2002 to 2004. Her Main Research Field is Acoustics and Communications.
Email: nokorany@hotmail.com

Onsy Abdel Alim is Currently Professor at the Department of Electrical Engineering (Communications and Electronics), University of Alexandria, Egypt since 1984. He Received his B. Sc. Eng in 1964 with Distinction Honors and his Ph.D. from Germany. in Acoustical Engineering with Distinction Honors. He honored as the Best Man of Communication in Egypt in 2007.
Email: onsy2066@hotmail.com

تأثير تخفيض عدد المعاملات التنبؤية الخطية على جودة الصوت من المشفر سلب باستخدام الكلمات العربية

[*1]نيرة عبدالحليم عيسى و[*2]نهى عثمان قرنى و[*3]انسى عبد العليم
الهندسة الكهربية كلية الهندسة جامعة الاسكندرية
[*1]eng_nayra158@outlook.com
[*2]nokorany@hotmail.com
[*3]onsy2066@hotmail.com

ملخص
الهدف من تشفير الكلام هو الوصول إلى أفضل جودة بأقل عرض للنطاق الترددي. في معدلات الوحدات الرقمية فوق 4 كيلوبت في الثانية، المشفرات الهجينة الخاصة بالكلام و الفائمة على أساس المشفرالمستحث بشفرة التوقع الخطى (CELP) يمكن أن تنتج نوعية صوتية جيدة. اما بالنسبة لمعدلات الوحدات الرقمية الأقل من 4 كيلوبت في الثانية يصبح من الصعب جدا تحقيق الجودة المطلوبة للصوت المشفرالمستحث بشفرة التوقع الخطى مبنى على اساس استخدام طريقة التحليل حسب التوليف.
في هذه البحث، يتم تنفيذ المشفر باستخدام برنامج ماتلاب. ويهدف البحث إلى إيجاد الحد الأدنى لعدد المعاملات التنبؤية الخطية التي تعطي جودة جيدة للمشفر، وبالتالي ينخفض معدل الوحدات الرقمية. و تم استخدام الكلمات العربية فى البحث . تتأثرجودة الترميز باللغات المستخدمة. ويتم تحليل اداء المشفر باستخدام اختبارات ذاتية وموضوعية مثل درجة فهم للكلام و متوسط نقاط الرأي (MOS) ونسبةالإشارة إلى الضوضاء (SNR).

الكلمات الرئيسية - تقليل عدد المعاملات التنبؤية الخطية، تنفيذ المشفرالمستحث بشفرة التوقع الخطى ، جودة الصوت باستخدام الكلمات العربية، التحليل حسب التوليف

Arabic Handwritten Recognition Using IoT Technology in Cloud Computing

Nada A. Shorim[*1], Norhan M. Eltopgy[*2], Sahar K. Mohamed[*3],Shehab Salah[*4], Taraggy M. Ghanim[*5], Ashraf M. AbdelRaouf[*6]

[*]*Faculty of Computer Science, Misr International University, Cairo, Egypt*

[1]nada.ayman@miuegypt.edu.eg,[2]Norhan120664@miuegypt.edu.eg,[3]sahar122284@miuegypt.edu.eg,
[4]shehabsalah25@gmail.com,[5]taraggy.ghanim@miuegypt.edu.eg, [6]ashraf.raouf@miuegypt.edu.eg

***Abstract*—Recognizing Arabic handwritten is a great challenge due to variations in its letters shapes according to its position in the word and handwriting styles. Internet of Things (IoT) is the inter-networking of connected devices embedded with electronics, software, and sensors to collect and exchange data. Cloud computing and IoT both serve to increase our daily efficiency, and they have a complimentary relationship. Recognizing Arabic handwritten using mobile device connected to cloud computing facilitate translating for non-Arabic speaking and finding locations on maps which are of great importance while visiting Arabic speaking countries. Our approach is the first to build a mobile app that propose a multi-phase hybrid classifier that works on the words geometric features. Our classifier is based on KNN, then passing a set of nearest neighbor votes to SVM. Training phase used a self-generated dataset and was tested on IFN/ENIT database. Our approach successfully achieves 83.04%Accuracy.**

***Keywords*—Arabic, Offline handwriting recognition, hybrid classifier, features extraction, IoT, Cloud Computing**

1 INTRODUCTION

Internet of Things (IoT)[1] nowadays is widely spreading using it and combining it with mobile phones apps and computing, IoT is a system of interrelated computing devices, mechanical and digital machines, objects embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data.

Cloud computing technology is also becoming popular due to time and storage constraints. Fast processing is satisfied with parallel processing done by different cloud services. This state of the art technology also provides huge storage space. Using cloud computing facility of Software-as-a-Service (SaaS) gives the opportunity to upload, share and execute developer's applications. We provide a mobile application that is portable, easy to use, fast processing time and provide storage space by using Internet of things (IoT) and cloud computing[2].

Handwritten character recognition is an important part of today's every mobile apps. Most of the mobile applications tend to use the handwritten as an alternative option for input data with the keyboard and voice. Arabic Handwriting character recognition has been one of the most important research topics for countries that use the Arabic alphabets. There are a lot of challenges concerning this field of language recognition, the characters shape changes according to its position in the word (isolated, start, middle, end), as Fig 1 shows samples of Arabic letters in their different location. Also, that Arabic is written from right to left. Arabic is a cursive language that most of its letters are connected when written. Handwritten has distinct challenge due to different handwriting styles[3].

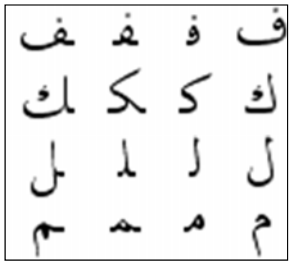


Figure 1: Samples of Arabic letters with their shapes in different locations

The approach proposed in this paper is recognizing the Arabic handwritten by calculating the geometric features of the word in the feature extraction phase and building a hybrid classifier based on artificial neural networks (ANN) and support vector machine (SVM) in the classification phase. Our mobile app is the only mobile app on Android platform (except that come from Google with the Android operating system) that is recognizing the Arabic handwritten using the mobile and cloud computing.

The remaining of this paper is organized as follows; section 2 explains the related works that are related to this field, and discuss their approaches. In section 3, our proposed approach is discussed to enhance the recognition of the Arabic handwritten. An experiment is drawn in terms of statistics to show the efficiency of our approach in section 4. And finally, in section 5, discussion and concluding remarks and the future enhancement work.

2 RELATED WORK

The recognition of Arabic handwritten is a very important research topic and for its importance it is very common topic of research. We are going to explain the related work of the research from two points of view, firstly, is concerning the mobile apps that are related to our topic. Secondly, the overall research work related to the Arabic handwritten recognition. It is explained with relation to the main phases of the handwritten recognition.

A. Related Applications

Some mobile apps are available freely on the Google play store for recognizing printed English letters. One of these applications is OCR_Text Scanner [4], it is a free app and is used for recognizing printed English text from images, no internet connection is required, camera option is available; it takes rate 3.7/5 from users' point of view. Another similar application is Image to Text [5] it is a free also and is used for recognizing printed image, it supports English, and some other languages but is verified mainly in English. It supports different file formats such as: PNG, JPG, GIF, TIFF, BMP image formats. It sometimes generates unknown errors. Finally, its user rate is 3/5.

B. Related Research

The related work in this part is explained depending on the different Arabic handwritten recognition phases which are preprocessing, segmentation, feature extraction and classification. Fig.2 explains the block diagram of the different phases of the Arabic handwritten character recognition.

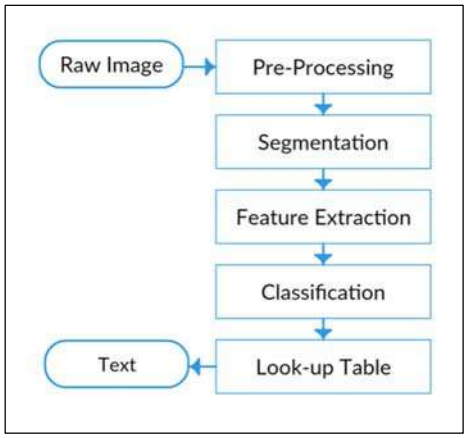


Figure 2: Block diagram of the Arabic handwritten recognition phases

1) Preprocessing

The preprocessing phase is an important stage in the Arabic handwritten recognition. It prepares the text image for the segmentation phase. The more success preprocessing phase, the more accurate recognition will be [6]. The approach proposed by AlKhateeb et al. [7] applied skew/slant correction and normalization on images. The Image was divided into frames and two types of features were used. Intensity features are used for training and structure-like features are used for re-ranking for improved accuracy a set of features and these features are extracted from each frame. Finally, they achieved a recognition rate of 83.55 after testing on the IFN/ENIT database using HMM.

AbdelRaouf et al. [8] proposed an approach that uses the basic preprocessing (binarization, thinning, slant and skew, baseline finding) then passing through the feature extraction. Haar-like features were used which is a simple method depending on the basic visual features of the objects. It uses gray-scale differences between rectangles to extract the object features then going to the classification stage. The Haar Cascade Classifier (HCC) is a machine learning approach that combines three basic components (Integral image, Haar-like features extraction and boosting of cascade classifiers). Their approach got a total accuracy 87% on MMAC corpus database [9].

2) Segmentation

Segmentation phase in the character recognition is the first phase to deal with the text itself. It is responsible for dividing the word to its letter or stroke contents [10]. Al-Hamad and Abu Zitar [11] presented an approach that uses Arabic Heuristic Segmentation to segment a word into primitives. These primitives might be a letter or part of a letter. The features then are converted to an artificial neural network Multi-layer perceptron (MLP) [12] for training and testing to validate the segmentation points. This system was tested on an AHD/AUST database containing 12,300 words and the IFN/ENIT database containing 26,400 and achieved segmentation rates of 95.66%.

Xiang et al. [13] used a sliding window technique to divide the image into frames and each frame is divided into four cells. Then distribution features and concavity features are extracted from each frame. They also used hidden Markov models (HMM) for training and recognition. Finally, they achieved an average recognition rate of 84.09% from testing on the IFN/ENIT database.

3) Feature Extraction

Feature extraction phase is to select the features that are going to differentiate between letters, and is divided into two types. Structural features are usually computed from a skeleton or a shape of the text image. Structural features remain more common for the recognition of Arabic letters. Statistical features are numerical measures computed over images or regions of images [14]. Lawgali et al. [15] computed highest value Discrete Cosine Transform (DCT) coefficients of each character as features. Artificial Neural Network (ANN) is used for feature extraction and classifications. The ANN has 3 layers (8 neurons as input layer, 40 neurons as a middle layer and 70 neurons as output layer) is used for feature extraction and classification. The classification is achieved in two steps (classification of the segmented characters and classification of the word). Finally, they achieved 90.73% after testing on IFN/ENIT database.

Surinta [16], used preprocessing schemes such as rescaling the image by preserving the aspect ratio and converting it from color to grayscale. The feature descriptors which we selected are the Scale Invariant Feature Transform (SIFT) [17] and Histograms of Oriented Gradients (HOG) [18] that extract the orientation histograms from the handwritten character grayscale images. The K-Nearest Neighbor (KNN) and support vector machine (SVM) classifier were used for classification. Finally, they achieved 99.07% on the THI-D10 dataset.

4) Classification and Databases

This phase is the final phase to recognition. It compares the features of the tested word with the features of the saved trained words, and decides which word is recognized [19]. Hussein et al. presented "ALEXU-WORD" database [20]. It is a new dataset for online Arabic handwriting recognition aiming to obtain a very large database of segmented letter images, to evaluate Arabic handwriting recognition systems. They used couple of windows descriptors and they are based on Histograms of Oriented Gradients (HOG) [18] and the Scale Invariant Feature Transform (SIFT) [17] descriptors. The (HOG) and (SIFT) are experimented with three classification algorithms, k-Nearest Neighbor (k-NN), Artificial Neural Networks (ANN) as they used the linear kernel with Support Vector Machines (SVM). Their technique competes IFN/NET database by 92.16 success using SIFT-based descriptor and artificial neural network (ANN).

During 2016, Haider et al. [21], used high-performance Graphical Processing Units (GPUs) which has 1200 cores which made the training and testing fast so they used Convolutional Neural Networks (CNNs) to train and test handwritten digits. CNN is used to extract robust features that help in justifying the class to which they belong to, and they use the last layer named softmax layer which used to minimize the errors. Finally, they achieved 95.7% accuracy on MNIST dataset.

3 OUR PROPOSED APPROACH

Our approach depends on building an Arabic handwritten recognition application that is run from a mobile device. This system must be able to use the detected words to implement the Google Software Development Kit (SDK) of Google map and Google translate. The main difficulty here is implementing the Hand written recognition system on a mobile for the its limited processing capabilities. The solution is to benefit from the huge processing capabilities of the Cloud computing.

As shown in Fig. 3, the proposed approach uses the mobile device to scan the Arabic handwritten image. The preprocessing phase of the Arabic handwritten application is implemented on the mobile device as it requires limited processing capabilities. The preprocessed image is then transferred to the cloud using IBM BlueMix Cloud platform to implement the remaining from the Arabic handwritten phases. The recognized words from the Arabic handwritten application in the cloud are sent back to the mobile device. The Google SDK of the Google map and / or Google translate is then run with the words returned from the cloud. If the words are an address then the Google map is run with the mentioned address, else the Google translate is run with the translation of the words to English language.

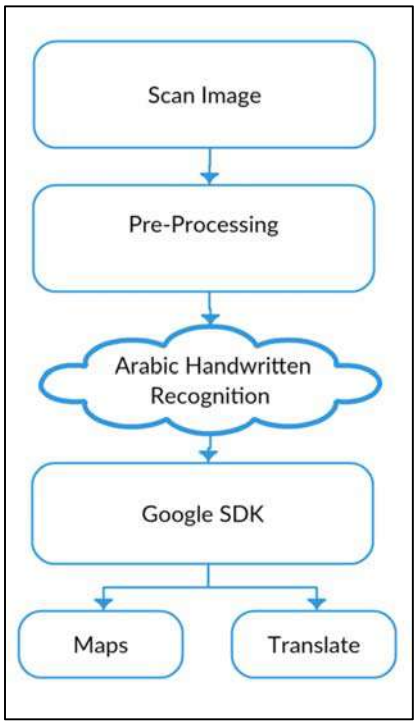


Figure 3: Block diagram of the proposed approach

The proposed Arabic handwritten character recognition approach includes the following phase; preprocessing, segmentation, feature extraction, clustering and classification. We are going to explain the followed stages in each phase in the proposed approach. Finally, we explain how these processing phases should be uploaded on cloud servers to reduce complexity and high processing on our limited mobile device.

In the preprocessing phase, each input image should be converted to binary, filtered from any additional noise and thinned to one pixel wide stroke in the segmentation phase, where the words are isolated from the background to concentrate computations on valuable pixels only. Some needed information is computed in this phase like number of Pieces of Arabic Words (PAW). Next, we compute a set of geometric features in the phase of feature extraction. Some of these features are used in clustering images into sets. Each set represents similar classes. Again, our set of geometric features is passed to our hybrid classifier to finally recognize the handwritten text, and convert it to editable form.

A. Preprocessing

The aim of preprocessing phase is to improve the quality of the scanned image. It is very important to try to improve the quality of the scanned image as much as possible for better recognition accuracy. In this phase, we aim to select algorithms that are good for preprocessing and at the same time with low complexity expensive.

1) Binarization and Noise Removal

Binarization is converting the grey scale image into black and white (binary) image. So, this stage converts any captured image to binary, defining only two possible values, 1 for background and 0 for required text. We applied the adaptive threshold mean algorithm [22] to convert images to binary. Adaptive threshold doesn't depend on fixed threshold value but it varies at each pixel location according to the neighboring pixel intensities. It is applied by constructing a kernel of size 15x15 pixels around each pixel *(x,y)*. The average *A(x,y)* of the pixels in this kernel is computed. The threshold *T(x, y)* is computed by subtracting an offset value from the local mean *A(x,y)*. This offset value is used for fine tuning the threshold value. Equation 1, shows the calculation of the threshold for every pixel, where *param1* is the offset value.

$$T(x, y) = A(x, y) - \text{param1} \quad (1)$$

Noise removal stage can be achieved using smoothing spatial filters that are used for blurring and for noise reduction. The arithmetic mean filter is a very simple one and suitable for the noise removal [23] is then applied on the captured binary image to remove any additive noise. Fig. 4 shows an example of applying the filter

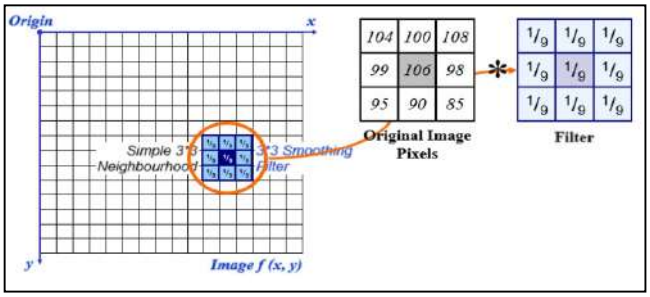


Figure 4: Example of applying arithmetic mean filter

2) Thinning

Thinning stage is responsible for remove the thickness of the text that came from using different types of pens. In this stage, we applied Zhang-Suen-thinning-algorithm [24] on our binary image after removing noise from it to convert it to one pixel wide text. This operation is important to be invariant to different writing tools and styles. Fig. 5 shows the Zhang applied thinning algorithm. Fig. 6 shows an example of an Arabic word before and after thinning.

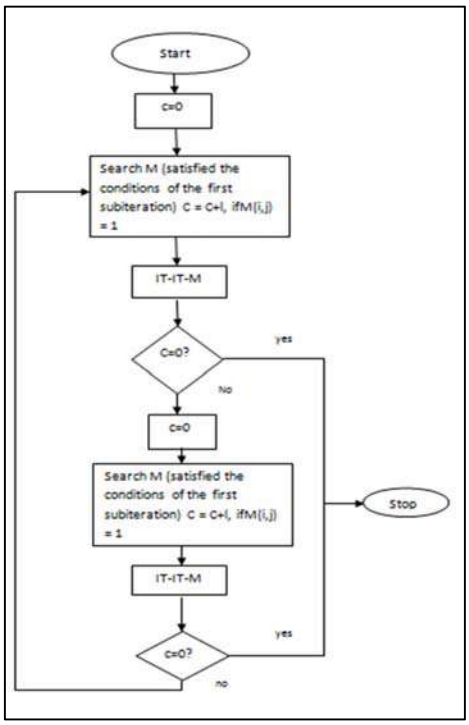


Figure 5: Flowchart of Zhang thinning algorithm applied [17]

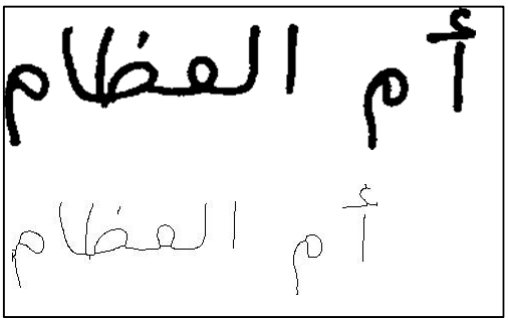


Figure 6: Example of Arabic word before thinning (up) and after thinning (down)

3) Filling Gaps (Dilation)

Dilation takes the union of copies of the structuring element centered at every pixel location in the fore ground. This stage is used to take the binary image after binarization, apply a (5 X 5) matrix on the image to fill the gaps occurred from the thinning which turn the binary shape in the image into one pixel wide lines. This stage is used to be applied on all the possibilities that will make a gap in the word, and this gap means that two connected pixels or one pixel have the same color of the background in the image. So, it will turn all these possibilities to the color of the word to decrease the percentage of making gaps in the word. Fig. 7 shows an example of filling the gap in the word.

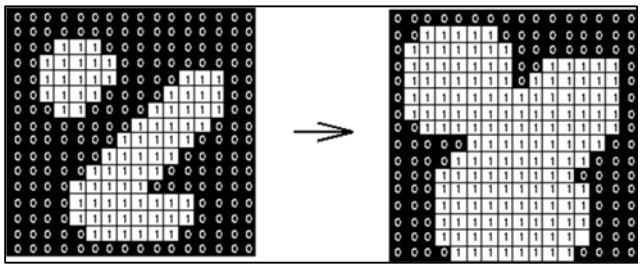


Figure 7: Example of filling gap (Dilation) process

4) End/Start Points

This stage is responsible for defining the start and end point in every stroke of the scanned image after thinning it [25]. It goes through all the pixels in the image to detect each pixel that has one or two neighbors, which means that this is an end or start point in the image, so after applying the median filter it detects all the results which are one or two pixels then count them. This stage is partially used in the feature extraction stage, as the start and end points are part of the features.

5) Number of holes

This stage is responsible for defining the number of holes in each word and label them using the connected component algorithm [26]. It deals with the binary image where (0 for black pixels and 1 for white pixels), and go around the whole image pixel by pixel to detect the entire connected component in the image. The algorithm by labeling each connected component (gave each connect component a number). The start point is (1, 1) and end with (n-1, n-1) and get the 8-Neighbors of each pixel. Then give each connected pixel (the pixels of the same color) a number (Labeling). Then looping once again from (1, 1) to (n-1, n-1) and get the minimum number of the 8-Neighbors and assigning it to the current pixel. Then we loop one time from (n-1, 1) to (1, n-1) and one time from (n-1, n-1) to (1, 1) and get the minimum number in the 8-Neighbors and assigning it to the current pixel. After the labeling stage, we subtract the background from it and get the average pixels number of all holes. Then we start to count if the hole has number of pixels more than the average so, it is a hole and it is counted with us, if the hole has number of pixels less than the average that means it's not a hole and may be noise, so do not count it with us. Fig.8 shows the flowchart of the applied algorithm.

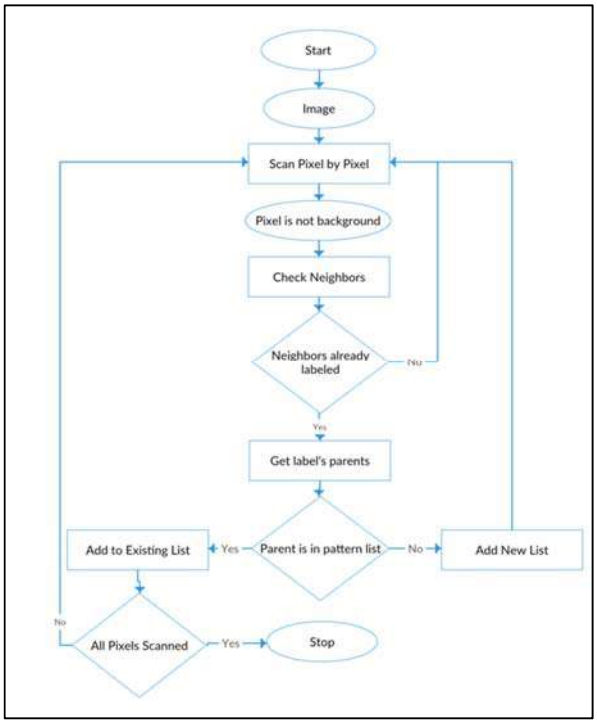


Figure 8: The flowchart of the number of holes and labeling

B. Segmentation

This phase is responsible for dividing the word into its sub-words, letters or strokes [27]. It goes through all the pixels in this image, to calculate the number of connected components in the image and give each connected component a specific label [28]. And count all these labels to count the number of PAWs, then cut them to make each region in external image.

C. Feature Extraction

This phase is responsible for extracting the important features from the text image and keeping them as a reference for that word. When in the phase of recognizing the text, it recalls these features and compares it with what was saved from the training dataset to decide whether it is the same word or not. These features are:

1) Geometric features:

Geometric features of text images are computed. It is based on dividing captured image into six zones, three horizontal and three verticals. A 5 x 5 kernel is used to compute the geometric features of each zone. These features are number of Horizontal, Vertical Lines, right Diagonals, left Diagonals, number of normalized Horizontal lines, normalized Vertical Lines, normalized right Diagonals and normalized left Diagonals.

2) Eccentricity[29]*:*

Eccentricity measures the shortest length of the paths from a given vertex v to reach any other vertex w of a connected graph. If we considered an imaginary ellipse drawn around the word, then the eccentricity is computed by equation (2)

$$\text{Eccentricity}=C/A \qquad (2)$$

where C is the distance from the center to foci and a is the distance from that foci to a vertex as shown in Fig.9. Foci equation (3) is given as:

$$C^2= A^2- B^2 \qquad (3)$$

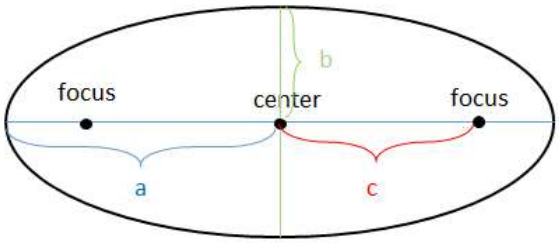


Figure 9: Calculation of the foci distance

3) Orientation[29]*:*

Orientation is the angle between the major axis of the ellipse and the X-axis. As shown in Fig. 10.

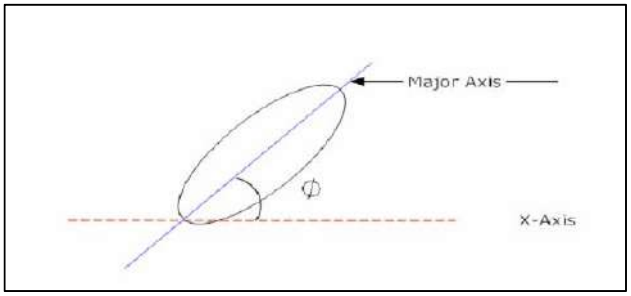


Figure 10: Calculation of the orientation angle

4) Extent[29]:

Extent is a ratio between the number of pixels and area of the ellipse.

D. Clustering

Clustering is the phase of separating the word into different clusters based on their feature extraction grouping. After features computation, the application groups training data to sets according to a range of these features; eccentricity, extent and orientation [29], number of holes, number of paws, start and end points. We use all the previous features to cluster the word to its group. The Applied algorithm is the K-Nearest Neighbor algorithm (KNN) which proved very high performance in the clustering phase.

E. Classification

Classification is the phase in the Arabic Handwritten character recognition to identify to which of a set of categories a new word belongs, on the basis of a training dataset containing words whose category membership is known. We apply two different algorithms to achieve better result, first the Support Vector Machine (SVM) then apply the Kohonen Neural Network algorithm.

1) Support Vector Machine (SVM)

SVM is one of the most powerful supervised classifiers. It has been used in pattern recognition in different fields and achieves high classification rates. A brief literature was done by Burges [30] or Cristianini and Shawe-Taylor [31]. Extensions of the binary classification to the multi-class situation are suggested in several approaches [32, 30]. We are cascading the KNN with SVM to achieve more accuracy. By taking the number of votes from the KNN as an input to the SVM to train and test on, then outputs its final decision.

The mathematical function used for the transformation is known as the kernel function. It is defined as Linear, Polynomial and Radial basis as shown in the following equations.

• Linear

$$F(x; w, b) =< w, x > +b \quad (4)$$

• Polynomial

$$K(x, z) = (< x, z > +1)\, p \quad (5)$$

• Radial basis function (RBF)

$k(x; y) = \exp(-||x- y||^2) / (2 *alpha^2)$ (6)

2) Kohonen Neural Network (KNN)

KNN is Self-organizing maps that classify data without supervision. First, all weights are initialized randomly to a value between 0 and 1. Second, random image is chosen at random and Euclidean distance is calculated between it and rest of the nodes using equation (7):

$$Dist = \sqrt{\sum_{i=0}^{i=n}(V_i - W_i)^2} \quad (7)$$

Initial radius is calculated and weights of nodes within it are adjusted according to equation (8)

$$Weight = Weight + LearningRate(FeatureVector - Weight) \quad (8)$$

Euclidean distance is calculated between the chosen image and all nodes and then the reduced radius and the learning rate are calculated. Repeat until reach max iteration. Fig. 11 shows the flow chart of KNN algorithm.

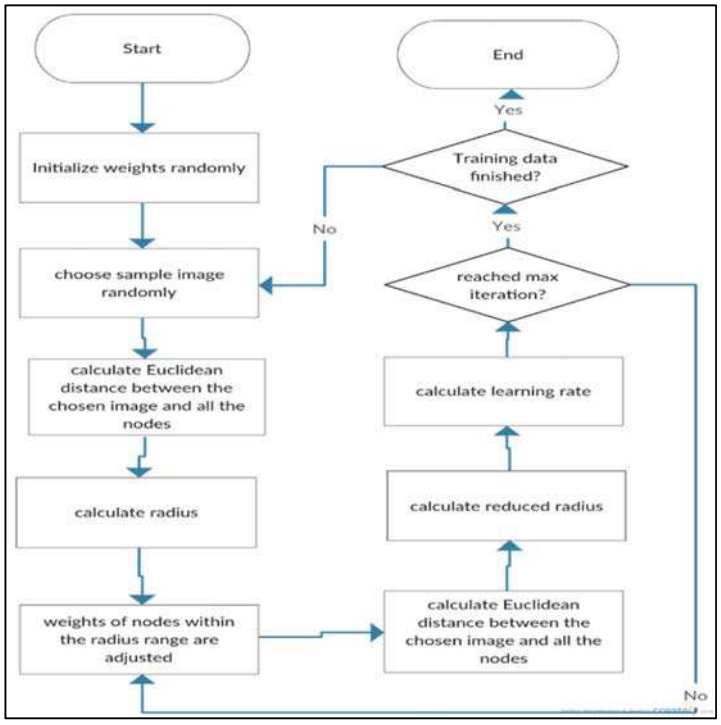


Figure 11: Flow chart of KNN algorithm

4 IMPLEMENTATION AND EXPERIMENT

This section explains mainly two parts of the proposed approach. The first one concerning the Google Android application and the parts that are implemented on the mobile device and that implemented on the cloud. The second part is concerning experimenting the Arabic handwritten recognition approach and measure its accuracy.

A. Implementation of the Google Android application

The mobile app design is based on simplicity as shown in Fig. 12. The figure shows the first start up activity, home activity. The user has two options; either to select camera button or view his history. Then, two different options are available, capturing any Arabic text or Arabic address. By choosing any of the options the camera is launched and user is ready to capture Arabic handwritten text.

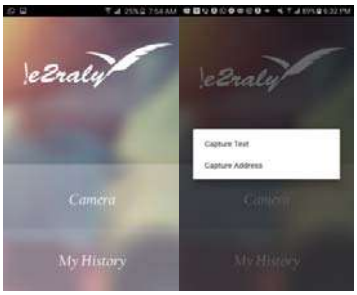


Figure 12: Home screen of mobile app

All captured images are automatically saved in one of the android device folders. Camera data (image) is converted to bitmap image, and data is downsized. We do prefer the second solution, to preserve data.

1) Preprocessing

The preprocessing phase is implemented on the mobile device using its limited capabilities. It is implemented using the OpenCV tool libraries. Implementing OpenCV functions on the android devices exceeds time limits required for any real-time application. The main thread of the android operating systems can't bear such high processing. We first created another thread to solve this problem, but another problem appears concerning the memory limitations, that forced the application to stop. Our tried second solution was to divide functions into independent sets. After the execution of each set, the memory is reset. This solution causes the pre-processing stage to take 2.5 to 3.5min. The preprocessing stages that are run on the mobile device are: Converting Image to grey scale, rotating image if it is landscape, sharpening image and converting it to binary. Fig. 13 shows the image scanned using the mobile app and the image after applying the preprocessing phase on it.

Figure 13: Scanned image and image after preprocessing

2) Handwritten recognition

The other phases of the system, features extractions, segmentation and classification, is done on IBM Blue Mix Cloud [33], it provides enough memory size and faster processing unit than the Mobile Device. Images are encoded to a smaller size before sending them to IBM Blue Mix cloud. We used base64 encoding Algorithm [34] to upload images on cloud faster. Images are sent to cloud using HTTP request. On Blue Mix we used java servlet class to get the image strings and decode them to an array of bytes.

After uploading images on the cloud, the recognition process starts. The process returns the string to the servlet class, the servlet class encode the string as JSON Object and return it to the mobile app, the mobile application takes the JSON Object and decodes it to string the return it to the main running thread.

3) Implementing SDK using recognized text

Users have two options, either to save recognition results to their history or translate the captured recognized text. Saving function is done if the user has an account, while the translate function is connected to Google server with the API key. The API key is requested from Google API's services.

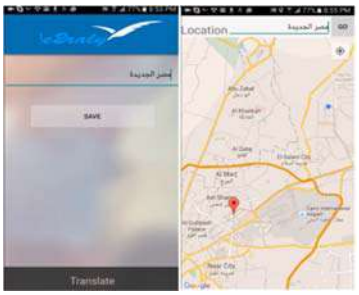


Figure 14: Running Google map app with the recognized word

Another option is capturing a handwritten address, recognizing it automatically; same steps as mentioned before, then redirecting text to Google map activity. With the help of Google maps and GPS, the user can get the location history of his captured address. The application also draws the path from user current location to the captured location then display the map to the user, as shown in Fig. 14.The user can access the saved data at any time by clicking on my_History button in the Home Activity, as shown in Fig.15 and by clicking on one of the list it will appear the larger image and the full text.

Figure 15: Recall saved image with recognized text

B. Arabic Handwritten recognition accuracy

Our proposed solution is building a hybrid classifier based on KNN and SVM. We first tried running each classifier alone on INF/NET database set (B), and we achieved 74.74% accuracy from KNN, and achieved 81.99 % from SVM as shown in Table(1). Then we cascaded both classifiers together and achieved 83.04% by using KNN 3 votes, which is the highest accuracy we've achieved from applying 100 votes and the results as shown in Fig. 16, the graph shows how the percentage increases and decreases according to number of votes taken from KNN

TABLE I
RECOGNITION ACCURACY BASED ON DIFFERENT ALGORITHMS

SVM	KNN
81.99%	74.74%

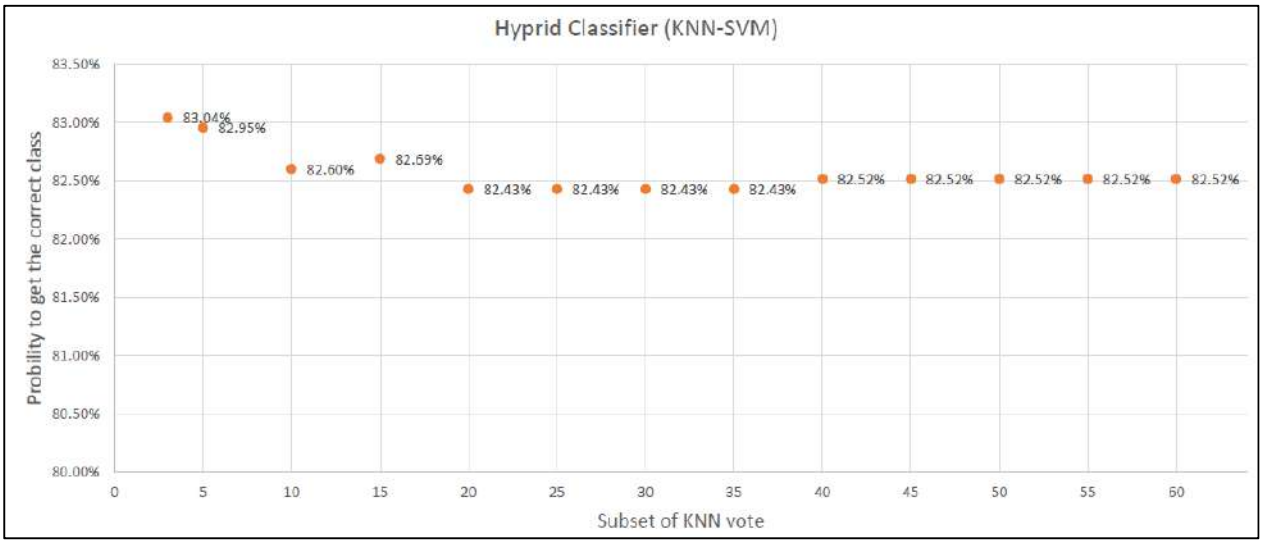


Figure 16: the relationship between KNN votes and the accuracy

5 CONCLUSION

So summing up our proposed System is a hybrid system that serves the Arabic language by first recognizing the handwriting then transforming it into editable text. The recognition system will be executed on IBM Blue Mix cloud instead of executing it on the mobile to increase the run time. On the mobile side after extracting the text, two options can be applied on it. Whether to translate it or if the captured image is address the user can use the maps option to allocate it from his location.

REFERENCES

[1] A. Kamilaris and A. Pitsillides, "Mobile Phone Computing and the Internet of Things: A Survey," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 885-898, 2016.

