الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

The Sixteenth Conference on Language Engineering (ESOLEC'2016)

December 7 - 8, 2016

Organized by

Egyptian Society of Language Engineering (ESOLE)

Under the Auspices of

PROF. DR. ABDELWAHAB EZZAT
President of Ain Shams University

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

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

Conference Cochairperson
PROF. DR. SALWA ELRAMLY

Faculty of Engineering –Ain Shams University

Cairo, Egypt

http://esole-eg.org

Conference Chairman

Prof. Dr. M.A. R.Ghonaimy

Conference Sponsors

Technical Program Committee:

Prof. Taghrid Anber, **Egypt**
Prof. I. Abdel Ghaffar, **Egypt**
Prof. M. Ghaly, **Egypt**
Prof. M. Z. Abdel Mageed, **Egypt**
Prof. Khalid Choukri, ELDA, **France**
Prof. Nadia Hegazy, **Egypt**
Prof. Christopher Ciri, LDC, **U.S.A**
Prof. Mona T. Diab, StanfordU., **U.S.A**
Prof. Ayman ElDessouki, **Egypt**
Prof. Afaf AbdelFattah, **Egypt**
Prof. Y. ElGamal**, Egypt**
Prof. M. Elhamalaway**, Egypt**
Prof. S. Elramly, **Egypt**
Prof. Hani Kamal**, Egypt**
Prof. A. A. Fahmy**, Egypt**
Prof. I. Farag, **Egypt**
Prof. Magdi Fikry, **Egypt**
Prof. Wafaa Kamel, **Egypt**
Prof. S. Krauwer, **Netherlands**
Prof. Bente Maegaard, CST, **Denmark**
Prof. A. H. Moussa, **Egypt**
Prof. M. Nagy**, Egypt**
Prof. A. Rafae, **Egypt**
Prof. Mohsen Rashwan**, Egypt**
Prof. H.I. Shaheen**, Egypt**
Prof. S.I. Shaheen**, Egypt**
Prof.HassaninM. AL-Barhamtoshy**, Egypt**
Prof. M. F. Tolba, **Egypt**
Dr.Tarik F. Himdi**, Saudi Arabia**

Organizing Committee

Prof. I. Farag
Prof. S. Elramly
Prof. H. Shahein
Dr. Mona Zakaria

Prof. H. M. Al-Barhamtoshi
Prof. M. Z. Hani Kamal
Dr. A. Passant Elkafrawy
Dr. Bassant Abdelhamid

Conference Secretary General

Prof. Dr. Salwa Elramly

The Sixteenth Conference on Language Engineering
Final Program

Wednesday 7 December 2016

9.00 - 10.00 Registration

10.00 - 10.30 Opening Session: Seminar Room, Central Biblioteque Building

10.30 - 11.00 **Session 1:**Seminar Room: **Invited Paper 1:Computational Linguistics**
Chairman**:** Prof. Dr. Ibrahim Farag

أثر العين الحلقي للفعل الثلاثي المضعف على الباب الصرفي لمضارعه: دراسة لغوية حاسوبية
ا.د/ وفاء كامل فايد
كلية الآداب – جامعة القاهرة

11.00 - 11.30 **Session 2:** Seminar Room: **Invited Paper 2: Speech Recognition I**
Chairman**:** Prof. Dr. Ibrahim Farag

Arabic Speech Recognition: Challenges and State of the Art
Sherif Abdou
Faculty of Computers and Information Technology, Cairo University, Giza, Egypt.

11.30 - 12.00 Coffee Break (Conference Center)

12.00 - 12.30 **Session 3** : Conference Center: **Invited Paper 3: Arabic OCR**
Chairman: Prof. Dr. Aly Aly Fahmy

Arabic Document Pre-processing and Layout Analysis
Hassanin M. Al-Barhamtoshy
King Abdulaziz University, Faculty of Computing, IT Department, Jedda, Saudi Arabia.

12.30 - 13.30 **Session 4**: Conference Center: **Natural Language Analysis**
Chairman: Prof. Dr. Aly Aly Fahmy

1. MASAR: A Morphologically Annotated Gold Standard Arabic Resource
Sameh Alansary
Bibliotheca Alexandrina, Alexandria, Egypt

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

2. Improving Alserag Arabic Diacritization System through Syntactic Analysis
Sameh Alansary
Bibliotheca Alexandrina, Alexandria, Egypt

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

13.30 - 14.30 **Session 5**: Conference Center: **Language Analysis for Classification**
Chairman: Prof. Dr. M. Younis El-Hamalawy

1. **Software and Hardware Implementation for Documents Classification using Self-Organizing Maps (SOM)**
Abdelfattah ELsharkawi *, Ali Rashed **, Hosam EldinFawzan*
[*]Department of Systems and Computer Engineering, Al-Azhar University,Egypt
[**]Department of Electrical and Computer Engineering, Faculty of Engineering Science, Sinai University, Egypt.

2. **Semantic Approach for Classification of Web Documents**
PassentElkafrawy, Dina ElDemerdash
Faculty of Science, Menofia University, Egypt.

14.30 - 15.30 Lunch

15.30 - 16.30 **Session 6:**Conference Center (Room A): **NLP for Information Retrieval**
Chairman**:**Prof. Dr. Hani Mahdi

1. **Building Topic-Language based Index for Multi-lingual Information Retrieval**
EbtsamSayed[*], Samir Elmougy[**], MostafaAref [***]
[*]Computer Science, Faculty of Computers and Information, Minia University, Minia, Egypt
[**]Computer Science, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt
[***]Computer Science, Faculty of Computers and Information, Ain Shams University,Cairo, Egypt

2. **Ambiguity Detection and Resolving in Natural Language Requirements**
Somaia Osama, Safia Abbas, Mostafa Aref
[*]Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

15.30 - 16.30 **Session 7:** Conference Center (Room B): **Speech Encryption**
Chairman : Prof. Dr. Ayman Bahaa

1. **A Combined DES and Elliptic Curve Cryptography Cryptosystem to Secure Audio Data**
Mohamed Ahmed Seifeldin[*], Abdellatif Ahmed Elkouny[**], Salwa Elramly[*]
[*]Electronics and Communication Department, Faculty of Engineering, Ain Shams University, Abbassia, Egypt
[**]Computer Science Department, Faculty of Engineering, Ahram Canadian University,6th October, Egypt.

2. **Speech Cryptosystem Based On Chaotic Modulation Technique**
Mahmoud F. AbdElzaher[*], Mohamed Shalaby[**], Yasser Kamal[**], Salwa Elramly[*]
[*]Department of Electronics and Electrical Communications, Ain Shams University,Cairo, Egypt
[**]Department of Computer Science, Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt

Thursday 8 December 2016

10.00 - 11.00 **Session 8:** Conference Center: **Speech Recognition II**
Chairman: Prof. Dr. Waleed Fakhre

1. **Robust Speaker Recognition Using Adaptive Hidden Markov Models**
   Aya S. Mostafa, Amr M. Gody, Tamer M. Barakat
   Department of Electricity, Faculty of Engineering, Fayuom University, Egypt.
2. **Enhancement Quality and Accuracy of Speech Recognition System by Using Multimodal Audio-Visual Speech signal**
   Eslam E. El Maghraby, Amr M. Gody, Mohamed H. Farouk
   Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt.

11.00 - 12.00 **Session 9:** Conference Center: **Natural Language Processing for Information Retrieval**
Chairman: Prof. Dr. Hassanin Al-Barhamtoushi

1. **A Proposed Arabic Text to Sign Language Translator**
   A. S. Elons*, A.Ali**, M. F. Tolba*
   * *Scientific Computing Department- Faculty of Computers and Information Sciences- Ain Shams University-Cairo-Egypt*
   ***Department of Electrical Electronics and Communication Engineering, Cairo University, Giza, Egypt*
2. **Data Preparation and Handling for Written Quran Script Verification**
   Mohsen A. Rashwan, Ali Ramadan, Hazem M. Safwat, Salah Ashraf, Hazem Mamdouh
   Department of Electrical Electronics and Communication Engineering,Cairo University, Giza, Egypt

12.00 - 12.30 Coffee Break

12.30 - 14.00 **Session 10**: Conference Center: **Natural Language Processing**
Chairman : Prof. Dr. Nadia Hegazi

1. **Towards Building CECA WordNet: A Domesticated Arabic-English Lexicon of Contemporary Egyptian Colloquial Arabic Words from Twitter**
   Bacem A. Essam*, Prof Dr. Mostafa M. Aref**
   * *English Language Department, Faculty of Al-Alsun, Ain Shams University*
   ** *Computer Science, Faculty of Computer Science and Information Sciences, Ain Shams University, Cairo, Egypt*
2. **A Rule Based Method for Adding Case Ending Diacritics for Modern Standard Arabic Texts**
   Sameh, Fashwan
   Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt.
3. **Lexical Growth in Child Egyptian Arabic:A Corpus Based Study**
   Heba Salama, Sameh Alansary
   Phonetics and linguistics Department, Faculty of Arts Alexandria University, Alexandria, Egypt.