[2] A. Botta, W. Donato, V. Persico, and A. Pescapé, "Integration of Cloud computing and Internet of Things" *Future Generation Computer Systems*, vol. 56, no. C, pp. 684-700, 2016.

[3] P. Ahmed and Y. Al-Ohali, "Arabic Character Recognition: Progress and Challenges " *Journal of King Saud University*, vol. 12, pp. 85-116, 2000

[4] Rishi Apps. (2017) .OCR Text Scanner (Version 1.6.9) [Mobile application software]. Available from: https://play.google.com/store/apps/details?id=com.offline.ocr.english.image.to.text.pro&hl=en (accessed on 20 August 2017).

[5] Pulsar studio. (2015). Image To Text (Version 1.5.2)[Mobile application software]. Available from: https://play.google.com/store/apps/details?id=ngapham.com.vnocr&hl=en (accessed on 20 August 2017).

[6] Khorsheed, M. S. (2002). "Off-Line Arabic Character Recognition – A Review." *Pattern Analysis & Applications* 5(1): 31-45

[7] J. H. Alkhateeb, J. Ren, J. Jiang and H. Al-Muhtaseb, "Offline Handwritten Arabic Cursive Text Recognition Using Hidden Markov Models and Re-ranking", *Pattern Recognition Letters*, vol. 32pp. 10811088, (2011)

[8] Ashraf AbdelRaouf, Colin A. Higgins, Tony Pridmore and Mahmoud I. Khalil. "Arabic character recognition using a Haar-Cascade Classifier approach (HCC)", *The Pattern Analysis and Applications Journal (PAA)*, May 2016, Volume 19, Issue 2, pp 411-426, first online: April 2015, DOI: 10.1007/s10044-015-0466-2.

[9] Ashraf AbdelRaouf, Colin A. Higgins, Tony Pridmore and Mahmoud Khalil. "Building a Multi- Modal Arabic Corpus (MMAC)" *The International Journal of Document Analysis and Recognition (IJDAR)* 13(4): 285-302, DOI: 10.1007/s10032-010-0128-2,(2010),

[10] Gillies, A., Erlandson, E., Trenkle, J., & Schlosser, S. (1999, April). Arabic text recognition system. In Proceedings of the Symposium on Document Image Understanding Technology (pp. 253-260).

[11] H. A. Al-Hamad and R. Abu Zitar, "Development of an Efficient NeuralBased Segmentation Technique for Arabic Handwriting Recognition", *Pattern Recognition*, vol. 43, no. 8, pp. 27732798, (2010).

[12] Lawgali, Ahmed. (2015). "A Survey on Arabic Character Recognition. International Journal of Signal Processing", *Image Processing and Pattern Recognition*. 8. 401-426. 10.14257/ijsip.2015.8.2.37.

[13] D. Xiang, H. Yan, X. Chen and Y. Cheng, Offline Arabic Handwriting Recognition System Based on HMM, In 3rd IEEE International Conference on Computer Science and Information Technology, vol. 1, pp. 526 529, (2010).

[14] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 712-724, May 2006.

[15] A. Lawgali, M. Angelova and A. Bouridane, A Framework for Arabic Handwritten Recognition Based on Segmentation, International Journal of Hybrid Information Technology, vol. 7, no. 5, (2014), pp. 413428.

[16] O. Surinta , M. F. Karaaba, Lambert R.B. Schomaker, M. A. Wiering "Recognition of handwritten characters using local gradient feature descriptors", Engineering Applications of Artificial Intelligence 45 ,405–414 (2015).

[17] David G. Lowe , Distinctive image features from scale-invariant keypoints , International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition (CVPR), 2005 The IEEE Computer Society Conference on, pages 886–893 vol. 1, Jun 2005.

[19] Abandah, G. A., Younis, K. S., & Khedher, M. Z. (2008, February). Handwritten Arabic character recognition using multiple classifiers based on letter form. In Proceedings of the 5th IASTED International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA'08) (pp. 128-133).

[20] M. E. Hussein, M. Torki, A. Elsallamy, and M. Fayyaz , ALEXU-WORD: A NEW DATASET FOR ISOLATED-WORD CLOSED-VOCABULARYOFFLINE ARABIC HANDWRITING RECOGNITION, Dec 27, 2014, Computer and Systems Engineering Department, Faculty of Engineering, Alexandria University

[21] Haider A. Alwzwazy1 , Hayder M. Albehadili2 , Younes S. Alwan3 , Naz E. Islam , Handwritten Digit Recognition Using Convolutional Neural Networks , International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 2, February 2016

[22] Mittal M, Sharma RK, Singh VP. Validation of K-means and Threshold based Clustering Method. International Journal of Advancements in Technology. 153–60 , Vol. 5 No. 2 (July 2014)

[23] The Scientist & Engineer's Guide to Digital Signal Processing (1997) by S. W. Smith, pp. 277-284.

[24] T. Y. Zhang ,C. Y. Suen , A fast parallel algorithm for thinning digital patterns , Communications of the ACM, v.29 n.3, p.239-242, March 1986

[25] Haraty, R. A., & Ghaddar, C. (2004). Arabic text recognition. Int. Arab J. Inf. Technol., 1(2), 156-163.

[26] Gupta, S., Palsetia, D., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2014, May). A new parallel algorithm for two-pass connected component labeling. In Parallel & Distributed Processing Symposium Workshops (IPDPSW), 2014 IEEE International (pp. 1355-1362). IEEE.

[27] Zeki, A. M., Zakaria, M. S., & Liong, C. Y. (2013). Segmentation of Arabic Characters: A Comprehensive Survey. Technology Diffusion and Adoption: Global Complexity, Global Innovation. IGI Global, 251-288

[28] Richard Szeliski. " Computer Vision - Algorithms and Applications", Springer-Verlag London, pp 235-271, DOI: 10.1007/978-1-84882-935-0. ,2011

[29] Eranna, K, and D Girishkumar. "Dimentional Object Extraction by Using Color Feature and KNN Classification." International Journal of Engineering Research & Technology 3.10 (2014).

[30] C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121–167, 1998.

[31] N. Cristianini and J. Shawe-Taylor. Support Vector Machines. Cambridge University Press, 2000

[32] J. H. Alkhateeb, F. Khelifi, J. Jiani and S. S. Ipson, A New Approach for Off-line Handwritten Arabic Word Recognition Using KNN Classifier, In IEEE International Conference on Signal and Image Processing Applications, pp. 191194. (2009)

[33] Cloud Platform: Cloud Infrastructure - IBM Bluemix. N.p., n.d. Web. <https://www.ibm.com/cloud-computing/bluemix/>.(accessed on 17 March 2017).

[34] Baldwin, R. G. (2014, July 27). Understanding Base64 Data. Retrieved from http://www.developer.com/java/other/article.php/3386271/Understanding-Base64-Data.htm

BIOGRAPHY

Nada A. Shorim Graduated from the Faculty of Computer Science, Misr International University (2016) with Honors.

Doing pre-masters in Software Engineering at Faculty of Computer Science, Cairo University. Now she is a teaching assistant in Faculty of Computer Science at Misr International University.

Norhan M. Eltopgy received a bachelor degree in computer science 2016 from Misr International University. Certified from Microsoft app factory assembly 5-1 line program, Egypt

Sahar K. Mohamed Graduated from the faculty of computer science at Misr International University 2016. Had an oracle certificate in (oracle Database, PL/SQL Forms, and reports). Now working as an oracle developer at Allianz Egypt.

Shehab Salah Graduated from the Faculty of Computer Science, Misr International University (2016). Android Developer certified from Google and holds Android Nano degree certificate from Udacity. Worked on some famous Android Mobile Applications such as Al-Tayyar Online, Hindawi Kotob, Nagwa Mathematics, Nagwa Studio, and Nagwa Tutoring. Creator of Vodafone sports portal (stadvodafone.com), and founder of ezfunction E-Learning. Now he is a Technical Lead at Appcorp for value-added services and Senior Android Developer at Nagwa E-Learning.

Taraggy M. Ghanim Finalizing her Ph.D. studies in Arabic Handwriting Recognition, from Ain Shams University, department of Computer Engineering. Graduated in 2006 and achieved her MSc. in 2012, Faculty of Computer Engineering, University of Ain shams, Cairo, Egypt. She worked as a teaching assistant then as an assistant lecturer in faculty of Computer Science, Misr International University. Her research interest is pattern recognition, image processing, and Bioinformatics.

Ashraf M. AbdelRaouf Achieved his PhD in 2012 from the School of Computer Science at the University of Nottingham, Nottingham, UK. Graduated from the Faculty of Engineering 1988. He studied a Computer Science Diploma in 1990 at the American University in Cairo (AUC). He worked in the software and IT industries. Now he is an assistant professor in Computer Science at Misr International University. In the IT business industry, he was working as a Chief Operating Officer (COO) at Cloudypedia, Egypt. Cloudypedia is a premium Google business partner. His research interest is pattern recognition specifically in character recognition, natural language processing, image processing, Bioinformatics, Medical imaging, Arabic linguistics and morphology. Other research interests include Data Structures and algorithms, computer graphics. He is an IEEE senior member since 2015.

التعرف على الكتابة العربية اليدوية باستخدام إنترنت الأشياء مع الحوسبة السحابية

ندى ايمن شريم[1]*, نورهان محمد التبجي[2]*,سهر خالد محمد[3]*,شهاب صلاح[4]*, ترجي محي غانم[5]*,أشرف محمد عبدالرؤوف[6]*

كلية الحاسبات و المعلومات, جامعة مصر الدولية, القاهرة, مصر

[1]nada.ayman@miuegypt.edu.eg, [2]Norhan120664@miuegypt.edu.eg, [3]sahar122284@miuegypt.edu.eg, [4]shehabsalah25@gmail.com,[5]taraggy.ghanim@miuegypt.edu.eg, [6]ashraf.raouf@miuegypt.edu.eg

ملخص -- التعرف على الكتابة بخط اليد للغة العربية هو تحد كبير بسبب الاختلافات في شكل الاحرف وفقا لموقعها في الكلمة وأيضا اختلاف أساليب الكتابة اليدوية. إنترنت الأشياء هو الربط بين الشبكات مع الأجهزة المتصلة بها والتي تحتوي على الالكترونيات والبرمجيات وأجهزة الاستشعار لجمع وتبادل البيانات. الحوسبة السحابية وإنترنت الأشياء يعملان على حد سواء لزيادة كفاءة الاستخدامات اليومية وعلاقتهما ببعض علاقة تكميلية. إن التعرف على اللغة العربية المكتوبة بخط اليد باستخدام جهاز محمول متصل بالحوسبة السحابية يسهل الترجمة لغير الناطقين بالعربية وإيجاد الاماكن المطلوبة على الخرائط وذلك يعتبر ذو أهمية كبرى لهم أثناء زيارة البلدان الناطقة باللغة العربية. نهجنا في هذا البحث هو بناء التطبيق على الجهاز المحمول الذي يقدم المصنف متعدد المراحل الهجين الذي يعمل على الميزات الهندسية للكلمات. المصنف الاساسي لدينا هو الشبكة العصبية (KNN)، ثم تمرير مجموعة من أقرب النقط المجاورة الى (SVM). ولقد استخدمت في مرحلة التدريب مجموعة البيانات التي تم إنشاؤها ذاتيا وتم اختبارها على قاعدة بيانات (IFN/ENIT). نهجنا يحقق دقة في العمل تساوي 83.04٪.

الكلمات الدلالية:- العربية، التعرف على نص مكتوب بخط اليد باللغة العربية، المصنف الهجين، استخراج الميزات، إنترنت الأشياء، الحوسبة السحابية

Mood Miner: Sentiment Mining of Financial Market

Hany Mohamed[*1], Ayman Atia[**2], Mostafa Sami[*3]

[*]*Faculty of Computer Science, Helwan University, Egypt*
[1]hany.abdelmawgood@its.ws

[**]*Faculty of Computer Science, Misr International University, Helwan University, HCI-LAB, Egypt*
[2]ayman@fcih.net

[***]*Faculty of computer science, Helwan University, HCI-LAB, Egypt*
[3]mostafa.sami@hotmail.com

***Abstract*—the sentiment classification of social content in relation to financial market has received an increasing interest from research community. Classification tells whether sentiment is bullish or breach, helping traders to check sentiment before trade decision. Machine learning (ML) techniques play a great rule for detecting sentiment of social content. In this research, our investigation leads that selection of inappropriate classifiers' features address issues in sentiment detection and reduces classification efficiency. This paper addresses the problem of sentiment classification accuracy by selecting appropriate features from training dataset using genetic algorithms, approved that it can provide improvement in classification accuracy(89.9%) rather than support vector machine technique(88.6%).This paper also shows that considering reputation of social content users has a high impact on evaluating overall sentiment score and thus gives a direction for financial trading, proved trading success factor raise from 63% to 82%.**

Keywords: Social Network, Sentiment mining, NLP, GENETICS.

1 INTRODUCTION

The World Wide Web has brought change to a point where it would be difficult to imagine a world not connected through online networks [1].Currently, people are being able to be interactive and active author for digital content through social media. People are get used to express their sentiment in social media towards daily events in different life areas whether sports, political, and so on. In [1], [2], [3], it is shown that if these sentiments well analyzed, it could help investors and decision makers. Sentiment analysis is defined as the process of detecting meaning or emotion or opinion of user's statement towards daily activities. Sentiment has clear impact against country economy or product listed in financial market. For instance, traders will have bad impression for currency if jobless report related to this currency is high. The same if a listed company declares increase in yearly profit, the traders think investing in it with expectation that unit price will be high. Sentiment analysis through social media is our scope of interest from the intention to help investors know how the trend is going for different capital investments whether currency pairs, commodities, and indices. The trend or sentiment in financial market takes either Bullish or Breach. Bullish means that investors interested to buy while breach means investors need to hold out from their investments. Different approaches in sentiment classification area are presented in [21][22][23][24], which is categorized based on lexical analysis or ML(Machine learning) techniques. Support vector machine (SVM) algorithms are used to perform best accuracy in sentiment analysis [7]. Sentiment analysis is marked as a text classification problem. Text classification is different than classification in other domains due to large number of features. Most techniques are depending on bag of word model to generate unique words. In social media, Most of generated features are irrelevant, redundant or noisy [27], this is because of users that may use or coin new abbreviations or acronyms, "It is cooooooool", "OMG :-(", are intuitive and popular in micro blogging, but considered as noisy features for text classification model due to informal words [29], and this is highlighted as main issue to be investigated through this research.

In machine learning, solutions are created to some problems by using training data, some search method is used to search over a class of candidate solutions to find an effective one called genetic algorithms (GA) [14]. GA is an algorithm which makes it easy to search in large space, by implementing the Darwinian selection to the problem. Only the best solutions will remain, thus narrowing the search space [12]. Standard steps of any GA technique starts by initializing encoded population as 'genes'. New solutions can be produced by 'mutating' members of the current population [14].Then Fitness functions are computed to select best solution. In our research we show that applying genetics can advance ML classification for sentiment and can resolve the issue of noisy feature selection.

At another point, sentiment detection is highly impacted by online reputation score which measures influence of brand or user reputation in social media [26]. In general, business utilizes such tools to know which areas should be improved. A large number of popular tools exist to measure score: Klout, Scale, My Web Career, and Peer Reach are samples. Detailed techniques in that area are out of scope.

This research proposes a system for mining social media to analyze sentiment and predict trends in financial market business domain. The main contribution of this paper is a new technique of mining sentiment by applying genetics to consider selected features of Training data in order to prepare accurate model. Once sentiment detected, our system will calculate bullish or breach score considering reputation of social network users that direct investors taking trade decision.

The rest of this paper is organized as follows, Section II discuss popular approaches in emotion detection area from lexical to machine learning approaches. While Section III presents our proposed method. Section IV shows results and comparative study. Finally, conclusion and future work listed in section V.

2 PAGE LAYOUT BACKGROUND AND RELATED WORK

The development of lexical resources for sentiment analysis has attracted attention of the computational linguistic community [8]. Although human effort needed to build corpus, they play great rule of detecting emotion through effective words. WordNet-Affect [9], Opinion Finder [10], Sentiword-net [11] are samples of syntactic-level resources for sentiment analysis. Table 1 shows examples of synsets representation in wordnet. WordNet-Affect is an extension of WordNet, in which effective concepts representing emotional state are individuated by synsets marked with thea-label emotion [9]. Simplicity in implementing such approaches is one advantage based on counting of Positive or negative words. However, lack of handling non grammatical linguistic is a disadvantage.

TABLE I
SYNSETS REPRESENTATION IN WORDNET

Labels	Examples
emotion	noun anger#1, verb fear#1
mood	noun animosisy#1, adjective amiable#1
trait	noun aggressiveness#1, adjective, competitive#1
cognitive,state dazed#2	noun confusion#2, adjective
physical,state	noun illness#1, adjective all in#1
hedonic,signal	noun hurt#3, noun suffering#4

Another traditional approach is called semantic orientation (SO). The main idea of SO is to calculate Semantic of each word based on the difference between its associations with positive and negative words. Point wise Mutual Calculation (PMI) is used to calculate based on equation 1:

$$PMI(x;y) = \log \frac{P(x,y)}{P(x)P(y)} = \log \frac{P(x|y)}{P(x)} = \log \frac{P(y|x)}{P(y)} \quad (1)$$

Where x, y refers to terms of feature distribution. Here, each word is defined based on percentage of its relation with position or negative emotion, Finally SO is calculated using equation 2:

$$SO(t) = \sum_{t' \epsilon V+} PMI(t,t') - \sum_{t' \epsilon V-} PMI(t,t') \quad (2)$$

An advantage of this technique is simplicity but handling negation and non-natural language considered high limitation. NAIVE BAYES, Maximum Entropy, and SVM are popular ML techniques that are used in sentiment classification due to accuracy. In [10], authors provide detailed comparison between different techniques, while they use Unigram, Bigram and Parts of speech as features to compare between different ML techniques. Their research lead that SVM has best accuracy (81.2%) based on Unigram features. However it lacks handling of negation because of relying on individual words.

Accuracy is calculated based on fitness relation F1 method, it is a harmonic mean of the precision and recall scores specially used in machine learning and information retrieval [11] as shown in equation in 3.

$$P(Precision) = \frac{TP}{TP+FP}$$
$$R(Recall) = \frac{TP}{TP+FN} \quad (3)$$
$$F1 = \frac{2.P.R}{P+R}$$

Precision is defined as the probability that (randomly selected) retrieved document is relevant. While Recall is the probability that (a randomly selected) relevant document is retrieved in a search.

3 PROPOSED FRAMEWORK

A. Step 1 : Crawling

Our research focuses on financial market as a case study for analyzing sentiment. System starts with collecting posts for specified capital investments that is manually annotated from Tradebird [13]. Tradebird is driven by a passionate community sharing trading ideas, news and opinions about the markets in real time. Sample of collected posts shown in figure 1, while investors express their opinion/expectation for EURUSD after critical event hold in USA for FED in subject related to interest rate. Rebroadcasting is removed from collected dataset. Each record contains the following: Author Name - No. of Followers - Tags Time – Post - Sentiment. Our system will work only on posts in English language only.

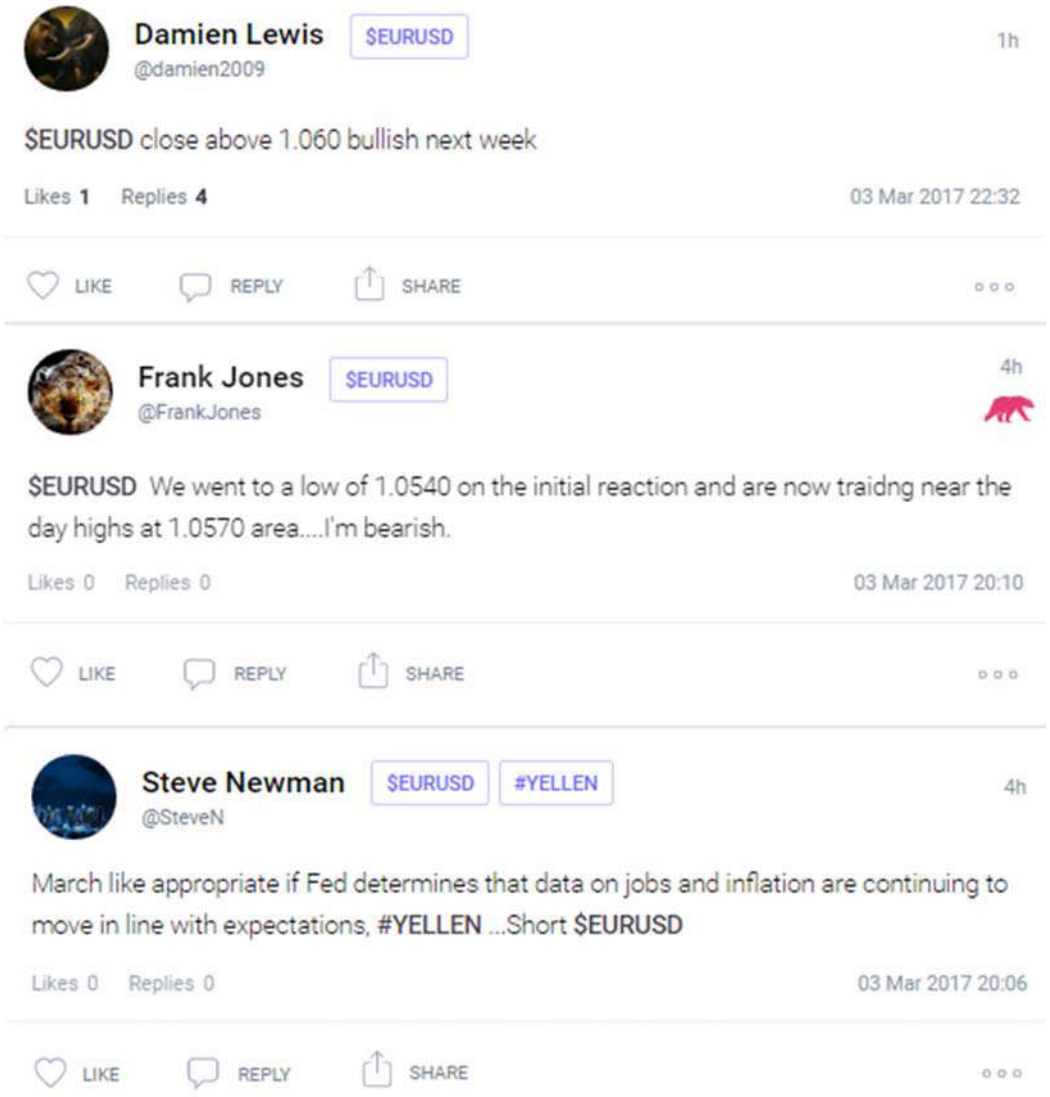


Figure 1Sample of Tradebird posts

B. Step 2: Text processing

Captured text is analyzed by common NLP (Natural language processing) techniques. We implement NLP techniques using python technology and its library Natural language toolkit (NLTK). NLTK is a leading plat form for building Python programs to work with human language data. It provides more than fifty corpora and lexical resources such as WorldNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum[28].The following steps are considered:

- **Stop words:** Removing Sarcastic words is important to avoid wrong predication or model building," . ?% $" are examples of stop words.
- **Stemming:** A processing interface for removing morphological affixes from words. For better accuracy, we use Python Lancaster Stemmer which is based on the Lancaster stemming algorithm.
- **Bag-Of-Words modeling:** Bag-Of-Words model is a representation in NLP for text, used to model frequencies or number of occurrence for each word in document.
- **TF-IDF:** This will be used to calculate weight of each word based on frequency in different posts, standard calculation is shown in equation 4.

$$idf_i = log(\frac{N}{n_i}) \qquad (4)$$

where N is number of posts collected, and n_i is number of posts that contain specified word.

C. Step 2: Model preparation

Feature extraction: In this step, we pick a list of words or features generated from previous step as an input for proposed prediction model. Removing noise features with high entropy build accurate and stable classification model.

Removing noise features: Features with high entropy build accurate and stable classification model. In our system, we pick high discriminative features generated from previous step and this can be applied using genetics. There exist two important aspects here for feature evaluation in general, Entropy and Information gain. Entropy is a common way in information retrieval area to measure impurity, while impurity refers to class distribution within dataset, High impurity leads to high classification accuracy. Entropy is calculated as the following:

$$Entropy = \sum -P_i Log_2 P_i \qquad (5)$$

Where Pi is the probability of class i, the higher Entropy leads to better accuracy and high information content. The next step is to check which of extracted features are considered the most important in our classification problem. IG (Information Gain) is a common way of doing this task. Standard equation is asfollowing:

$$IG = entropy(parent) - [averageentropy(children)] \qquad (6)$$

Based on GA, we apply the filtering mechanism
- Chromosome consists of list of features captured from training data set.
- Construct best mixture of best features lead to high accuracy, which is objective of our technique.
- Fitness refers to accuracy of predicted model, calculated as mentioned in equation 3.

Our proposed genetic algorithm for feature selection is the following:

1. Pick high repeated features based on bag of words techniques
2. Construct chromosome representation of features (x1, x2,....xn)
3. Compute fitness of constructed chromosome
4. Mutate other feature with lower fitness chromosome
5. Compute fitness of generated chromosome
6. If child has more fitness than parent, then replace parent with child
7. Go to step 3

The above steps can be repeated multiple times according to state of collected dataset and controlled by number of generation parameter specified by user, the high number of rounds leads to better accuracy but impact on running time. In this research, we apply three rounds of filtration to verify proposed method on collected database.

D. Step 3: Evaluating Trade decision

Now, the emotion is extracted per post whether it is bullish or breach or neutral. However, this is not final step that can lead financial market investors whether to buy or sell. As mentioned in introduction section that sentiment classification is impacted by people reputation, people has high rank in financial market let his words has terrible reputation. Reputation measurement has multiple factors, while we consider here only weight of author based on number of followers, and a number of users interacts with his posts as a proof of concept. Handling other factors is considered as future work for our research.

TABLE II
ACCURACY PERCENTAGE FOR DIFFERENT TECHNIQUES

Size/Method	NAIVE	SVM	Proposed
1,000	83.5%	86%	79%
5,000	87%	89.2%	87.5%
10,000	86.4%	88.6%	89.7%

TABLE III
PASSED AND FAILED TRADES ACCORDING TO SCORING APPROACH

	Normal	Proposed
Passed	63	82
Failed	37	18
Total	100	100
Probability	63%	82%

4 EVALUATION

a. Experimental Setup

Sample of published posts taken at 03-Mar-2017 for EURUSD currency pair is shown in figure 3. For evaluation, we present financial dataset composed of 10,000 records that help upcoming researches in evaluating their technique, divided into 80 percent for training and 20 percent for testing.

b. Results

In this section, we compare between three different techniques: NAIVE Bayes, SVM, and our proposed one. Table 2 shows results for different dataset size. As Shown in table 2, our method works better for large data set and get high accuracy (F1) for prediction, while SVM is better for small size of dataset. The main reason of correlation between database size and accuracy is detection of noisy features increased by enlarging dataset size. Next step is to evaluate scoring approach which help investors to take decision based on common sentiment (Bearish or bullish). Table 3 shows result on 100 trades of EURUSD pair currency on March-2017, passed refereed to trades closed with profit while Failed refereed to trades closed with loss. The result shows that our approach increases probability of succeed trading.

5 CONCLUSIONS AND FUTURE WORK

Through our paper, we propose novel technique of evaluating sentiment in social media for financial market business area. Our approach targets achieving more accuracy in model prediction by removing noise features and this can be done using genetics. As shown in table 2, our method has more accuracy for large dataset in addition to succeed trading probability using score approach that considers rank of authors and content reputation. In the future, our technique will be applied on dataset of other areas rather than financial to evaluate efficiency, in addition Big Data techniques to be applied to get high performance.

REFERENCES

[1] Hany Mohamed, Ayman Ezzat, Mostafa Sami, the Road to Emotion Mining in Social Network. International Journal of Computer Applications Pages 41-47, August 2015.
[2] Darren Quinn, Liming Chen, Maurice Mulvenna, Social Network Analysis: A survey, International Journal of Ambient Computing and Intelligence, Volume 4 Issue 3, Pages 46-58, September 2012.
[4] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka, Compositionality Principle in Recognition of Fine-Grained Emotions from Text, Proceedings of the Third International ICWSM Conference, 2009.
[5] Stuart Koschade, A Social Network Analysis of Jemaah Islamiyah: The Applications to Counter terrorism and Intelligence, Journal, Studies in Conflict & Terrorism, Issue 6, Volume 29, 2006.
[6] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining strengths, emotions and polarities for boosting twitter sentiment analysis, in: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM, ACM, New York, USA, 2013.
[7] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford, Technical report, Stanford Digital Library Technologies Project, 2009.
[8] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, Siddharth Patwardhan, Opinion Finder: A system for subjectivity analysis, HLT-Demo '05 Proceedings of HLT/EMNLP on Interactive Demonstrations, Pages 34-35, October 2005.
[9] Strapparava and A. Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon.
[10] Andrea Esuli and Fabrizio Sebastiani, SentiWord-Net: A High-Coverage Lexical Resource for Opinion Mining, Institute of science and technology,Technical Report 2007.
[11] Catak, F.O.U.: Genetic algorithm based feature selection in high dimensional text dataset classification. In: WSEAS Transactions on Information Sciences and Application, vol. 12, Pages 290-296, 2015.
[12] Amna Asmi, Tanko Ishaya , Negation Identification and Calculation in Sentiment Analysis, The Second International Conference on Advances in Information Mining and Management, IMMM2012.
[13] Weihui Daia, Dongmei Hanb, c, Yonghui Daib, Dongrong Xud, Emotion Recognition and Affective Computing on Vocal Social Media, Volume 52, Issue 7, Pages 777-788, November 2015.
[14] Elvis Saravia, Carlos Argueta, Yi-Shin Chen, EmoViz: Mining the World's Interest through Emotion Analysis, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
[15] Olivier Janssens1, Maarten Slembrouck, Steven Verstockt, Sofie Van Hoecke, Rik Van de Walle, Real-time Emotion Classification of Tweets, ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.
[16] Bin Liua, Jingyuan Zhangb, Qiang Liuc, Han Lid, Mingliang Zhange, Rui Qiuf, Jingyang Zhaog, Reading-Weibo: A Sina Weibo Oriented Data Mining System, International Industrial Informatics and Computer Engineering Conference, 2015.
[17] Aamera Z. H. Khan, Mohammed Atique, V. M. Thakare, Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis, National Conference on Advanced Technologies in Computing and Networking ATCON-2015.
[18] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu, Empa Tweet: Annotating and Detecting Emotions on Twitter, Human Language Technology Research Institute University of Texas at Dallas.
[19] J. Shapiro. Genetic Algorithms in Machine Learning: Machine Learning and Its Applications, Advanced Lectures. Pages 146-168, 2001.
[20] Shivhare, S. N., Khethawat, S., Emotion detection from text, International Journal of Engineering Research and Development , Vol. 11, Pages 23-34, 2012.
[21] Bollen, J., Pepe, A., Mao, Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. ICWSM11, Barcelona, Spain, 2011.
[22] Gerald Petz Micha , Karpowicz, Opinion Miningon the Web 2.0 Characteristics of User Generated Content and Their Impacts, Springer, Pages 35-46, volume 7947, 2013 .
[23] Lee H, Choi YS, Lee S, Park I, towards unobtrusive emotion recognition for affective social communication. In: Consumer communications and networking conference (CCNC), IEEE,IEEE, Pages 260-264, 2012.
[24] Daniel Preotiuc- Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. Trendminer: An architecture for real time analysis of social media text. In Proceedings of the ICWSM Workshop on Real-Time Analysis and Mining of Social Streams, Dublin, Ireland, 2013.
[25]A tool for measuring SMEs' reputation, engagement and goodwill: A New Zealand exploratory study, 2017.
[26] Buck, Christian, Kenneth Heafield, and Bas Van Ooyen., N-gram counts and language models from the common crawl., Proceedings of the Language Resources and Evaluation Conference, 2014.
[27] Bird, S., and Loper, E. NLTK: The natural language toolkit. In Proc. of 42nd Annual Meeting of the Association for Computational Linguistics,2004.

[28] Hu, X., Tang, L., Tang, J., Liu, H., Exploiting social relations for sentiment analysis in micro blogging, In Proceedings of the sixth ACM international conference on Web search and data mining. Pages 537-546, 2013.

BIOGRAPHY

Hany Mohamed is PhD student at Faculty of Computers and Information, Helwan University, Egypt. He received his master degree from Arab academy, Egypt 2008. He received his BSc. degree from the department of Computer Science, Zagazig University, Egypt 2003.

Ayman Atia is Assistant professor at HCI-LAB, Department of Computer Science, Faculty of Computers and information systems, Helwan University, Egypt. He received his PhD from the University of Tsukuba, Japan 2011. He received the BS and MS degrees from the Department of Computer Science, Helwan University, Egypt in 2000 and 2004, respectively. He is a current member of the IEEE Computer Society. Dr. Ayman is Co-Founder of HCI-LAB, and head of the interaction group. His work includes finding new interaction techniques for large display and finding abnormal behaviour for driving vehicles, theft detection and software engineering frameworks.

Mostafa Sami M. Mostafa is a professor of Computer Science**,** Faculty of Computers and Information, Helwan University. Ex-Dean of Students Affairs; Ex-Chairman of Computer Science Department and member of the HCI research lab in the same faculty. Ex-Dean of Faculty of Information Technology, MUST University, 6th of October, Egypt. He has been graduated 1967 as computer engineer from Military Technical College, Cairo, Egypt, then joined the teaching staff in the same institution as teaching assistant. He has received his Master of Science (1978) and Philosophy Doctorate (1980) degrees from University of Paul Sabatier, Toulouse, France. He had supervised and awarded more than 65 MSc. and 20 PhD theses. He has more than 65 publications in different conferences and journals. His major interesting research tracks are System Modelling and Design, Software Engineering, Cloud Computing, Wireless Sensor Network and Bio-medical Computing.

التنقيب عن الرأي الغالب في مواقع التواصل الاجتماعي المتعلق بالاسواق المالية

*[1]هاني محمدو **[2]أيمن عطية و ***[3]مصطفي سامي
*** كلية الحاسبات والمعلومات – جامعة حلوان*
**** كلية الحاسبات والمعلومات – جامعة مصر الدولية – جامعة حلوان*
***** كلية الحاسبات والمعلومات – جامعة حلوان*

[1]hany.abdelmawgood@its.ws
[2]ayman@fcih.net
[3]mostafa.sami@hotmail.com

ملخص

يتلقي تصنيف الرأي العام علي مواقع التواصل الاجتماعي فيما يتعلق بالسوق المالية اهتماما متزايدا في البحث العلمي. ويفسر التصنيف ما إذا كانت المشاعر إيجابية أو سلبية، مما يساعد رواد الاعمال للتحقق من الاراء قبل إتخاذ أي قرار. ويقوم مجال تعلم لغة الاله بدور للكشف عن تلك الاراءفي محتوي التواصل الاجتماعي. ويقوم هذا البحث بدور كبير في معالجة وتصنيف الرأي الغالب بشكل ادق ومعالجة المشاكل الناتجة من ذلك التصنيف. وتقوم فكرة هذا البحث علي طريق اختيار الخصائص المناسبة من مجموعة بيانات التدريب باستخدام الخوارزميات الجينية، حيث تم التفحص على أنها يمكن أن توفر تحسنا في دقة التصنيف (89.9٪) بدلا من (88.6٪)، وتم التوصل في هذا البحث ايضا الي ان الاستعانه بمدي شهرة المستخدمين علي مواقع التواصل الاجتماعي له تأثير كبير على تقييم النتيجة النهائيه للتصنيف، وبالتالي إعطاء اتجاه للتداول المالي، وقد أثبتت ارتفاع نسبة التداول من 63٪ إلى 82٪.

الكلمات الدلالية: الشبكات الاجتماعية، التنقيب عن الرأي العام، علم الوراثه، معالجة اللغة

SimAll: A Flexible Tool for Text Similarity

Wael H. Gomaa[*1], Aly. A. Fahmy[**2]

[*]*Computer Science Department, Faculty of Computers and Information, Beni-Suef University, Egypt*

[**]*Computer Science Department, Faculty of Computers and Information, Cairo University, Egypt*

[1]wael.goma@fcis.bsu.edu.eg

[2]aly.fahmy@cu.edu.eg

***Abstract*—Measuring the similarity between texts is an essential component in various natural language processing tasks. Many text similarity tools are available in different manners and techniques. These techniques include string, corpus and knowledge similarity approaches. The goal of our new tool SimAll is to combine the features of each approach in a single tool. Different types of preprocessing modules, fusion methods and similarity levels are included. SimAll supports both Arabic and English languages. Case studies of using SimAll in short answer grading task are also presented in this paper.**

***Keywords:*Text Similarity, Semantic Similarity, Short Answer Grading.**

1 INTRODUCTION

The use of text similarity plays an increasingly important role in natural language processing (NLP) tasks such as short answer grading, automated essay scoring, questions generation, question answering, information retrieval, text classification, document clustering, topic detection, topic tracking, machine translation, text summarization and others. Measuring similarity between pair of words is a basic part of text similarity which is then used as a primary level for sentence, paragraph and document similarities. Existing work on determining text similarity is broadly classified into three major groups: string-based, semantic-based and hybrid-based [1]. String-based approach operates on string sequences and character composition to compute similarities and can be categorized into two groups: character-based and term-based. Words are similar semantically if there is any semantic relation between them like antonym, homonym, polysemy and synonymy. Semantic similarity includes two different approaches: corpus-based and knowledge-based algorithms. Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. Knowledge-Based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. Hybrid-based approach combines multiple similarity algorithms from the previously explained methods. An excellent and more detailed overview of text similarity measures can be found in [1]. Collecting these different similarity algorithms into a single tool is a challenge. In this paper we introduce SimAll tool; the basic idea of SimAll is to define abstractly a unified method for all text similarity algorithms. The tool focuses on applying multiple similarity measures separately and in combination. Different types of preprocessing modules like stop word removing and stemming are included for both Arabic and English languages. Similarities between words, sentences, paragraphs and documents are available. We have tested the proposed tool in a short answer grading task.

The remainder of this paper is organized as follows: Section 2 presents the most accurate text similarity tools. Section 3 explains the proposed tool and its features. Section 4 shows the results of applying SimAll tool in short answer grading module, and finally, Section 5 presents the conclusion of the research.

2 TEXT SIMILARITY TOOLS

A. String-Based Similarity

This section presents the most popular string similarity tools SimMetrics, SecondString, SimPack, AlignAPI, SimString and FLAMINGO package.

SimMetrics [2] is an open source extensible java library containing numerous similarity metrics. A similarity metric is an algorithm that considers two input strings and returns a measure of their similarity. Similarity measures come from a variety of disciplines, including statistics, DNA analysis, artificial intelligence, information retrieval, information integration and databases. SecondString [3] is a java library for developing and evaluating a wide assortment of string matching algorithms, those algorithms based on the edit distance technique and other matching algorithms. It also provides tools for combining multiple string-based algorithms. SimPack [4] is a java package for measuring the similarities between concepts in ontologies; it also supports other applications such as source code similarity and hierarchically-structured data similarity. The Alignment API [5] is an implementation for storing, finding, sharing

andimproving alignments; it also includes many services such as generating processing outputs and tests. SimString [6] is a library for finding strings in a database whose similarity with a query string is greater than a given threshold; finding not only identical but similar strings. This task refers to approximate string matching. SimString is implemented using different programming languages like C++, Python and Ruby. FLAMINGO [7] is a C++ tool for finding approximate string matching; it includes different algorithms for approximate selection queries and selectivity estimation. Table I shows the string-based algorithms in each tool.

TABLE I
STRING-BASED ALGORITHMS

Similarity Tools	String-Based Algorithms
SimMetrics	Hamming, Levenshtein, Needleman-Wunch,Smith-Waterman, Gotoh, City Block, MongeElkan,Jaro, Jaro Winkler, SoundEx, Matching Coefficient, Dice, Jaccard, Overlap Coefficient, Euclidean, Cosine, Hellinger, Skew divergence, Tau,tf-idf, N-gram
SecondString	N-gram, Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunch, Monge-Elkan, Jaccard, Cosine w/tf-idf
SimPack	Levenshtein, Jaccard, Dice, Cosine w/tf-idf, City Block, Euclidean, Cosine Overlap
Alignment API	N-gram, Levenshtein, Jaro, Jaro-Winkler, Needleman
SimString	Cosine w/tf-idf, Dice, Jaccard, Overlap Coefficient
FLAMINGO	Cosine w/ tf-idf, Dice, Jaccard, Levenshtein

B. Corpus-Based Similarity

Corpus-Based similarity is a semantic similarity measure that determines the similarity between words according to information gained from large corpora. In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). It is used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. The most widely used corpus-based similarity measures were explained in [1]. In this section, we will focus on only two packages extracting DIStributionally similar words using CO-occurrences (DISCO) [8] and A SEManticsimILARity Toolkit (SEMILAR) [9].

DISCO is a java package that focuses on distribution similarity; Distributional similarity between words assumes that words with similar meaning occur in similar context. Large text collections are statistically analyzed to get the distributional similarity. DISCO has two main similarity measures DISCO1 and DISCO2; DISCO1 computes the first order similarity between two input words based on their collocation sets. DISCO2 computes the second order similarity between two input words based on their sets of distributionally similar words. DISCO supports many languages such as Arabic, Czech, Dutch, English, French, German, Italian, Russian and Spanish. It also contains DISCO Builder which allows creating a database of similar words from a text corpus. Table II shows the word spaces in both Arabic and English languages.

SEMILAR is a java toolkit that offers users, researchers and developers, easy access to fully-implemented semantic similarity measures in a single package through both a GUI-based interface and a library. The core component of SEMILAR is a set of text-to-text semantic similarity methods. All methods are implemented to handle both unidirectional similarity measures as well as bidirectional similarity measures. It has many important features like easy data management, preprocessing, lexical and syntactic feature extraction, visualization and performance reports. Besides automated ways for assessing the semantic similarity of texts, the toolkit offers facilities for manual assessment by experts. This component is called SEMILAT, the SEManticsimILarity AnnotationTool. Table III shows the corpus-based measures in SEMILAR package.

TABLE III
WORD SPACES

Language	Word Space Name	Corpus Size
Arabic	ar-general-20120124	188 million tokens
English	enwiki-20130403-sim-lemma-mwl-lc	1.9 billion tokens
English	enwiki-20130403-word2vec-lm-mwl-lc-sim	1.9 billion tokens

TABLE III
SEMILAR CORPUS-BASED MEASURES

Measure	Model Size
Latent Semantic Analysis (LSA)	Wiki Models From 127 MB to 2.64 GB TASA Models From 60 MB to 185 MB
Latent Dirichlet Allocation (LDA)	17.5 MB
Pointwise Mutual Information (PMI)	1.06 GB

C. Knowledge-Based Similarity

Knowledge-based similarity is a semantic similarity measure that determines the degree of similarity between words using information derived from semantic networks. WordNet is the most popular semantic network in the area of measuring Knowledge-Based similarity between words. Knowledge-based similarity measures can be divided roughly into two groups: measures of semantic similarity and measures of semantic relatedness. Semantically similar concepts are deemed to be related on the basis of their likeness. Semantic relatedness, on the other hand, is a more general notion of relatedness, not specifically tied to the shape or form of the concept. In other words, Semantic similarity is a kind of relatedness between two words. It covers a broader range of relationships between concepts that includes extra similarity relations such as is-a-kind-of, is-a-specific-example-of, is-a-part-of, is-the-opposite-of. There are six measures of semantic similarity; three of them are based on path length: Path (path), Leacock &Chodorow (lch) and Wu & Palmer (wup). The other three measures are based on information content: Resnik (res), Lin (lin) and Jiang &Conrath (jcn) [10]. Knowledge-based measures are available for English language only. Table IV shows the most widely used tools that support knowledge-based similarity measures.

TABLE IV
KNOWLEDGE-BASED MEASURES

Package	Programming Language
WordNet::Similarity [11]	Perl
WS4J [12]	Java
NLTK [13]	Python
SEMILAR [8]	Java

3 SIMALL

The main feature of SimAll is collecting all discussed types of similarity into a single tool. SimAll is implemented using java programing language. 61 algorithms are included; these algorithms are derived from four previously explained packages: SimMetrics, SEMILAR, Disco and WS4J. All similarity measures are normalized to output similarity value between 0 and 1. The other features are applying preprocessing modules, supporting different granularity of similarity methods, combining multiple measures from different similarity categories and enabling users to build their model using different machine learning algorithms.

SimAll supports both Arabic and English languages. Tokenization, part of speech tagging and parsing are available using Stanford NLP package [14]. Khoja's [15] and Porter's [16] stemmers are used for Arabic and English languages respectively. Stop word removing module is also available using predefined stop word lists.

SimAll supports different similarity levels: word to word, sentence to sentence, paragraph to paragraph and document to document. Furthermore, any number of text pairs can be arranged into two column spread sheet to apply any similarity measures separately or in a combination.

Combining similarity measures is applied by defining a new fusion function; the user should configure three steps to create his/ her fusion function. The first step is to choose the similarity algorithm(s) and assign manually a similarity scale or weight for each algorithm; the scale is a floating value between 0 and 1 which reflects the relative importance of each algorithm in a combination function. The second step is to set the operators among the selected similarity measures; the available operators are MAX, AVERAGE and SUM. This means that the final similarity value will be obtained by selecting the maximum similarity value of all selected algorithms, average similarity values among all selected algorithms or the sum of each obtained similarity values. While using the SUM operator; the final similarity value may exceed 1according to the scaling of each algorithm. The final step is to select the preprocessing method; predefined set of preprocessing modules are available as discussed in this section.

A useful option is added to SimAll which is enabling users to train a predefined similarity values or any target classes to build a new model or to define the scaling weight of each algorithm automatically. WEKA [17] machine learning algorithms are ready to use within our tool.

4 SHORT ANSWER GRADING (SAR)

Automatic Scoring (AS) systems address evaluating a student's answer by comparing it to model answer(s). AS technology handles different types of students' responses such as writing, speaking and mathematics [18]. Writing assessment comes in two forms: Automatic Essay Scoring (AES) and Short-Answer Scoring. Speaking assessment includes low and high entropy spoken responses, while mathematical assessments include textual, numeric or graphical responses. AS Systems are easily implemented for certain types of questions, such as Multiple Choice, True–False, Matching and Fill-in-the-Blank. Implementing an automatic scoring system for questions that require free text answers is more difficult because students' answers require complicated text understanding and analysis. Gomaa and Fahmy tested the task of SAR using SimALL through three articles [19,20,21].

In [19] short answer grading task is handled from an unsupervised approach which is bag of words. The used data set contains 81 questions and 2273 student answers about data structure course in English language. The proposed model tested through three stages: The First stage was measuring the similarity between model answer and student answer using 13 String-Based algorithms. Six of them were Character-based and the other seven were Term-based measures. The best correlation values achieved using Character-based and term-based were 0.435 and 0.382 using N-gram and Block distance respectively. The Second stage was measuring the similarity using DICSO1 and DISCO2 Corpus-based similarity. Disco1 achieved 0.465 correlation value using the max overall similarity. The Third stage was measuring the similarity by combing String-based and Corpus-based measures. The best correlation value 0.504 was obtained from mixing N-gram with Disco1 similarity values.

In [20] the authors focused on applying multiple similarity measures separately and in combination. Many aspects were introduced that depend on translation to overcome the lack of text processing resources in Arabic, such as extracting model answers automatically from an already built database and applying K-means clustering to scale the obtained similarity values. Additionally, this research presented the first benchmark Arabic data set that contains 610 students' short answers together with their English translations. Questions presented in the data set cover one chapter of the official Egyptian curriculum for Environmental Science (ES) course. This research presents a system that automatically scores each student's answer (for 610 answers) with 536 different runs: 256 of the runs used String-Based Similarity, 64 used Corpus-Based Similarity, and the other 216 used Knowledge-Based Similarity measures. For each run, the Pearson Correlation Coefficient (r) and the Root Mean Square Error (RMSE) were computed. The best r and RMSE values were 0.83 and 0.75 respectively.

In [21] a new short answer benchmarking data set called Philosophy was presented; it contains 50 questions with 12 answers per each with total number of 600 answers. Model answer for each question is divided to set of elements; each element may contain Section(s) and Sub Section(s) with certain mark for each. Assigning a certain mark for each section and subsection helped in scoring either by comparing student Answer to model Answer as a whole or partially and finally providing useful feedback to students depending on the description of each Section and Sub Section. Fourteen String-Based and two Corpus-Based similarity algorithms were experimented through two models. The first model (Holistic Model) measured the similarity between the complete form of student answer and model answer without dividing the student answer and ignoring the partition scheme of model answer. The second model (Partitioning Model) automatically divided student answer into set of sentences using sentences boundary detection templates based on regular expression, then it mapped each sentence to the highest similarity element of model answers. Partitioning model achieved better results than holistic model in all cases although simple sentence boundary detection templates were used. Combining multiple similarity measures enhanced both the correlation and the error rate values. An interesting research point was to benefit from the combination of similarity algorithms in reducing the total required time of measuring the automatic mark. Applying String-based measures to map each sentence in student answer to each element in model answer obtained similarity reduced the elapsed time to the sixth which is considered real achievement. Also this combination paved the way to multithreading approach which accordingly decreased the elapsed time. Providing students with useful feedback was introduced; this module was performed by selecting four thresholds according to K-means clustering. These thresholds defined the range of each type of feedback comments. The accuracy of feedback module was promising especially for the two extremes "Wrong Answer" and "Correct Answer" types.

5 CONCLUSIONS

The different similarity approaches are explained in this article. Text similarity algorithms are classified into three categories: string-based, corpus-based and knowledge-based. Also the most widely used similarity tools are described. A new tool named SimAll is introduced; the tool offers many useful features such as: grouping 61 different similarity algorithms into a single tool, enabling users to use these algorithms separately or in a combination, supporting both Arabic and English languages, applying different preprocessing modules, handling different granularity of similarity methods and building customized models based on machine learning. SimAll was used in three different case studies of short answer grading task.

REFERENCES

[1]Gomaa, W. H., &Fahmy, A. A. (2013). A survey of text similarity approaches. International Journal of Computer Applications, 68(13), 13-18.

[2] Chapman, S. (2009). SimMetrics. URL http://sourceforge.net/projects/simmetrics/.

[3] Cohen, W. W., Ravikumar, P., & Fienberg, S. (2003). URL http://secondstring.sourceforge. net.

[4] Bernstein, A., Kaufmann, E., Kiefer, C., &Bürki, C. (2005). Simpack: A generic java library for similarity measures in ontologies. University of Zurich.

[5] Euzenat, J. (2004). An API for ontology alignment.In The Semantic Web–ISWC 2004 (pp. 698-712).Springer Berlin Heidelberg.3. Cohen, W. W., Ravikumar, P., & Fienberg, S. (2003).

[6] Okazaki, N., &Tsujii, J. I. (2010, August). Simple and efficient algorithm for approximate dictionary matching.In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 851-859).Association for Computational Linguistics.
[7] A. Behm, R. Vernica, S. Alsubaiee, S. Ji, J. Lu, L. Jin, Y. Lu, and C. Li. UCI Flamingo Package 4.1, 2010.
[8] Kolb, P. (2008). Disco: A multilingual database of distributionally similar words. Proceedings of KONVENS-2008, Berlin.
[9] Ştefănescu, D., Rus, V., Niraula, N. B., &Banjade, R. (2014). Combining Knowledge and Corpus-B to-Word Similarity Measures for Word.
[10] Pedersen, T., Patwardhan, S., &Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration papers at hlt-naacl 2004 (pp. 38-41).Association for Computational Linguistics.
[11] Pedersen, T., Patwardhan, S., Michelizzi, J., & Banerjee, S. (2005). Wordnet:: similarity. 2008-06-03]. http://www.similarity, source forge, net.
[12] Shima, H. (2013). WS4J-WordNet Similarity for Java.URL https://code.google.com/p/ws4j/.
[13] Bird, S. (2006, July). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72).Association for Computational Linguistics.
[14] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., &McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).
[15] Khoja, S., & Garside, R. (1999). Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University.
[16] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.
[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.
[18] Gomaa, W. H., &Fahmy, A. A. (2011). Tapping Into The Power of Automatic Scoring. In the eleventh International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC'2011).
[19] Gomaa, W. H., &Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. International Journal of Advanced Computer Science and Applications (IJACSA), 3(11).
[20] Gomaa, W. H., &Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. Computer Speech & Language, 28(4), 833-857.
[21] Gomaa, W. H., &Fahmy, A. A. (2014). Arabic Short Answer Scoring with Effective Feedback for Students. International Journal of Computer Applications, 86(2), 35-41.

BIOGRAPHY

Dr.Wael Hassan Gomaa: Currently working as a lecturer, Beni-Suef University. His research interest is in NLP, text mining, machine and deep learning. He worked as a manager for MIS Unit, Beni-Suef University. He obtained PhD degree from Faculty of Computers and Information, Cairo University in the field of Automatic Assessment under supervision of Prof. AlyFahmy. He received his BSc and Master degrees from Faculty of Computers and Information, Helwan University.

Prof.AlyFahmy: He is the former Dean of Faculty of Computers & Information – Cairo University. His research interest is in Artificial Intelligence Topics such as natural language processing, data and text mining, deep learning and information retrieval. Prof. Aly has a number of publications. He obtained B.Sc in June 1972, Computer Engineering Department Military Technical College (M.T.C) Excellent with Honor Grade. DPL - Diploma: General Purpose Simulation, June 1973, Computer Department, Military Technical College (M.T.C), M.Sc in Logical Database Systems 1976, Computer Department, E.N.S.A.E, Toulouse, France. Ph.D in Artificial Intelligence Control of Automatic Deductions for Logic Based Systems 1979, Computer Department, The Centre of Research and Studies, Toulouse,

France(C.E.R.T) under the supervision of H. Gallaire (Ex Vice President and Chief Technical Officer of Xerox Corporation) and J.M. Nicolas.

أداة مرنة لقياس تشابه النصوص

[1]*وائل حسن جمعة ، [2]**على على فهمى

* قسم علوم الحاسب ، كلية الحاسبات و المعلومات ، جامعة بنى سويف

[1]wael.goma@fcis.bsu.edu.eg

** قسم علوم الحاسب ، كلية الحاسبات و المعلومات ، جامعة القاهرة

[2]aly.fahmy@cu.edu.eg

ملخص-يعتبر قياس التشابه بين النصوص عنصر أساسى فى مختلف مهام معالجة اللغات الطبيعية. تتوافر الآن العديد من أدوات قياس التشابه بين النصوص التى تختلف فى طرق القياس و التقنيات المستخدمة، فبعض هذه الأدوات يعتمد على قياس تكرار الحروف و الكلمات بين الجمل و البعض الأخر الأكثر تقدما تعتمد على قياس تشابه معانى الكلمات والجمل.الهدف من هذا البحث هو تقديم أداة جديدة تجمع بين كل مميزاتو تقنيات أدوات قياس التشابه. تقدم الأداة المقترحة طرق مختلفة لتجهيز النصوص قبل إجراء عملية قياس التشابه، كما تعرض أساليب متنوعة لدمج طرق القياس المختلفة. تدعم الأداة المقترحة اللغتين العربية والإنجليزية. يقدم البحث بعض نتائج التصحيح الآلى لإجابات الطلبة عن الأسئلة المقالية فى مناهج مختلفة بإستخدام الأداة المقترحة.

A Survey on Mental Illness Detection using Language via Social Media Networks

Eman Hamdi[*1], Sherine Rady[**2], Mostafa Aref[*3]

[*]*Computer Science Department, Faculty of Computer and Information Science, Ain Shams University*
Abbassia, Cairo, Egypt
[1]emanhamdi@cis.asu.edu.eg
[3]mostafa.m.aref@gmail.com

[**]*Information Systems Department, Faculty of Computer and Information Science, Ain Shams University*
Abbassia, Cairo, Egypt
[2]srady@cis.asu.edu.eg

***Abstract*—Mental illness is any disease that affects person's thoughts, behavior, and feelings. It is a serious threat on people and society as well. Due to the risk of mental illness, the researchers seek to discover original methods to collect informative data representing how the person thinks and behaves. They found the person's language to be one of the most essential factors to define the susceptibility of a mental illness. Recently, social media networks are considered as sources of data that can be used to collect a fair combination of a person's language and to realize his behavior. The purpose of this survey is to shed light on recent researches of detecting a mental illness using language analysis on social media networks with summarized details.**

Keywords: **Natural Language Processing, Social Media Networks, Mental Illness**

1 INTRODUCTION

Many millions suffer from mental illness conditions such as depression, Post-Traumatic Stress Disorder (PTSD), anxiety disorder and phobia; however, a lot of them don't obtain appropriate curing. That's because the recognition of people who suffers from a mental illness became a challenge. The problem is those battling with a condition don't explicitly ask for help, they are afraid of social participating, or they can't recognize their symptoms. That causes a lack of data which is one of the main obstacles in front of the researchers in this field and raises the need for a reliable source of informative data [1], [2].

Fortunately, proven by psychology researches, Language is a fundamental key to detect a mental illness condition. It indicates how a person thinks, feels, interacts with others and captures any transformation in a person's mental state [3], [4].Here comes the role of social media networks as a great source of textual data along with other personal data such as age, gender, and connections. Social media networks as Facebook and Twitter become an alternative window on people's emotions, interactions, behavior and even mood swings through the day. Given those information, many researchers investigate the usage of social media networks to detect a mental illness [5].

In this paper, an overview has been done for the recent researches on using social media networks language to detect a mental illness condition per various measurements. Then a conclusion will be given to help in improving further researches in this field.

2 OVERVIEW ON MENTAL ILLNESS DETECTION RESEARCH

In this section, six works are discussed. The aim was to detect a specific mental illness based on language analysis. The potential of language analysis to indicate behavioral and emotions changes was proven. The methods are discussed based on dataset creation, feature extraction techniques, classification methods, and experimental results. They are [1], [6], [7], [8], [9], and [10].

De Choudhury et al. [1] used Twitter data to detect clinical depression in individuals. Crowd sourcing was used to get a collection of Twitter users who reported that they were diagnosed with depression. Using the patients' social media postings over a year after the report of being depressed, Behavioral attributes were measured relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. Then, behavioral cues were leveraged to build a statistical classifier that provided estimates of the risk of depression before the reported onset.

Coppersmith et al. [6] measured PTSD using Linguistic Signals extracted from tweets. A log-linear classifier was trained on the linguistic features to determine the differences in language usage between PTSD users and the general population.

In [7], Detection of depression, PTSD, and control users -who have neither of those conditions- was done using their Twitter data. Binary logistic regression classifiers with Elastic Net regularization were trained to classify three types of

users: Depressed vs. Control, Depressed vs. PTSD, and PTSD vs. Control. The classifiers were trained using features such as: age, gender, personality, emotions, and textual features.

In [8], statistical and machine learning methods were used on data collected from an online depression community to distinguish posts made in low versus high valence mood, in different age categories and in different degrees of social connectivity.

Brideanne et al. [9] examined whether the level of concern for a suicide-related post on Twitter could be determined based on the content of the post, as judged by human coders and then replicated by machine learning. Using human coded data, machine learning methods were applied to develop a text classifier that could automatically distinguish tweets into three categories of concern. suicide-related tweets were classified as 'strongly concerning', 'possibly concerning' and 'safe to ignore'.

In [10], the results of a web-based questionnaire results were used as ground truth data for measuring degree of depression of Japanese Twitter users. Features were extracted from the activity histories of Twitter users. Using those features, SVM was tested estimating the presence of active depression. Experiments showed that features obtained from user activities can be used to predict depression of users with an accuracy of 69%, topics of tweets are useful features, and approximately two months of observation data were enough for detecting depression.

3 MENTAL ILLNESS DETECTION ASPECTS

In this section, we are going to discuss the aspects used in the above works. These aspects include dataset creation, feature extraction techniques, classification methods, and experimental results as follows:

A. Dataset Creation

De Choudhury et al. [1] applied crowd sourcing to collect labels that were taken as ground truth data on the presence of depression. It's an efficient mechanism to get access to behavioral data from a varied population. It is less time consuming and inexpensive [11]. Then, Amazon's Mechanical Turk interface was used to design human intelligence tasks (HITs) wherein who participated were asked to take a standardized clinical depression survey. A final dataset consisting of 476 users was formed, it contained 171 users who were positively depressed (positive class); and 305 users who were negatively depressed (negative class). Then, a total number of 2157992 Tweets were collected from the users' profiles.

While Coppersmith et al. [6] accessed a huge collection of data from the Twitter keyword streaming API, wherein keywords were picked to concentrate on health topics. A regular expression was used to look up statements in which the user self-reported being diagnosed with PTSD. Next, the 3200 most recent tweets posted by each user were retrieved via the Twitter API. After filtration, 244 users were marked as positive examples of having PTSD. This process was repeated for a randomly selected group and after filtration, 5728 users as negative examples.

Preotiuc-Pietroet et al. [7] built a dataset of Twitter users who self-reported to suffer from a mental illness, specifically depression and PTSD. This dataset was originally initiated in [12]. The reports are collected by searching a huge Twitter data for disclosures using a regular expression. Selected users were filtered manually and then all their most recent tweets were continuously crawled using the Twitter Search API. The final set consisted of 370 users with PTSD, 483 with depression and 1104 control users; each one of them had 3400.8 messages on average.

In [8], data was crawled from *depression.livejournal.com* which offers a list of 132 predefined moods for bloggers to tag to their posts. Three sub-corpuses were created based on mood valence, age and social connectivity of users as follows: Valence based corpus consisted of 400 posts, age based corpus consisted of a set of 500 young users and a set of 500 old users. Social connectivity based corpus contained three corpora which were built based on the extreme numbers of followers, friends, and community membership.

In [9], a collection of suicide-related tweets was used. It was extracted using Twitter's Application Program Interface (API). Tweets were stored in a data coding tool developed by the Common wealth Scientific and Industrial Research Organization (CSIRO). Also, human coding was used to determine the level of concern within the suicide-related tweets. As a result, 14,701 tweets matched the suicide-related search terms divided as: 2000 (14%) were randomly selected for human coding. A total of 9% (n = 178) of data was discarded or known, and thus excluded. When data sets A and B were combined, 14% (n = 258/1822) were coded as 'strongly concerning', 57% (n = 1030/1822) 'possibly concerning', 29% (n = 534/1822) were coded as 'safe to ignore'.

In [10], crowd sourcing was applied to collect labels that were taken as ground truth data on the presence of depression. Questioners were taken by Japanese-speaking volunteers then the activity histories were collected through the Twitter application programming interface (API). The result was, data about 209 experiment participants (male: 121; female: 88) aged 16 to 55 (mean: 28.8 years; standard deviation: 8.2 years) were analyzed and at most 3,200 tweets were collected for each participant.

B. Features Extraction Techniques

The first work [1] depended on five major features: user engagement, language, emotions, linguistic style, and ego network. User Engagement included five measures which are driven from: the volume of posts per day, the proportion of replies per day, the fraction of retweets from a user per day, the proportion of links/question-centric posts per day, and insomnia index. Language was defined based on two features: depression lexicon and Antidepressant usage. Emotions included four measures: positive affect, negative affect, activation, and dominance. That was determined using Linguistic Inquiry and Word Count (LIWC) [13], wherein emotion categories were scientifically validated to perform well for determining affect in Twitter. Linguistic style: LIWC was used to determine 22 specific linguistic styles. Ego network: an egocentric social graph was constructed to determine the strength of connections of a user on twitter. It's based on the user, friends and friends of friends on the entire network.

The second work to be discussed [6] used four features: Tweets as words and strings of characters and LIWC output. Words and characters from tweets were used to train a unigram language model (ULM) and a character language model (CLM). LIWC was used to calculate the proportion of tweets that scored positively by each LIWC category. These proportions are used as a feature vector.

While in [7] four major features were used: age, gender, personality, emotions and textual features. Age, gender and personality were obtained using Automatic personality assessment introduced in [14]. Emotions: expressions were characterized per two measures; affect (from positive to negative) and intensity (from low to high). Textual features: 64 different categories were built based on LIWC, including different parts-of-speech, topical categories and emotions. Thereby, each user was represented as a distribution over these categories. Also 1-3 grams and 2-3 grams were used. To reduce dimensionality, they used a set of 2,000 clusters introduced in [14] obtained by applying a popular Bayesian probabilistic modeling tool which is Latent Dirichlet Allocation (LDA) [15].

In [8], two features were used to characterize blog posts: topic and language style. Topics were extracted using LDA as 50 topics were used. Language style is captured using the LIWC, it returned 68 psycholinguistic categories, such as linguistic, social, affective, cognitive, perceptual, biological, relativity, personal concerns, and spoken.

In [9], three representations of features were used. First each tweet was represented as a vector of features using the Scikit-Learn toolkit [16]. In which all words occurred in the dataset became features. Second, the weighting Term Frequency weighted by Inverse Document Frequency (TFIDF) was used instead of the simple frequency. Third, a variant of the TFIDF was used as an attempt to remove words with little information that occurred above a threshold for document frequency such as words like 'the' and 'of' would be removed from the feature. Those representations were referred to as "freq", "TFIDF", and "filter".

In [10], three major sets of features were used: the frequencies of words in a tweet, the topics of the tweets, and the ratio of positive and negative words in the tweets. Additionally, some features from [de] were used, as the user's timing of tweets, frequency of tweets, average number of words, retweet rate, mention rate, ratio of tweets containing a uniform resource locator (URL), number of users being followed, and number of users following are used as features independent of the content of the tweet.

C. Classification Methods

De Choudhury et al. [1] investigated supervised learning to build classifiers to predict depression in the two classes; positive and negative. Principle component analysis was used to avoid over fitting. After comparing different binary classifiers, the best performing classifier was Support Vector Machines (SVM) with a radial basis function (RBF) kernel [1], [17].

Coppersmith et al. [6] depended mostly on n-grams models as three methods were applied: unigram language model (ULM). Also, a log linear regression model was trained using the proportion of tweets by each user that scored positively by each of LIWC categories of PTSD.

Preotiuc-Pietro et al. [7] trained a binary logistic regression classifier with Elastic Net regularization. The model was trained using variable combinations of features which lead to achieve variable results depending on what features were used in training.

In [8], the least absolute shrinkage and selection operator (Lasso) [18] was performed to do logistic regression using topics or LIWC categories as features. Positive and negative weights were assigned to features indicating the importance of each feature in the prediction. Ten-fold cross validations were produced on the training data to estimate the prediction model and accuracy is used to evaluate the performance of the classification.

In [9], two machine learning algorithms for text classification were tested: SVM and Logistic Regression.

In [10], Machine learning classifiers were constructed with SVM for estimating the presence of active depression, and their classification accuracy was evaluated by 10-fold cross validation.

D. Experimental Results

In [1], using dimension-reduced features with SVM was the best performing model that yielded an average accuracy of ~70% [1].

Second, the classifiers used in [6] were evaluated via leave-one-out cross validation setting in both a balanced and a non-balanced dataset. That validation provided maximum training data while evaluating every user in turn. Various operating points were obtained by changing the threshold of the classification. The best result was obtained from using ULM, then CLM and finally LIWC.

Finally, in [7], while Gender is weakly predictive of any illness, Age is highly predictive for PTSD classification. Personality alone resulted in very good mental predictive accuracies, reaching over 0.8 receiver operating characteristic-area under the curve (ROC AUC) for classifying depressed vs. PTSD, but Average affect and intensity achieve modest predictive performance. The highest predictive performance was reached using textual features. It obtained 0.819 ROC AUC for classifying Depressed vs. PTSD, 0.859 for classifying Depressed vs. Control, and 0.917 for classifying PTSD vs. Control.

In [8], the best is with the classification of blog posts into low and high valence mood using LIWC categories as predictors, achieved an accuracy of 78.32%. The result in other classifications achieved an accuracy of approximately 60%.

In [9], the training set consisted of 90% of data points and the testing set was the remaining 10%. Two machine learning algorithms for text classification were tested: SVM and Logistic Regression. The classifiers were tested with each variant of the feature space ("freq", "tfidf", and "filter").Using cross-validation, the average accuracy when the training set is divided into 10 "folds" or subsets was assessed. The total number of tweets used in the training and testing was 1820: Set A=829 (training: 746, testing: 83) and Set B=991 (training: 891, testing: 100).The accuracy was raised when sets A and B were combined resulted in 76% accuracy. A precision score of 80%was found for 'strongly concerning', 76% for 'possibly concerning' and 75% for 'safe to ignore'.

In [10], three models were used with the SVM classifier as the 10-topic model (Model 1); a model using the features positive, negative, tweet frequency, RT, URL, followee, and follower (Model 2); and a 10-topic model including the features positive, negative, tweet frequency, RT, URL, followee, and follower with the highest estimation accuracy from the previous section (Mode 3). Using 10-fold cross validation to evaluate accuracy, the highest accuracy was achieved when using tweets from the recent 8 weeks, and the presence of active depression can be estimated by Model 3 with 69% accuracy.

The final comparison between the six works is shown in Table 1.

TABLE I
FINAL COMPARISON BETWEEN THE SIX INTRODUCED WORKS

Title	**Task**	**Dataset Creation**	**Feature Extraction Techniques**	**Classification Methods**	**Experimental Results**
Predicting Depression via Social Media (2013)	Predicting depression	Tweets Crowd sourcing to collect data labels	User engagement Language Emotions Linguistic style Ego network	SVM	Accuracy: 70%
Measuring Post Traumatic Stress Disorder in Twitter (2014)	Detecting PTSD	Tweets Regular expression to find self-reports	Linguistic and textual features	Log Linear Regression	Unigram Language Model>Character Language Model>LIWC
The Role of Personality, Age and Gender in Tweeting about Mental Illnesses (2015)	Classifying depressed (D), PTSD and control users (C)	Tweets Regular expression to find self-reports	Age Gender Personality Emotions Textual Features	Logistic Regression	Area Under the Curve: 0.859 for C vs. D 0.917 for C vs. PTSD 0.819 for (vs. PTSD

Effect of Mood, Social Connectivity and Age in Online Depression Community via Topic and Linguistic Analysis (2014)	Examining the effect of mood, social connectivity and age on the depressed users' online messages	Posts from an online depression community	Topics Language Style	Logistic Regression	Accuracy: 78.32% predicting low and high valence 60% for the other models
Detecting suicidality on Twitter (2015)	Examining the level of concern of suicide tweets using human coders and machine learning	Tweets	Word Frequency TFIDF Filtered TFIDF	Logistic Regression SVM	Accuracy: 80% for 'strongly concerning' 76% for 'possibly concerning' 75% for 'safe to ignore'
Recognizing Depression from Twitter Activity (2015)	Detecting depression	Tweets Crowd sourcing to collect data labels	Frequencies of words Topics The ratio of positive and negative words Metadata	SVM	Accuracy: 69%

4 CONCLUSION

Upon discussing the six works, Language is found to be a major component in understanding others in means of thinking, feeling, acting, and communicating, which is very useful to detect any transformation in peoples' minds. Social media networks are very powerful data source and can help in many fields of research. Over the presented methods, the creation of the dataset was fair if the reports are real and accurate, which was proven by the applied results.

The use of statistical models to develop features vector leaded to good results, but that's when given a very specified and even average-sized dataset. Also, the used classifiers were SVM and linear regression classifiers, which have limitations if working on larger amount of data, different types of words, and especially close classes which are hard to be distinguished from each other. That led us to many open questions. How to ensure that the data collected is real? Would real life surveys help to solve this problem? What if the dataset was much bigger and diverse? Will statistical models provide good accuracy or the investigation of more advanced feature extraction algorithms is required? Finally, what about the language itself? The presented methods processed on a specific language and can't handle other languages, will advanced machine learning tetchiness help in developing a model which is not data or language oriented? Pursuing the answers to these questions will provide many exciting opportunities for mental health, natural language processing and machine learning researches.

References

[1] De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz, *"Predicting Depression via Social Media,"* Proceedings of the Seventh International Conference on Weblogs and Social Media, pp. 1-10, 2013.

[2] Lu, Cheng-Yu, William WY Hsu, and Jan-Ming Ho, *"Event-Level textual emotion sensing based on common action distributions between event participants,"* Advanced Research in Applied Artificial Intelligence, Springer Berlin Heidelberg, pp. 427-436, 2012.

[3] Coppersmith, Glen, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell, *"CLPsych 2015 shared task: Depression and PTSD on Twitter,"* The Conference of the North American Chapter of the Association for Computational Linguistic, pp. 31-39, 2015.

[4] Resnik, Philip, William Armstrong, Leonardo Claudino. and Thang Nguyen, *"The University of Maryland CLPsych 2015 shared task system,"* The Conference of the North American Chapter of the Association for Computational Linguistic, pp. 54-60, 2015.

[5] Coppersmith, Glen, Mark Dredze, Craig Harman, and Kristy Hollingshead, *"From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses,"* CLPsych@ HLT-NAACL, pp. 1-10, 2015.

[6] Coppersmith, Glen, Craig Harman, and Mark Dredze, *"Measuring Post Traumatic Stress Disorder in Twitter,"* Proceedings of the Eighth International Conference on Weblogs and Social Media, pp. 579-583, 2014.