4. **نحو إنجاز قاعدة بيانات للاسم في اللغة العربية**

دكتورة فطوم القرييش

أستاذة التعليم الثانوي التأهيلي- عضوة مع فريق ابتكارات بالمدرسة المحمدية للمهندسين بالرباط -المغرب.

14.00 - 15.00 Lunch

15.00 - 16.00 **<u>Session 11</u>**:Conference Center: **Students' Workshop**
Chair Committee: Prof. Dr. Nadia Hegazy, Prof. Dr. M.Elhamalawy, Prof. Dr. Hassanin Al-Barhamtoushi

16.00 - 16.30 Closing session

Program at A Glance

	Day	Time	Location	Subject	Chairman
Opening Session	Wednesday	10.00 - 10.30	Seminar Room		
Session 1		10.30 - 11.00		**Computational Linguistics**	Prof. Dr. Ibrahim Farag
Session 2		11.00 - 11.30		**Speech Recognition I**	Prof. Dr. Ibrahim Farag
Coffee Break		11.30 - 12.00	Conference Center		
Session 3		12.00 - 12.30		**Arabic OCR**	Prof. Dr. Aly Aly Fahmy
Session 4		12.30 - 13.30		**Natural Language Analysis**	Prof. Dr. Aly Aly Fahmy
Session 5		13.30 - 14.30		**Language Analysis for Classification**	Prof. Dr. M. Younis El-Hamalawy
Lunch		14.30 - 15.30			
Session 6		15.30 - 16.30	Room A	**NLP for Information Retrieval**	Prof. Dr. Hani Mahdi
Session 7		15.30 - 16.30	Room B	**Speech Encryption**	Prof. Dr. Ayman Bahaa
Session 8	Thursday	10.00 - 11.00	Conference Center	**Speech Recognition II**	Prof. Dr. Waleed Fakhre
Session 9		11.00 - 12.00		**Natural Language Processing for Information Retrieval**	Prof. Dr. Hassanin Al-Barhamtoushi
Coffee Break		12.00 - 12.30			
Session 10		12.30 - 14.00		**Natural Language Processing**	Prof. Dr. Nadia Hegazi
Lunch		14.00 - 15.00			
Session 11		15.00 - 16.00		**Students' Workshop**	Prof.Dr. Nadia Hegazy, Prof.Dr. M.Elhamalawy, Prof. Dr. Hassanin Al-Barhamtoushi
Closing session		16.00 - 16.30			

Seminar Room: Central Biblioteque Building, **Conference Center:** Main Building, **Room A/B:** Conference Center

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

أعضاء الجمعية من المؤسسات

1- مركز نظم المعلومات – كلية الهندسة – جامعة عين شمس

2- معهد الدراسات والبحوث الإحصائية – جامعة القاهرة

3- مركز الحساب العلمى – جامعة عين شمس

4- الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحرى

5- أكاديمية أخبار اليوم

6- معهد بحوث الإلكترونيات

7- معهد تكنولوجيا المعلومات

8- مكتبة الإسكندرية

9- المعهد القومى للاتصالات (NTI)

10- الشركة الهندسية لتطوير نظم الحاسبات (RDI)

11- الهيئة القومية للاستشعار من بعد و علوم الفضاء

12- كلية الحاسبات و المعلومات جامعة قناة السويس

13- دار التأصيل للبحث و الترجمة

أهداف الجمعية

1- الاهتمام بمجال هندسة اللغويات مع التركيز 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.

2- متابعة التطور فى العلوم والمجالات المختصة بهندسة اللغة

3- التعاون مع الجمعيات العلمية المماثلة على المستوى المحلى والقومى والعالمى.

4- 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.

5- إنشاء مجلة علمية دورية للجمعية ذات مستوى عال لنشر البحوث الخاصة بهندسة اللغة وكذلك بعض النشرات الدورية الإعلامية الأخرى بعد موافقة الجهات المختصة.

6- عقد ندوات لرفع الوعى فى مجال هندسة اللغة

7- 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

8- إنشاء مكتبة تتاح للمهتمين بالموضوع تشمل المراجع وأدوات البحث من برامج وخلافه.

9- خلق مجال للتعاون وتبادل المعلومات وذلك عن طريق تهيئة الفرصة لعمل بحوث مشتركة بين المشتغلين فى نفس الموضوعات.

10- تقييم المنتجات التجارية أو البحثية والتى تتعرض لعملية ميكنة اللغة.

11- رصد الجوائز التشجيعية للجهود المتميزة فى مجالات هندسة اللغة.

12- إنشاء فروع للجمعية فى المحافظات.

الجمعية المصرية لهندسة اللغة

THE EGYPTIAN SOCIETY
OF LANGUAGE ENGINEERING

المؤتمر السادس عشر لهندسة اللغة

7-8 ديسمبر 2016

جمهورية مصر العربية-القاهرة

ينظم المؤتمر

الجمعية المصرية لهندسة اللغة

تحت رعاية

الأستاذ الدكتور/ عبد الوهاب عزت
رئيس جامعة عين شمس

الأستاذ الدكتور/محمدأيمن عاشور
عميد كلية الهندسة -جامعة عين شمس

رئيس المؤتمر
الأستاذ الدكتور/ محمد أديب رياض غنيمى

مقرر المؤتمر
الأستاذ الدكتور / سلوىحسين الرملى

مكان عقد المؤتمر : كلية الهندسة - جامعة عين شمس

Table of Contents

IX. Students Workshop

بسم الله الرحمن الرحيم

أثر العين الحلقي للفعل الثلاثي المضعف على الباب الصرفي لمضارعه: دراسة لغوية حاسوبية

أ. د. وفاء كامل فايد

أستاذة اللسانيات
قسم اللغة العربية، بكلية الآداب، جامعة القاهرة
الجيزة - القاهرة – جمهورية مصر العربية
wafkamel@yahoo.com

فاتحة:

كانت دراستي عن (تراكب الأصوات في الفعل الثلاثي الصحيح)(1)منطلقاً مهما في هذا البحث؛ إذ خلَصَت إلى عدد من القواعد التي تحكم تآلف الأصوات العربية وتنافرها، وهو ما يشير إلى أن وراء السلوك اللغوي التلقائي للعربية نسقا ضمنيا يحدد النماذج المقبولة وغير المقبولة، وهذا ما أتاح للعربي القديم أن يقبل منها ما هو جدير بالقبول، ويعرض عمّا سواه.

واتجهت بتفكيري إلى الفعل الثلاثي المضعف: هل يخضع لذلك النسق الضمني الذي يفعل فعله في تمييز المقبول من غير المقبول؟ وهل يكون تصرف الفعل المضعف على باب صرفي بعينه راجعا إلى سيطرة هذا النموذج المختزن في العقل العربي؟ وهل يكون لأحياز أصوات هذا النوع من الأفعال ومخارجها أثر في اتجاه الفعل للتصرف على باب صرفي بعينه ؟ وهل يمكن أن نتبين مدى ارتباط أحياز الأصوات ومخارجها بالباب الصرفي للمضعف؟

أسئلة راحت تلح على تفكيري فحاولت البحث عن إجابة لها، ورأيت أن أعتمد القاموس المحيط للفيروزابادي في رصد جميع الأفعال الثلاثية المضعفة به، متوخية بذلك أن يكتسب البحث طابع الاستقصاء؛ كي يخلص من دراسة المعطيات الشاملة إلى صورة واضحة محددة المعالم، يمكن أن تؤدي إلى تحليل دقيق، يفضي بنا إلى تلمس الطريق إلى إجابات شافية لتلك التساؤلات، وقد تساعدنا في معرفة بعض القواعد التي تزيح الغموض عن هذا الجانب، وتوضح لنا أثر تجاور صوتَيْ الفعل المضعف فيتصرفه على باب صرفي بعينه، ومدى ارتباط أحياز أصوات المضعف ومخارجها بالباب الصرفي للفعل.

في عدد من بحوثي السابقة دَرستُ أثر تجاور صوتي الفعل الثلاثي المضعف على بابه الصرفي(2) ، ورصدتُ عددا من القواعد التي تربط بين أصوات الفعل الثلاثي المضعف واتجاهه إلى التصرف على باب صرفي بعينه. وسجلت أن العلاقة بين صوتي الفعل المضعف وبابه الصرفي تمثلت في مظهرين: أولهما تنافر صوتي الفعل، وثانيهما اتجاه الصوتين إلى التصرف على باب صرفي دون غيره. ومن ثم فقد ارتضيت تقسيم هذه الدراسة إلى ثلاثة محاور:

- يرصد أولها القواعد الحاكمة لتنافر صوتي الفعل المضعف(3).
- ويبحث الثاني أثر فاء الفعل المضعف في تصرفه على باب صرفي بعينه(4).
- ويبحث الثالث أثر عين المضعف ولامه في اتجاهه إلى التصرف على باب صرفي بعينه.

وتعالج هذه الدراسة المحور الثالث منها، وهو أثر العين الحلقي للفعل الثلاثي المضعف في تصرف مضارعه على باب صرفي بعينه، وهو جانب لم يدرس من قبل، فيما أعلم.

وقد ارتضيت معالجة تصنيف الدراسة للأصوات العربية الصامتة إلى أقسام وفقا لأحياز الأصوات، كما وردت عند الخليل، مع ربط هذه الأحياز والمخارج وتسمياتها بما أورده اللغويون المحدثون من عرب وأجانب.

أهداف البحث :

يهدف هذا البحث إلى تلمس الإجابة عن التساؤلات الآتية:

1- هل يؤثر مخرجالعينالحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه ؟

2- هل يؤثر حيزالعينالحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه ؟

3- هل يؤثر اتفاق صفات صوتي الفعل الثلاثي المضعف، أو اختلافها، في ورود الفعل على باب صرفي بعينه ؟

4- هل يمكن تلمس بعض القواعد التي تحكم أثر العين الحلقي على الباب الصرفي للفعل الثلاثي المضعف ؟

عينة البحث:

اعتمدت الدراسة القاموس المحيط للفيروزابادي؛ لاستخراج الأفعال الثلاثية الصحيحة التي وردت به(5) ؛ لغزارة مادته مع اختصاره، ولحرصه على ضبط حروف كلماته بالشكل، إلى جانب التزامه بتحديد الباب الصرفي لأفعاله بربطها بأوزان الأفعال المعروفة. وقد استقصت الدراسة الأفعال الثلاثية الصحيحة المضعفة به، واتخذتها عينةً للبحث.

خطوات البحث

استقصت الدراسة الأفعال الثلاثية الصحيحة المضعفة التي وردت بالقاموس المحيط، وسجلتها مع تصريفاتها في جدول خاص قام عليه البحث: جدول رقم (3).

وحين وردت بعض الأفعال بالقاموس المحيط بصيغة الماضي دون المضارع(6) ، استكمل مضارعها من لسان العرب لابن منظور، ثم من تاج العروس للزبيدي؛ حرصا على التثبت من الباب الصرفي.

ومن الأفعال المرصودة في الجدول رقم (3) رصدت الدراسة تصرف المضعف الثلاثي حين يكون أحد الأصوات الحلقيةعيناولاما له، وتتغير أصوات فائه: جدول رقم (4).

المصطلحات:

قبل عرض نتائج البحث يلزم تحديد منظومة المصطلحات المستخدمة فيه؛ حتى لا يحدث لبس في المفاهيم، وهي:

المخرج(7) Point of articulation :

هو النقطة التي يلتقي فيها عضوان من أعضاء النطق ليمر هواء الزفير بينهما ويحدث الصوت.

الحيز(8) Range of articulation :

مساحة تشتمل على أكثر من مخرج، وتكون المخارج فيها متقاربة.وتشترك الأصوات التي تنتمي إلى حيز واحد عادة في خصائص جامعة.

الصوت المجهور(9) :Voiced

الصوت المجهور صوت يكون معه الوتران الصوتيان متقاربين، بحيث يسبب اندفاع هواء الزفير من الحنجرة– مارا خلالهما - تذبذبا منتظما شديدا في الوترين الصوتيين.

الصوت المهموس(10) Voiceless :

صوت يكون معه الوتران الصوتيان متباعدين، بحيث يمر هواءالزفير في منطقة الحنجرة، دون اهتزازللوترين الصوتيين.

الصوت الشديد (الانفجاري/ الوقفي)(11):Plosive (stop)

صوت ينتج عن التقاء تام لحظي بين عضوين من أعضاء النطق، يوقف تيار الهواء في الفم عند نقطة

الالتقاء، ويتبعه تسريح سريع وفوري لهواء الزفير.

الصوت الرخو (الاحتكاكي): Fricative

صوت ينطق بحدوث تقارب شديد بين عضوي النطق، ينشأ عنه تضييق لممر الهواء عند نقطة المخرج، وحدوث حفيف أو احتكاك مسموع(12).

الأصوات المتوسطة(13)**(الموائع)**Liquids:

أصوات تنطق بالتقاء عضوين من أعضاء النطق التقاء تاما، ولكن النفس يجد له مسربا إلى الخارج، فيمر الهواء دون أن يحدث أي نوع من الصفير أو الحفيف المسموع، ويندرج تحتها الصوامت الآتية: اللام والميم والنون والراء، ثم العين- وفقا لرأي سيبويه(14).

الإطباق(15): Velarization :

ظاهرة يرتفع فيها مؤخر اللسان إلى الحنك الأعلى، آخذا شكلا مقعرا؛ مما يزيد من حجم تجويف الفم، ويضيقمن حجم تجويف الحلق أثناء إخراج الصوت، فيسمع الصوت مفخما. والأصوات المطبقة(16)أربعة، هي الصاد والضاد والطاء والظاء.

الانفتاح(17): Non velarization

الانفتاح ضد الإطباق، ويكون تجويف الفم مع الصوت المنفتح أقل منه مع نظيره المطبق. ويندرج تحت الأصوات المنفتحة كلُّ الأصوات غير المطبقة.

الاستعلاء(18):

ارتفاع اللسان إلى الحنك الأعلى، سواء أصاحبه إطباق أم لا. وتضم أصوات الاستعلاء كلا من: الخاء والغين والقاف، مع الصاد والضاد والطاء والظاء.

الاستفال(19):

ضد الاستعلاء، وهو انخفاض اللسان في الفم. ويندرج تحت الأصوات المستفلة كلُّ الأصوات غير المستعلية.

الصوتان المستطيلان :

هما الشين والضاد، ويجمعهما حيز واحد(20) ، كما تجمعهما صفة الاستطالة(21).

التفشي(22):Hushing

يمثله صوت الشين، ويتم النطق به مع ارتفاع مقدم اللسان بصورة تسمح بحدوث احتكاك زائد (هشيش).

التردد:Trill

يمثله صوت الراء، وينطق بطريقة يُحدث فيها طرف اللسان سلسلة من عمليات غلق لحظية، تتخللها عناصر حركية صغيرة(23).

الخيشومية (الأنفية):Nasalization

تتصف الأصوات بهذه الصفة حين يغلق تجويف الفم، ويهبط الحنك الرخو، مع السماح بمرور الهواء عن طريق الأنف(24) ، وهما صوتا الميم والنون في العربية.

الجانبية:Lateralization

يمثلها صوت اللام، ويلتصق اللسان فيه مع الجزء الأوسط من أصول الأسنان، على حين تسمح حافتا

اللسان الجانبيتان للهواء بالانطلاق إلى الخارج، وأحيانا يكون ممر الهواء الجانبي من جانب واحد فحسب(25).