[7] Preotiuc-Pietro, Daniel, et al., *"The role of personality, age and gender in tweeting about mental illnesses,"* The Conference of the North American Chapter of the Association for Computational Linguistic, pp. 21-30, 2015.

[8]Dao, Bo, Thin Nguyen, Dinh Phung, and Svetha Venkatesh, "Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis," In International Conference on Web Information Systems Engineering, pp. 398-407, Springer, Cham, 2014.

[9] O'Dea, Bridianne, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen, "Detecting suicidality on Twitter," Internet Interventions 2, no. 2 (2015), pp. 183-188.

[10] Tsugawa, Sho, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki, "Recognizing depression from twitter activity," In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3187-3196, ACM, 2015.

[11] Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng, *"Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks,"* In Proceedings of the conference on empirical methods in natural language processing, pp. 254-263, Association for Computational Linguistics, 2008.

[12] Coppersmith, Glen, Mark Dredze, and Craig Harman, *"Quantifying mental health signals in Twitter,"* In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 51-60, 2014.
[13] Pennebaker, James W., Martha E. Francis, and Roger J. Booth, "Linguistic inquiry and word count: LIWC 2001," Mahway: Lawrence Erlbaum Associates 71, no. 2001.
[14] Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah et al., *"Personality, gender, and age in the language of social media: The open-vocabulary approach,"* PloS one 8, no. 9 (2013): e73791.
[15] Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research 3, no. Jan (2003): 993-1022.
[16] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research 12, no. Oct (2011): 2825-2830.
[17] Duda, Richard O., Peter E. Hart, and David G. Stork, 2000, Pattern classification, 2nd Edition, Wiley.
[18] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani, "Regularization paths for generalized linear models via coordinate descent," Journal of statistical software 33, no. 1 (2010): 1.

BIOGRAPHY

Eman Hamdi is a Teaching Assistant and M.Sc. student at the Faculty of Computer and Information Sciences of Ain Shams University in Cairo, Egypt. B.Sc. in Computer Science, Ain Shams University, Cairo, Egypt.

Sherine Rady is an Assistant Professor at the Faculty of Computer and Information Sciences of Ain Shams University in Cairo, Egypt. Ph.D. from University of Mannheim in Germany. M.Sc. in Computer and Information Sciences from Ain Shams University. B.Sc. in Electrical Engineering (Computer and Systems), Ain Shams University, Cairo, Egypt.

Mostafa Aref is a professor and Vice Dean for Society Service & Environmental Development at the Faculty of Computer and Information Sciences of Ain Shams University in Cairo, Egypt. Ph.D. in Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. in Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. in Electrical Engineering-Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, Egypt.

دراسة عن إكتشاف الأمراض العقلية بإستخدام اللغة من خلال مواقع التواصل الإجتماعي

*إيمان حمدي ، **شيرين راضي، *مصطفى عارف
*قسم علوم الحاسب، كلية الحاسبات والمعلومات، جامعة عين شمس
العباسية، القاهرة، مصر
[1]emanhamdi@cis.asu.edu.eg
[3]mostafa.m.aref@gmail.com
**قسم نظم المعلومات كلية الحاسبات والمعلومات، جامعة عين شمس
العباسية، القاهرة، مصر
[2]srady@cis.asu.edu.eg

***ملخص*—المرض العقلي هو أي مرض يؤثر على أفكار الإنسان، تصرفاته ومشاعره. إنه يشكل تهديد على الأشخاص والمجتمعات أيضاً. وبسبب خطورة المرض العقلي، يسعى الباحثون لإكتشاف طرق جديدة لجمع معلومات مفيدة تعبر عن طريقة تفكير الشخص وتصرفاته. وجد الباحثون أن لغة الشخص هي من أهم العوامل الرئيسية لتحديد قابلية الإصابة بمرض عقلي. مؤخراً، تعتبر مواقع التواصل الإجتماعي مصادر للبيانات من الممكن أن تستخدم لجمع مزيج عادل من لغة الشخص ولملاحظة سلوكه. الهدف من هذه الدراسة هو إلقاء الضوء على الأبحاث الحالية في تحديد المرض العقلي بإستخدام تحليل اللغة على مواقع التواص الإجتماعي، مع تلخيص التفاصيل.**

***الكلمات الدلالية:*الكشف عن الأمراض العقلية، مواقع التواصل الإجتماعي**

Cueing Conspiratorial Ideation in the Egyptian Tweets using Web-as-Corpus

Bacem A. Essam[*1], Prof Dr. Mostafa M. Aref [**2]

[*] *English Language Department, Faculty of Al-Alsun, Ain Shams University*
[**] *Computer ScienceDepartment, Faculty of Computer Science and Information Sciences*
Ain Shams University, Cairo, Egypt
[1]literaryartrans@gmail.com
[2]mostafa.aref@cis.asu.edu.eg

***Abstract*— This paper uses linguistic cues to detect the elements of conspiracy in a large-scale corpus of Egyptian Tweets (2012-2017). The study quantitatively identifies the hypothesized enemies in the Egyptian society through corpus-driven evidence. Using linguistic and corpus tools to identify the aspects of the socio-political theory of conspiracy represents a measurable and retrievable way for detecting a social phenomenon. The results suggest that the Egyptian society describes two types of enemies or schemers: intra-societal and inter-societal adversaries. The perceived conspiracy, at the country level, is politically, theologically and historically-driven. The inside schemers, however, are defined only on a political basis.**

***Keywords*—** Linguistic cues, Conspiratorial Ideation, Arabic ontology, Web-as-Corpus.

1 INTRODUCTION

Conspiracy theories allege that multiple actors are intentionally plotting to accomplish malevolent goals. Adopting a conspiracy theory stance psychologically emanates from the individual desire to be secured within a group or the desire to project a positive image of the social group [1]. Latent psychopathology, biased cognitions, psychological stress, anxiety and individual differences in traits (such as thinking styles, political cynicism, and self-esteem) are promoted as predisposing factors of adopting conspiratorial ideation, too. This paper evaluates the conspiratorial ideation by the Egyptian folks on twitter both at the intrasocietal and intersocietal levels. Thus, it investigates the conspiracy elements in which 'Egypt', as a country, is instantiated as a victim of a plot, and 'the Egyptians', as a society, are depicted as a schemed-against party. For doing so, two lists of all countries and conspiracy-theory-ridden words are concordanced in individual tweets (2012-2017).

2 THEORETICAL BACKGROUND

The controversial notion of conspiratorial ideation, or conspiracism, is variably defined among social sciences and, even more, within the same domain. Despite the non-consensual definitions, conspiracy, as a concept, is featured as a 'hidden hand' scheme. Identifying the 'hidden hand' and the type of scheme are the main concern of conspiracy theories. The scheming hidden hand is usually believed in by most of the folk in a society and it is always linked either to indoor plotters or outside enemies [1].

Conspiracist ideation is associated with negative health, sociopolitical, and environmental consequences. Thus, adopting conspiracy theories helps conspiracist to regulate levels of acute stress by providing simplified, causal explanations for such distressing events. Recently manifested as a major subcultural phenomenon, Conspiracism endorses skepticism regarding the reality of perceived (mis)information. Popular contemporary examples include the theory that the September 11 attacks were planned and carried out by elements within the American government [2]. Adopting the idea that authorities are engaged in motivated deception of the public and disseminating a blind-hostility-conception of a given group/nation, which is grounded in the stereotypy of the victimized subject, are central to all conspiracy theories.

Highlighting that the Middle East is espousing conspiracism more than other nations, Gray [3] explains that the Middle Eastern gap between leaderships and society stems not just from the crack between policy rhetoric and policy implementation. However, the gap is a manifestation of minority governments, which remain a feature of some Arab

states and the opaque neo-patrimonial networks that several leaders create around themselves to reinforce their positions and enhance their reach into the institutions and social forces of politics. Examples include the reigning power (a) in Syria, where an Alawi leadership controls a majority Sunni population; (b) in Bahrain, where a Sunni royal family controls a majority Shi'a population, and in effect; and (c) in Lebanon, where a consociationalist quasi-democracy means that no one group – Sunni, Shiite, Christian, Druze, or others – controls government, even if the Christians and Sunnis have disproportionate power and influence over it.

3 RELATED WORKS

The present study conducts a linguistic corpus-driven analysis of the conspiracy theory. The study venture semantic cueing of conspiracim into a mold similar to detecting deception linguistically. Using linguistic cues to elicit psychological and cognitive information is a verified method. For instance, Linguistic Inquiry and Word Count (LIWC), as a transparent text analysis a probabilistic system, is created and validated. LIWC is that correlate words to psychologically meaningful categories for analyzing what the cognitive load words may imply. Either content words, generally nouns, regular verbs, adjectives and adverbs, or function words, pronouns, prepositions, articles, conjunctions and auxiliary verbs, integrating words into an utterance reveals linguistic features. Such features help investigators to analyze and conclude much about the non-disclosed message [4-7]. Jensen et al [8] demonstrate that quantity, complexity, certainty, immediacy, diversity, specificity, affect are significant linguistic cues which can pinpoint the deceptive language.

Hancock, Curry, Goorha, and Woodworth [9] expanded these findings to study lying within pairs of participants over instant messenger. They found that the people being deceived, the partners of the participants lying, changed their language in response to receiving lies. There, a higher total word count and more sense words, elaborating on the description of the deception scenario, are observed. Motion, exclusion, and sense words all indicate the degree to which an individual. Using LIWC, Newman et al. [10] have predicting deception from linguistic styles. However, LIWC, in spite of scoring marvelous results, runs into several problems. These problems emanate from separating words from the context they live in. Thus, irony, sarcasm, and idioms, with their due essentiality, are difficult to be measured. Given this shortcoming, this study has replaced recruiting LIWC analysis by running a corpus-driven supervised analysis of the Conspiratorial Ideation in the present Egyptian Tweets.

4 METHODOLOGY

A. Data Collection

GOOGLE is used to compile a representative corpus using a search query as follows:

*https://www.google.com.eg/search?q=**country1**+**egypt**+**site**:http://**twitter**.com/&lr=**langar**&hl=en-EG&asqdr=all &tbs=lr:lang1ar&**year**=**2012:2017**.*

After normalization and cleaning of the extracted tweets, a corpus of 3,442,640 tokens and 18,050,423 word types was compiled. All original tweets are correlated to a random sample of ordinary individuals. Moreover, organizational Twitter IDs are identified. Their tweets are then omitted because they profess an institutional voice. The data is normalized and cleaned to be uploaded onto Sketch Engine [11] for further processing.

B. Data Processing Software

After compiling a corpus of the collected snippets, the data is analyzed using Sketch engine, an online tool for processing corpora, where concordance lines have been extracted. In order to attribute the driven data to the framework of the conspiracy theory, inclusion criteria are defined, for every tweet, to imply intentional deception, view a general action and to promote a topic justification/subjective interpretation/drawn conclusion.

This paper assigns, therefore, a biphasic method. First, the concordance lines are annotated, for defining the Egyptian conflicts with other countries, and conspiratorial words are labeled. The concordance of conspiracy-related countries is further categorized to capture the linguistic cues

denoting the conspiracy theory and determine the type of the presupposed 'scheme'. Second, the collected specific conspiracy-laden words are enriched using synonyms. That is to say, for detecting the 'enemy within' and the 'enemy outside', a concordancing lexicon-based list of linguistic cues is applied corpus to identify the 'enemy within' parties and the type of scheme, at the intersocietal level.

5 Results and Discussion

Phase one included annotating the contemporary description of the international relation of Egypt in the studied tweets. The annotation was primarily labeled as union, conflict or not applicable. By sorting out the irrelevant data, conflict-expressing tweets were then analyzed to define the alleged subject of the conflict, which should justify the conspiratorial ideation, the domain or the channel that inherit such a conflict as well as the hypernym of such a domain (Table 1).

TABLE I

TWEETERS' CONSPIRATORIAL IDEATION AT THE COUNTRY LEVEL

Country (Enemy Outside)	**Alleged Subject of Dispute**	**Domain**	**Hypernym**
Angola	*Islamophobia*	Religion	Belief
Iran	*Islamophobia*	Religion	Belief
Algeria	*Shiite*	Religion	Belief
Denmark	*Islamophobia*	Religion	Belief
The Netherlands	*Islamophobia*	Religion	Belief
Yemen	*Houthis*	Religion	Belief
Germany	*Internal affair*	Democracy	Social relation
Italy	*Internal affair*	Democracy	Social relation
Uganda	*The River Nile*	Geopolitics	Social relation
Saudi Arabia	*Island of Tiran & Sanafir (Straits of Tiran)*	Geopolitics	Social relation
Turkey	*Ikhwan*	Theopolitics	Social relation
Uzbekistan	*Internal affairs*	Theopolitics	Social relation
Qatar	*Internal affairs*	Theopolitics	Social relation
Somalia	*Turkey*	Military	Force
Eritrea	*Military base*	Military	Force
Israel	*Military Invasion*	Military	Force
Jordan	*Israel*	Military	Force
Iraq	*Daesh*	Military	Force
Libya	*Daesh*	Military	Force
Spain	*Muslim Caliphates*	History	Noesis
Albania	*Albanian Kings of Egypt*	History	Noesis
China	*Commodities*	economy	Human Activity
The United Arab Emirates	*Funding Ethiopia's Dam*	Economy	Human Activity
Sudan	*Fund Appeal*	Economy	Human Activity

The commonest disputes were linked to (Sunni) Islamophobia, internal affair, the River Nile, straits of Tiran, Ikhwan, Turkey and military threats. By categorizing these subjects, domains of interest have turned to be Geopolitics, Theopolitics, religion, history and military interventions. All hypernyms of such domains did belong to abstraction, psychological features. The only exception was dispute over economic encounters. The hierarchical hypernym of such an activity was human activity. That is to say, the Egyptian folks care the most, in communicating internationally, about the integrity of beliefs and guard the earning of living. Figure 1 visualizes the described imminent threats at the folkloric perception level.

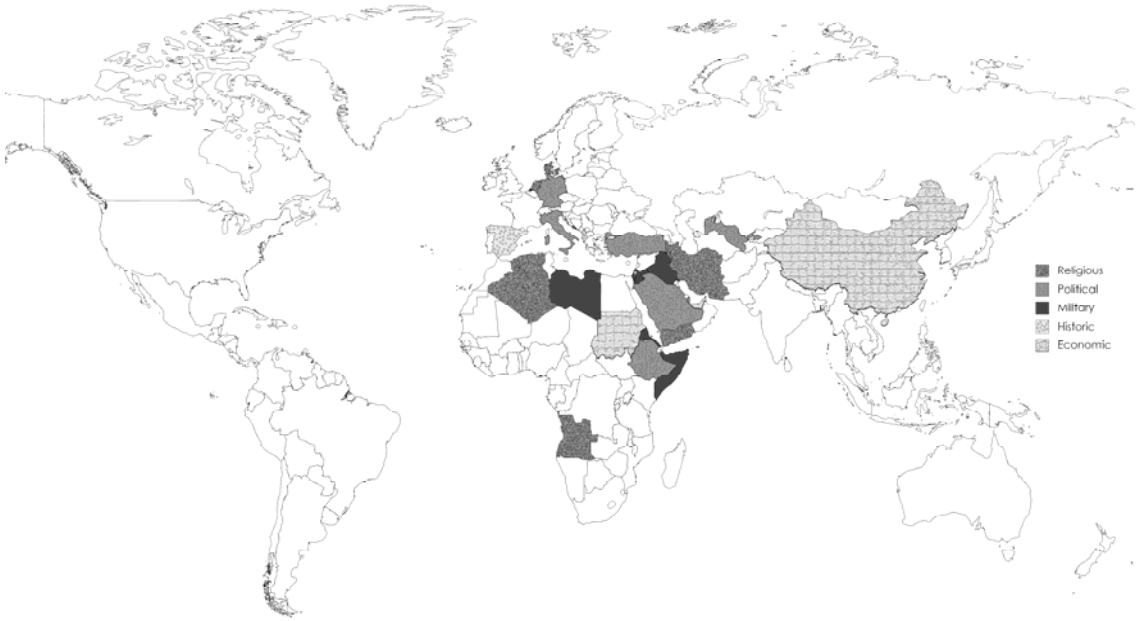


Figure 1: Mapping the 'enemy outside' according to the ideation of the Egyptian tweeters (2012-2017).

Similar to the recognized multiple enemies are the multiple types of scheme or conspiracy. The Egyptian society, as reflected in the corpus, distinguishes political, historical, religious and military schemes from each other. Categorizing enemies depends on the type of conflict or conspiracy, whether in the real world or in the imaginary one. The assumed conspiracy between Egypt and Albania dates back to the era of Muhammad Ali. Still, Egyptian tweets blame Albania for conspiring against the Egyptian history and modern civilization. Some conspiracies, marked in the Egyptian society, are related to virtual conflicts with all Arabs. For instance, Egyptian Twitter users, offended by the Spanish celebration of the end of the Arab ruling, identify Spain as a conspirator against Egypt, as an Arab country, and portray the element of this claimed scheme. Although there evinces no shred of evidence of conspiring against Egypt, it is difficult to dissuade a conspiracist from such a belief.

More frequently than the conspiratorial ideation of the historical background, some of the deeply-rooted conspiratorial ideations in the Egyptian society have political and military bases, such as the Israeli conspiracy. Recently, Libya, Syria and Iraq are represented as schemers against Egypt. Furthermore, this classification is not related to these neighboring countries themselves. This 'enemy outside' labeling is geopolitically-driven because such countries represent a seeding territory for the ramifying ISIS terrorist group.

Banking on some inductive semantic words that connote the meaning of conspiracy, such as – مؤامرة – اشتغالة/ات – افتكاسة هري-تعريص- تحوير-خديعة- حوارات- كذب-مكر, bootstrapped tweets were manually annotated. Results how that the alleged conspiracism at the intersocietal level, political corruption, caused by Islamist, opportunist, governor-citizens hiatus, lack of moral codes, which imposed injustice and immorality, as well as the inability to trust the Other's motivation, either because of their vague acts or because of the perceiver's inadequate perspicacity, are main routes upon which the Egyptian conspiratorial ideation orbits. The following are representative concordance lines that demonstrate so.

الإخوان محل ثقة لأن عندهم خبرة سياسية كبيرة!..طلعت اشتغالة
صورت فيديو بالكام الامامية عشان يطلع صغير وارفعه .. فووووووجئت ان الفيديو ٣ دقايق ٢٠٠ ميجا .. فانا كدا اللي اترفعت..طلعت..اشتغالة ..
اشتغالة اتعب دلوقتي عشان تستريح التصحيح : "اتعب دلوقتي و بعدين هاتتعود ع التعب"..طلعت..اشتغالة
اشرب كوكاكولا انما البيبسى مضر وبيجيب هشاشه فى العظام...افتكاسة
انا ابنى متربى احسن تربية امال مين ولاد الكلب اللى بنقابلهم فى حياتنا دول ؟..وش..اشتغالة
انا معجب بيكى علشان تويتاتك حلوة ....اشتغالة..ساذجة
انت بتحاول تبسطهم بحاجه انت مش بتحبها .. ليه متبسطش نفسك بحاجه انت بتحبها و هم بيكرهوها..الدنيا..افتكاسة
بديع بيقول لا نملك الا سلطان الحب طب كان فين وانت بتعمل انقلاب على عاكف المرشد السابق مستعملتش معاه الحب لية ....اشتغالةغبية
الجزيرة: إسرائيل تفتح الملاجئ في بلدات حدودية مع لبنان عقب إطلاق الصواريخ اشتغاله ولا كلب المجوس هيعمل حاجة اتعلمنا..طلعت..اشتغالة
جسر الملك سلمان هيخلى شرم اللى هى مدينه سياحية تتحول لسفاجا تانيه واشبه بقرية بضائع كبيرة غير إن مافيش ولا حتى اعتماد لفلوس المشروع #هرى

طلعت..اشتغالة..الحب فمصر عباره عن .. حوار تاني خاااالص اليومين دول .. تنتهى باشتغالة
حوراتى فاشلة عصب عنى كل اما آجى أفتكس إفتكاسة تفكس منى
خليك مصدق تسريبات..يابنى..اشتغالة
الدكتور يجي في اخر محاضره في السنه ويقلك سمحوني لو قصرت مع حد واستمعت بالتدريس معاكم ...الخ وكلنا نلبس في الامتحان قراطيس !!!....اشتغالة
علي جمعة : ميزو كاذب ويستحيل أن يكون المهدي المنتظر ..المهدي المنتظر كما جاء في كتب السيرة لابد أن يكون من أبناء القوات المسلحة....افتكاسة
علي فكرة الجو مش حر كلب الجو جميل بس انتوا اللي مش عارفين تتعاملوا معاه ..شكلها..افتكاسة
و عندنا اتسمت كدة بسبب ضم كلمتين "استخبي و غمي" في كلمة واحدة ..مش..افتكاسة..اوى Hide and seek ...عندنا إسمها استغماية عندهم
غالبا محدش بيكون مؤدب أوي ... غير لما يبقى قليل الأدب أوي..بتبقى..افتكاسة
في ناس كدا زي الموبيل الصيني رخيصه اووووي وصوتهم عالي ..بتبقى..افتكاسة
قالك الجيش مش هيسمح بإن دم مصري يسيل..شكلها..اشتغالة
قالها ممكن نتعرف :-! قالتلو اتنيل انا اختك ياموكوس....افتكاسة
قيادي بداعش يدعو مقاتليه للهجرة لمصر ..أجمد..اشتغالة
كنت محتاج حد يديني كورسات #تعريص علي طريقه #احمد-موسي او #مصطفي-بكري
لم يتم التحفظ على اموال ابو تريكة .. ونظامنا الفاشل يستعمل نفس اسلوب مبارك..طلعت..اشتغالة
ليه لا نتفق مع اثيوبيا ان تبني خط انابيب ضخم يصب المياه من اثيوبيا الي مصر بالتوازي مع بناء سد النهضة..بس..افتكاسة
ليه ما يكونش مباركة سد النهضة فى اثيوبا احد بنود الاتفاق اللى وقعه مرسى مع اسرائيل لمصلحة حماس ؟ ..شكلها..افتكاسة
ما هى الجرثومه التى لم يقضى عليها ديتول من بين 99.9 ؟ ..ولا دى..افتكاسة
نفسى فى مره اقابل بوجى وطمطم واسألهم ازاى قرد مجوز ارنبه اشتغاله
و الله لأحبك موت و أحب برج الحوت ..هري
واشنطن تدين فرض حالة الطوارئ بمصر وتعتبره انتكاسة لثورة يناير!!!..دى..افتكاسة؟
وسيقف التاريخ كثيرا عند افتكاسة الافتكاسات..شكلها..افتكاسة
وطول مالناس فى توهه ...مزاج الشيطان عالى ..شكلها..افتكاسة
وقصة "تبادل الجواسيس" ومكر و دهاء بلحه .. يا عيني بيحاولوا يألفوا أي حاجه لحفظ ماء الوجه ..طلعت..افتكاسة
وقومنا عاملين ثورة...افتكاسة
وكأن الحياة لا تهتم بتلك الايام التي تضيع من العمروما الدنيا الا افتكاسة كبيرة ونسمع سلام ...شكلها..افتكاسة
يا منجي نجي من #المهالك اقوله #اسلك يقولي مش #سالك #هري كتير #تحوير كلام في # اقوله اسلك يقولك مش سالك..هري
يعني اي اتنين بيحبوا بعض ويرجعو يقولو احنا اخوات هو الشاي بلبن ينفع يرجع ساده..شكلها..افتكاسة
يعني جيش وشرطة ومدرعات ثقيلة في الشوارع وحظر تجول وقانون طوارئ ومش عارف يأمن مبنى عسكري , ده يا اما وكسة يا اما اشتغالة
يكون القضاء مهنة سامية ، والطب الشامخ؟ تمنياتى بهرى سعيد

All in all, table 2 classifies the 'enemy within' according to the cognition of the Egyptian tweeters. The Egyptian mass media and the high-order authorities are accused of intentionally deceiving the public. Next to these, the Egyptian citizens are themselves blamed for being the absolute threat on Egypt. Theocrats, Daesh (ISIS), pro-regime opportunists and moral decay are also blamed for the devastating dystopia we live in.

TABLE 2

TWEETERS' CONSPIRATORIAL IDEATION AT THE INTEROCIETAL LEVEL

The conspiring enemy	**Percentage (%)**	**Translation**	**Categorization**
الاعلام المصرى	21	Mass Media	Communication channel
الجنرال بلحة	17	The General	Field Marshal/President
المصريين	12	The Egyptians	Citizens
الاخوان	10	Ikhwan	Heterodoxy
الدقنجية	9	Theocrats	Theologians
داعش	6	Daesh (ISIS)	Militia
الدولجية	5	Pro-regime	Deceiver
الرشاوي	5	Bribe	Vice
الواسطة	4	Mediation	Vice
الخونة	3	Traitors	Deceiver
حمضين	3	Hamdeen	Political figure
بتوع السبوبة	2	Opportunists	Improbity
فتي و هري	2	Delusion	Vice
خالد على	1	Khaled Ali	Political figure

Characteristic to linguistic cueing is the use of stylistic and syntactic patterns, which may provide insights into a person's thinking style, such as complexity of thought. For example, "think," "understand," "realize" are casual keywords that describe a person's developed creed. In our studied corpus, the phrase بدات اقتنع ان has proven effective in cueing conspiracy-based thoughts.

بدات اقتنع ان العسكر هما سبب الخراب اللي احنا فيه واللي ساعدهم عواجيز بلدنا اللي اتربوا علي الخوف والجبن
بدات اقتنع ان الشعب المصري بأه مسخره واضحوكة العالم

بدات اقتنع ان الاخلاق والاحترام كدبه تبلها وتشرب ميتها او تلفها وتلبسها
بدات اقتنع ان ان اللي بيحصلنا طول اليوم هو اللي بيخلينا نشخر واحنا نايمين
بدات اقتنع ان الناس فترات وبتخلص .. بس
بدات اقتنع ان اتعب انهارده واتعب بكره وهترتاح لما تموت
بدات اقتنع ان ليس بالاحتلال فقط تتأخر الدول
بدات اقتنع ان احنا عايشين في مجتمع هايف لاقصي حدود كل همه فولانة اتقفشت اداب وعلان ابوه مسكة بيشرب حشيش
بدات اقتنع ان انا هعيش حياتي مره واحده بس فاعمل كل اللي انا نفسي فيه طالما مش بعمل حاجه غلط
بدات اقتنع ان إن سيساوية فرعون مفيش فيهم فايدة ختم الله علي قلوبه
بدات اقتنع ان الانقلاب جي لخراب مصر
بدات اقتنع انت لازم تحب نفسك بالقدر الكافي علشان متحتاجش محبه حد
بدات اقتنع انه مفيش حد يستحق انك تثق فيه لانهم كلهم هيطلعوا عالم زبالة في الآخر
بدات اقتنع ان ف ناس تستاهل فعلا
بدات اقتنع ان الأوضة الضلمه اللى كانت عقاب لينا زمان..بقت دلوقتى المكان المفضل اللى بنحب نقعد فيه..
بدات اقتنع ان البلد دى عايزة دين ام قلة ادب
بدات اقتنع ان الحاجة الوحيده الكويسه فى حياتنا هى الولا حاجه
بدات اقتنع ان الحب ما بيخبتش ع الباب
بدات اقتنع ان الدنيا مسرح كبير واهم دور فيها البلياتشو
بدات اقتنع ان الدنيا مش علي كيف حد
بدات اقتنع ان الرومانسيه كهن ولسلسه
بدات اقتنع ان الشعب المصرى مااااات اكلينيكيا والأحرار يحاولون معه بالصدمات الكهربائية ولكن مصمم على الموت
بدات اقتنع ان الشعب المصري عايز الحرق بجاز وسخ طول ما هو ساكت علي العسكر
بدات اقتنع ان العربجه هي الحل في كل حاجه
بدات اقتنع ان غالبية الشعب بيعشق الذل والمهانة
بدات اقتنع ان لن يحك جلدك الا ظفرك ...ما تحاولش حتى لو جبت هراشه
بدات اقتنع ان اللى يكون كداب وابو لمعه ويشاغل فبنات الجامعه يبقى الواد ده شنكوتي

The deduced linguistic cues denoting conspiracy are syntactically and semantically instantiated. Semantically, as مؤامرة – اشتغالة/ات – افتكاسة – هري-تعريص- تحوير-خديعة- حوارات- كذب-مكر, which are frequently recruited, represent key markers of the conspiracy theory. At the syntactic level, conditional structures are identified as scaffolding of conspiracy elements. To demonstrate, the 'enemy inside' is typically instantiated in the syntactic pattern [Syntactic cues (if only)/ مصر هاتبقى أحسن لو]. Ironically, the filler of the 'enemy inside' slot is, in many cases, 'the Egyptians'. The boundaries between the victim of the conspiracy, supposedly Egypt and the Egyptians, and the plotter, recently the Egyptians, is blurring in the society. Thus, the 'enemy inside' moves beyond the traditional blaming of institutions or governments, in a society, to a more contradictory, and even irrational, level in which the members of the one society are hypothesized to scheme against each other.

مصر هتبقي احسن لو خلصنا م الدقنجية والدولجية ف طلعة واحدة
مصر هتبقي احسن لو ربنا ريحنا من الخونة والاخوان وحمضين وخالد على وبتوع السبوبة
مصر هتبقي احسن لو مفيهاش مصريين
مصر هتبقي احسن لو الاخوان سافروا تركيا
مصر هتبقي احسن لو إتمحت من ع وجهه البسيطة و نجيب حلواني جديد يشتغل ع نضيف
"مصر هتبقى أحسن لو... خِلِصنا من "الجنرال بلحة" وظُلمه
مصر هتبقي احسن لو تم القضاء علي داعش
مصر هتبقي احسن لو الواسطة و الرشاوي اختفوا من مؤسساتنا
مصر هتبقي احسن لو الشعب المصري اختفى من علي وجه البسيطة اكيد
مصر هتبقي احسن لو بطلنا فتي و هري و اشتغلنا زي خلق الله
مصر هتبقي احسن لو قفلنا كتاب التعريص صفحة الاعلام المصرى
مصر هتبقي احسن لو الجيش صدق وعمل جهاز الكفته ال مخبيه من االأمركان

Further removing the border lines between the 'enemy inside' and the 'enemy outside', many subordinating syntactic patterns link schemers at the country level to schemers at the intersocietal level. The 'enemy outside' is believed to be supported by the 'enemy inside' forming a whole of multiple enemies. This shifts the danger of powerful 'outside enemy' to the allied forces of 'within-outside union'.

6 CONCLUSIONS

the linguistic corpus-based detection of conspiratorial ideation is proved to be successful in capturing the schemer and the scheme aspects of the conspiracy. The linguistic cues denoting the conspiracy are most salient in the semantic and syntactic levels. The conspiracy perceived in the Egyptian society is multidimensional. It has political, religious, historical and military aspects. However, the most significant level is the religious one. Most of the outside enemies,

present in the cognition of the Egyptian society, are plotting against Egypt for religion-based reasons. The inside plotters, however, are believed to conspire against Egypt for political reasons. Egypt is believed to be sandwiched by several enemies that wish to destroy the integrity of the Egyptian history, economy, social life and armed forces. This assumed inside and outside conspiracy reflects a state of insecurity regarding both the individual position in the society and the group position among other societies.

REFERENCES

[1] Swami, Viren, Martin Voracek, Stefan Stieger, Ulrich S. Tran, and Adrian Furnham. "Analytic thinking reduces belief in conspiracy theories." *Cognition* 133, no. 3 (2014): 572-585.
[2] M.J. Wood, K. M. Douglas, and R. M. Sutton. "Dead and alive: Beliefs in contradictory conspiracy theories." *Social Psychological and Personality Science* 3, no. 6 (2012): 767-773.
[3] M. Gray. "Explaining conspiracy theories in modern Arab Middle Eastern political discourse: some problems and limitations of the literature." *Critique: Critical Middle Eastern Studies* 17, no. 2 (2008): 155-174.
[4]A. Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Icwsm* 10, no. 1 (2010): 178-185.
[5] Y. Tausczik, and J. W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29, no. 1 (2010): 24-54.
[6] J. W. Pennebaker, and L.A. King. "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology* 77, no. 6 (1999): 1296.
[7] Chung, Cindy K., and James W. Pennebaker. "Using computerized text analysis to assess threatening communications and behavior." *Threatening communications and behavior: Perspectives on the pursuit of public figures* (2011): 3-32.
[8] Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication." *Discourse Processes* 45, no. 1 (2007): 1-23.
[9] M. L. Newman, J. W. Pennebaker, Diane S. Berry, and Jane M. Richards. "Lying words: Predicting deception from linguistic styles." *Personality and social psychology bulletin* 29, no. 5 (2003): 665-675.
[10] M. Jensen, B. Elena Bessarabova, B. Adame, J. K. Burgoon, and s. M. Slowik. "Deceptive language by innocent and guilty criminal suspects: The influence of dominance, question, and guilt on interview responses." *Journal of Language and Social Psychology* Volume 30, no. 4 (2011): 357-375.
[11] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, and Vít Suchomel. "The Sketch Engine: ten years on." *Lexicography* 1, no. 1 (2014): 7-36.
[12] A. Graf. *Orientalism and Conspiracy: Politics and Conspiracy Theory in the Islamic World*. IB Tauris, 2010.

BIOGRAPHY

B. A. Essam is a graduate of Faculty of Arts. He studies a Master Degree in Translation at Al-Alsun, Ain Shams University, Cairo, EGYPT.

Mostafa Aref is a professor of Computer Science and Vice Dean for Society Service & Environmental Development, Ain Shams University, Cairo, Egypt. Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. of Electrical Engineering - Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

TRANSLATED ABSTRACT

الاستدلال على الفكر التأمرى فى التغريدات المصرية من خلال استخدام الانترنت كذخيرة

باسم عبدالله عصام و أ.د. مصطفى محمود عارف
**قسم اللغة الأنجليزية، كلية الألسن، جامعة عين شمس*
***قسم الحاسبات، كلية المعلومات و الحاسبات*
[1]literaryartrans@gmail.com
[2]mostafa.aref@cis.asu.edu.eg

ملخص:
يعتمد هذا البحث على استخدام الإشارات اللغوية للكشف عن عناصر المؤامرة في ذير معاصرة من تغريدات المصريين (2012-2017) حيث يتم تحديد الأعداء المفترضة في المجتمع المصري كميا. فاستخدام الأدوات اللغوية لتحديد جوانب النظرية الاجتماعية والسياسية للتآمر أداة موضوعية قابلة للاسترجاع . وتشير النتائج إلى أن المجتمع المصري يصف نوعين من الأعداء أو المتآمرين: الخصوم داخل المجتمع وخارجه. والفكر التآمرى، على الصعيد الدولى، ذات بعد سياسي ودينى وتاريخي. ومع ذلك، فإن المخططات الداخلية لا تميز إلا التوجه السياسي.

الكلمات الدلالية: استدلال لغوى – الفكر التآمرى – الأنطولوجيا العربية – التواصل الاجتماعى- علم الذخائر اللغوية

Modern Standard Arabic Grammar Extraction from Penn Arabic Treebank Using Natural Language Toolkit

Amira Abdelhalim[1], Sameh Alansary[2]
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
[1]Amira.Abdelhalim@yahoo.com
[2]s.alansary@alexu.edu.eg

***Abstract*—this paper present a methodology forrule based bottom up parsing technique for Modern Standard Arabic (MSA) in Context Free Grammar (CFG) formalism in Phrase Structure Grammar (PSG) representation, where the grammar is automatically extracted from an annotated corpus. The used grammatically annotated corpus is Penn Arabic Treebank(PATB)and for algorithm implementation Natural Language Processing Toolkit (NLTK).Furthermore, the extracted CFG is further transformed into Probabilistic Context Free Grammar (PCFG) that could be used in a hybrid approach, which is also calculated automatically. The parser showed that automatic extraction of grammar improved the grammar building phase in both coverage of structures and time needed, but still needs further manual constrains addition. Automatic extraction of grammar is able to enhance rule based grammar parsers and it will enable linguists to start a competitive challenge in front of statistical parsing.**

Keywords: **Observational Based Grammar - Automatic Grammar Extraction- Rule Based Grammar – Enhancing Arabic Grammar Parsing**

1 Introduction:

Parsing is responsible of determining the syntactic structure of an expression. Syntactic parsing is a vital step in any Natural Language Processing (NLP) application.

Many attempts have been proposed to the study of syntactic structure analysis and generation, but only some of them have been proposed to Arabic. Syntax is concerned with describing the logical sequence of sentence units. Syntactic analysis process has been defined as ―the process of analyzing a sequence of tokens to determine its grammatical structure with respect to a given formal grammar. Parsing is used to refer to the process of building automatically syntactic analysis of sentences according to a given grammar (Grune and Jacobs, 1990).The parsing transforms input text into a data structure, usually a tree, which is suitable for later processing and which captures the implied hierarchy of the input, where different grammatical frameworks have been proposed (Al-Daoud and Basata, 2009).