ويوضح الجدول التالي مصطلحات الخليل، التي استخدمها البحث، والمصطلحات العربية والانجليزية الحديثة المقابلة لها: جدول رقم (1)

جدول رقم (1) : المصطلحات المستخدمة في البحث، كما وصفها الخليل وسيبويه

المصطلح قديما	حديثا	مقابله الأجنبي	تفسير المصطلح	الأصوات المندرجةتحته
المخرج	المخرج	Point of articulation	نقطة التقاء عضوي النطق ليمر هواء الزفير بينهما ويحدث الصوت.	كل الصوامت
الحيز		Range of articulation	مساحة تشتمل على أكثر من مخرج، وتكون المخارج فيها متقاربة.	
صوت مجهور	مجهور	Voiced	صوت يتقارب معه الوتران الصوتيان، بحيث يسبب اندفاع هواء الزفير من الحنجرة – مارا خلالهما - تذبذبا منتظما شديدا فيهما.	أ- ع- غ- ق- ج – ض- ل- ن- ر- ز- ط- ظ– د- ذ- ب- م
صوت مهموس	مهموس	Voiceless (Unvoiced)	صوت يكون معه الوتران الصوتيان متباعدين، بحيث يمر هواء الزفير في منطقة الحنجرة، دون اهتزازللوترين الصوتيين.	هـ- ح- خ- ك- ش- ص- س- ت- ث- ف
صوت شديد	انفجاري (وقفي)	Plosive (stop)	صوت ينتج عن التقاء تام لحظي بين عضوي النطق، يوقف تيار الهواء في الفم عند نقطة الالتقاء، ويتبعه تسريح فوري لهواءالزفير.	ق- ك- ج- ط- ت د- ب- أ
صوترخو	احتكاكي	Fricative (Lax)	صوت ينطق بحدوث تقارب شديد بين عضوي النطق، ينشأ عنه تضييق ممر الهواء عند نقطة المخرج، وحدوث حفيف أو احتكاك مسموع.	هـ- ح- غ- خ- ش- ص- ض- س- ث- ف- ز- ظ- ذ
صوتمتوسط	الموائع	Liquids	ينطق بالتقاء عضوين من أعضاء النطق التقاء تاما، ولكن النَفَس يجد له مسربا إلى الخارج، فيمر الهواء دون أن يحدث صفيرا أو حفيفا مسموعا.	ل- م – ن- ر + (ع) عند سيبويه
صوت مستطيل			صوتان هما الشين والضاد، ويجمعهما حيز واحد، وصفة الاستطالة.	ش- ض

تابع جدول رقم (1) : المصطلحات المستخدمة في البحث، كما وصفها الخليل وسيبويه

المصطلح قديما	حديثا	المقابل الأجنبي	تفسير المصطلح	الأصوات المندرجة تحته
صوت مطبق	مفخم	Velarized	يرتفع فيه مؤخر اللسان إلى الحنك الأعلى آخذا شكلا مقعرا؛ مما يزيد حجم تجويف الفم، ويضيق حجم تجويف الحلق أثناء إخراج الصوت، فيسمع الصوت مفخما.	ص- ض- ط- ظ
صوت منفتح	مرقق	Non-velarized	يكون تجويف الفم مع الصوت المنفتح أقل منه مع نظيره المطبق.	الأصوات غير المطبقة
المستعلي			ارتفاع اللسان إلى الحنك الأعلى، سواء أصاحبه إطباق أم لا.	خ- غ- ق- ص ض- ط- ظ
مستفل	منخفض		الاستفال ضد الاستعلاء، وهو انخفاض اللسان في الفم.	غير المستعلية
الصوت المتفشي		Hushing	صفة لصوت الشين، الذي يتم النطق به مع ارتفاع مقدم اللسان بصورة تسمح بحدوث احتكاك زائد (هشيش).	ش
المكرر	ترددي	Trill (Sonorant)	صفة لصوت الراء، الذي ينطق بطريقة يُحدث فيها طرف اللسان سلسلة من عمليات غلق لحظية، تتخللها عناصر حركية صغيرة.	ر (من أصوات الرنين)
خيشومي	أنفي	Nasal	تتصف الأصوات بهذه الصفة حين يغلق تجويف الفم، ويهبط الحنك الرخو، مع السماح بمرور الهواء عن طريق الأنف.	ن- م (من أصوات الرنين)
منحرف	جانبي	Lateral	يلتصق اللسان فيه مع الجزء الأوسط من أصول الأسنان، في حين تسمح حافتا اللسان الجانبيتان للهواء بالمرور إلى الخارج، وأحيانا يمر الهواء من جانب واحد فحسب.	ل (من أصوات الرنين)

وقد اتبعت الباحثة ترتيبالخليل(26) للأصوات الصامتة، كما ورد في كتاب (العين)، وأضافت إليه الهمزة بترتيب سيبويه، فقسمت الصوامت إلى المجموعات الآتيـة:

1- **أصوات الحلق** (أ[ʔ]- هـ [h]– ع[ʕ]- ح [ħ] -غ[ɣ]- خ[x](27)):

ويضم حيزها ثلاثة مخارج: أولهامخرج صوتين من أقصى الحلق، هما الهمزة والهاء(أ[ʔ]- هـ [h]()(28)، والثاني مخرج صوتين من وسط الحلق، هما العين والحاء (ع[ʕ]- ح [ħ])،والثالث مخرج صوتين من أدنى الحلق، هما الغين والخاء (غ[ɣ]- خ[x]).

2- **صوتا اللهاةوالحنك الأعلى**: (ق [q] - ك [k]):

وهذان الصوتان يجمعهما حيز واحد(29)، وهما القاف اللهوي، ثم الكاف من أقصى الحنك.

3- **الأصوات الشجرية**(30):(ج [ʤ]- ش [ʃ] - ض [d’]).

4- **الأصوات الأسلية**(31):(ص[sʕ] - س[s] - ز[z]).

5- **الأصوات النطعية**(32):(ط[dʕ] - ت[t] - د [d]) .

6- **الأصوات اللثوية**(33): (ظ [ðʕ] - ث[θ] – ذ[ð]).

7- **الأصوات الذلقية**(34) : (ر[r] - ل[l] - ن[n]).

8- **الأصوات الشفهية** : (ف [f] - ب [b] - م [m]) .

للملاحظة:

1- لما كان المضعف في صيغة الماضي يختلط فيه كل بابين من الأبواب الآتية:

1- (نصَر) مع (كَرُم).

2- (ضرَب) مع (حسِب يحسِب)، بكسر السين فيهما، بمعنى: ظن.

3- (فتَح) مع (عَلِم).

لاتحادهما في صيغة المضارع. ولما كان القاموس المحيط يكتب ماضي الفعل المضعف للغائب-غالبا- فقد اكتفت الدراسة بالأبواب الثلاثة: (نصر) و(ضرب) و(فتح) للمضعف. على أنها نبهت في الحاشية على صيغة الفعل الذي نص القاموس على أنه يتصرف على باب آخر، أو نسَبَه القاموس إلى ضمير الرفع فظهر بابه الصرفي من صيغة الماضي.

2- رمزت الدراسة في جداول البحث لكل من الأبواب الصرفية برقم خاص هو:

(1) = نصر. (2) = ضرب. (3) = فتح.

والجدول التالي يحدد أحياز الصوامت الصحيحة ومخارجها وصفاتها كما وردت عند الخليل، وعند علماء اللغة المحدثين من عرب وأجانب: جدول رقم (2)

جدول رقم (2):مخارج الأصوات الصحيحة وصفاتها كما وردت عند الخليل

الحيز عند الخليل	المصطلح الحديث	المصطلح الأجنبي	المخرج	الصوت	صفات الصوت
حلقي	أصوات الحلق	laryngeal	(أقصى الحلق)	الهمزة	انفجاري مجهور منفتح مستفل
				هـ	احتكاكي مهموس منفتح مستفل
		Pharyngeal	وسط الحلق (من البلعوم)	ع	متوسط مجهور منفتح مستفل
				ح	احتكاكي مهموس منفتح مستفل
			أدنى الحلق (الحنك اللين)	غ	احتكاكي مجهور منفتح مستعل
				خ	احتكاكي مهموس منفتح مستعل
لهوي	طبقي (الحنك الرخو)	Uvular	اللهــاة	ق	انفجارى مجهور منفتح مستعل
		Velar	الحنك الاعلى	ك	انفجارى مهموس منفتح مستفل
شجري	غاري (حنكي)	Palatal	شجر الفم	ج	مزجى مجهور منفتح مستفل
				ش	احتكاكى مهموس منفتح مستفل متفش مستطيل
				ض	احتكاكى مجهور مطبق مستعل مستطيل
أسلي	صفيري (أسلي)	Sonorant (sibilant)	الأسلة	ص	احتكاكى مهموس مطبق مستعل
				س	احتكاكى مهموس منفتح مستفل
				ز	احتكاكى مجهور منفتح مستفل

تابع جدول رقم (2): مخارج الأصوات الصحيحة وصفاتها كما وردت عند الخليل

الحيز عند الخليل	المصطلح الحديث	المصطلح الأجنبي	المخرج	الصوت	صفات الصوت
نطعي	أسناني	Dental	نطع الغار	ط	انفجارى مجهور مطبق مستعل
				ت	انفجارى مهموس منفتح مستفل
				د	انفجارى مجهور منفتح مستفل
لثوي	بين أسناني	Alveolar (interdental)	من بين الأسنان	ظ	احتكاكى مجهور مطبق مستعل
				ث	احتكاكى مهموس منفتح مستفل
				ذ	احتكاكى مجهور منفتح مستفل
ذلقي	الأصوات المتوسطة (الموائع)	Alveolar	ترددي/ مكرر	ر	متوسط مجهور منفتح مستفل مكرر
			جانبي	ل	متوسط مجهور منفتح مستفل جانبى
			خيشومي	ن	متوسط مجهور منفتح مستفل خيشومى
شفهي		Labial	شفهي اسناني	ف	احتكاكى مهموس منفتح مستفل
			شفتاني	ب	انفجارى مجهور منفتح مستفل
			شفتاني انفي	م	متوسط مجهور منفتح مستفل خيشومى