2 Parsing Approaches:

Three main approaches are recognized for parsing: the linguistic rule based approach, statistical approach and hybrid mixture of the two. The first linguistic approach uses lexical knowledge and language rules in order to parse a sentence. It is very promising approach but requires huge amount of work and time. On the other hand, statistical approaches are based on statistics and probabilistic models. It is based on the frequencies of occurrences that are automatically derived from corpora. It is known for fast development that saves time and effort but still has many challenges due to the complexity of language infinite identity type, reflecting human mind. The third hybrid approach integrates both of them, taking advantage of grammar rules robustness and statistical models fastness. This paper extracts grammar automatically, for both rule based parsing in CFG and PCFG and uses the set of grammar rules from them on further data.

3 Formal LanguageCFG and Rewrite Rules:

In both mathematics and linguistics a formal language is a set of strings of symbols that may be constrained by rules that are specific to it. It is used as an agreed language to describe some knowledge of certain kind of data or to define the relationship between elements (linguistic data) and their representation formally. For linguistics, one of

the commonly used formal languages is called CFG. CFG consists of a set of rewrite rules with certain categories of terminal and non-terminal symbols defined by the linguist of the form A --> B, where A belongs to the set of non-terminals and B belongs to the set of terminal or non-terminal symbols. CFG defines a formal relationship between a set of possible texts and their representations. Using this language with any linguistic representation (dependency, phrase structure or feature based) is able to supply a representation of sentences using these rewrite rules.

It is used to describe or define the sentences, whereas the representation combined with a certain linguistic theory used as a procedure or instructions to be followed. This bundle of rules as well as the chosen approach of sentence representation is called Generative Grammar.

(Sarkar, 2011) illustrated parsing issues, CFG as one, and stressed the important point that using CFG for the syntactic analysis of natural language is very problematic. The grammar of natural languages is far too complicated than just listing a set of rules; he described it as being similar to an acquisition problem. He also highlighted the second problem of resolving ambiguity such as recursive rules.

However, this limitation has been reinforced by the addition of augmentation and features to rules. The sub categorization features of the categories may also be added between brackets V [transitive] and sequence is represented by order and sentence position by dash [- NP]. Sub categorization features is added to CFG as appropriate restriction formal representation added to represent context.

4 Basic Search and Matching Strategies for Parsing:

Two basic approaches of Top-down and Bottom-up parsing, as other approaches are based on them. The start point of handling the data is the first basic decision that needs to be taken in the parsing process. In top-down parsing, the process starts from the most abstract point, in our case study of syntactic structure of PSG, it is the S and directs towards the lowest level building the structure reaching words. On the other hand, in the bottom-up approach the parsing starts at the lower level, which is words, and attempts to build upwards. In most real applications, the top-down approach is commonly used with statistical parser whereas bottom-up is used with rule-based applications. The recursive rules of rule-based applications with a large grammar and many potentially ambiguous sentences predicts along with top down approach an infinite variety of possible structures. On the other side, using bottom-up approach makes it possible to parse the hypothesis list faster as it goes upwards testing through a defined set of restricted categories. Suppose the proposed grammar contains the following set of rules that are written in terms of categories, taking into consideration that the lexicon also contains the words with their features attached,

a. S -> NP - VP
b. S -> NP – VP – PP
c. NP ->Det – N
d. NP ->Det – Adj – N
e. NP ->Pron
f. NP ->Det – Adj – N – NP
g. NP ->Det – N – PP
h. PP -> Prep – NP

Looking at the number of possibilities using top-down technique along with these possibilities embedded in the recursive rules the number of predicted structures is enormous even before consulting any word in the lexicon. Bottom-up process less possibilities but does not consider a backtrack solution.

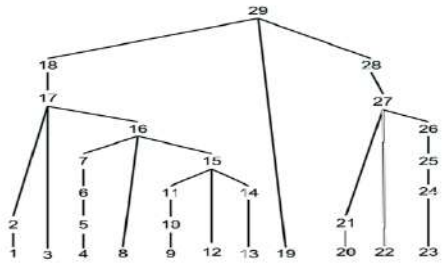


Figure 1: Bottom-Up Parsing

5 Intuition Based Vs. Observation Based Grammars:

In order to formulate rules two main approaches have to be discussed, intuition-based grammars and observational grammars (Aarts, 1991). The intuition based grammar was adopted by Chomsky, it is based on constructing sentences and introspection. The second is based on actual texts taken as evidence to draw conclusions as corpus linguists do.

Corpus provides empirical data and in case conjoined with a computational tool, it addresses issues that were previously intractable, as not only it allows for quantitative analysis, but also investigation of structures embedded in real discourse (Biber et al., 1998). Corpus have had opened new areas of research in grammar. It facilitates the study of a single grammatical construction and obtains information about the usage of different grammatical constructions and uses this information as the basis for writing a reference grammar (Meyer, 2004).

6 Grammar Development Strategies:

The rule-based grammar is usually built either with Manual Grammar Development, toy grammar, that needs a skilled human team with a solid experience and knowledge in both theoretical linguistics and grammar formal representation. The major problem is time and consistency of each rule represented. That‘s why different Grammar Development Environment of software systems offer grammar writers incremental input, grammar editing, browsing, searching and tracing or debugging. The other approach is Automatic Grammar Induction which is based on Treebank‘s, as the linguistic intuition is externalized into the annotation of the Treebank and the grammar. It is a fast and cheap method (Kakkonen, 2007).

7 Related Works:

Some trials concentrated on rule based parser such as (Ouersighni et al., 2001) used Affixs Grammars over Finite Lattices formalism to build Arabic morpho-syntactic analyzer. (Othman et al. 2003) used Unification Based Grammar formalism. Some trials also concentrated on statistical parser such as (Tounsi et al., 2009) developed a parser that learns from Penn Tree Bank (PTB) the functional labels to use it in Lexical Functional Grammar formalism. (Ben Fraj, 2016) used PTB as a learning data in order to extract most common trees for syntactic interpretation of new sentences with accuracy 89,85%. (Diab et al., 2007) used machine learning methods for tokenization and part of speech (POS) tagging and base phrase chunking, it used 10% of the PAT corpus with F-score of 96.33%.

As for Arabic and CFG(Al Taani, Msallam and Wedian, 2012) used CFG for designing a top down parser for simple Arabic sentences with specific domain. They developed a precise description of Arabic grammatical sentences to feed their parser with. The parser starts with word classification, rule identification then parsing. They mentioned that it showed effective results for MSA sentences. They used simple sentences both verbal and nominal from real documents, but for a specific domain with accuracy 70%. (Alqrainy et al., 2012) implemented a parser that checks Arabic sentence grammatical structure well-formedness. Their top-down parser scored average accuracy rate of 95%.

It is obvious that each trail whether statistical or rule based has its own formalism, parser and even evaluation metric; which causes comparison difficulty to researchers.

8 The PATB Corpus:

Treebanks are a collection of syntactically annotated sentences of a large amount of corpora. PATB is considered the most usable Treebank that uses PSG and also available for Arabic. It is a syntactically annotated Treebank‘s that is vital for training parsers as well as finding constructions for any syntactic study, specifically development of grammar based parsers.

The PTB project started in 2001 at the Linguistic Data Consortium and University of Pennsylvania. It offers two types of conventions, the original constituency and a converted dependency representation in the Columbia Arabic Treebank (CATiB), for many languages including Arabic. It consists of 23,611 parse annotated sentences from Arabic newswire text in MSA. It is one of the most significant transitions of Arabic NLP as many researches and tools for morphology and syntax, data-driven or rule based depended on it as a standardized source of annotated data (Maamouri&Bies et al., 2004). Many of the significant Arabic NLP is based on it, the morphological analysis, disambiguation, POS tagging and tokenization (Habash et al., 2005).

The version used is part three, version one that consists of, basically 600 stories from Al Nahar News Agency, referred to as ANNAHAR. The stories are specified with a DOC ID along with date. The average number of words per story is 567 and total word token is 340,281.
The corpus is first annotated with Tim Buckwalter’s lexicon and morphological analyzer to generate a list of candidate POS tags for each word. The second step is manual choice from candidate tag (lexical category) along with inflectional features and gloss and then automatic clitic separation and then parsing annotation of constituent structure along with functional function categories for each non-terminal node. The main files that are vital are the “.sgm” file that contains the raw corpus, the “.tree” file that has the parsed annotated corpus.

Features and Their Annotation:

Main Features of Penn Treebank Constituent tags:

S	sentence
NP	noun phrase
VP	verb phrase
PP	prepositional phrase
SBAR	S-bar (subordinate clause, complementizer or WH- and sentence)
SBARQ	S-bar that is a question
SQ	S that is a question
NX	noun head in certain complex coordination contexts
PRN	parenthetical
PRT	particle
QP	quantity phrase (multi-word numbers)
ADJP	adjective phrase
ADVP	adverb phrase
FRAG	fragment
WHNP	WH- noun phrase
WHPP	WH- prepositional phrase
WHADJP	WH- adjective phrase
WHADVP	WH- adverb phrase
CONJP	conjunction phrase (multi-word conjunction)
INTJ	interjection
NAC	Not-A-Constituent (mostly rightward moved conjuncts with conjunction)
UCP	Unlike-Coordinated-Phrase (dominates coordination of NP and PP, e.g.)
X	unknown, technical problem, etc.

9 Grammar Extractions and Parsing:

NLTK Framework:

NLTK is a python platform for building and testing NLP applications. It provides easy to use libraries based on Object Oriented Model of programming. Its libraries are organized into packages of modules, classes and functions that are easily used for different purposes such as classification, stemming, and tagging, parsing and semantic reasoning. It also offers a powerful API documentation.

It is used in this paper as the platform that is responsible for reading the parsed corpus, extracting CFG and CFG augmented with Features grammar, calculated probability for PCFG augmented with features productions generation, drawing parsed trees representation, generating files of written extracted grammar both rule based and probabilistic grammars and testing these extracted files on further data.

Algorithm:

The CFG class first identifies the non-terminal symbol as an object and then expands it to the right hand side. It accepts a feature structure object, a grammatical category along with its features description in CFG representation, which is used in a feature based grammar and equivalent to CFG but all non-terminals are feature-struct non-terminal of feature based grammar in CFG augmented with features. This feature structure is important to represent annotated data grammar of a parsed corpus. The grammar production maps a single symbol on the left to sequences on the right.

It can construct a probabilistic production by creating another new object from the given start state and a set of probabilistic productions. It takes the featured CFG productions and return featured PCFG production. A featured PCFG consists of a start state and a set of productions with probabilities. The set of terminals and non-terminals is implicitly specified by the production. Any given left hand must have a probability that sum to 1.

Table 1: Algorithm for Extracting Grammar with NLTK

Read Arabic Penn Treebank	**Preprocessing.**
	Create a Parsed Corpus Reader.
	Iterate over Parsed Sentences.
	Draw Sentences Trees (to check reader and sentences).
	Extract Grammar Rules and Lexical Rules.
	Write Them in a File.
	Print Number of Sentences and Tokens.
PCFG	**Split Data To Training and Testing**
	Extract CFG Grammar and Lexicon
	Add Probability on CFG
	Write PCFG To a File
Testing Grammar:	**Open and Read Extracted PCFG File** **Open and Read CFG File**
	Create a Probability Parser to use PCFG Extracted Grammar For Parsing Test Sentences. **Create a CFG Parser and Read Extracted FCFG Grammar For Parsing Test Sentences.**

The Training Phase:

The training phase involves the usage of the parsed corpus to extract the grammar rules along with their features. The corpus is divided into a training set and testing set. For training, a preprocessing phase is performed where each annotated sentence is copied manually to a file, each sentence in a separate line. The combination of the features and categories allows the training corpus to learn allocation of each word in the sentence as grouping of sequence of labels, both features and categories, in the most probable syntactic group. The extracted grammar is saved to a file.

Rule: S -- > (C1) (C2)
C1-- > (W1, W2)
C2 -- > (W3, W4)
W -- > terminal word as written in raw text

Where S represents the sentence, C represents the constituent category non- terminal label and W represents the word category along with their features. The first three rules are called grammar rules, whereas the last is a lexical rule.

Examples of Extracted Rules:

Grammar productions (start state = S)
NP-SBJ-1 -> -NONE-
PP -> PREP NP
VP -> IV3MS+IV+IVSUFF_MOOD: I NP-SBJ-3 NP-OBJ NP-ADV
NP-OBJ-2 -> -NONE-
PUNC -> '-LRB-'
S -> CONJ VP PUNC
S -> VP
IVSUFF_DO:1S -> 'ny'
NOUN+NSUFF_FEM_SG+CASE_INDEF_NOM -> 'mfAj>p'
NEG_PART -> 'lm'
NP-OBJ -> DET+NOUN+CASE_DEF_ACC
NEG_PART -> 'lA'
S -> NP-TPC-1 VP

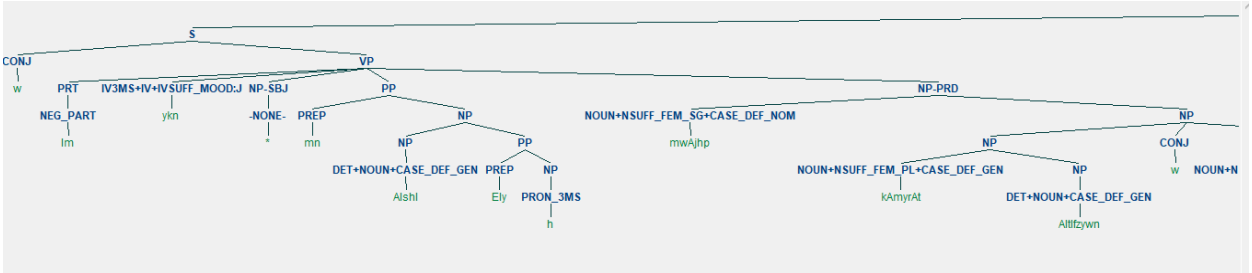


Figure 2: Extracted Grammar and Lexical Items for a Part of a Sentence

Calculate PCFG:

NLTK could be used as well for constructing probabilistic models. Internally the library generates a descriptive extraction vector for each word by its morphological features, both its category along with features. The vector is completed by the appropriate syntactic class of the non-terminal constituent label. Each vector represent the corpus in a tabular way which consists of the words sequence, represented in terms of features, and the return at the end of it (vector 1: Det N ?, NP).

The Testing Phase:
The testing corpus also contains a set of annotated sentences and same set raw un-annotated each in a line. Both extracted PCFG and CFG grammar are used to analyze it. The parsing is generated in a file along with the tracing of the steps.

10 Conclusions:

Automatic extraction of grammar improved rule based parsing in terms of coverage and time needed. Bottom Up approach does not permit backtrack, Top-down may result in better solutions as it permits backtrack. Ambiguity will be further refined with the addition of manual constrains, that could be studied from the as a benefit of automatic extraction from the Treebank.

BIOGRAPHIES:

Amira Abdelhalim: Teacher Assistant, Faculty of Arts, Phonetics and Linguistics Department, Alexandria University. She got her MAwith excellent degree on "A Formal Approach to Modern Standard Arabic Syntax: A Corpus Based Study" in 2016.Her main areas of interest are corpus based studies, Arabic morphology, Arabic syntax, Arabic semantics, Machine Learning Techniques and Language Modeling.

Dr. SamehAlansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.

Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

References

[1] Aarts, J. (1991). Intuition Based and Observation-Based Grammar. In Aijmer

[2] Al-Daoud, E., &Basata, A. (2009). A framework to automate the parsing of Arabic language sentences. Int. Arab J. Inf. Technol., 6(2), 191-195.

[3] Alqrainy, S., Muaidi, H., &Alkoffash, M. S. (2012). Context-Free Grammar Analysis for Arabic Sentences. International Journal of Computer Applications,53(3), 7-11.

[4] Al-Taani, A. T., Msallam, M. M., &Wedian, S. A. (2012). A top-down chart parser for analyzing arabic sentences. Int. Arab J. Inf. Technol., 9(2), 109-116.

[5] Ben Fraj, F., Ben Othmane-Zribi, and Ben Ahmed, M., 2010, ―Parsing Arabic Texts Using Real Patterns of Syntactic Trees‖ The Arabian Journal for Science and Engineering, Volume 35, Number 2C.

[6] Biber, D., Conrad, S., &Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

[7] Diab, M. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In 2nd International Conference on Arabic Language Resources and Tools (Vol. 110).

[8] Grune, D., & Jacobs, C. Parsing Techniques–A Practical Guide. 1990. VU University. Amsterdam.

[9] Habash, N., &Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics (pp. 578-580).

[10] Kakkonen, T. (2007). Framework and resources for natural language parser evaluation. University of Joensuu.

[11] Maamouri, M., Bies, A., Buckwalter, T., &Mekki, W. (2004, September). The pennarabictreebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (pp. 102-109).

[12] [12] Meyer, C. F. (Ed.). (2002). English corpus linguistics: An introduction. Cambridge University Press.

[13] Othman, E., Shaalan, K., &Rafea, A. (2004, September). Towards resolving ambiguity in understanding arabic sentence. In International Conference on Arabic Language Resources and Tools, NEMLAR (pp. 118-122).

[14] Ouersighni, R. (2001, July). A major offshoot of the DIINAR-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts. In ACL 39th Annual Meeting (pp. 9-16).

[15] Tounsi, L., Attia, M., & van Genabith, J. (2009). Parsing Arabic using treebank-based LFG resources.

استخلاص قواعد النحو والتحليل التركيبي لجمل اللغة العربية المعاصرة آليا بإستخدام عينة لغوية من Penn Arabic Treebank وبناء محلل نحوياستخدام NLTK

Amira Abdelhalim[1], Sameh Alansary[2]
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
[1]Amira.Abdelhalim@yahoo.com
[2]s.alansary@alexu.edu.eg

ملخص—تقدم هذه الورقة البحثية نظاما للتحليل النحوي الآلي للغة العربية المعاصرة. تعتمد المنهجية على بناء محلل نحوي يقوم بإستخلاص القواعد النحوية للجملة آليا من خلال مدونة لغوية موسومة بتحليل تركيبي للجمل. المدونة المستخدمة هي **Penn Arabic Treebank** والأداة الحاسوبية لبناء المحلل النحوي واستخلاص القواعد لبناء معجم آلي للقواعد والكلمات هي **NLTK**. وقد أدى بناء معجم الكلمات والقواعد إلى سرعة البناء وشمولية المركبات الممثلة ولكن لازالت القواعد تحتاج لإضافة قيود لفك اللبس التركيبي للجمل المتشابهه في أجزاء منها من الوحدات التركيبية.

A Formal Grammar for Describing Modern Standard Arabic Structures

Marwa Saber [1], Sameh Alansary [2]

Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
[1]marwa.saber@bibalex.org
[2]sameh.alansary@bibalex.org

***Abstract*—the structure of the sentence in Modern Standard Arabic (MSA) could be described through outlining a hypothesis, and then it should be tested on authentic data in an automatic processing environment. In this paper, the data is syntactically analyzed by Affix Grammar over Finite Lattices (AGFL) formalism. The output has been evaluated using a version of the processed corpus annotated manually. The method that is adopted for the evaluation of the results is the precision and recall. Precision was 0.90 and recall was 0.83. The result was compared to the output of Stanford Parser.**

Keywords: AGFL, Parsing, MSA, Syntax, Computational Syntax, Computational Linguistics, NLP, Syntactic Analysis

1 INTRODUCTION

In the framework of generative linguistics (Universal Grammar – UG), [1] has been elaborating an interpretative description of sentence structure in Modern Standard Arabic (MSA). He adapted by means of specific raising rules the basic unmarked VSO (verb – subject – object) sequence in order to accommodate for sentence structures in MSA following a clear SVO (subject – verb - object) sequence. Positive in this approach is the attempt to extent the scope of universal grammar applicability with language facts from other natural languages. Negative is the identification of the SVO order typology with the nominal sentence structure in MSA in general. Instead of [1] IP-structures we rather prefer to present a single IP-structure description in which the slash represents alternatives.

This paper will present two different modules: the first module is responsible for parsing Arabic syntactic structures and transform Arabic sentences to the trees structure using binary relation based on X-bar theory with the AGFL formalism. The general design for this module is that "it starts by composing small trees for the small phrases in the sentence and combining these small trees together to form a bigger tree". While building the syntactic trees for the 85 Arabic sentences many linguistic issues have been faced, some will be described in the following sub-subsections. The second module is the semantic module: This module is responsible for mapping the syntactic rules with their equivalent semantic relations.

The NL Reference Corpus (NC) consists of about 500 words of Arabic text that are compiled from the Arabic Wikipedia. The corpus is intended to be representative of the contemporary standard use of the written Arabic language; it includes documents from various genres and domains. It is segmented into sentences and tagged for POS with a dictionary of 500 words, besides all required linguistic attributes are assigned to each word. The total number of distinct sentences is 80 sentences. This paper is concerned with the 80 annotated sentences with sentence length < 12 words. Moreover, the frequency of each structure is documented. Since that, these sentences are compiled from the Wikipedia, so their coverage rate is high. Thus, the opportunity of the occurrence of different structures is extremely high, as a result the data is considered more robust. The semantic analysis of the corpus is supplemented in this phase.

2 LINGUISTIC DESCRIPTION OF THE STRUCTURES

The data consists of different types of phrasal categories. The definiteness of nouns can be used as a cue in determining the noun phrase boundaries. In Arabic, the definiteness can by edafa or by the definite article "ال" 'the'. In nominal phrases, the topic 'mobtadaa' should be definite; therefore, if the noun was not definite by a definite article, it should be definite by edafa. After determining the topic, the predicate should be determined. The predicate could be a single noun or clausal; clausal predicates are noun phrase, verb phrase, prepositional phrase and adverbial phrase. The topic and the comment together form the noun phrase boundary, which means that, it is a complete sentence that does not permit further modification. The researcher intends to vary the structures of the different constituents of the sentences. For example the subject of the sentences in the data was presented in different structures, also the object and the other functions in the sentences varies in its structure. Hence, these different structures were implemented in the developed grammar in the AGFL formalism.

A. The linguistic description of the subject in the corpus

The subject is the constituent that separates the verb from its complement in the VSO structure. The patterns of the subjects in the selected data are diverse. The following subsection will introduce the different patterns of the subject in details:

1) Null subjects

It is one of the most vexing issues. They result from pro-drop clauses in which there is no lexical subject. Instead, the inflection of the verb indicates the gender, number, and person of the pronominal subject. The direct object often appears adjacent to the verb and there is no need to perform re-ordering during the automatic analysis. During the syntactic analysis, the pronouns are projected directly to their maximal projection, which is the noun phrase (NP); they do not have specifiers or complements. Consider the subject in sentence (1) and its representation in figure 1.

2) A pronoun (PRO)

1) نعتمد نحن في حياتنا العصرية كثيرا علي الوجبات الجاهزة والسريعة.

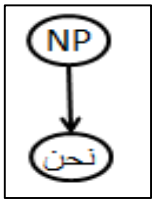


Figure 1: The representation of the subject "نحن"

3) Clausal subject: Different types of structures that can function as the subject have appeared in the selected corpus. The following are the types of structures:

- Proper nouns (PPNs): Proper nouns are projected directly to their maximal projection, which is the noun phrase (NP). Consider the sentence in (2):

2) خدم بنيامين نيتنياهو العالم العربي من حيث يدري أو لا يدري

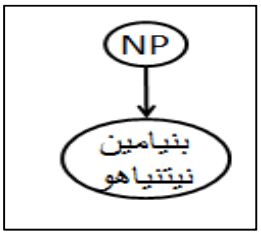


Figure 2: The representation of the subject "بنيامين نيتنياهو" of the sentence in (3).

The proper names do not have specifiers or complements [2] and [3] as shown in figure 2.

- Noun phrase (NP) that consists of ARTICLE (ART) + NOUN as in sentence (3):

4) المشروع يضم مسجدا وقسما للادارة والاستقبال ومركزا تجاريا ومطعما ومركزا للخدمات.

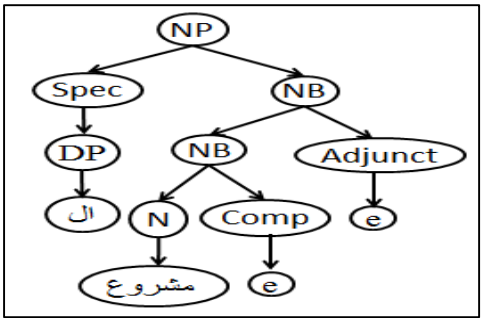


Figure 3: The representation of the subject "المشروع"

In figure 3, the article "ال"'the' is projected to its maximal projection; the determiner phrase (DP) to constitute the specifier of the whole noun phrase. The noun "مشروع" 'project'is combined with the empty complement (e) and is projected to the intermediate projection noun phrase (NB).Then; it is combined with the empty adjunct to form a bigger NB. Hence, it is

combined with the specifier "ال" 'the' (DP) to form the maximal projection noun NP; the definite article has been analyzed as the head of a Determiner Phrase (DP) to which the nominal head raises and incorporates [1].

- Noun modified by another noun phrase (N+NP) (مضاف ومضاف إليه) as in sentence (4):

5) مطار ابو ظبي يخدم نحو 3ملايين مسافر سنويا.

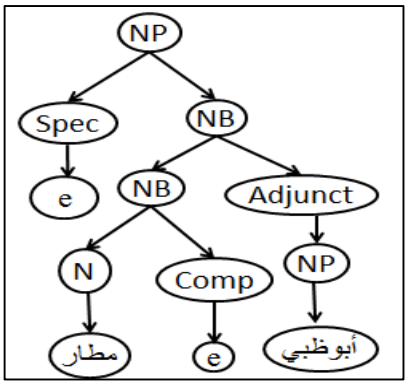


Figure 4 :The representation of the subject "مطار أبوظبي"

In figure 4, the noun "مطار" 'airport'has no complement, so it is combined with an empty one to form an intermediate projection (NB) [4], which is combined with the directly projected noun phrase "أبو ظبي" 'Abu Dhabi' form a bigger intermediate projection (NB) "مطار أبو ظبي" 'Abu Dhabi airport'which is combined with an empty specifier to form a noun phrase (NP).

- Noun phrase that consists of (Noun modified by noun phrase(NP) that consists of (noun +NP) and a prepositional phrase (N + NP + PP)) as in sentence (5):

6) شركة مدينة نصر للاسكان تمسكت بحقها.

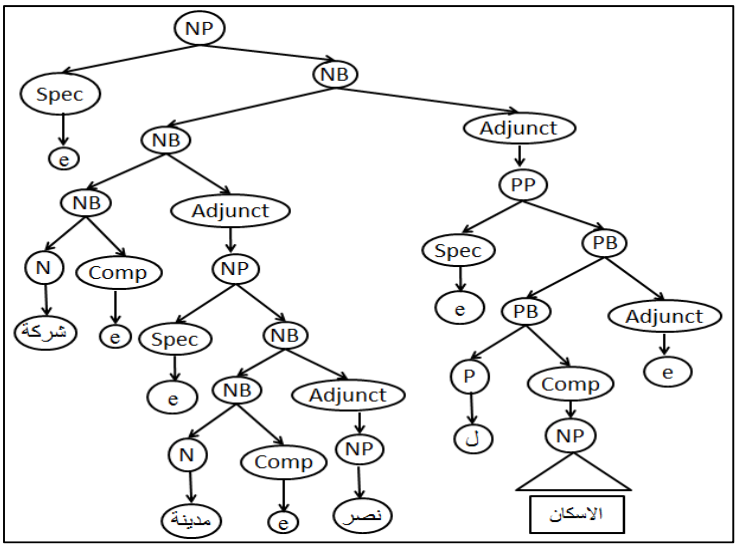


Figure 5 :The representation of the subject "شركة مدينة نصر للاسكان"

In figure 5, the noun "شركة" 'company'has no complement, so it is combined with an empty one to form an intermediate projection; a noun phrase N-bar (NB), which is combined with the adjunct "مدينة نصر" 'Nasr city' which is a noun phrase consisting of noun "مدينة"+ proper noun "نصر"to form a bigger intermediate projection (NB) "شركة مدينة نصر" 'Nasr city company'. The projected (NB) is combined with the prepositional phrase (PP)"للاسكان" 'for housing'that is considered its adjunct to form a bigger intermediate projection (NB) which is combined with an empty specifier to form a noun phrase (NP).

- Noun phrase that consists of DET +Noun modified by two adjectival phrases (DET + N + JP + JP)) as in sentence (6):

7) «البنك الاهلي التجاري»حقق أرباحا صافية العام الماضي.

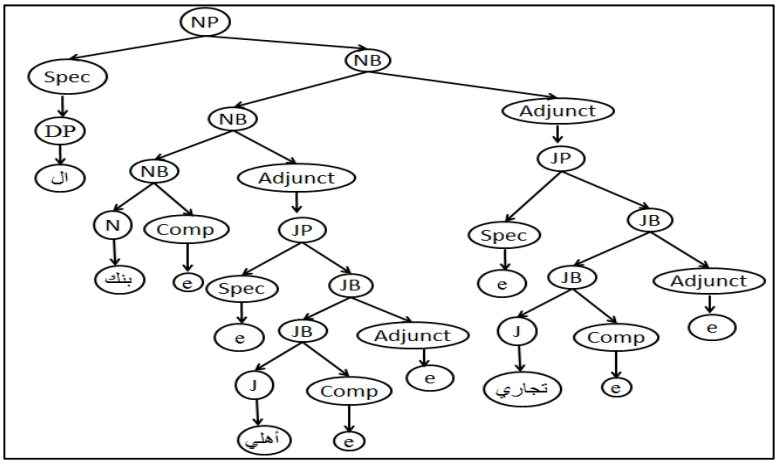


Figure 6: The representation of the subject "البنك الأهلي التجاري"

In figure 6, the noun "بنك" 'bank'has no complement, so it is combined with an empty one to form an intermediate projection (NB) which is combined with the adjectival phrase (JP) "أهلي" 'National' that is considered as its adjunct to form a bigger intermediate projection (NB) which is combined with another adjectival phrase (JP) "تجاري" 'commercial' to form a bigger intermediate projection (NB) which is combined with the specifier "ال" 'the' to form a noun phrase (NP).

- Noun phrase that consists of (Noun modified by noun phrase(NP) and two adjectival phrases (N + NP + JP + JP)) as in sentence (7):

8) حققت شركةالنقل المتحدةالمحدودة مبيعات بلغت 43مليون ريال.

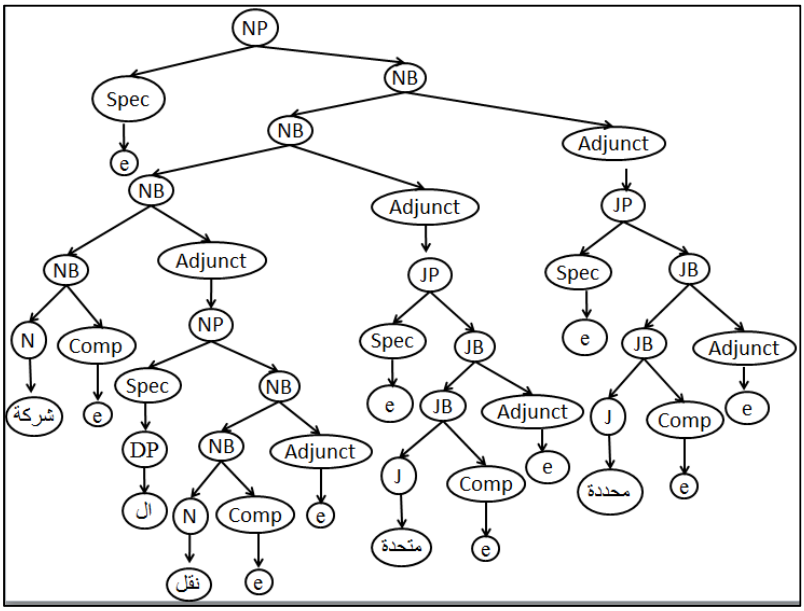


Figure 7: The representation of the subject "شركة النقل المتحدة المحدودة"

In figure 7, the noun "شركة" 'company'has no complement, so it is combined with an empty one to form an intermediate projection noun phrase N-bar (NB), which is combined with the adjunct "النقل" 'transport' which is a noun phrase (NP) that consisting of article "ال"+ noun "نقل"to form a bigger intermediate projection (NB) "شركة النقل" 'transport company'. The

projected (NB) is combined with the adjectival phrase (JP) "متحدة" 'united' that is considered its adjunct to form a bigger intermediate projection (NB) which is combined with another adjectival phrase (JP) "محددة" 'limited' to form a bigger intermediate projection (NB) which is combined with an empty specifier to form a maximal projection noun phrase (NP).

B. The linguistic description of the object in the corpus

In this section, the researcher will introduce the different structures of phrases that may occur as complements of the selected verbs in the data. All analyses provided in this thesis represent the deep representations of the sentences, because the researcher decided that it is more appropriate to identify the arguments of the predicate out of the deep representations. In general, a triangle under a phrasal node means that the further structure is not shown, because it is irrelevant to the point being discussed.

9) ينظم اتحاد غرف التجارة والصناعة والزراعة في البلاد العربية ندوة عن[1]

"The federation of chambers of commerce, industry and agriculture organizes a symposium in the Arabcountries."

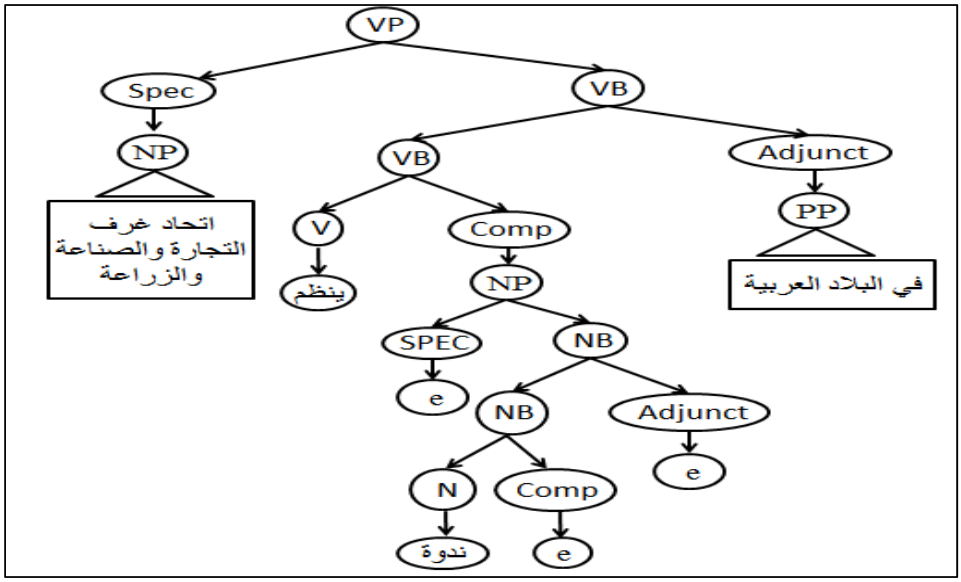


Figure 8: The detailed representation of the complement "ندوة" in sentence (8).

After the analysis of the sentence in (8), the sister node of the verb "ينظم" 'organize' is analyzed as a noun phrase, consisting of a noun "ندوة" 'symposium'that is projected to its intermediate projection (NB) when it is combined with an empty complement and then combined with an empty adjunct to be projected to a higher intermediate projection- noun phrase double bar (NB) which is finally combined with an empty specifier to form the maximal projection noun phrase (NP) "ندوة" 'symposium' as in figure 8.

10) ينظم مركز القاهرة الاقليمي للتحكيم التجاري الدولي مؤتمرا قانونيا دوليا

"The Cairo regional center for international commercial arbitration organizes an international legal conference"

[1] The preposition عنwill be omitted in the processing phase which will be detailed in section 5

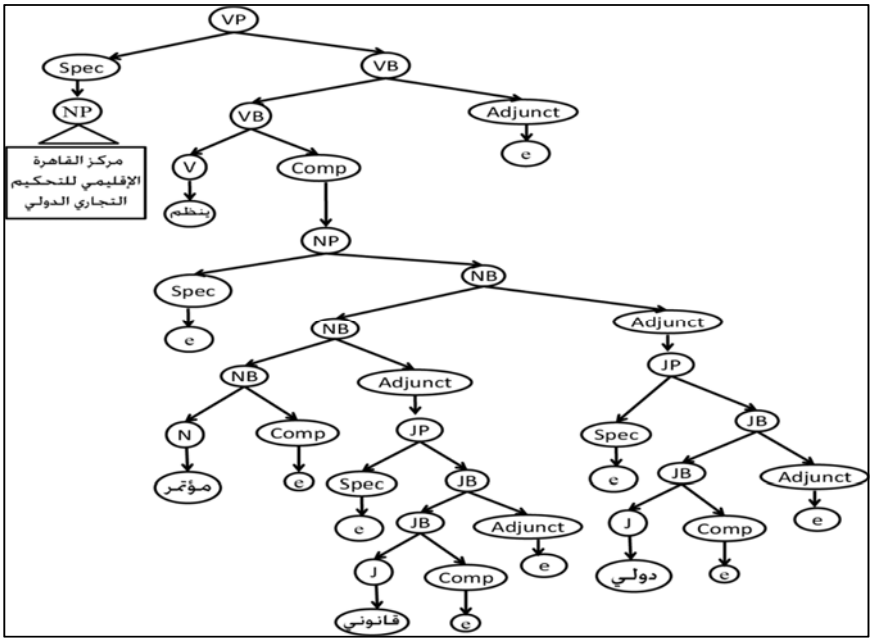


Figure 9: The detailed representation of the complement "مؤتمرا قانونيا دوليا" in sentence (9).