جدول رقم (3)
الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																									
	الحلق						اللهاة		الأحياز الوسطية												الذلقية			الشفهية		
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
همزة	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
هاء	*	*	–	–	–	–			–	–	*	–	–	–	–	–	–	–	–	–	–		–			
عين	–	*	–	–	–	**						–	–	–				–		–		–	–	–		
حاء	*	–	–	–	–	–		–			–					–		–	–		–					
غين	–	–	–	–	–	–	–	–	–		–		–	–	–	–	–	–	–		–	–	–			–
خاء	–	–	–	–	–	–	–										–	–	–	–						–
قاف	–						–	–							–	–		–	–	–						
كاف					–	–	–	–	–						–			–		–			–			

تابع جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حيز أصوات فاء المضعف																									
	الحلق						اللهاة		الأحياز الوسطية												الذلقية			الشفهية		
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلة			نطع الغار			اللثة			حروف الذلاقة			الشفتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
جيم					–		–		–			(1)			–	–		(2)								
شين										–	–	–	–	–		–		–	(3)			–				
ضاد						–		–		–	–	–	–	–	–	–		–	–	–		–				
صاد											–	–	–	–	–	–		–	–	–						
سين											–	–	–	–		–		–	–	–				–		
زاي										*		–	–		–	–	–	–	–	–						
طاء		–						–	–		–	–	–		–	–	–	–		–	–			–		
تاء											–		–		–	–	–	–	–	–						
دال														-	–	–	–	–	–	–						

تابع جدول رقم (3)

الأفعال المضعفة الثلاثية التي لا تقع في العربية بسبب تنافر أصواتها

عين الفعل ولامه	حـــيـز أصـوات فـــاء المـضـعـف																									
	الحلـــق						اللـهاة		الأحـيـاز الوسـطـيـة												الذلقية			الشفهية		
	أقصاه		وسطه		أدناه		لهاة	حنك	شجر الفم			الأسلـة			نطع الغار			اللثـة			حروف الذلاقة			الشفـتان		
	أ	هـ	ع	ح	غ	خ	ق	ك	ج	ش	ض	ص	س	ز	ط	ت	د	ظ	ث	ذ	ر	ل	ن	ف	ب	م
ظاء	–	–			–		–				–	–	–	–	–	–		–	–	–	–		–			
ثاء		–				–				–	–	–	–	–	–	–		–	–	–						
ذال	–		–								–	–	–	–	–	–	–	–	–	–						–
راء																					–	–	–			
لام																					–	–	–			
نون												–				–			–			–	–			
فاء					*											–			–					–	–	–
باء																		–						–	–	–
ميم			–															–						–	–	–

جدول رقم (4)

تقسيم المضعف الثلاثى الصحيح وفقا لأحياز فاء الفعل مع عينه ولامه الحلقي

حيز الفاء	مخرج الفاء	فـاء المضعف	عين المضعف ولامه الحلقي					
			الهمزة	الهاء	العين	الحاء	الغين	الخاء
الحلق	أقصى الحلق	الهمزة	—	نصر	—	نصر	—	—
		الهاء	—	فتح	نصر	—	—	—
	وسط الحلق	العين	—	—	—	—	—	—
		الحاء	—	—	—	—	—	—
	أدنى الحلق	الغين	—	—	—	—	—	—
		الخاء	—	—	ضرب	—	—	—
الحنك واللهاة	اللهاة	القاف	—	نصر(1)	نصر	نصر	—	—
	الحنك الأعلى	الكاف	—	ضرب	نصر(2)+ ضرب	—	—	ضرب
شجر الفم	منفتح مجهور	الجيم	—	—	نصر	نصر	—	نصر
	منفتح مهموس	الشين	—	—	نصر+ ضرب	نصر+ ضرب+ فتح	نصر	نصر
	مطبق	الضاد	—	نصر(3)	نصر	—	—	نصر
الأسلة	مطبق	الصاد	—	—	—	ضرب	—(4)	ضرب
	منفتح مهموس	السين	—	—	—	نصر+ ضرب	—	ضرب
	منفتح مجهور	الزاي	—	—	—	نصر	—	نصر+ ضرب

(1) القاموس" (قهقه): اشتد ضحكه كقة، أو قهّ: قال في ضحكه قه". ويلحظ أنه حكايةِ صوت. ولم يرد الفعل (قهّ) في تاج العروس، ولا في اللسان.
(2) التاج (ك ع ع)" كع يكِع بالكسر على القياس، حكاه سيبويه، وقال: هو أجود، ويكُع (بالضم) حكاه يونس، وهو قليل ... جبن وضعف".
(3) جاء بالقاموس: " ضهَّه: شاكله وشابهه، لغة في ضاهاه". ولم يرد الفعل (ضهّ) في لسان العرب، ولا في تاج العروس.
(4) ورد الفعل (صغّ) في القاموس المحيط، ولم يرد في لسان العرب. وجاء بالتاج: " صغَّ : أهمله الجوهري".

تابع جدول رقم (4)

تقسيم المضعف الثلاثى الصحيح وفقا لأحياز فاء الفعل مع عينه ولامه الحلقي

حيز الفاء	مخرج الفاء/ صفته	فاء المضعف	عين المضعف ولامه الحلقي					
			الهمزة	الهاء	العين	الحاء	الغين	الخاء
نطع الغار	مطبق مجهور	الطاء	—	—	نصر	نصر	—	نصر
	منفتح مهموس	التاء	—	—	نصر	—	—	نصر
	منفتح مجهور	الدال	—	—	نصر	نصر	—	—
اللثة (بين الأسنان)	مطبق	الظاء	—	—	—	—	—	—
	منفتح مهموس	الثاء	—	—	ضرب*	—	—	—
	منفتح مجهور	الذال	—	—	—	نصر	نصر	—
ذولق اللسان	مكرر	الراء	—	—	—	—	—	نصر
	جانبي	اللام	—	ضرب	—	ضرب	—	نصر
	خيشومي	النون	—	—	—	نصر+ ضرب	—	ضرب
الشفتان	شفهي أسناني	الفاء	—	فتح	—	نصر+ ضرب	نصر	ضرب
	شفهي	الباء	—	نصر	ضرب	فتح	نصر	نصر
	شفهي خيشومي	الميم	—	فتح	—	نصر+ ضرب+ فتح	—	—

* التاج: " تعَّ يتِعّ بالتاء والثاء جميعا: ضرب".

● الأصوات الحلقية عينا ولاما للمضعف

أولا: مع الأصوات الحلقية فاءً:

- تتنافر الأصوات الحلقية بعضها مع بعض: فلا يقع أحدها فاءً والآخر عينا ولاما للفعل الثلاثي المضعف[35].

ثانيا: مع صوتي اللهاة والحنك الأعلى فاءً:

1. لا يقع (القاف) و(الكاف) فاءً للمضعف مع (أ) أقصى الحلقي الانفجاري المجهور المنفتح المستفل.
2. يتصرف (القاف) اللهوي الانفجاري المجهور المنفتح المستعلي- فاءً- مع (الهاء) أقصى الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (نصر)، في الفعل (قهّ)[36].
3. يتصرف (الكاف) الحنكي الانفجاري المهموس المنفتح المستفل- فاءً- مع (الهاء) أقصى الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (ضرب)، في الفعل (كهّ)[37].
4. يتصرف (القاف) اللهوي- فاءً- مع صوتي وسط الحلق (ع- ح) على (نصر)، في الفعلين (قعّ، قحّ).
5. لا يقع (الكاف) الحنكي الانفجاري المهموس المنفتح المستفل فاءً مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل.
6. لا يقع (القاف) اللهوي الانفجاري المجهور المنفتح المستعلي- فاءً- مع أدنى الحلقيين الاحتكاكيين المنفتحين المستعليين (غ- خ).
7. لا يقع (الكاف) الحنكي الانفجاري المهموس المنفتح المستفل فاءً مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي.
8. يتصرف (الكاف) الحنكي الانفجاري المهموس المنفتح المستفل- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (ضرب)، في الفعل (كخّ).