In figure 9, the noun "مؤتمر" 'conference'has no complement, so it is combined with an empty one to form an intermediate projection (NB), which is combined with the adjectival phrase (JP) "قانوني" 'legal' that is considered as its adjunct to form a bigger intermediate projection (NB), which is combined with another adjectival phrase (JP) "دولي" 'international' to form a bigger intermediate projection (NB) which is combined with an empty specifier to form a noun phrase (NP) "مؤتمر قانوني دولي" 'international legal conference'. Here the *complement* is *preceded* by a subject that is a noun phrase " مركز القاهرة الاقليمي للتحكيم التجاري الدولي" 'Cairo international center for international commercial arbitration'.

3 FORMAL DESCRIPTION OF SENTENCES

The AGFL formalism for the syntactical description of natural languages belongs to the family of two level grammars: a context free grammar that is augmented with set-valued features for expressing agreement between syntactic categories. The formalism is suited for describing morphological structure and syntactic structure, case distinction and agreement and finite semantics [5],[6] and [7]. The AGFL parser-generator compiles grammars written in the AGFL formalism into compact and efficient parsers, reads sentences from a file, or prompting the user for input, the output consists either of decorated parse-trees in one of several formats, or of strings as specified by the transduction.

The LEXIGEN lexicon system makes it possible to connect large lexical databases to your grammars in an efficient way. The lexicon system is fully integrated with the parser generator [5], [6] and [7].

Affix grammars can be seen as the formalization of a notion of a Context Free (CF) grammar extended with features. It supports two kinds of applications: linguistic applications: where all analyses of a given utterance are to be found and Information Retrieval and Natural Language Front ends applications: the most likely analysis has to be found for consecutive segments of a running text[8] and [9] .

Affix grammars can be seen as a two level attribute grammar formalism as follows [10]:

The first level: A syntax rule without a second level is just a CF rule; it consists of context-free syntax rules, rewriting non-terminals to terminals or to other non terminals.

RULE sentence: subject, verb ------ one production
RULE subject: Personal pronoun; Noun phrase
------ two productions

The second level: adds parameters (called affixes) to the first-level rules. The characteristic property of AGFL is that the affixes are finite set-valued. The two levels are combined by using affixes as parameters to the non-terminals in the syntax rules.
Consistent substitution rule: all occurrences of an affix within one rule obtain the same value.

- Rule sentence : subject (NUMBER) , verb (NUMBER)

- Affix grammar is more compact than the context free grammar

GRAMMAR cfg.
ROOT sentence.
RULE sentence: subject_plural, verb_plural.
RULE sentence: subject_singular, verb_singular.
RULE subject_plural: "نحن".
RULE subject_singular: "أنا".
RULE subject_plural: "أنت".
RULE subject_singular: "أنت".
RULE verb_plural: "مشوا".
RULE verb_singular: "يمشي".

A nonterminal symbol may be productive to a smaller or larger degree, in the sense that it: always (RULE), sometimes (OPTION) or never (CONDITION) generates terminal symbols. A rule that never generates a symbol may enforce some condition on its affix values by failing or succeeding depending on those values.

Toy: VP (NUMBER, PERSON), NP (NUMBER, PERSON).
NP (NUMBER, PERSON): [DET], Noun (NUMBER, PERSON).

A rule consists of a left-hand-side, followed by a single colon, followed by a right-hand-side. The left-hand-side of a rule consists of a nonterminal symbol, the head, optionally followed by a list of affixes expressions enclosed between brackets. The right-hand-side of a rule consists of one or more alternatives, separated from one another by semicolons. An alternative is a (possibly empty) list of members, separated by comma's. A member is either a terminal symbol or it is a call, which looks just like a left-hand-side. Nonterminal symbols can be written in small or large letters, spaces can be used to enhance readability. A terminal symbol is written as its representation enclosed between quotes. An affix expression is either a nonterminal affix (which is then termed an affixes variable), or it consists of one or more terminal affixes separated from one another by the set union-operator I.

The developed grammar was built to deal with the different structures faced during the description of the corpus mentioned in section 1. The following figures 10, 11, 12, 13, 14, 15 and 16 are examples of the automatically parsed sentences using the AGFL formalism:

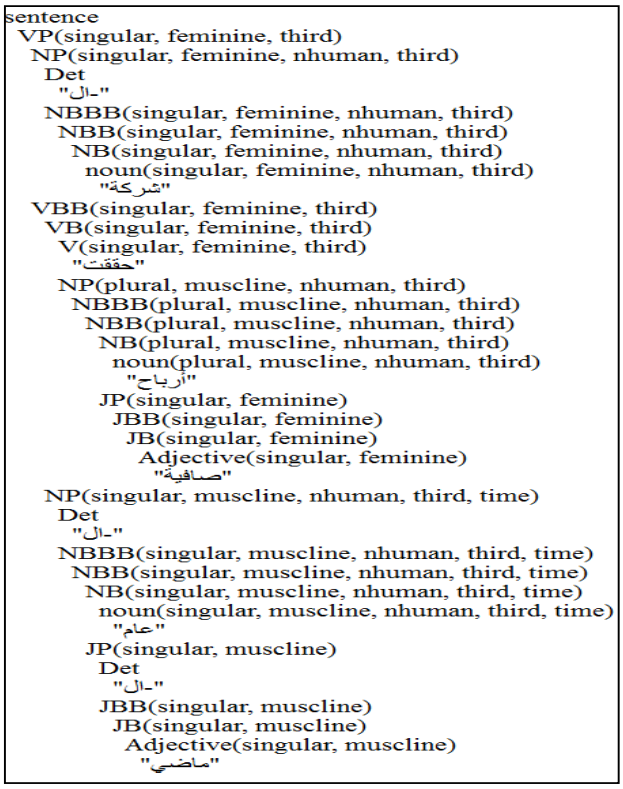


Figure 10: The representation of the phrase " حققت الشركة أرباحا صافية العام الماضي"

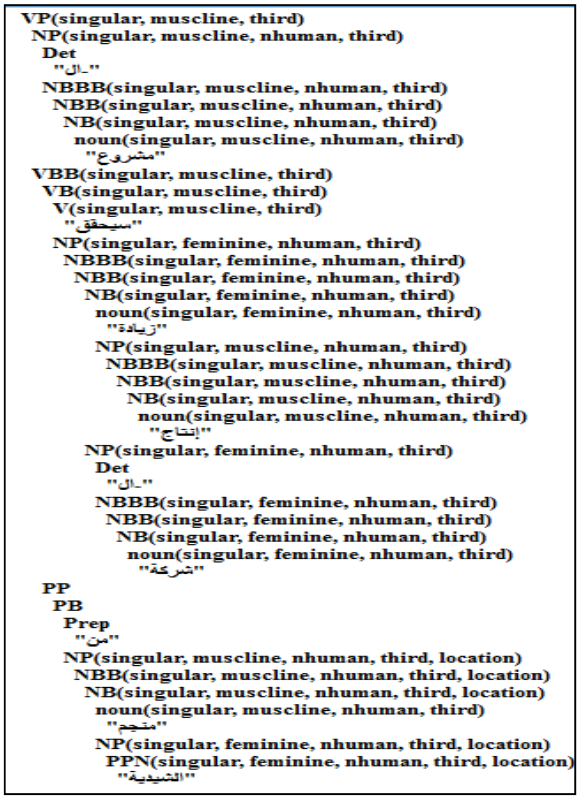


Figure 11: The representation of the phrase " سيحقق المشروع زيادة إنتاج الشركة من منجم الشيدية"

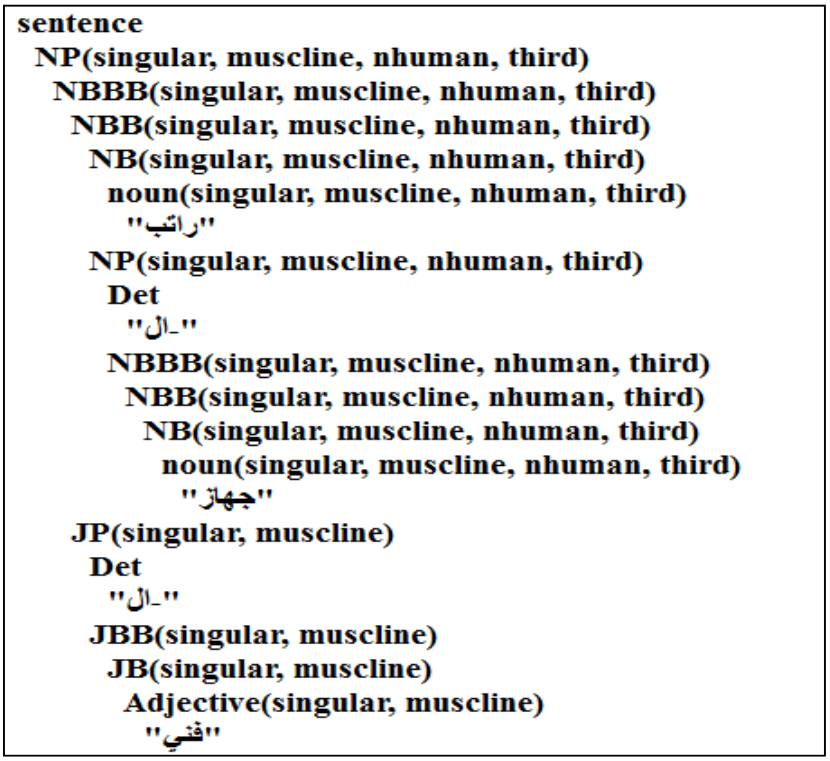


Figure 12: The representation of the phrase "راتب الجهاز الفني"

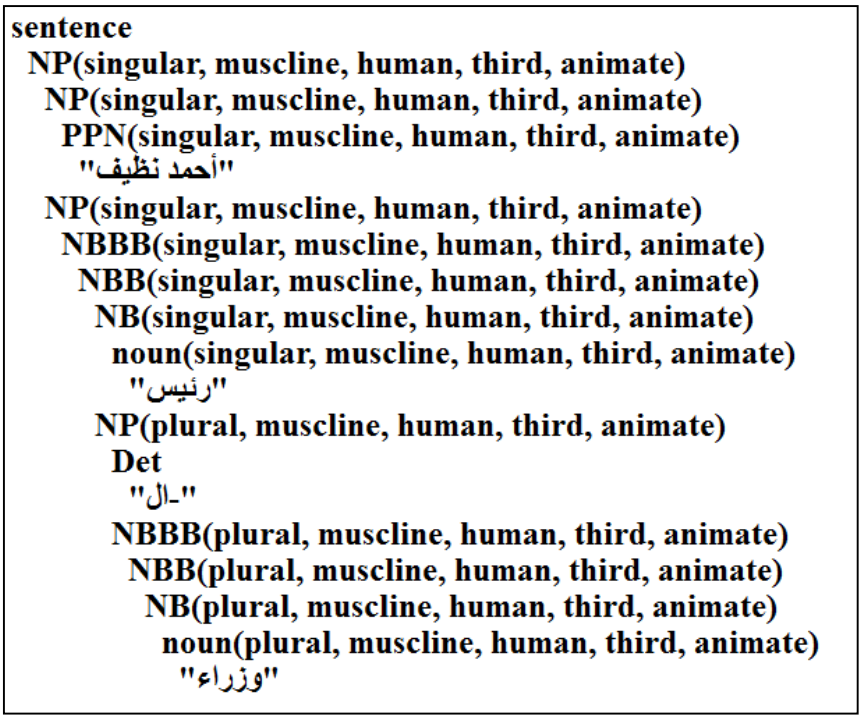


Figure 13: The representation of the phrase "أحمد نظيف رئيس الوزراء"

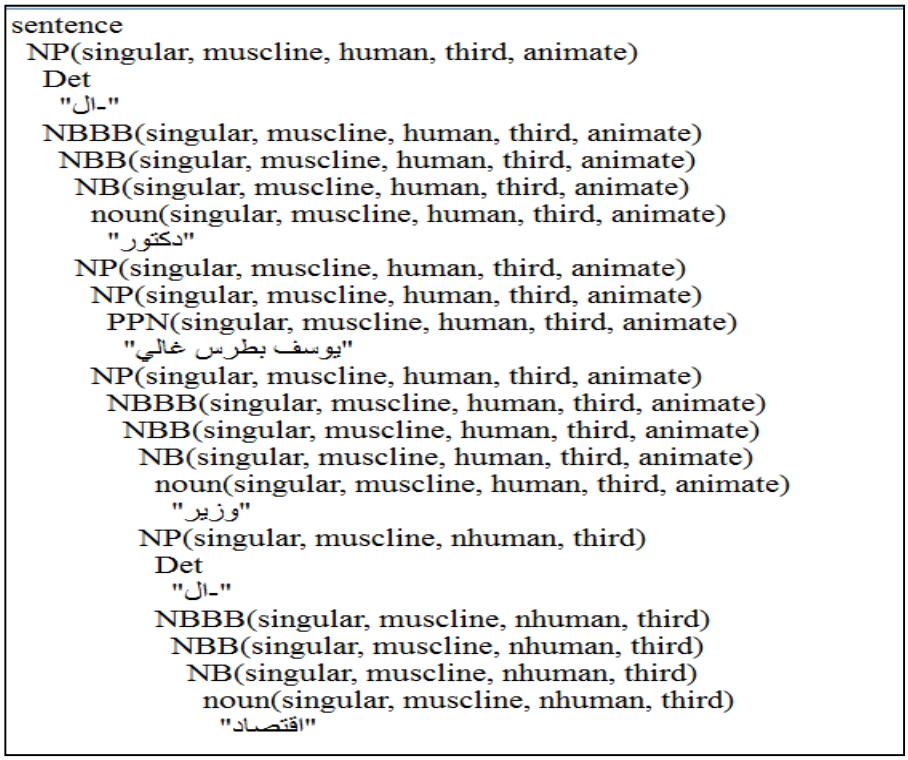


Figure 14: The representation of the phrase "الدكتور يوسف بطرس غالي وزير الاقتصاد"

Figure 15: The representation of the phrase " نظم أحمد نظيف رئيس الوزراء مؤتمرا"

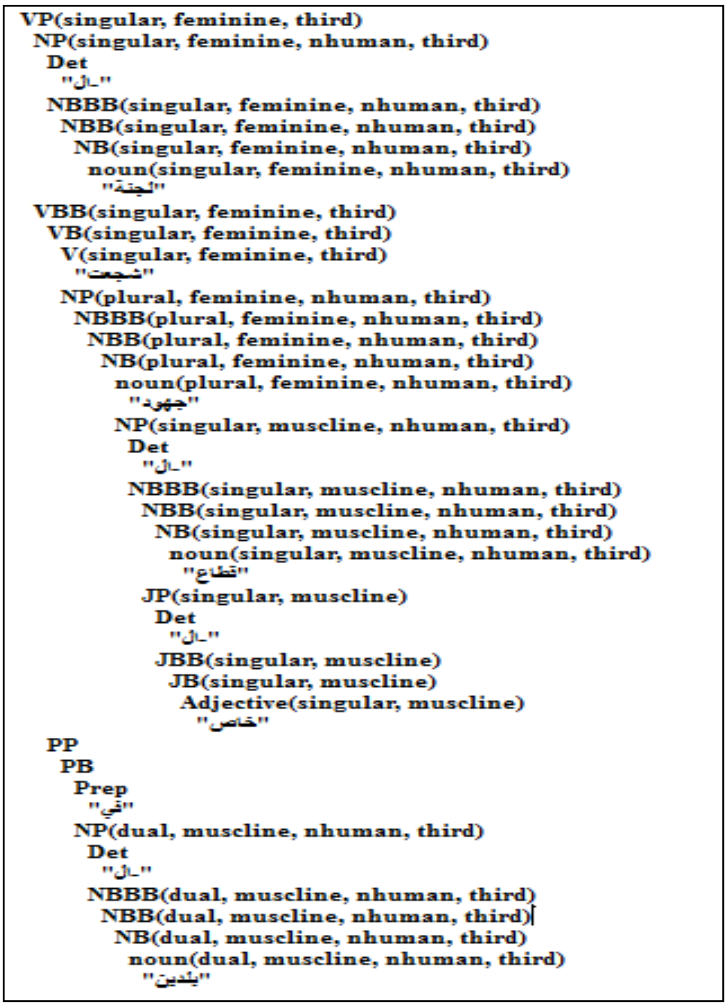


Figure 16: The representation of the phrase "شجعت اللجنة جهود القطاع الخاص في البلدين"

4 SYNTAX SEMANTICS INTERFACE

The second module of the developed grammar is the semantic module: This module is responsible for mapping the syntactic roles with their equivalent semantic relations.

It was faced with many challenges. For instance, the subject of the verb with the syntactic role verb specifier "VS" could be mapped to three different semantic relations depending on the syntactic and semantic classification of the verb. If the verb is transitive or intransitive (unergtive), the subject will be the doer of the verb; therefore, it will be mapped with the agent as in "جرى الولد" 'the boy ran'. However, if the transitive or intransitive verb carries the semantic feature stative, which describes a state of being, then the syntactic subject will be mapped to the experiencer "exp" as in "شعر الولد" 'the boy feels'. While, if the verb is intransitive (unaccusative), the subject will be mapped to the object "obj" as in "يذوب الجليد" 'The ice melts'.

5 EVALUATION

The output has been evaluated using a version of the processed corpus annotated manually. The total number of the sentences was 80 sentences. The results were evaluated against a manually analyzed data. The method that is adopted for the evaluation of the results is the precision and recall. Precision is the number of correct results divided by the number of all returned results, it was 0.90. Recall is the number of correct results divided by the number of results that should have been returned, it was 0.83. A

result is considered "returned" when: the output is a tree (i.e., all the words are interlinked).Also, the output has been compared with the output of Stanford parser[2].Figures 17 and 18 show a sentence output from both Stanford and current proposed system which indicates that the current system have more details in affixes of word forms and more details in the syntactic layers:

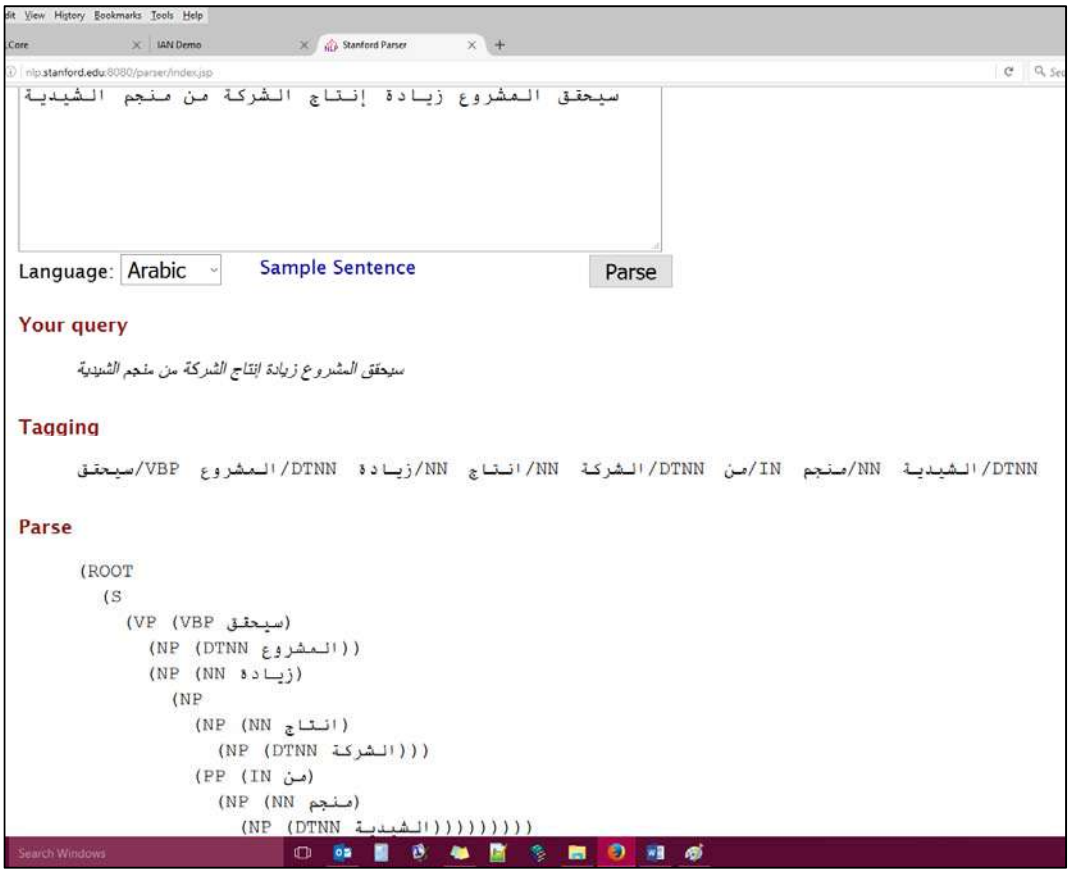


Figure 17: The output of Stanford for sentence "سيحقق المشروع زيادة إنتاج الشركة من منجم الشيدية"

[2]A natural language parser is a program that works out the grammatical **structure of sentences**, for instance, which groups of words go together (as "phrases") and which words are the **subject** or **object** of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the *most likely* analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s.

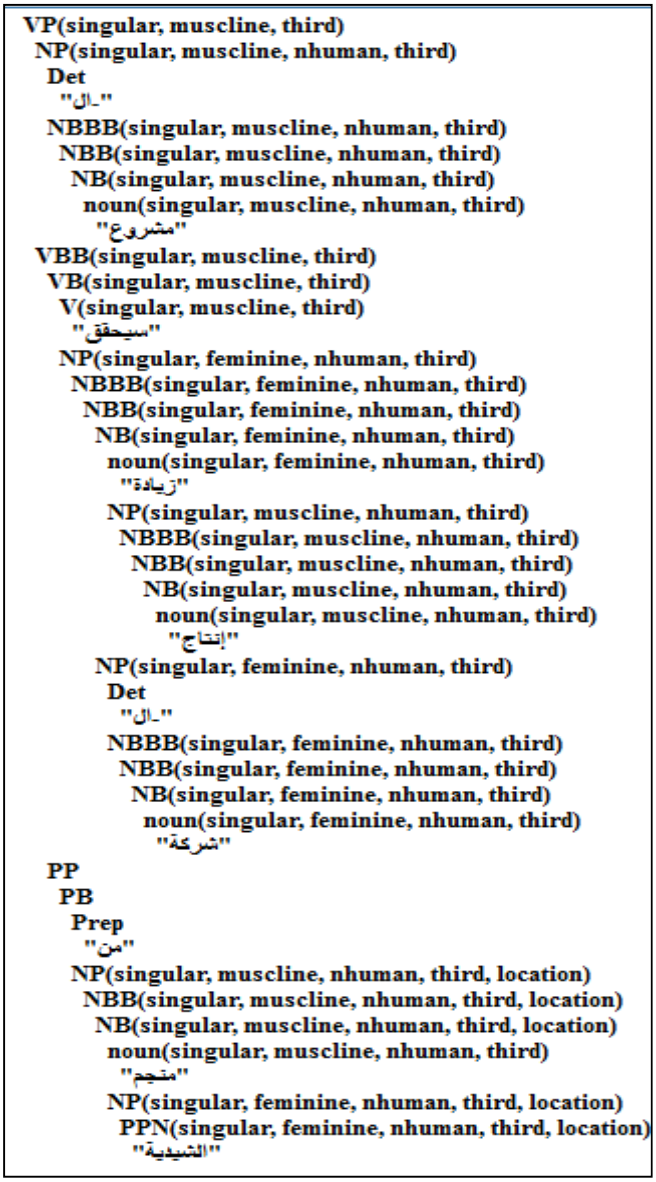


Figure 18: The output of the current system for the phrase "سيحقق المشروع زيادة إنتاج الشركة من منجم الشيدية"

6 CONCLUSIONS

The task of sentence understanding requires a variety of different types of knowledge; morphological, syntactic and semantic knowledge. The syntactic structures of some Arabic sentences have been analyzed based on the X-bar theory. The researcher has considered different structures in Arabic and demonstrated how they were analyzed. The problem of analyzing the different nominal phrases such as noun and its adjectival modifiers, genitive construction and phrases that include apposition has encountered.

REFERENCES

[1] F. Abdelkader, "Issues in the Structure of Arabic Clauses and Words", Dordrecht: Kluwer, 1993.

[2] S. Alansary, M. Nagi and N. Adly, "Generating Arabic Text: the Decoding Component in an Interlingual System for Man-Machine Communication in Natural Language", 6th International Conference on Language Engineering, Cairo, Egypt, 2006.

[3] S. Alansary, "A Formalized Reference Grammar for UNL-based Machine Translation between English and Arabic", 24th international conference on computational linguistics (COLING), Mumbai, India, 2012.

[4] C. Vicente, "Syntax of Modern Arabic Prose", 3 Vols. Bloomington: Indiana University Press, 1974.

[5] D. Everhard, "A Formal Approach to Arabic Syntax: The Noun Phrase and the Verb Phrase." PhD Nijmegen University. Nijmegen: Luxor, 1992.

[6] D. Everhard. "Distinct(ive) Sentence Functions in Descriptive Arabic.Linguistics", in: Parkinson, Dilworth (ed.): Perspectives on Arabic Linguistics XV. Amsterdam: John Benjamins.

[7] D. Everhard. "A Formal Approach to Arabic Syntax: The Sentence". Amsterdam: Rodopi.

[8] H. Halteren, "Excursions into Syntactic Databases", University of Nijmegen, [Published: Amsterdam: Editions Rodopi Language and computers: Studies in Practical Linguistics, Volume 21, 1997.

[9] K. Kees, "Affix Grammars", in: Peck, J (ed.): Algol 68 Implementation. Amsterdam: North-Holland. Pp. 95-105, 1971.

[10] K. Kees, "Affix Grammars for Natural Languages", in Albas, H. and B. Melichar (eds.): Attribute Grammar Applications and Systems, SLNCS, 545, Springer, pp. 469-484, 1992.

BIOGRAPHY

Marwa Saber Selim Arafat: Head of Grammar Development unit, Arabic Computational Linguistics Center. Bibliotheca Alexandrina, Alexandria, Egypt.

She graduated from Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Egypt. She is Computational Linguistics researcher. Her MA thesis is in the "The Automatic Extraction of the syntactic arguments of the Arabic verbs in modern standard Arabic". Her main areas of interest are Arabic morphology, syntactic parsing of MSA, semantic analysis, lexicography. She has Experience in summarization, machine translation and working on Interlingua-based Machine Translation Systems.

She obtained the UNL certificates; CLEA250, CLEA750, CUP250, and CUP500. She attended the X UNL School organized by the UNDL foundation at Bibliotheca Alexandrina (7-11 October 2012).

She is Developer at UNDL foundation, Geneva- Switzerland. She is a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

She attended many conferences such as Language Engineering Conferences, Ain-Shams University, Cairo-Egypt, (2006,2007,2008,2009,2010, 2011,2012 and 2013), http://www.esole.org (Presence), Arabic Language Technology International Conference (ALTIC), Bibliotheca Alexandria, Alexandria-Egypt 2011, http://www.altec-center.org/conference (Presence), Human Language Technology for Development Conference (HLTD 2011), Bibliotheca Alexandrina, Alexandria, Egypt, May 2 - 5 2011 and The fourth international Arabic linguistic symposium (ALS), Cairo, Egypt.

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.

He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society -

USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

TRANSLATED ABSTRACT

النحو البنيوي لوصف تراكيب اللغة العربية المعاصرة

مروة صابر[1]، سامح الأنصاري[2]

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الإسكندرية، مصر

[1]*marwa.saber@bibalex.org*

[2]sameh.alansary@bibalex.org

ملخصـــهناك طريقة لوصف تركيب الجملة في اللغة العربية المعاصرة عن طريق وضع افتراضات واختبارها على عينة لغوية مجمعة بعناية من أجل التوثيق في وجود بيئة معالجة آلية.العينة اللغوية المستخدمة في هذه الورقة محللة نحويا باستخدام صيغة AGFL، وقد تم تقيم النتائج بالاعتماد على عينة لغوية محللة آليا وقد وصلت نسبة الدقة إلى 90% .

Query Expansion for Arabic Information Retrieval Model: Performance Analysis and Modification

Ayat Elnahaas[*1], Nawal Alfishawy[**2], Mohamed Nour[*1], Gamal Attiya[**2], Maha Tolba[**2]
[*]*Department of Research Informatics, Electronics Research Institute, Cairo, Egypt*
[1]eng_ayatelnahas@yahoo.com; [1]mnour@eri.sci.eg
[**]*Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Egypt*
[2]nelfishawy@hotmail.com; [2]gamal.attiya@yahoo.com; [2]maha_saad_tolba@yahoo.com

Abstract- **Information retrieval aims to find all relevant documents responding to a query from textual data. A good information retrieval system should retrieve only those documents that satisfy the user query. Although several models were developed, most of Arabic information retrieval models do not satisfy the user needs. This is because the Arabic language is more powerful and has complex morphology as well as high polysemy. This paper first investigates the most recent Arabic information retrieval model and then presents two different approaches to enhance the effectiveness of the adopted model. The main idea of the proposed approaches is to modify and/or expand the user query. The first approach expands user query by using semantics of words according to an Arabic dictionary. The second approach modifies and/or expands user query by adding some useful information from the pseudo relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out the non-related textual words. The adopted retrieval model and the two proposed approaches are implemented, tested, compared, and evaluated considering Arabic document collection. The obtained results show that the proposed approaches enhance the effectiveness of the Arabic information retrieval model by about 15% to 35%.**

Keywords: **Arabic Documents, Indexing, Vector Space Model, Query Expansion, Semantics, and Relevance Feedback.**

1 INTRODUCTION

Information retrieval is one of the most important research areas in information technology. The main objective is to match and retrieve the most relevant documents to the user query. Therefore, a good information retrieval system should retrieve only those documents that satisfy the user needs.

Generally, an information retrieval system contains several modules mainly: document collection, query processing, matching operations and query performance [1]. Figure 1 shows the main modules of an information retrieval system [2]. Document collection and representation involves an important process called indexing. The indexing process associates a document with a descriptor represented by a set of features automatically derived from the content. It also optimizes the query performance and improves the response time by sorting terms in an interested file structure. Moreover, a number of processing tasks can take place during the indexing phase similar to the query processing which further improves the performance [3-5]. Document-query matching aims to estimate the relevance of a document to the given query. Most information retrieval models compute a relevance score. This score is used as a criterion to rank the list of documents retrieved to the user in response to the query. That is, the results of matching between the user query and the index terms are posted based on a Ranking method.

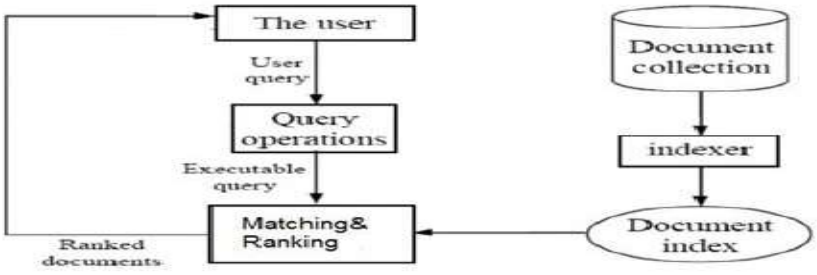


Figure1: The Main Modules of an Information Retrieval System [2]

A natural language query specifies the user's information need in a sentence. Representing the user need involves query formulation using terms expressed by the user and/or additive information driven by iterative query improvements like relevance feedback. The query/ user need is parsed and compiled into an international form. In case of textual retrieval, query terms are generally preprocessed to select the index objects. The query representation involves one-step or multi-step query formulation driven by iterative query improvements [6-8]. The querying stage involves many themes including query preprocessing, removal of stop-words, query expansion, and others. The query expansion expands the query with similar terms and then retrieves another set of documents using expanded query [6, 9]. Moreover, information retrieval implements a basic term matching for identical terms. The document-query matching is known as query evaluation for estimating the relevance of documents to the given query. The information retrieval system employs some ranking methods based on mathematical bases to exploit some properties found in the document collection. Matching between the query keywords and index terms may be exact matching, partial matching, or intelligent matching [10].

Although several models were developed [11-17], most of Arabic information retrieval models do not satisfy the user needs. This is because the Arabic language is known with its powerful and complex morphology as well as its high polysemy. This paper first investigates an Arabic information retrieval model and then presents two different approaches to enhance the effectiveness of the model. The focus is concerned with the modification and/or expansion of the user query. The first approach expands user query by using semantics of words according to an Arabic dictionary. The second approach modifies and/or expands user query by adding some useful information from the pseudo relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out the non-related textual words. The proposed approaches are implemented, tested and evaluated using some measurable criteria such as precision, recall, and F-measure. In addition, the obtained results are compared with that obtained by the most recent adopted Arabic retrieval model for Arabic document collection [2].

The rest of this paper is organized as follows. Section 2 presents a literature survey for related work. Section 3 presents an adopted information retrieval model. Section 4 presents the proposed approaches and describes the query expansion using semantics of keywords and relevance feedback. Section 5 presents the simulation experimental results and discussions while section 6 concludes this work.

2 RELATED WORK

Regarding the information retrieval systems/ models, several research efforts were presented by a lot of researchers [11-17]. In [11], the authors mentioned that any information retrieval model can be represented by four attributes: D, Q, F, and R. D is the set of documents in the document collection. Q is the set of queries representing the user needs. F is concerned with classical document representation, queries, and their relationships. R is a ranking function $R(q_i,d_j)$ which affiliates a real number with a query q_i € Q and a document representation d_j € D. In [12], the authors mentioned that the Boolean model is one of the oldest information retrieval models. That model uses the set theory and/or Boolean algebra. The user Query can be represented by a set of keywords connected together logically by a set of connections like AND, OR, and NOT. The AND operator produces the set of documents of both sets. The OR operator produces a document set that is bigger than or equal to the document sets of any of the single terms. The NOT operator is used to avoid retrieving a document containing a specific keyword. In [13], the authors discussed the vector space model that represents the documents and queries as vectors in a multidimensional space. To assign a numeric score to a document for a query, the model measures the similarity between the query vector and the document vector. The angle between two vectors is used as a measure of divergence between the vectors. The cosine angle is used as the numerical similarity. If the cosine angle has the value '1' it means the vectors are identical while the vectors are orthogonal if the cosine angle has the value '0'. The vector space model is good as it attempts to rank documents by some similarity values between the user query and each document. In [14, 15], the authors discussed the probabilistic model of information retrieval which relies on the nation that each document has a certain probability of being relevant to a query. The documents that are most likely to be relevant and useful to the user are ranked by a decreasing order of probability. For two events A and B, the joint event of both events occurring is described by the joint probability P(A, B). The conditional probability P(A|B) expresses the probability of event A given that B occurred. Probabilistic information retrieval models include classic Probabilistic models, language models and the relevance model. All those models have variants that incorporate word dependence.

In [16], the authors conducted the process of developing ontology for Arabic Blogs retrieval. The authors mentioned that semantic search engines provide searching and retrieving resources related to the user's need. The authors proposed a model for representing Arabic knowledge in the computer technology domain using ontologies. The model was concerned with elicitation user's information needs. Ontologies play a vital role in supporting

information search and retrieval process of Arabic blogs on the web. In [17], the authors presented an enhanced Arabic information retrieval approach. The focus was on the effectiveness of using the list of stop-words and light stemming of Arabic. The authors used the vector space model as a popular weighting scheme in their work. Their work aims at combining the stop-words list with light stemming to enhance the performance and compare their effects on retrieval. The authors tested their adopted approach using the Arabic news consortium dataset. In [9], the authors discussed the concept of query expansion for improving the process of Arabic information retrial. The query expansion was based on the similarity of terms. The authors employed the expectation-maximization algorithm for selecting the relevant terms and weeding out the non-relevant ones. They tested performance of the adopted algorithm using INFILE test collection. The experiments indicate good performance of precision and recall for the used query expansion method.