ثالثا: مع الأصوات الشجرية فاءً:

1) لا تقع الأصوات الشجرية (ج-ش- ض[38]) فاء مع صوتي أقصى الحلق (أ-هـ).
2) يتصرف (الجيم) الشجري الانفجاري المجهور المنفتح المستفل- فاءً- مع وسط الحلقيين:(ع- ح) المنفتحين المستفلين على باب (نصر)، في الفعلين (جعّ ، جحّ).
3) يتصرف (الضاد) الشجري الاحتكاكي المجهور المطبق المستعلي- فاءً- مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل على باب (نصر)، في الفعل (ضعّ).
4) يتصرف (الشين) الشجري الاحتكاكي المهموس المنفتح المستفل المتفشي- فاءً- مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل على بابي (نصر) و(ضرب) في الفعل (شعّ)[39].
5) يتصرف (الشين) الشجري الاحتكاكي المهموس المنفتح المستفل المتفشي- فاءً- مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل على الأبواب (نصر، ضرب، فتح)، في الفعل (شح)[40].
6) لا يقع (الضاد) الشجري الاحتكاكي المجهور المطبق المستعلي فاءً مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل.
7) يتصرف الشجريان المجهوران (ج- ض)- فاء- مع (ع) وسط الحلقي المجهور على باب (نصر).
8) لا يقع الشجريان المجهوران (ج- ض) فاءً للمضعف مع (الغين) أدنى الحلقيالمجهور المستعلي.
9) يتصرف (الضاد) الشجري الاحتكاكي المجهور المطبق المستعلي- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (نصر)، في الفعل (ضخّ).

10) يتصرف (الجيم) الشجري الانفجاري المجهور المنفتح المستفل- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (نصر)، في الفعل (جخّ).

11) يتصرف (الشين) الشجري الاحتكاكي المهموس المنفتح المستفل المتفشي- فاءً- مع (غ- خ) أدنى الحلقيين الاحتكاكيين المنفتحين المستعليين على باب (نصر)، في الفعلين (شغّ، شخّ).

رابعا: مع الأصوات الأسلية فاءً:

1- لا تقع الأسليات (ص- س- ز) فاء للمضعف مع (أ- ه) أقصى الحلقيين المنفتحين المستفلين.

2- لا تقع الأسليات (ص- س- ز) فاءً للمضعف مع (العين) وسط الحلقي المجهور المنفتح المستفل.

3- يتصرف (الصاد) الأسلي الاحتكاكي المهموس المطبق المستعلي- فاءً- مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (ضرب)، في الفعل (صحّ).

4- يتصرف (الزاي) الأسلي الاحتكاكي المجهور المنفتح المستفل- فاءً- مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (نصر)، في الفعل (زحّ).

5- يتصرف (الصاد) الأسلي الاحتكاكي المهموس المطبق المستعلي- فاءً- مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي على باب (نصر)، في الفعل (صغّ)[41].

6- يتصرف الأسليان الاحتكاكيان المهموسان (ص، س)- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (ضرب)، في الفعلين (صخّ- سخَّ).

7- لا يقع الأسليان الاحتكاكيان المنفتحان المستفلان (س، ز) فاءً مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي.

خامسا: مع الأصوات النطعية فاءً:

1. لا تقع الأصوات النطعية الانفجارية (ط- ت- د) فاء للمضعف مع (الهمزة والهاء) أقصى الحلقيين المنفتحين المستفلين.

2. يتصرف النطعيان الانفجاريان المجهوران (ط- د)- فاءً- مع صوتي وسط الحلق (ع- ح) على باب (نصر)، في الأفعال:(طعّ، طحّ)، (دعّ، دحّ).

3. يتصرف (التاء) النطعي الانفجاري المهموس المنفتح المستفل- فاءً- مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل على باب (نصر)، في الفعل (تعّ).

4. لا يقع (التاء) النطعي الانفجاري المهموس المنفتح المستفل فاءً مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل.

5. لا تقع الأصوات النطعية الانفجارية (ط- ت- د) فاء للمضعف مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي.

6. يتصرف (الطاء) النطعي الانفجاري المجهور المطبق المستعلي-فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (نصر)، في الفعل (طخّ).

7. لا يقع (التاء) النطعي الانفجاري المهموس المنفتح المستفل فاءً مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي.

8. يتصرف (التاء) النطعي الانفجاري المهموس المنفتح المستفل- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (نصر)، في الفعل (تخّ).

9. لا يقع (الدال) النطعي الانفجاري المجهور المنفتح المستفل فاءً للمضعف مع أدنى الحلقيين الاحتكاكيين المنفتحين المستعليين (غ- خ).

سادسا: مع الأصوات اللثوية فاءً:

1) لا تقع الأصوات اللثوية الاحتكاكية (ظ- ث- ذ) فاءً للمضعف مع (الهمزة والهاء) أقصى الحلقيين المنفتحين المستفلين.

2) لا يقع (الظاء) اللثوي الاحتكاكي المجهور المطبق المستعلي فاءً مع أي من أصوات الحلق:(أ- هـ-ع- ح- غ- خ).

3) لا يقع اللثويان المجهوران (ظ ، ذ) فاءً مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل.

4) يتصرف (الثاء) اللثوي المهموس المنفتح المستفل- فاءً- مع (العين) وسط الحلقي المجهور المنفتح المستفل على باب (ضرب)، في الفعل (ثعّ).

5) لا يقع (الثاء) اللثوي الاحتكاكي المهموس المنفتح المستفل فاءً مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل.

6) يتصرف (الذال) اللثوي الاحتكاكي المجهور المنفتح المستفل- فاءً- مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل- فاءً- على باب (نصر)، في الفعل (ذحّ).

7) لا يقع (الثاء) اللثوي الاحتكاكي المهموس المنفتح المستفل فاءً مع (الغين والخاء) أدنى الحلقيين الاحتكاكيين المنفتحين المستعليين.

8) يتصرف (الذال) اللثوي الاحتكاكي المجهور المنفتح المستفل- فاءً- مع (الغين) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي على باب (نصر)، في الفعل (ذغّ)[42].

9) لا تقع الأصوات اللثوية الاحتكاكية (ظ- ث- ذ) فاء للمضعف مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي.

سابعا: مع الأصوات الذلقية فاءً:

1. لا تقع الذلقيات (ر- ل- ن) فاء مع (الهمزة) أقصى الحلقي الانفجاري المجهور المنفتح المستفل.
2. لا يقع الذلقيان المتوسطان المجهوران المنفتحان المستفلان: (الراء) المكرر، و(النون) الخيشومي فاءً مع (الهمزة والهاء) أقصى الحلقيين المنفتحين المستفلين.
3. يتصرف (اللام) الذلقي المتوسط المجهور المنفتح المستفل الجانبي- فاءً- مع الحلقيين المهموسين الاحتكاكيين المنفتحين المستفلين: (هـ، ح) على باب (ضرب)، في الفعلين: (لهّ)[43]، (لحّ).
4. يتصرف (اللام) الذلقي المتوسط المجهور المنفتح المستفل الجانبي- فاءً- مع (خ) الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (نصر)، في الفعل (لخّ).
5. لا يقع (اللام) الذلقي المتوسط المجهور المنفتح المستفل الجانبي- فاءً- مع (ع) وسط الحلقي المتوسط المجهور المنفتح المستفل، في حين يتصرف (اللام) - فاءً- مع (ح) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (ضرب)، في الفعل (لحّ).
6. لا تقع الأصوات الذلقية المجهورة (ر- ل- ن) فاء للمضعف مع الحلقيات المجهورة(أ- ع- غ).
7. لا يقع (الراء) الذلقي المتوسط المجهور المنفتح المستفل المكرر فاءً مع الحلقيات المستفلة من أقصى الحلق ووسطه: (أ- هـ- ع- ح)، ولا مع (الغين) أدنى الحلقي المستعلي المجهور.
8. يتصرف المتوسطان المجهوران المنفتحان المستفلان غير الخيشوميين: المكرر(ر) والجانبي (ل)- فاءً- مع (خ) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على (نصر)، في الفعلين (رخّ، لخّ).
9. لا يقع (النون) الذلقي المتوسط المجهور المنفتح المستفل الخيشومي فاءً مع الحلقيين المجهورين من وسط الحلق وأدناه (ع- غ).

10. يتصرف (النون) الذلقي المتوسط المجهور المنفتح المستفل الخيشومي- فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (ضرب)، في الفعل (نخّ).

ثامنا: مع الأصوات الشفهية فاءً:

1- تتنافر الأصوات الشفهية الثلاثة (ف- ب- م)- فاء- مع الهمزة أقصى الحلقي المجهور.