3 ADOPTED ARABIC INFORMATION RETRIEVAL MODEL

In 2016, an adopted Arabic information retrieval is developed [2]. The authors discussed the main challenges of Arabic query expansion using Word-Net and association rules. They mentioned that they are able to exploit Arabic word-Net to improve the retrieval performance. Their obtained results on a sub-corpus from the Xinhua collection showed that the automatic selection method is significant and improves the performance of information retrieval systems. The adopted Arabic information retrieval model [2] involves important themes mainly: preprocessing, document collection and indexing, user query, and matching operations.

A. Preprocessing

The preprocessing steps are done on the document terms before building the index and on the user query before matching process. The preprocessing should be done first to gain the benefit of speeding-up the retrieval time [18, 19]. The preprocessing steps involve tokenization, removal of stop-words and stemming.

1) Tokenization

Tokenization; in natural language processing; means splitting text into tokens. A token is the smallest unit of text that may be a word, a punctuation mark or a multi-word expression. The separator between two adjacent words may be a white space or punctuation marks. Tokenization is an important step for most natural language processing tasks [37]. In this work, Lucene Arabic tokenizer is used during the implementation of this stage http://www.apache.org/licenses/LICENSE-2.0. Figure 2 shows an example of a document title before and after tokenization.

أهم المركبات المسموح بها في الزراعة العضوية لمقاومة الأمراض والحشرات

(a) A document title before tokenization

أهم,المركبات,المسموح,بها,في,الزراعة,العضوية,لمقاومة,الأمراض,و,الحشرات

(b) The document title after tokenization

Figure 2: A document title tokenization

2) Removal of Stop-words

Removal of stop-words means rejecting the useless words like preposition, pronoun, specifiers, modifiers, and other tools. Examples of the stop words are: - من، إلي، علي، في، هو، هي، الذي . Such words frequently occur in Arabic documents. These words don't give any hint for the content of their documents. In information retrieval systems, stop-words should be eliminated (by referring to a stop-word list) from the query text and from the set of index terms [18, 20]. Figure 3 shows the tokens of a document title after removing the stop-words.

أهم، المركبات، المسموح، الزراعه، العضويه، مقاومة، الأمراض، الحشرات

Figure 3: The tokens of the document title, in Figure 1, after removal of stop-words

3) Stemming

The stemming process is very important for Arabic information retrieval. Stemming aims at reducing all of the inflectional derivational variants of words into a common form called the stem. A word stem can be obtained by removing all the affixes attached to the word. The words sharing some root or stem can increase the matching of documents to the user query. Stemming can reduce the index size and improve the performance of the retrieval process. Figure 4 shows the tokens of a document title after stemming.

أهم، مركب، مسموح، زراع، العضوي، مقاوم، الأمراض، الحشرات

Figure 4: A Document Title after Stemming

There are several types of stemmers. Examples of Arabic stemmers are: light stemmer (light 10), Khoja stemmer, Porter stemmer, and others. In this work, Porter stemmer is used during the implementation of this stage [18, 21-23]. For more details about the Porter stemmer mechanism, the reader can refer to the website https://tortous.org/mortim/porter.stemmer.

B. Document Collection and Indexing

Indexing is the process of choosing a term or a number of terms that can represent what the document contains. In other words, after doing the preprocessing steps on the chosen document collection, the index can be built. Each document is represented by a set of important terms, which were taken from the document title. Such terms are weighted and stored in an index (as index terms) without any repetition. The index contains document number, terms, frequency/weight in addition to other useful information such as the number of documents that contain each term. Figure 5 shows an Arabic example of a part of the index mapping [24-26]. The index terms will be matched against the query keywords.

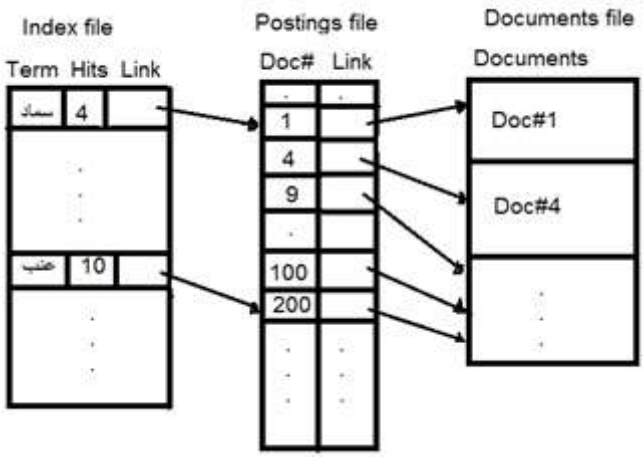


Figure 5: Arabic Example of a Part of the Index Mapping

C. User Query

The querying stage is handled exactly like the document. That is, the preprocessing steps; tokenization, removal of stop-words, and stemming are done on the input user query. The user query may be a word, phrase, or sentence containing a set of keywords. If the query is one word, the stemming operation only can be done. If the query contains a set of words, it should be preprocessed (tokenization, stop-words removal, and stemming). In this work, several queries are presented and processed. Some of the queries contain only one keyword while others contain two keywords, three keywords, and four keywords respectively. Table1 shows some examples of the user independent queries while Table 2 contains examples of some related queries.

Table 1: Examples of User Independent Queries

User Query	No. of Keywords
التين	1
زراعة الخضروات	2
صادرات مصر من القمح	3
أهمية البلح وطرق تجفيفة	4

Table 2: Examples of Some Related Queries

User Query	No. of Keywords
العنب	1
محصول العنب	2
الجديد فى محصول العنب	3
أالجديد فى إنتاج محصول العنب	4

D. Matching and Ranking

The matching process is done between the query keywords and document terms. To facilitate the matching process, a matching model is used. In this paper, the Vector Space Model (VSM) is used for the matching operation [18, 20, 27-32].
The VSM is an algebraic model where it uses non-binary weights that are assigned to the index terms of documents and queries. The document set D is represented as follows:-

$$D = \{d_1, d_2, d_3 ... d_N\} \quad (1)$$

where, d_j is the document number j, and N is the number of documents in the dataset collection.
Any document d_j is represented by a set of terms' weights as follows: -

$$d_j = \{w_{1j}, w_{2j}, w_{3j} ... w_{mj}\} \quad (2)$$

where, w_{1j} is the weight of the term i in the document j. The weight of term i in document j can be calculated using the term frequency (tf) and inverse document frequency (idf). So,

$$w_{ij} = tf_{ij} * idf_i \quad (3)$$

where, the term frequency tf_{ij} is the number of occurrence of term i in the document j and idf_i is the inverse document frequency of term i.

$$idf_i = log_2 \frac{N}{n_i} \quad (4)$$

where, n_i is the total number of occurrence of item i in all documents.
Documents can be retrieved and ranked by matching the query vector versus the document vector to compute the score or similarity. The retrieved documents are ranked according to the similarity to the user query [33-36].

$$sim(d_j q_i) = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (5)$$

where, $sim(d_j, q_i)$ is the similarity between document j and query q_j, w_{ij} is the weight of term i in document j, and w_{iq} is the weight of term i in query q.

4 PROPOSED APPROACHES

This section presents two new efficient approaches to enhance the effectiveness of the most recent adopted Arabic information retrieval model [2]. The main idea of the proposed approaches is to modify and/or expand the user query by using semantics of words in the first approach and using some useful information from the pseudo relevance feedback in the second approach.

A. Query Expression using Semantics of Keywords

Query expansion means adding extra new terms to the keywords of the initial query. Since the input user query has the significant effect on the document retrieval, hence the user query may be modified and/or expanded to retrieve more relevant documents. The addition of new terms should take place prior the initial search.

It is known that Arabic is one the Semitic languages. Arabic has a rich set of vocabularies. Arabic language is polysemous as the same word may have several meanings. Moreover, the Arabic language has a different morphological structure for its wide range of derivations [2, 4]. By searching the dictionary for the meaning of an Arabic keyword, more than one meaning may be found. This is the case for the majority of Arabic words. This means that each query keyword has multiple synonyms/meanings.

In this paper, the first proposed approach expands user query by using semantics of words. In this case, the synonyms or semantics of the query keywords can be obtained by referring to either the Arabic Word-Net or Arabic dictionary. In the first approach, semantics of the query keywords are chosen according to an Arabic dictionary. Expanding the query to include more or extra keywords will improve the performance of the retrieval model as it presents more relevant documents to the user.

To illustrate the query expansion method, let Q be the set of queries entered separately from the user, where $Q = \{q_1, q_2, q_3, \ldots..q_r\}$. Each query q_r has a set of m keywords. That is, $q_r = \{k_1, k_2, \ldots.k_m\}$, where k_i is the query keywords which represents the user needs and $1 \leq i \leq m$. By searching the dictionary for the meaning of each keyword, a list S_{ki} of n synonyms associated to the keyword k_i may be found, i.e., $S_{ki} = \{S_{i1}, S_{i2}, S_{i3}, \ldots, S_{in}\}$. Each list S_{ki} contains the number of synonyms associated to a keyword k_i in the query and $1 \leq i \leq m$. This means that the number of synonyms' lists of a query q_r equals the number of keywords in the initial use query. That is, $S(q_r) = \{S_{k1}, S_{k2}, \ldots..S_{km}\}$, where $S(q_r)$ is the set of lists.

Figure 6 shows the associated synonyms of query keywords. From Figure 6, each query keyword k_i has multiple synonyms/meanings S_{ij} where $1 \leq i \leq m$ and $1 \leq j \leq n$. Moreover, it is not necessary for all query keywords to have the same number of corresponding meanings. For this reason, we focus here on using only one meaning which is the commonly used one. The chosen meaning is taken based on its strong relation with the keyword. That is, the number of query keywords after expansion becomes the double of the original one. To illustrate that concept, some simple examples are given in Table 3. The query expansion can extract the equivalent terms of query keywords from the relation between the concepts or meanings such as:

(زراعة، فلاحة)، (طماطم، بندورة)، (كرنب، ملفوف)، وهكذا.

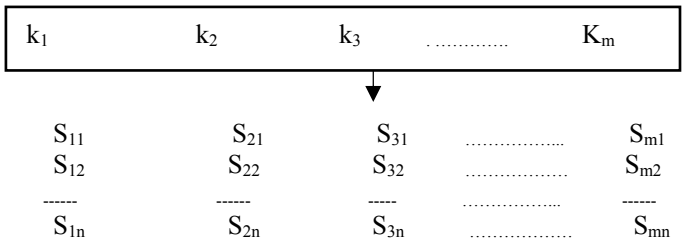


Figure 6: Query Keywords and their Semantics

Table 3: User Query Expansion using Synonyms/Semantics

Initial Query	Expanded Query
زراعة العنب	زراعة، فلاحة ، العنب ، الكرم
تسميد الكرنب	تسميد، تخصيب ، الكرنب ، الملفوف
إنتاج البلح	إنتاج، البلح، التمر

B. Query Expansion using Relevance Feedback

As mentioned above, query expansion aims to add extra terms or more information to clarify the user query. The query expansion helps in matching more additional documents. In this paper, the second proposed approach modifies and/or expands user query by adding some useful information from the pseudo/user relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out

the non-related textual words. The idea is going to keep track of those terms that should be added to the query and those should be eliminated.

The process of query expansion by the principle of user relevance feedback may be described as follows:

1) The original keywords of the user query, after doing the preprocessing operations, are matched against the index terms. The retrieved documents are presented from the highest to lowest values depending on the similarity values.

2) The retrieved documents should be analyzed to monitor and identify their terms' descriptors. This is important to add those terms appeared in the relevant documents to the original user query and also to eliminate those terms describing the retrieved irrelevant documents.

i) A maximum threshold value (max_{th}) of documents similarities should be defined. This means that the terms' descriptors for only those retrieved documents with similarity values ≥ mac_{th} will be chosen to be added to the original query keywords.

Let S_1 be the set that collects all relevant retrieved documents that satisfy the threshold condition max_{th}.

$$S_1 = \{ d_1, d_2, \ldots\ldots\ldots max_{th}\} \quad (6)$$

ii) A minimum threshold value (min_{th}) of documents similarities is defined. This means that the terms' descriptors for only those retrieved documents with similarity values ≤ min_{th} will be eliminated from the query. Let S_2 be the set that gathers all non-relevant retrieved documents and the min_{th} condition is satisfied

$$S_2=\{ d_1, d_2, \ldots\ldots\ldots d_y\} \quad (7)$$

3) The query can be expanded by adding the terms of the selected relevant documents from S_1 and also eliminating those terms of the chosen irrelevant documents from S_2. That is

$$q_{exp} = q_{user} + \sum_{di \in s_1} d_i^j - \sum_{dj \in s_2} d_j^j \quad (8)$$

5 SIMULATION RESULTS AND DISCUSSION

This section presents several experimental to evaluate the performance of the proposed approaches. To do so, the adopted information retrieval model [2] and the proposed approaches are implemented and tested considering a dataset in the agriculture field. The performance is evaluated using some measurable criteria such as precision, recall, and F-measure.

A. Simulation Environment

The proposed approaches are implemented using JAVA programming language besides Lucene APIS, which is a powerful searching library, using an HP-Labtop with a processor 2.5 GHZ, and Windows-7 operating systems. The approaches are coded in JAVA and supported by the Apache software foundation.

B. Document Collection Dataset

To check the efficiency of the proposed approaches against the adopted information retrieval model [2], they are operated and tested using a chosen document collection as a test-bed. The documents in the dataset are acquired from different Arabic websites mainly http://www.kenanaonline.net/page/Agriculture and http://www.zeraiah.net/index.php/baydar. The test-bed documents are in the agriculture field. It contains four hundred documents. Each document has a document title and contents. Each document is represented by a set of important terms, which were taken from the document title. Such terms are weighted and stored in an index (as index terms) without any repetition. The index terms will be matched against the query keywords.

C. Performance Metrics

The performance is evaluated using some measurable criteria such as precision, recall, and F-measure. These criteria are defined as follows [19-20].

$$precision = \frac{number\ of\ the\ relevant\ retrieved\ documents}{number\ of\ the\ retrieved\ documents} \quad (9)$$

$$Recall = \frac{number\ of\ the\ relevant\ retrieved\ documents}{number\ of\ the\ relevant\ documents} \quad (10)$$

$$F-measure = \frac{2(Recall*precision)}{Recall\ |\ precision} \quad (11)$$

D) Experimental Results

Several experiments are done to test and monitor the performance of the adopted information retrieval model and the proposed approaches. Four categories of quires are adopted with five different queries for each. The query categories have one keyword, two keywords, three keywords, and four keywords respectively. The queries in Figures 7, 8, 9, and 10 are independent. The queries in Figure 11 are related to each other, i.e., the keyword of query#1 exists in query#2. The two keywords of query#2 exist in query#3 and the three keywords of query#3 exist in query#4. This is also the case for other queries in Figures 12, and 13 respectively.

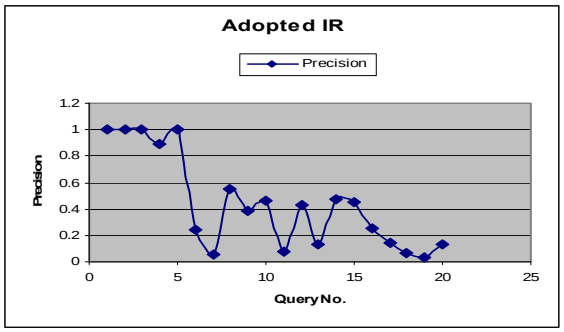


Figure 7a: Precision for Adopted IR

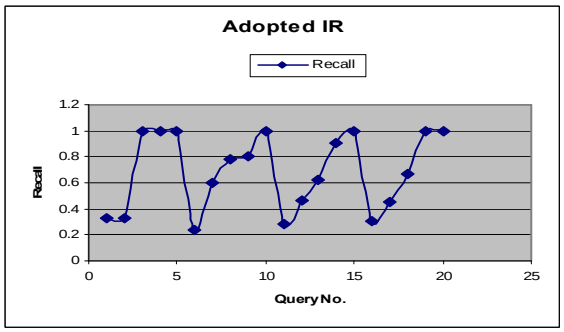


Figure 7b: Recall for Adopted IR

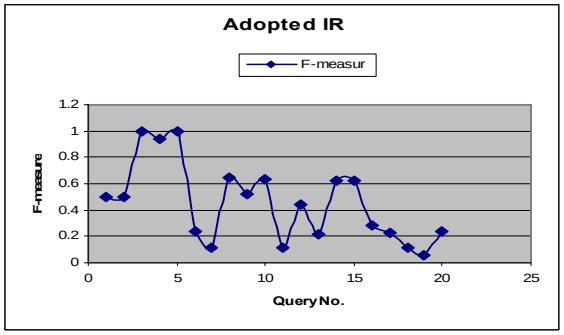


Figure 7c: F-measure for Adopted IR

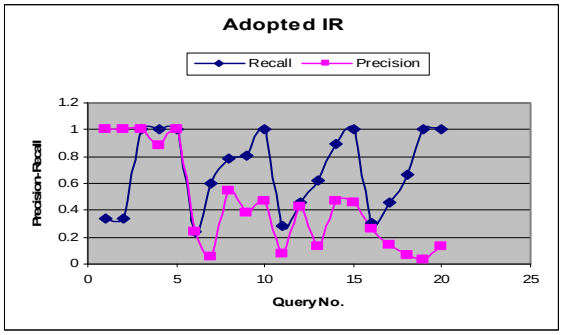


Figure 7d: Precision-Recall for Adopted IR

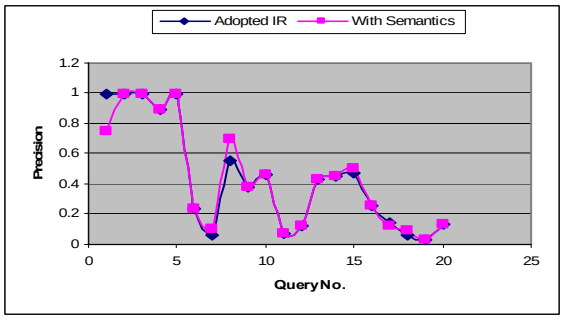


Figure 8a: Adopted IR and Keywords' Semantics

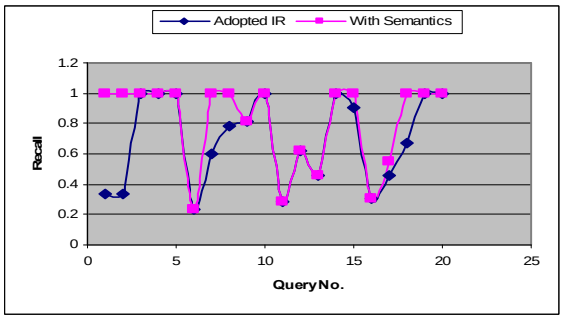


Figure 8b: Adopted IR and Keywords' Semantics

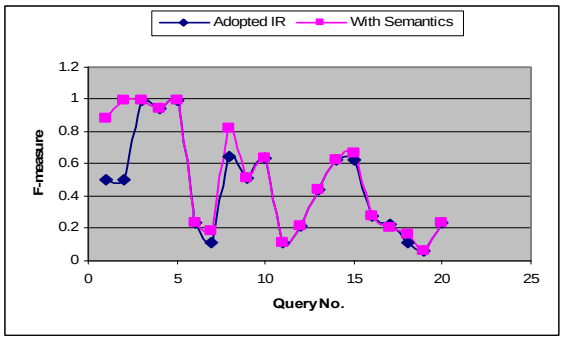


Figure 8c: Adopted IR and Keywords' Semantics

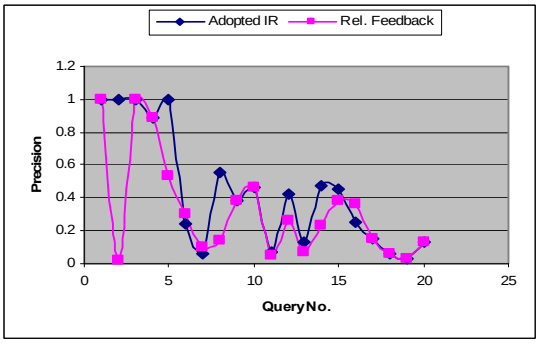


Figure 9a: Adopted IR and Relevance Feedback

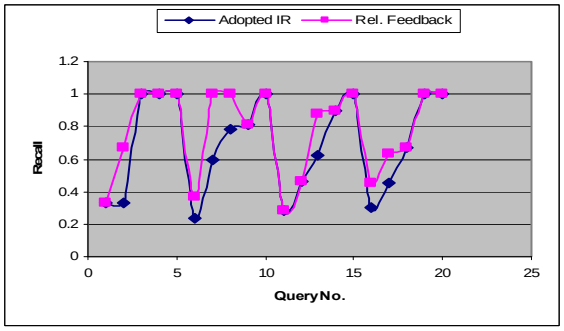


Figure 9b: Adopted IR and Relevance Feedback

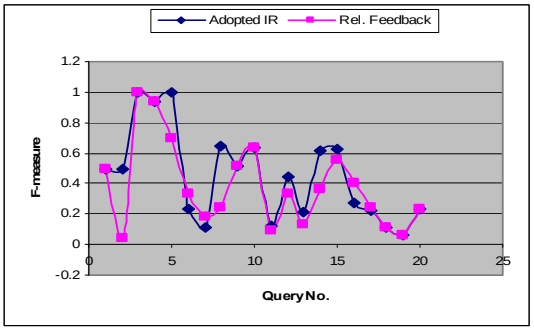


Figure 9c: Adopted IR and Relevance Feedback

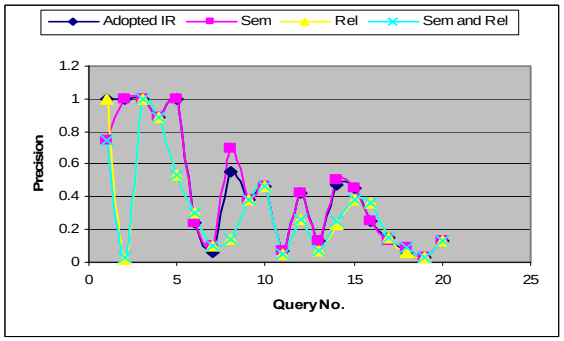


Figure 10a: Adopted IR, Sem, Rel, and Both

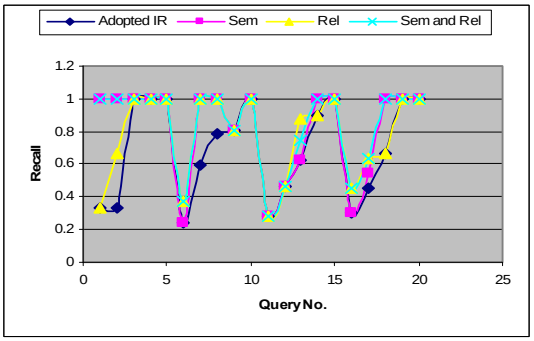


Figure 10b: Adopted IR, Sem, Rel, and Both

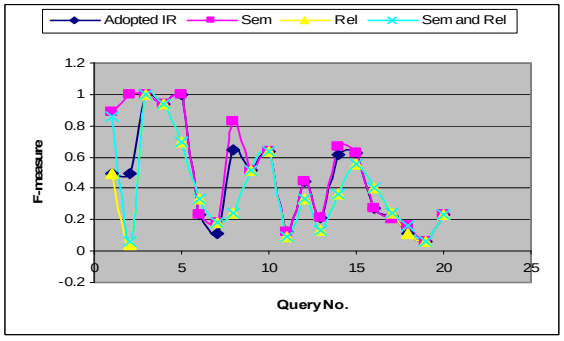


Figure 10c: Adopted IR, Sem, Rel, and Both

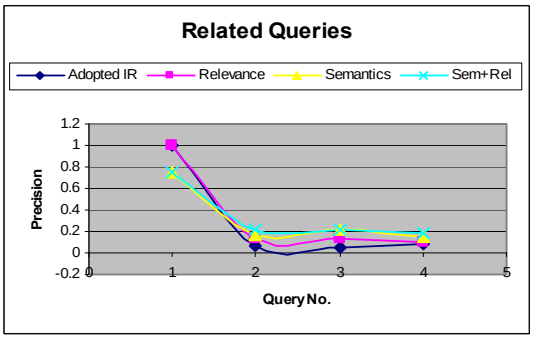


Figure 11a: Precision for Related Queries

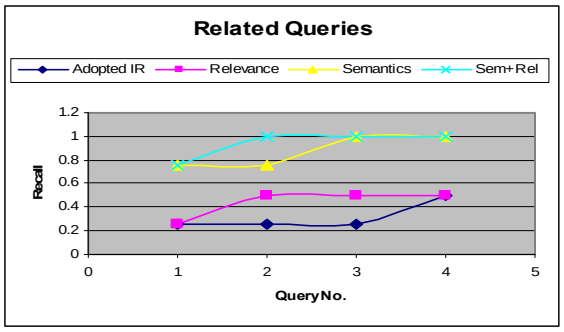


Figure 11b: Recall for Related Queries

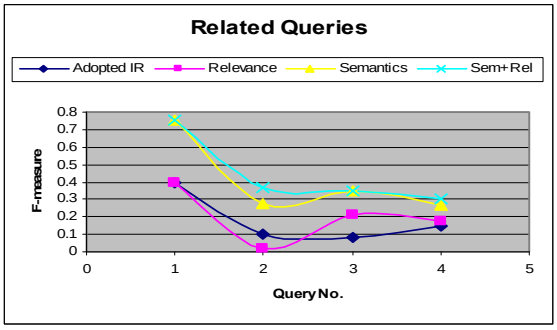


Figure 11c: Precision for Related Queries

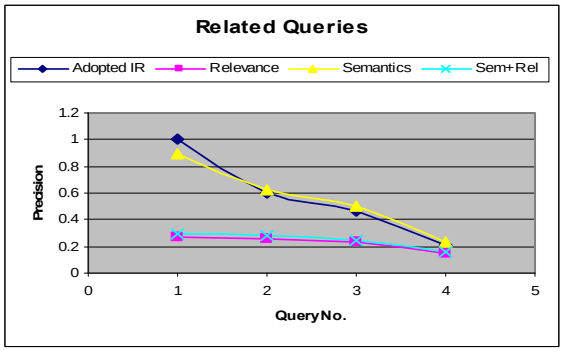


Figure 12a: Recall for Related Queries

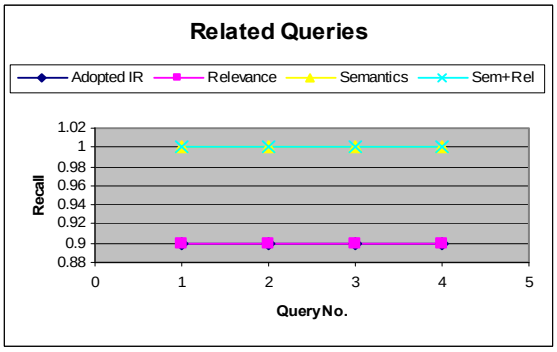


Figure 12b: Precision for Related Queries

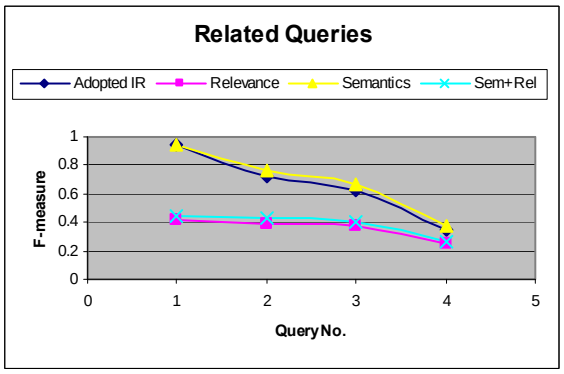


Figure 12c: F-measure for Related Queries

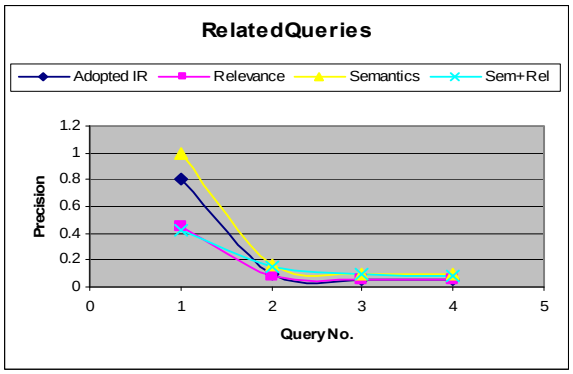


Figure 13a: Precision for Related Queries

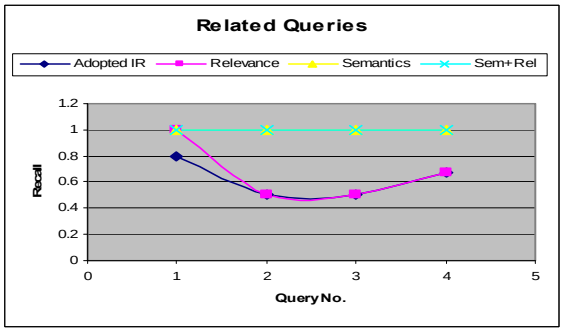


Figure 13b: F-measure for Related Queries

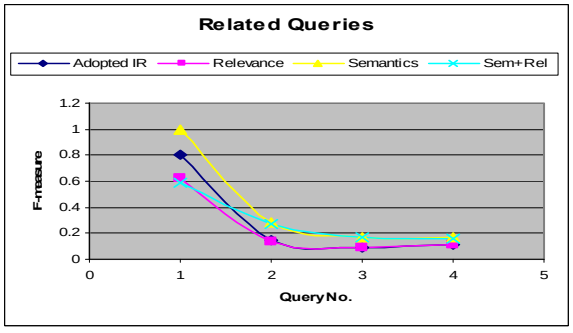


Figure 13c: Precision for Related Queries

From the experimental results, the values of precision, recall, and F-measure for each query are different from those of the other queries either in the same query category or in other categories. This means that the concept type where a query is asking for is significant and this is clear in all experiments. If the number of query keywords changes, the values of precision, recall, and F-measure are also changing. This is clear in Figures 7, 8, 9, and 10 respectively. In Figures 7, 8, 9, and 10, a query has one keyword for the first five queries, two keywords for the second five queries, three keywords for the third five queries, and four keywords for the forth five queries respectively. This means that the number of query keywords has a direct effect on the retrieval performance. If the precision value changes then the recall and F-measure values are also changing. This happened in the majority of the test-bed queries. We don't have a guarantee to say that increasing values always appear in a linear or a nonlinear form.

The values of precision, recall, and F-measure for the adopted retrieval model with semantics are different from those corresponding values without semantics. This means that the modified approach for the query expansion using semantics of keywords has a positive effect. This is clear in all query categories in Figures 7 and 8 respectively.
A precision value; regardless the query concept and query keywords; is always increasing or in some cases remains fixed compared to the adopted retrieval model without semantics. This is because the number of relevant retrieved documents is increasing or sometimes remaining unchanged. The values of precision, recall, and F-measure are better for the modified approach than those corresponding values of the adopted retrieval model without semantics. This is clear in Figures 8a, 8b, and 8c respectively. The improvement in performance for the modified approach using semantics of keywords ranges from 8% to 27% depending on the number of query keywords as well as the query concept.

The values of precision, recall, and F-measure for the test-bed queries for the modified approach using relevance feedback are better than their corresponding values of the model without modification. This is clear in Figures 9a, 9b, and 9c respectively. The improvement of performance is slightly better than that model without modification.
The values of precision, recall, and F-measure for the modified approach using both semantics of keywords and relevance feedback are better than those without any modification. This is clear in Figures 10a, 10b, and 10c respectively. The improvement values for the adopted experiments are ranging from 15% to 34% depending on the query concept and query category. Moreover, the performance of the retrieval model is better modified using keywords' semantics than that using only relevance feedback. In other words, combining both the relevance feedback and semantics makes slightly change in precision, recall, and F-measure compared to that one using only semantics of keywords. This is clear in Figures 10a, 10b, and 10c respectively. The improvement values are in the range of 3% to 13%.

Moreover, three different experiments with four related queries per each are also implemented and run as shown in Figures 11, 12, and 13 respectively. The experiments are tested and compared among the performance of the adopted retrieval model and the two modified approaches for query expansion using semantics of keywords, relevance feedback, and both. From the experimental results shown in Figures 11, 12, and 13 respectively, it is shown that the values of precision, recall, and F-measure are better for the modified approaches than those corresponding values of the adopted information retrieval without modification.

6 CONCLUDING REMARKS

In this research work, the most recent adopted information retrieval model was investigated and analyzed. In addition, two new efficient approaches are developed to enhance the effectiveness of the recent model. The adopted model is modified by expanding the queries using semantics of keywords and/or relevance feedback. The models are implemented and tested using an Arabic document collection test-bed. From the practical results, the representation and formulation of a user query plays an important role in the performance of the information retrieval model. The query expansion increases the number of retrieved relevant documents. The obtained results showed that the values of precision, recall, and F-measure for the two modified approaches are better than that without modification. The query expansion using word semantics improve the performance by about 27% compared to the original model. While, the query expansion using relevance feedback improve performance by about 14%. Finally, combining both the semantics of keywords and query relevance feedback for expanding the user queries outperforms the adopted retrieval model without modification. The hybrid query expansion using the two modifications improves the performance by 15% to 35%.