2- يتصرف (الفاء) الشفهي الاحتكاكى المهموس المنفتح المستفل- فاءً- مع (الهاء) أقصى الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (فتح)، في الفعل (فهّ).

3- يتصرف (الباء) الشفهي الانفجاري المجهور المنفتح المستفل- فاءً- مع (هـ) أقصى الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (نصر)، في الفعل (بهّ).

4- يتصرف (الميم) المتوسط المجهور المنفتح الخيشومي- فاءً- مع (هـ) أقصى الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (فتح)، في الفعل (مهّ).

5- لا يقع (الفاء) الشفهي الاحتكاكى المهموس المنفتح المستفل فاءً للمضعف مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل.

6- يتصرف (الباء) الشفهي الانفجاري المجهور المنفتح المستفل - فاءً- مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل على باب (ضرب)، في الفعل (بعّ).

7- يتصرف (الباء) الشفهي الانفجاري المجهور المنفتح المستفل- فاءً- مع (الحاء) وسط الحلقي الاحتكاكي المهموس المنفتح المستفل على باب (فتح)، في الفعل (بحّ).

8- لا يقع (الميم) الشفهي المتوسط المجهور المنفتح المستفل الخيشومي فاءً للمضعف مع (العين) وسط الحلقي المتوسط المجهور المنفتح المستفل.

9- يتصرف الشفهيان غير الخيشوميين (الفاء والباء) مع (الغين) أدنى الحلقي المجهور المستعلي على باب (نصر)، في الفعلين: (فغّ، بغّ).

10- يتصرف (الفاء) الاحتكاكى المهموس المنفتح المستفل- فاءً- مع (غ) أدنى الحلقي الاحتكاكي المجهور المنفتح المستعلي على باب (نصر)، في الفعل (فغّ)، في حين يتصرف (الفاء) - فاءً- مع (الخاء) أدنى الحلقي الاحتكاكي المهموس المنفتح المستعلي على باب (ضرب)، في الفعل (فخّ).

11- يتصرف (الباء) الشفهي الانفجاري المجهور المنفتح المستفل - فاءً- مع (غ- خ) أدنىالحلقيين الاحتكاكيين المنفتحين المستعليين على باب (نصر)، في الفعلين (بغّ، بخّ).

12- لا يقع (الميم) الشفهي المتوسط المجهور المنفتح الخيشومي فاءً للمضعف مع الحلقيات المجهورة: (الهمزة، والعين، والغين).

13- لا يقع (الميم) الشفهي المتوسط المجهور المنفتح الخيشومي فاءً للمضعف مع (غ- خ) أدنى الحلقيين الاحتكاكيين المنفتحين المستعليين.

● أحوال تصرف فاء المضعف مع الأصوات الحلقية عينا ولاما للمضعف

أولا: مع الأصوات الحلقية فاءً:

- سبقت الإشارة إلى تنافر الأصوات الحلقية بعضها مع بعض.

ثانيا: مع صوتي اللهاة والحنك الأعلى فاءً:

1. إذا اختلف اللهوي- فاءً- مع أقصى الحلقي في الجهر والاستعلاء تصرفا على (نصر).
2. إذا اتفق صوت اللهاة والحنك- فاءً- مع أقصى الحلقي في الهمس والاستفال تصرفا على (ضرب).
3. يتصرف اللهوي المجهور المستعلي- فاءً- مع صوتي وسط الحلق (ع ، ح) على (نصر).
4. إذا اختلف اللهوي- فاءً- مع وسط الحلقي في الاستعلاء تصرفا على (نصر).
5. إذا اتفق الحنكي- فاءَ- مع أدنى الحلقي في الهمس واختلفا في الاستعلاء تصرفا على (ضرب).

ثالثا: مع الأصوات الشجرية فاءً:

1. تتصرف الأصوات الشجرية الثلاثة- فاء- مع الأصوات الحلقية على باب (نصر) دائما.
2. يتصرف الشجري الانفجاري المجهور المنفتح- فاءً- مع صوتي وسط الحلق: (ع،ح) على (نصر).
3. إذا اتفق الشجري الاحتكاكي- فاءً- مع وسط الحلقي في الجهر واختلفا في الإطباق والاستعلاء تصرفا على (نصر).
4. إذا اختلف الشجري- فاءً- مع أدنى الحلقي في الجهر والاستعلاء، واتفقا في الاحتكاك والانفتاح، تصرفا على (نصر).
5. إذا اتفق الشجري– فاء- مع أدنى الحلقي في الاحتكاك والهمس والانفتاح، واختلفا في الاستعلاء تصرفا على (نصر).
6. إذا اختلف الشجري- فاءً- مع أدنى الحلقي في الجهر والإطباق، واتفقا في الاحتكاك والاستعلاء، تصرفا على (نصر).

رابعا: مع الأصوات الأسلية فاءً:

1. إذا اتفق الأسلي المهموس- فاءً- مع وسط الحلقي وأدناه في الاحتكاك والهمس، واختلفا في الإطباق أو الاستعلاءتصرفا على (ضرب).
2. اذا اتفق (ز) الأسلي- فاءً- مع وسط الحلقي في الاحتكاك والانفتاح والاستفال، واختلفا في الجهر تصرفا على (نصر).
3. إذا اتفق (ص) الأسلي- فاءً- مع (غ) أدنى الحلقي في الاحتكاك والاستعلاء، واختلفا في الجهر والإطباق تصرفا على (نصر).

خامسا: مع الأصوات النطعية فاءً:

1. لا تتصرف الأصوات النطعية الثلاثة- فاءً- مع الأصوات الحلقية إلا على باب (نصر).
2. يتصرف (ط- د) النطعيان الانفجاريان المجهوران- فاءً- مع وسط الحلقيين (ع-ح) على (نصر).
3. إذا اختلف (ت) النطعي- فاءً- مع وسط الحلقي في الانفجار والجهر، واتفقا في الانفتاح والاستفال تصرفا على (نصر).
4. إذا اختلف (ت) النطعي- فاءً- مع أدنى الحلقي في الاحتكاك والاستعلاء، واتفقا في الهمس والانفتاح، تصرفا على (نصر).
5. إذا اختلف (ط) النطعي- فاءً- مع أدنى الحلقي في الجهر والإطباق، واتفقا في الاستعلاء تصرفا على (نصر).

سادسا: مع الأصوات اللثوية فاءً:

1. لا يتصرف (ث) اللثوي المهموس- فاءً- إلا مع وسط الحلقي المجهور، ويتصرفان على (ضرب).
2. يتصرف (ذ) اللثوي المجهور- فاءً- مع وسط الحلقي المهموس، المتفق معه في الاحتكاك والانفتاح والاستفال، على باب (نصر).
3. يتصرف (ذ) اللثوي- فاءً- على باب (نصر) مع أدنى الحلقي المختلف في الاستعلاء، ويتفقان في الاحتكاك والجهر والانفتاح.

سابعا: مع الأصوات الذلقية فاءً:

1. يتصرف (ل) الذلقي الجانبي المجهور- فاءً- مع مهموسي أقصى الحلقي ووسطه (هـ،ح) على(ضرب). (أثر صفة الفاء -اختلاف مخرج العين)
2. يتصرف (ل) الذلقي المجهور- فاء- مع (خ) المهموس المستعلي من أدنى الحلق على باب (نصر).
3. إذا اتفقالذلقي- فاءً- مع أدنى الحلقي في عدم الخيشومية تصرفا على (نصر).
4. إذا اختلف الذلقي- فاءً- مع أدنى الحلقي في الخيشومية تصرفا على (ضرب). (أثر صفة الفاء)

ثامنا: مع الأصوات الشفهية فاءً:

1. إذا اتفق (ف) الشفهي- فاءً- مع أقصى الحلقي في الاحتكاك والهمس والانفتاح والاستفال تصرفا على (فتح).
2. إذا اختلف (ب) الشفهي- فاءً- مع أقصى الحلقي في الاحتكاك والجهر، واتفقا في الانفتاح والاستفال تصرفا على (نصر).
3. إذا اختلف (م) الشفهي- فاءً- مع أقصى الحلقي في الاحتكاك والجهر والخيشومية، واتفقا في الانفتاح والاستفال تصرفا على (فتح).   (أثر صفة الخيشومية)
4. إذا اتفق (ب) الشفهي- فاء- مع وسط الحلقي في الجهر والانفتاح والاستفال تصرفا على (ضرب).
5. إذا اختلف (ب) الشفهي- فاء- مع وسط الحلقي في الجهر، واتفقا في الانفتاح والاستفال تصرفا على (فتح).
6. إذا اختلف (ف) الشفهي- فاءً- مع أدنى الحلقي في الهمس والاستعلاء ، واتفقا في الاحتكاك والانفتاح تصرفا على (نصر).
7. إذا اختلف (ف) الشفهي- فاءً- مع أدنى الحلقي في الاستعلاء، واتفقا في الهمس والاحتكاك والانفتاح تصرفا على (ضرب).
8. يتصرف (ب) الشفهي غير الخيشومي المجهور- فاءً- مع أدنى الحلقيين المستعليين على(نصر).