REFERENCES

[1] Ghaith AbdulSattar Alkubaisi, "Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge" M.Sc. thesis, Department of Computer Science, Faculty of Information Technology, Middle East University, Amman, 2013.
[2] Ahmed Abbache, Farid Meziane, Ghalem Belalem, and Fatma Bellredim, "Arabic Query Expansion using Word-Net and Association Rules", The International Journal of Intelligent Information Technologies, Vol. 12, No. 3, pp. 51-64, July-September 2016.
[3] Soner Kara, O¨zg¨ ur Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Cicekl and Ferda N. Alpaslan, "An Ontology-based Retrieval System Using Semantic Indexing", https://etd.lib.metu.edu.tr/upload/12612110/in dex. pdf., Downloaded in 2016.
[4] Emad Elabd, Eissa Alshai, and Hatem Abdulkader, "Semantic Boolean Arabic Information Retrieval", The International Arab Journal of Information Technology, Vol. 12, No. 3, pp. 311-316, May 2015.
[5] Eissa Mohammed Mohsen Alshari, "Semantic Arabic Information Retrieval Framework", M.Sc. Thesis, Information Systems Department, Faculty of Computers and Information, Menoufiya University, 2014.
[6] Fatiha Boubekeur, and Wassila Azzoug, "Concept-Based Indexing in Text Information Retrieval", The International Journal of Computer Science and Information Technology (IJCSIT), Vol. 5, No. 1, pp. 119-136, Feb. 2013.
[7] Miriam Fernandez, Ivan Cantador, Vanesa Lopez, David Vallet, Pablo Castells, and Enrico Motta, "Semantically Enhanced Information Retrieval: An Ontology-based Approach", Downloaded in 2017 from http:/ /www.elsevier.com/ locate/websem.
[8] Komal Shivaji Mule, and Arti Waghmare, "Improved Indexing Technique For Information Retrieval Based On Ontological Concepts", The International Journal of Computer Applications (IJCA) and the National Conference on Advances in Computing (NCAC), pp. 5-20, 2015.
[9] Khaled Shaalan, Sinan Al-Shaikh, and Farhad Oroumchian, "Query Expansion Based on Similarity of Terms for Improving Arabic Information Retrieval", A Technical Report Presented to the University of Wollongong in Dubai, The 7th IFIP TC12 International Conference on Intelligent Information Processing, Springs, Heidelberg, pp. 167-176, 2012.
[10] Fausto Giunchiglia, Uladzimir Kharkevich, and Llya Zaihrayeu, "Concept Search", Downloaded in 2016 from the http://eprints .biblio.unitn.it/1434/1/037.pdf.
[11] Djoesd Hiemstsa, "Information Retrieval Models", Downloaded in 2016 from http://wwwhome.cs.utwente.nl/~hiemstra/papers/IRModelsTutorial-draft.pdf.
[12] W. B. Croft, "Knowledge-based and Statistical Approaches to Text Retrieval", The IEEE Expert, Vol. 8, No. 2, pp. 8-12, 1993.
[13] Amit Singhal, "Modern Information Retrieval: A Brief Overview", Downloaded from in 2016 from http://www.gib.fi.upm.es/sites/default/files/irmodeling.pdf.
[14] Jelita Asian, "Effective Techniques For Indonesian Text Retrieval", Ph.D. Thesis, the School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, March 2007.
[15] K. Tamsin Maxwell, "Term Selection in Information Retrieval", Ph.D. Thesis, Institute for Communicating and Collaborative Systems, University of Edinburgh, January 2014.
[16] Lilac Al-Safadi, Mai Al- Badrani, and Meshaal Al- Junidey, "Developing Ontology for Arabic Blogs Retrieval", International Journal of Computer Applications, Vol. 19, No. 4, pp. 41-46, April 2011.
[17] Jaffar Atwan, Mosnizah Mohd, and Ghassan Kanaan, "Enhanced Arabic Information Retrieval Light Stemming and Stop-Words", M-CAIT2013, CCIS378, Springer Verlag Berlin Heidelbeg, pp. 219-228, 2013.
[18] Jaffar Atwan, Mosnizah Mohd, Hasan rashadeh, and Ghassan Kanaan, "Semantically Enhanced Pseeudo Relvence Feedback For Arabic Information Retrieval", Journal of Information Science, Vol. 42, No. 2, pp. 246-260, 2016.
[19] Mohamed Wedyan, Basim Alhadidi, and Adnan Alrabea, "The Effect of Using an Arabic Information Retrieval System", International Journal of Computer Science Issues, Vol. 9, No. 6, pp. 431-435, November 2012.
[20] Waseem Alnomima, Ibrahim F. Moawad, Rania Elgohary and Mostafa Atef, "Ontology Based Query Expansion for Arabic Text Retrieval", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, pp. 223-230, 2016.
[21] Cheng-Hui Huang, Jian Yin, and Dong Han, "An Improved Text Retrieval Algorithm Based on Suffix Tree Similarity Measure", Springer-Verlag Heidelberg, ICICA, 2010, pp. 150-157, 2010.
[22] Susan Dumais, Edward Cutrell, J. J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins, "A System for Personal Information Retrieval and Reuse", SIGIR, Toronto, Canada, 2003.
[23] Mohamed A. Abdelhadi, Tirveedula Gopi Krishna, and Ghassan Kanann, "A New Developed Model for Arabic Information Retrieval System Based on Knowledge Base System", International Journal of Emerging Research in Management and Technology, Vol. 2, No. 11, pp. 1-7, November 2013.
[24] Roi Blanco Gonzalez, "Index Compression for Information Retrieval System", Ph.D. Thesis, University of A Corunna, 2008.
[25] Kolikipogu Ramakrishna and B. Padmaja Rani, "Study of Indexing Techniques to Improve the Performance of Information Retrieval in Telguw Language", The International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 1, pp. 1-10, January 2013.
[26] Moon Soo Cha, So Yeon Kim, Jae Hee Ha, Min-June Lee, Young-June Choi, and Kyng Ah Sohn, "Topic Model Based Approach for Improved Indexing Content based on Document Retrieval", International Journal of Networked and Distributed Computing, Vol. 4, No. 1, pp. 55-64, January 2016.
[27] Yang Wei, Jinmao Wei, Zhenglu Yang, and Yu Liu, "Joint Probability Consistent Relation Analysis For Document Representation", Springer International Publishing Switzer Land, LCNS 9642, pp. 517-532, 2016.
[28] Stefan Pohl, "Boolean and Ranked Information Retrieval For Biomedical Systematic Reviewing", Ph.D. Thesis, Department of Computer Science and Software Engineering, University of Melbourne, Victoria, Australia, Feb. 2012.
[29] Xiao Wei, Jun Zhang, Daniel DajunZeng, and Qing Li, "A Multi-Level Text Representation Model Within Background Knowledge Based on Human Cognitive Process For Big Data Analysis", Cluster Computing, Vol. 19, pp. 1475-1487, 2016.
[30] Yang Wei, Jinmao Wei, and Hengpeng Xu, "Context Vector Model for Document Representation: A Computational Study", Springer International Publishing Switzerland, NLPCC 2015, LNAI 9362, pp. 194-206, 2015.

[31] Leemon Baird, and Donald H. Kraft, "A New Approach for Boolean Query Processing in Text Information Retrieval", Downloaded From the Internet in 2017 From the Website http://leemon.com/papers/2007bk.pdf.
[32] S. Ruban, S. Behin Sam, Lenita Veleza Serrao, and Harshitha, "A Study and Analysis of Information Retrieval Models", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, No. 7, pp.1-7, October 2015.
[33] Tareq Z. Ahram, "Information Retrieval Performance Enhancement Using the Average Standard Estimator and the Multi-Criteria Decision Waited Set of Performance Measures", Ph.D. Thesis, Department of Industrial Engineering and Management Systems, College of Engineering and Computer Science, University of Central Florida Orlando, Florida, 2008.
[34] Simon Jonassen, and Svein Erik Bratsberg, "Improving the Performance of Pipelined Query Processing with Skipping—and its Comparison to Document Wise Partitioning", Downloaded From the Internet in 2017 From the Website https://link.springer.com/chapter/10.1007/978-3-642-35063-4_1.
[35] Balwinder Saini, Vikram Singh, and Satish Kumar, "Information Retrieval Models and Searching Methodologies: Survey", Downloaded in 2016 From the Website https://www.researchgate.net/ publication/274837522.
[36] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "An Introduction to Information Retrieval", Cambridge University Press, 2009.
[37] Michael McCandless, Erik Hatcher, and Otis Gospodnetic, "Lucien in Action", The 2nd Edition, 2005, Downloaded From the Internet in 2017 From the Website http://www.apache.org/licenses/LICENSE-2.0

Biography:

Prof. Dr. *Nawal El-Fishawy* received the Ph.D degree in mobile communications, Faculty of Electronic Eng., Menoufia University, Menouf, Egypt, in collaboration with Southampton University in 1991. Now she is the head of Computer Science and Engineering Dept., Faculty of Electronic Eng. Her re-search interest includes computer communication networks with emphasis on protocol design, traffic modeling and performance evaluation of broadband networks and multiple access control protocols for wireless communications systems and networks. Now she directed her research interests to the developments of security over wireless communications networks (mobile communications, WLAN, Bluetooth), VOIP, and encryption algorithms. She has served as a reviewer for many national and international journals and conferences.

Prof. Dr. Mohamed Nour Elsayed is a professor of computer engineering at the Electronics Research Institute, Cairo. He was graduated from the Computer Department at the Faculty of Engineering, Ain Shams University in 1980. He obtained his M.Sc. and Ph.D. in 1987 and 1993 respectively. He taught more than twenty-years ago at the American University in Cairo (AUC) as a part-time instructor. He taught also five years ago at Princess Nourah University (Former Riyadh University), Riyadh, KSA. He was the head of the Department of Research Informatics as well as the Vice-President of the Electronics Research Institute, Cairo. He is an IEEE member and a reviewer of some national and international computer journals. The areas of his interest include; but not limited to; high performance computing, computational linguistics, and artificial intelligence applications. ***Dr.Gamal M. ATTIYA*** graduated in 1993 and obtained his M.Sc. degree in computer science and engineering from Menoufia University, Egypt, in 1999. He received PhD degree in computer engineering from University of Marne-La-Vallée, Paris-France, in 2004. He is currently associate professor at Computer Science and Engineering department, Faculty of Electronic Engineering, Menoufia University, Egypt. His main research interests include distributed computing, allocation and scheduling, cloud computing, Big Data analysis, computer networks and protocols.

Dr. Maha Saad Tolba is a lecturer of Computer Engineering at the Faculty of Electronic Engineering, Menofia University. She was graduated from the Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University in 1997. She obtained her M.Sc. and Ph.D. in 2006 and 2011 respectively. The areas of her interest include; but not limited to; computer networks, information security, and information technology.

Eng. Ayat Elnahaas is a research assistant at the Department of Research Informatics, Electronics Research Institute, Cairo. She was graduated from the Faculty of Electronic Engineering, Menofia University in 2013. Currently; she is working in her M.Sc. in the area of Arabic text processing. The research areas of her interest are: computational linguistics, and information technology.

تحليل آداء وتعديل نموذج استرجاع المعلومات العربية إعتمادا على تمديد استفسار المستخدم

آيات النحاس*[1]، نوال الفيشاوى**[2]، محمد نور*[1]، جمال عطية**[2]، مها طلبة**[2]
*قسم بحوث المعلوماتية، معهد بحوث الإلكترونيات
القاهرة- جمهورية مصر العربية
[1]eng_ayatelnahas@yahoo.com; [1]mnour@eri.sci.eg
قسم هندسة وعلوم الحاسب- كلية **الهندسة الإلكترونية
منوف- جامعة المنوفية- جمهورية مصر العربية
[2]nelfishawy@hotmail.com; [2]gamal.attiya@yahoo.com; [2]maha_saad_tolba@yahoo.com

الملخص العربى:
تهدف عملية استرجاع المعلومات إلى إيجاد الوثائق والنصوص المناسبة والتى تلبى رغبات استفسار المستخدم. يهدف هذا العمل البحثى إلى تحليل وتدقيق أحد النماذج لاسترجاع المعلومات، ومن ثم سيتم عرض العناصر الأساسية لذلك النموذج مثل تجميع الوثائق، تمثيل استفسار المستخدم، عمل الفهرسة، وكذا عمل المضاهاة. ومن منطلق أن اللغة العربية هى أحد اللغات الهامة فى اللغات الطبيعية التى يتعامل بها العالم، فإنه قد تم استخدام مجموعة من الوثائق العربية لاختبار آداء ذلك النموذج االذى تم إختياره. وعلى ذلك فإن هناك بعض العمليات سيتم إجراؤها لتسهيل عملية المضاهاة بين الكلمات الدالة لاستفسار المستخدم مع العناصر التى تصف كل وثيقة أو نص عربى، ومن أمثلة تلك العمليات: عملية تجزئة النص العربى إلى كلمات Tokens، استبعاد الكلمات التى لاتؤثر فى عملية استرجاع النصوص العربية Removal of Stopwords، وكذا إيجاد أصل الكلمة العربية Stemming بعد تجريدها من أحرف الزيادة سواء القبلية أو البعدية.
سيقدم هذا العمل أيضا إقتراحين لتعديل النموذج المستخدم رغبة فى تعزيز كفاءتة وتحسين آدائة. هذا ويعتمد التعديل على تمديد استفسار المستخدم. فالمقترح الأول يقوم بتمديد الكلمات الدالة لاستفسار المستخدم وذلك من خلال إضافة كلمات جديدة تعتمد على المعانى الدلالية للكلمات الأصلية لاستفسار المستخدم عن طريق الاستعانة بالقاموس العربى للحصول على معانى تلك الكلمات Semantics. ويقوم المقترح الثانى بتمديد الكلمات الدالة فى استفسار المستخدم بإضافة بعض الكلمات الدالة المصاحبة لبعض النصوص المسترجعة التى يراها المستخدم متوافقة مع استفسارة الأصلى، وأيضا يقوم هذا المقترح باستبعاد أى كلمات مصاحبة لبعض النصوص المسترجعة التى لايرحب بها المستخدم، وهذا ما يطلق عليه Relevance Feedback.
إضافة لما تقدم فإن نموذج استرجاع المعلومات سيتم تطبيقة واختباره وتقييمة قبل ويعد التعديلين المشار إليهما سابقا. هذا وسيتم تقييم آداءالنموذج والتعديلات التى ستجرى عليه من خلال عدد من المعايير مثل معيارالدقة، إعادة الاسترجاع، ومقياس F أو مايعرف باسم Precision, Recall, and F-measure.
هذا وتشير نتائج التجارب التى تم إجراؤها إلى أهمية عملية الحصول أصل الكلمات العربية، استبعاد الكلمات غير المؤثرة، وكذا أهمية نموذج الفراغ المتجهى. ويعتبر آداء النموذج باستخدام التعديل الأول أفضل من آداء النموذج الأصلى الذى تبنتة الدراسة بحوالى 27%، بينما وصلت نسبة التحسن إلى مايقارب 14% باستخدام التعديل الثانى مقارنة أيضا بالنموذج الأصلى. ومما تجدر الإشارة به هنا هو أن نسبة التحسن كانت أفضل بضم التعديلين سويا والتى وصلت فى حدود 15% إلى 35% مقارنة بالنموذج الأصلى.

الكلمات الدالة: الوثائق العربية، آعمال الفهرسة، نموذج المتجة الفراغى، تمديد استفسار المستخدم، المعانى الدلالية للكلمات الدالة، التغذية الخلفية المناسبة.

دور السياق فى صياغة المعنى فى الترجمة

أسماء جعفر عبد الرسول

مدرس مساعد بقسم اللغة الفرنسية، كلية الآداب، جامعة المنوفية

gmasmaa@yahoo.com

الملخص العربى :

سنتناول فى هذا البحث دور السياق فى صياغة المعنى عند الترجمة حيث أن السياق هو العامل المحدد للمعنى فى الجملة. قد يكون للجملة الواحدة خارج سياقها معانٍ متعددة. وبالتالى، فإننا سوف ندرس فى هذا البحث أنواع السياق المختلفة : السياق اللغوى والسياق الثقافى. أولاً : السياق اللغوى : أى أن هناك بعض المعانى التى لا يتحدد معناها إلا بوجودها داخل سياقها سواء هذا السياق يتمثل فى الجملة أو النص العام. وسوف نتكلم فى هذا الصدد عن المشترك اللفظى، والمعنى الخطأ، والمعنى المضاد، ونظرية تقارب المعانى. ثانيا : السياق الثقافى : وهذا السياق تتدخل فيه عوامل عدة منها الحصيلة المعرفية للمترجم، والتطويع، والتعريب، والاشتقاق، وتوليد ألفاظ جديدة لكى تساهم فى تطور اللغة الهدف. ومن جانب آخر، يجب على المترجم أن يأخذ فى اعتباره القارئ الذى سيستقبل هذه الترجمات لكى يحدد المنهجية التى ستقوم عليها عملية الترجمة. وبناءً على ذلك، سوف نطرح أمثلة للتوضيح.

***الكلمات المفتاحية* : السياق اللغوى، السياق الثقافى، المشترك اللفظى، الحصيلة المعرفية للمترجم، التعريب.**

الملخص باللغة الفرنسية :

Cette recherche porte sur l'étude du rôle du contexte dans le processus de traduction. Nous traiterons le contexte d'après deux échelles : le contexte linguistique et le contexte culturel. Premièrement, le contexte linguistique dont l'objectif est de réaliser une cohérence conformément à des éléments intertextuels ou plus précisément trouver des correspondances a posteriori. Nous étudions, à ce sujet, la polysémie, le contresens, le faux-sens, le glissement sémantique. Deuxièmement, le contexte culturel qui consiste à dégager le sens conformément à des éléments intertextuels et paratextuels. Nous étalerons, à cet égard, le bagage cognitif du traducteur, l'adaptation, l'étymologie et l'arabisation.

Les mots clés : le contexte linguistique, le contexte culturel, le bagage cognitif du traducteur, la polysémie et l'arabisation.

1 السياق اللغوى

1.1 *نظرية تقارب المعانى* ***Le glissement sémantique***

وهذه النقطة وثيقة الصلة بالحقل المعجمى، حيث أن كل كلمة هى جزء من حقل معجمى كبير تبدأ من الجزء إلى الكل والعكس صحيح، ونستطيع أن نطبق هذه النقطة عند دراسة الكناية، والاشتقاق، إلخ... ونأخذ مثال على هذه النقطة [1] :

Ex1 : «*Le Vicomte de Bragelonne*». DUMAS (Alexandre).

«عودٌ على بدءٍ». نجيب الحداد.

عند ترجمة هذا العنوان لإحدى الروايات محل الدراسة، نلاحظ أن الحداد لم يقم بترجمتها حرفياً ولكنه قام بترجمتها وفقاً للسياق النصى ووفقاً لعوامل أخرى قد تكون اقتصادية، لأن ترجمة هذه الرواية كانت فى النصف الأخير للقرن التاسع عشر وكانت منشورة آنذاك فى جريدة الأهرام، وبالتالى قد تكون هذه الترجمة بهدف شد انتباه القارئ فى هذه الحقبة الزمنية لشراء الجريدة.

***Ex2** : «La sœur du vétéran avait épousé **un expéditionnaire** du ministère de l'intérieur». SUE (Eugène), L'Orgueil, p. 3.*

«وكان يعزُّها كمعزتِه لابن شقيقتهِ النسيب الوحيد لهُ واسمهُ "اوليفيه ريموند" ابن لاحد **مستخدمى** وزارة المالية». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد1161، 1881.

فى هذا المثال، نلاحظ أن المترجم قام بترجمة الكلمة المظللة وفقاً لحقلها المعجمى الواسع، فى حين أن معناها يتحدد وفقاً للسياق بأنها : «**ناسخ أو كاتب النسخ**»، وقد وجدنا هذه الترجمة فى قواميس هذه الفترة [2]. ومما هو جدير بالذكر، أن **أمين** ذكر فى قاموسه، أن لفظ «**المستخدم**» كان معروفاً قديماً بأنه **: «الموظفون ويسمون أيضاً المستخدمين، وكان يسمّى العوام الواحد منهم "ابن عيشه"، أى أنه خاضع للوظيفة التى عليها قوام معاشه»**[3] ومن هنا نلاحظ أن الكلمة اكتسبت معانٍ أخرى عبر العصور مما يدل على تطور اللغة العربية، والدليل على ذلك، أننا فى عصرنا الحالى، نستخدم لفظ «الموظف» بدلاً من لفظ «المستخدم».

1.2 المشترك اللفظى La polysémie

يُعرف المشترك اللفظى على أن للكلمة الواحدة معانٍ متعددة. وهذه النقطة هى التى توضح حرفية المترجم وفهمه الجيد للسياق. نستعرض بعض الأمثلة لتوضيح هذه النقطة :

Ex1: «*L'Orgueil*». SUE (Eugène).

***«عزة النفس»**. ديمترى أفندى خلاط.*

عند البحث عن هذه الكلمة فى قاموس فرنسى عربى، نجد أن لها معانٍ متعددة يختلف كل منها حسب سياقه : « **زهو، كبرياء وتكبر، عجرفة، غطرسة، عُجب، خيلاء، شعور بالكرامة وموضع فخر**» [4]. ولكن وفقاً لمعنى النص ولفهم المترجم لهذه الرواية فإنها تعطى المعنى الذى نقله المترجم.

Ex2 : «Vous le voyez, Herminie, ‒ dit mademoiselle de Beaumesnil, ‒combien il y a de **raisons** pour que nous nous aimions». SUE (Eugène), *L'Orgueil*, p. 83.

«فقالت ارنسة؟؟- أترين يا ارمنا وفرة **الأسباب** الموجبة لتبادل الصداقة بيننا». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1254،1881.

Ex3 : «Vous **avez raison**, Ernestine». SUE (Eugène), *L'Orgueil*, p. 120.

«قالت ارمنا – **الحق معك** يا عزيزتى». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1903، 1881.

Ex4 : «– Soit, madame, je vous accorde jusqu'à demain midi ; je viendrai savoir votre réponse… et si elle est conforme à la **raison**… à la véritable affection maternelle… je vous devancerai de quelques instants chez Herminie, afin de me trouver chez elle alors de votre arrivée». SUE (Eugène), *L'Orgueil*, p. 130.

«قال- موافق... سأفسح لك أَجل الجواب الى غد وفى منتصف النهار أعود اليك وان رأيت جوابك منطبقا على **أحكام العقل** صادرا عن حقيقة الحنو الوالدى فحينئذ أسبقك الى دار ارمنا لانتظر قدومك». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1324، 1882.

نجد أن هذه الأمثلة تنصب على كلمة واحدة، ولكن معناها تحدد وفقا للسياق فنجد أن فى كل مثال من الأمثلة السابقة، كاتب النص يقصد معنى بعينه، وقد وُفق المترجم عند ترجمة كل هذه الأمثلة الخاصة بهذه النقطة.

من جانب آخر، نستطيع توضيح أن هناك بعض الكلمات والألفاظ التى يتم ترجمتها بمعنى واحد سواء داخل السياق أو خارجه ومن هنا سنقوم بدراسة **المعنى الخطأ والمعنى المضاد.**

1.3 ***المعنى الخطأ Le faux-sens***

ينتج المعنى الخطأ من عدم فهم النص فهماً صحيحاً من قبل المترجم، أو نتيجة لغموض بعض الأجزاء فى النص، إلا أننا فى المثال الآتى، نجد أن النص لا يحتوى على غموض، بل إنه نصاً واضحاً، وعندما قمنا بالبحث فى معاجم هذه الفترة، وجدنا أن المعنى لم يختلف عن المعنى الذى نستخدمه فى عصرنا الحالى.

Ex1 : « (…), il était venu passer **un semestre** à Paris». SUE (Eugène), *L'Orgueil*, p. 4.

«التمس رخصة **خمسة عشر** يوماً ليزور خالهُ فى باريس». حضرة ديمترى أفندى خلاط، *عزةالنفس*، الاهرام، عدد 1161، 1881.

إن معنى هذه الكلمة سواء داخل السياق أو خارجه كما هو موجود فى المعاجم : «**ستة أشهر أو نصف العام»**. ومن هنا نستطيع أن نستنتج أن هذه الترجمة سقطت سهواً من المترجم، أو أن هذه الترجمة لم تأخذ حقها فى المراجعة الكافية.

1.4 ***المعنى المضاد Le contresens***

ينتج المعنى المضاد فى بعض الأحيان من عدم نقل الإيحاء المقصود فى النص الأصلى على الوجه الصحيح، والمثال الآتى يوضح هذه النقطة :

Ex1 : «C'est que je me défie de D'Artagnan. **Il n'est pas à Fontainebleau** comme vous l'avez pu remarquer, et d'Artagnan n'est jamais absent ou oisif impunément. Aussi maintenant que mes affaires sont faites, je vais tâcher de savoir quelles sont les affaires que fait d'Artagnan». DUMAS, *Vicomte de Bragelonne II*, p. 5.

«وانهُ خائف من دارتانيان واقامتِه فى فونتنبلو ويخشى أن يكون لذلك شأن ولكنه سيكتشف خفاياهُ ويستطلع أمرهُ». نجيب الحداد، *عود على بدء*، الأهرام،4463، 1892.

من خلال قراءة النص كاملاً، نجد أن المترجم أعطى المعنى المضاد لهذه الجملة، ولم يتضح له القاعدة النحوية المتعارف عليها : «نفى النفى إثبات». وبناءً على ذلك، نجد أن المترجم قد أخل بالمعنى المراد فى النص الأصلى، وسنقوم بترجمتها على النحو التالى : «**لم يقم فى فونتنبلو أو غادر فونتنبلو».**

2. السياق الثقافى

2.1 ***الحصيلة المعرفية للمترجم Le bagage cognitif***

يجب على المترجم أن يمتلك حصيلة معرفية بجانب حصيلته اللغوية : وهذا يعنى أن ثقافة المترجم تلعب دوراً هاماً فى فهم النص وتوصيل المعنى المراد فى النص الأصلى للقارئ الذى سيتلقى هذه الترجمة. وكما أشرنا فى بداية بحثنا أن هذه الترجمات كانت منشورة فى جريدة الأهرام فى النصف الثانى من القرن التاسع عشر حيث من الممكن أن يكون القارئ لهذه الجريدة ليس قارئاً متخصصاً أو يقوم بالقراءة من أجل مهمة علمية أو مثقفاً، فمن الممكن أن يكون قارئاً بهدف التسلية، أى أن المترجم يجب عليه توضيح المعلومات التى قد لا يعرفها هذا القارئ أو أنها تختلف عن ثقافته. وبالتالى سوف نعطى أمثلة فى هذا السياق كالآتى :

Ex1 : «- Molière.

- Ah ! oui, Molière, Molière». DUMAS, *Vicomte de Bragelonne II,* p. 489.

« والى جانبه موليير الشاعر المضحك المشهور صاحب الروايات البديعة الذى لم يوجد لهُ نظير من قبله ولا من بعده فى كل ما صنعت جميع رجال الدنيا من الروايات الهزلية المضحكة والذى هو ابو المرسح الفرنساوى لا يزال يدعى باسمه ويعيد لهُ فيه الى اليوم». نجيب الحداد، *عود على بدء*، الأهرام،4644، 1893.

بالنسبة لهذا المثال، يتضح لنا أن المترجم قام بإعطاء نبذة عن **موليير** لتعريفه للقارئ الذى لا يعرفه. ومن هنا نجد أن المترجم قام بدور المعلم الذى يقوم بإضافة معلومة. وفى الوقت نفسه، أعطى هذا الكاتب الفرنسى حقه، لأنه يعتبر أيقونة كبيرة فى الأدب الفرنسى.

Ex2 : «La polka terminée, Herminie, qui tenait le piano depuis le commencement de la soirée, fut entourée, (…), et surtout invitée pour une foule de contredanses». SUE (Eugène), *L'Orgueil*, p. 80.

« ولما انتهت اغانى **البولكا (نوع من الرقص)** احدق الجميع بارمنا الجالسة وراء البيانو يتداعونها الى الرقص معهم». حضرة ديمترى أفندى خلاط، *عزةالنفس*، الاهرام، عدد 1881،1251.

نلاحظ فى هذا المثال أن المترجم لجأ إلى أسلوب التوضيح لكى يوضح للقارئ العربى هذا النوع من الرقص الذى قد يكون غريباً عنه. وبالتالى، نجد أن المترجم قد أزال الغموض الذى قد يواجه القارئ عند محاولته فهم النص.

2.2 التطويع L'adaptation

هناك عدة أسباب تجعل المترجم يلجأ إلى **التطويع** ومن بينها : **أولاً**، عندما لا يكون هناك معادل فى اللغة الهدف. **ثانياً**، أن يكون النص الأصلى به معلومات لا تتفق مع العادات والتقاليد الخاصة بالثقافة الهدف. **ثالثاً**، أن يكون التطويع هو إحدى سمات الترجمة فى العصر الذى نُشرت فيه كما هو الحال فى دراستنا.

Ex1 : «**Fasse le ciel qu'elle me pardonne**». SUE (Eugène), *L'Orgueil*, p. 136.

«**وعسى الله يغفر لى ما تقدم من ذنبى**». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1336، 1882.

من خلال هذا المثال، يتضح لنا أن المترجم قام بعملية **التطويع** من خلال استبدال الصيغة الدينية الخاصة بالدين المسيحى بصيغة تتناسب مع ثقافة القارئ العربى. وبناءً على ذلك، نستطيع تفسير سبب لجوء المترجم إلى عملية الأسلمة من خلال : أن اللغة العربية هى لغة القرآن إلى جانب أن هذه الترجمة لها وقع إيجابى على القارئ العربى حتى لا يعطيه أى شعور بالغرابة.

Ex2 : «- Un duel… avec vous ? - s'écria M. de Mornand qui, dans le premier emportement de la colère, avait oublié la position exceptionnelle du bossu, **et qui seulement alors songeait à tout ce qu'il pouvait y avoir de ridicule pour lui dans une pareille rencontre**, aussi reprit-t-il : - Un duel avec vous, monsieur ? Mais…». SUE (Eugène), *L'Orgueil*, p. 14.

«ولما سمع دى مورنان تحريضهُ على المبارزة بردت سورة غضبهِ وصار يقول المثل راحت السكرة وجاءَت الفكرة ودرى **ان مركزهُ على شفا جرف هارٍ** ومبارزتهُ اذا ما اعدمتهُ النفس وافقدتهُ الحس تزرى بقدرهِ وتحط بمنزلتهِ لان خصمهُ عنيد خبير بأبواب القتال وهذا النزال غير سجال..... ». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1165، 1881.

فى هذا المثال، قام المترجم بعملية **التطويع**، إلا أننا لا نتفق مع المترجم فى اللجوء إلى هذه الصيغة لأنه يستطيع ترجمة المعنى المراد من النص الأصلى دون أى صعوبة ويعطى نفس المعنى.

2.3 ***التعريب والاشتقاق***

إن من أهم الأدوار التى يجب على المترجم القيام بها هى **عمليتى التعريب والاشتقاق،** لأن هاتان العمليتان تمثلان الوظيفة الأساسية للمترجم وتدل على إبداعه إلى جانب دورها فى إثراء لغتنا العربية. ومن خلال الأمثلة التالية، سنجد أن المترجم لم يقم بنقل المعنى الدقيق للكلمة ولكن ترجم الكلمة وفقاً لحقلها المعجمى الواسع.

Ex1 : «M. de Maillefort et la jeune fille étaient depuis quelque temps en **voiture**, lorsque, un instant arrêtée par un embarras de **charrettes**». SUE (Eugène), *L'Orgueil*, p. 137.

«ريثما الامير سائر فى **مركبة** مع ارمنا اعترض مسيرهما **عربة** واقفة منقلبة فى الطريق». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام،عدد 1336، 1882.

Ex2 : «À peine la première voiture était-elle sortie de la cour, qu'une très belle **berline**, largement armoriée, y entra». SUE (Eugène), *L'Orgueil*, p. 150.

«ثمَّ أعقبت الاولى **المركبة الثانية** وليست أقل زهاءَ منها». حضرة ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد 1359، 1882.

من خلال المثالين السابقين، نجد أن المترجم من خلال ترجمته لم ينقل الصورة الذهنية لكل عربة من هذه العربات ولكنه اكتفى بترجمتها فقط ب «**المركبة أو العربة**»، وبالتالى، فإن القارئ المستقبل لهذه الترجمة لم يتلقى نفس المعنى الذى تلقاه القارىء الأصلى للرواية الأصلية. ولذلك، نستطيع القول، بأن ترجمة «***la voiture***» ب «**مركبة**» لا يمثل أدنى مشكلة لأن هذا هو المقابل الطبيعى لها. ولكن فيما يخص «***charettes***»، فإننا نجد أنه كان يجب على المترجم ترجمتها ب «**عربة نقل**» أو تعريبها ب «**كارتة**» أو ترجمتها بالتسمية التى كانت منتشرة فى هذه الفترة ألا وهى «**عربة كارو**».

أما فيما يخص النوع الثانى من العربات وهو «***la berline***»، فإننا نستطيع القول بأن المترجم كان يستطيع ترجمتها ب «**عربة ألمانية الصنع**»، ولكن فى هذه الحالة، كان لابد أن يصطحب هذه الترجمة بشرح لها سواء داخل النص أو من خلال الحاشية فى نهاية الصفحة.

وسنذكر مثالاً أخراً ولكنه يتناول الحقل المعجمى «للكرسى»، وفيما يلى هذا الحقل :

Ex3: «Et M. de Maillefort rapprocha **son fauteuil** du **canapé** où la baronne était assise». SUE (Eugène), *L'Orgueil*, p. 48.

«وقرَّب **كرسيهُ** من **مقعد** البارونة». ديمترى أفندى خلاط، *عزة النفس*، الاهرام، عدد1194، 1881.

ومن خلال ترجمة المترجم للفظتى «***fauteuil et canapé***» ب «**كرسيه ومقعد**»، نستطيع أن نذكر أن المترجم لم يضع يديه على المعنى الصحيح لكلتا اللفظتين. وبناءً على ذلك، فإننا نرى أن المترجم كان يجب عليه ترجمة «***canapé***» ب «**مقعد**»، وكان هذا يستدعى وصفاً دقيقاً منه لهذا المقعد حتى لا تختلط المفاهيم على القارىء المستقبل للترجمة وهذا الوصف هو «**مقعد خشبى ذو يدين وظهر عالٍ مبطن ومغطى بالقماش**». أو كان يستطيع ترجمتها ب «**الأريكة**»، وهذه الأخيرة تنتمى للغة الحبشية كما ذكر **السيوطى** [5] فى كتابه ***الاتقان فى علوم القرآن***، ولكن **العنيسى** [6] لم يتفق مع هذا الرآى ووضح أنها تنتمى للغة اليونانية.

ومن هنا نجد أن دراسة أصل الكلمات تعتبر مبحثاً مهماً وإثراء للغتنا العربية.

الخاتمة

من خلال دراسة النقاط السابقة، اتضح لنا أن المترجم لا يكتفى بمعرفته بلغته الأم فقط ولكن يجب أن يكون على دراية موسوعية باللغة الأجنبية التى يترجم منها. وبالتالى فإن المترجم، يجب عليه أن يمتلك الملكة اللغوية والملكة الثقافية أيضاً ولاغنى عنهما فى عملية الترجمة. وعلاوة على ذلك، نستطيع القول أننا نتفق مع عملية التطويع فى بعض الأحيان عندما تكون هى الحل الوحيد أمام المترجم، ولكننا لا نؤيد الإفراط فيها فى أحايين

كثيرة طالما أن المعنى المراد يمكن نقله من النص الأصلى إلى النص الهدف بكل سهولة دون حدوث أى غموض أو سوء فهم للقارئ المتلقى للترجمة. وبناءً على ذلك، قمنا بدراسة عمليتى التعريب، والاشتقاق.

قائمة المراجع :

[1] Les exemples de cette recherche sont des extraits des «***l'Orgueil***» de Eugène **Sue** et «***Le Vicomte de Bragelonne***» de Alexandre **DUMAS** et leur traduction vers l'arabe dans la deuxième moitié de 19ème siècle.

[2] «Expéditionnaire : مبيِّض، ناسخ», *Dictionnaire français-arabe*, Joseph J. Habeiche, 2ème édition, éditeur J.C. Lagoudakis, Alexandrie, 1896, P. 286. Et «Expéditionnaire : commis, copiste كاتِب النُسَخ، نَاسِخ», *Dictionnaire français-arabe*, P. J.—B. Belot S. J., 4ème édition revue et corrigée, seconde partie, imprimerie catholique, 1913, P. 482.

[3] أمين (أحمد)، ***قاموس العادات والتقاليد المصرية***، مطبعة لجنة التأليف والترجمة والنشر، الطبعة الأولى، 1953، ص386.

[4] ***ALMANHAL, dictionnaire français-arabe***, Dr. Souheil Idris, Dar El-Adab, Beyrouth, Liban, 2006, p. 850.

[5] السيوطى (جلال الدين أبو الفضل عبد الرحمن بن ابى بكر)، ***الاتقان فى علوم القرآن***، نسخة مصورة، تحقيق مركز الدراسات القرآنية، المملكة العربية السعودية، بدون تاريخ، ص 942.

[6] العنيسى (طوبيا)، ***نبذة فى أصول الألفاظ السامية كالعربية والسريانية التى دخلت فى اللغات الايتالية والاسبانية والافرنسية والانكليزية واليونانية واللاتينية وبالعكس مع ثلثة ملاحق تشتمل على كلمات دخيلة من اليونانية إلى السريانية وعلى تفسير اعلام اعجمية مستعملة فى العربية وعلى تفسير أصول ألفاظ تتألف منها أسماء قرى ومدن وعلى فوائد جمة***، طبع بنفقة حضرة الأباتى يوسف الخازن النائب العام ورئيس مدرسة الرهبانية الحلبية اللبنانية المارونية برومية، 1909،ص 12.

السيرة الذاتية

أسماء جعفر عبد الرسول حاصلة على ليسانس آداب وتربية عام 2007 ثم على ليسانس آداب عام 2009، حصلت على الماجستير فى الترجمة وتعمل حالياً مدرس مساعد بكلية الآداب، جامعة المنوفية. لقد شاركت بالبحث المعنون «الاختلاف الثقافى» فى المؤتمر الدولى للترجمة الذى أقيم فى المجلس الأعلى للثقافة والمجلس القومى للترجمة فى نوفمبر 2016. ومن جانب آخر، شاركت بالبحث المعنون «حركة الترجمة وتأثيرها على الأدب العربى» فى ملتقى العلاقات الثقافية الفرنسية-المصرية والذى أقيم فى المجلس الأعلى للثقافة يومى 21 و22 مايو 2017. ولها بحث منشور فى مجلة كلية الآداب، جامعة المنوفية، تحت عنوان «إعادة قراءة الروايات الفرنسية المترجمة إلى العربية فى الصحافة المصرية فى الفترة من 1881 حتى 1893. دراسة فى ترجمة الثقافة». وعضوة فى عدة نشاطات تابعة للجمعية المصرية لأساتذة اللغة الفرنسية، وقد حضرت الكثير من المؤتمرات والندوات واللقاءات على هامش هذه الجمعية. وقد حصلت على شهادة تفيد بإجادتها للمستوى اللغوى *Delf B2* من وزارة التربية والتعليم الفرنسية.

الملخص باللغة الإنجليزية

Abstract: This research is concerned with the study of the role of context in the translation process. It handles context according to two scales: the linguistic context and the cultural context. In the first place, the objective of the linguistic context is to realize coherence according to intertextual elements or more exactly to find a posteriori correspondences. In this regard, the study examines polysemy, contradiction and semantic nuances. Secondly, the cultural context consists of examining the sense according to intertextual and paratextual elements. In this respect, the study explores the cognitive reporting of the translator, adaptation, derivation, neologism and coined expressions and arabization. The readership or audience of a translation should always be taken into account. Examples of these concepts and strategies are provided.

Keywords: the linguistic context, the cultural context, the cognitive repertoire of the translator, the polysemy and the arabization