خاتمة

بهذا يكون البحث قد حقق أهدافه بالإجابة عن التساؤلات الأربعة الواردة في مطلعه:

1)فقد تبين أثر **مخرج**العينالحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه.

2)كما تبين أثر **حيز**العينالحلقي للفعل الثلاثي المضعف في ورود الفعل على باب صرفي بعينه.

3)واتضح أثر اتفاق **صفات العين**الحلقي لهذا الفعل، أو اختلافها، مع صفات فائهفيتصرفه.

4)كما أمكن **تلمس قواعد**تحكم تآلف صوتي الفعل الثلاثي المضعف ،**علىالنحو التالي:**

الاتجاهات العامةلتصرف المضعف الثلاثي، حلقي العين واللام
أثر مخارج صوتي المضعف، وصفاتها على الباب الصرفي

❖ أثر مخرج صوتي المضعف مع صفتهما في تصرف الفعل على الباب الصرفي:

1) يتصرف (ق) اللهوي- فاءً- مع (ع- ح) وسط الحلقيين على باب (نصر).

2) يتصرف (ك) الحنكي المهموس- فاء- مع المهموسين من أقصى الحلقوأدناه(هـ- خ) على (ضرب).

3) يتصرف (ج) الشجري- فاءً- مع (ع- ح) وسط الحلقيين على باب (نصر).

4) يتصرف (ش) الشجري- فاء- مع (غ- خ) أدنى الحلقيين على باب (نصر).

5) يتصرف (ض)- فاء- مع (ع) وسط الحلقي المجهور على باب (نصر).

6) يتصرف الأسليان المهموسان (ص- س)- فاءً- مع (خ) أدنى الحلقي المهموس على (ضرب).

7) يتصرف (ث) اللثوي المهموس- فاء- مع (ع) على باب (ضرب).

8) يتصرف (ذ) اللثوي المجهور- فاء- مع (ح) على باب (نصر).

9) لا يقع (ر) الذلقي المجهور فاءً إلا مع (خ) أدنى الحلقي المهموس المستعلي، ويتصرفان على (نصر).

❖ أثر حيز فاء المضعف مع مخرج عينه ولامهعلى الباب الصرفي:

1. تتنافر أصوات الأحياز التالية- فاءً- مع صوتي أقصى الحلق (أ- هـ):

أ- الأصوات الشجرية (ج- ش- ض)(44).

ب- الأصوات الأسلية (ص- س- ز).

ت- الأصوات النطعية (ط- ت- د).

ث- الأصوات اللثوية (ظ- ث- ذ).

2. تتصرف الشجريات (ج- ش- ض)- فاءً- مع (خ) أدنى الحلقي المهموس على باب (نصر).

3. تتصرف النطعيات الانفجارية (ط- ت- د)- فاءً- مع (ع) وسط الحلقي المتوسط على (نصر).

❖ أثر صفة الجهر أوالهمس في صوتي المضعف على بابه الصرفي:

1. يتنافر الشجريان المجهوران (ج- ض)- فاء- مع (غ) أدنى الحلقي المجهور، في حين يتصرف (ش) الشجري المهموس- فاء- مع (غ) أدنى الحلقي المجهور على باب (نصر).

2. لا يقع (ض) الشجري المجهور المطبق فاء للمضعف مع (ح) وسط الحلقي المهموس، في حين يتصرف (ض)- فاء- مع (ع) وسط الحلقي المجهور على باب (نصر).

3. يتصرف (ص) الأسلي المطبق المهموس– فاءً- مع الحلقيين المهموسين (ح- خ) على (ضرب).

4. يتصرف الأسليان المهموسان (ص- س)- فاءً- مع (خ) أدنى الحلقي المهموس على (ضرب)، في حين يتنافر الصوتان– فاء- مع (غ) أدنى الحلقي المجهور.

5. يتصرف النطعيان المجهوران (ط- د)- فاءً- مع (ح) وسط الحلقي المهموس على (نصر)، في حين لا يقع (ت) النطعي المهموس- فاءً- مع (الحاء) المهموس.

6. يتصرف (ت) النطعي المهموس المنفتح - فاء- مع (خ) أدنى الحلقي المهموس على (نصر)، في حين لا يقع (د) النطعي المجهور المنفتح فاء للمضعف مع (الخاء) المهموس.
7. لا يقع (ذ) اللثوي المجهور– فاء- مع (ع) وسط الحلقي المجهور، في حين يتصرف الذال مع (ح) وسط الحلقي المهموس على باب (نصر).
8. لا يقع (ن) الذلقي المجهور الخيشومي فاءً مع (غ) أدنى الحلقي المجهور، في حين يتصرف (ن) مع (خ) أدنى الحلقي المهموس على (ضرب).
9. يتصرف (ل) الذلقي المجهور- فاء- مع مهموسي أقصى الحلق ووسطه (ه- ح) على (ضرب).
10. يتصرف (ب) الشفهي المجهور- فاءً- مع (ع) وسط الحلقي المجهور على باب (ضرب)، في حين يتصرف (ب) – فاء- مع (ح) وسط الحلقي المهموس على باب (فتح).

❖ أثر صفة الإطباق أو الاستعلاء في صوتي المضعف على بابه الصرفي:

1) يتنافر (ق) اللهوي المستعلي- فاء- مع أدنى الحلقيين المستعليين (غ- خ)، في حين يتصرف (ق)– فاءً- مع (ع- ح) وسط الحلقيين المستفلين على (نصر).
2) لا يقع (ق) اللهوي المستعلي- فاء- مع (خ) أدنى الحلقي المستعلي، في حين يتصرف (ك) الحنكي المستفل- فاءً- مع (خ) المستعلي على باب (ضرب).
3) يتصرف (ذ) اللثوي المجهور المستفل- فاء- مع (ح) وسط الحلقي المهموس المستفل على (نصر)، ولا يقع (ذ) فاء للمضعف مع (خ) أدنى الحلقي المهموس المستعلي.
4) يتصرف (ل) الذلقي- فاء- مع (ه- ح) الحلقيين المهموسين المستفلين على (ضرب)، في حين يتصرف اللام- فاء- مع (خ) الحلقي المهموس المستعلي على باب (نصر).
5) يتصرف (ب) الشفهي المستفل- فاءً- مع (غ- خ) أدنى الحلقيين المستعليين على باب (نصر).

❖ أثر صفة الخيشومية في أحد صوتي المضعف على بابه الصرفي:

1. يتصرف (ن) الذلقي الخيشومي المجهور- فاء- مع (خ) أدنى الحلقي المهموس على (ضرب).
2. يتصرف (ب) الشفهي المجهور- فاء- مع صوتي أدنى الحلق (غ- خ) على باب (نصر)، في حين لا يقع (م) الشفهي الخيشومي المجهور فاء مع هذين الصوتين.

* * * * * * * *

كلمة أخيرة عن أهمية هذه الدراسة

كانت هذه الدراسة بحثا في البنية الصوتية للفعل الثلاثي المضعف، وهي دراسة تنتمي إلى مجال الصوتيات المعجمية Lexical Phonology ، وهو مبحث لساني موضوعه البحث في الأسس والقواعد الصوتية (الفونولوجية) التي تحكم تكوّن الوحدات المعجمية. وتتوزع موضوعاته بين الصوت (الفونولوجيا)، وبنية الكلمة (الصرف)، والوحدة المعجمية متحققةً (المعجم). وقد ربطت الدراسة بين تعالق مجموعة أصوات الحلق عند تحققها في الصيغة الصرفية لبنية الفعل الثلاثي المضعف.

والدراسة، إذ تنتهج نهجا تجريبيا في مقاربتها وإجراءاتها، إنما ترنو إلى تحقيق تلك النتائج النظرية في مجال الصوتيات المعجمية.

ويمكن الاستفادة من نتائج الدراسة في كل من اللسانيات التطبيقية واللسانيات الحاسوبية كما يلي:

أولا: في اللسانيات التطبيقية Applied linguistics :

❖ **في الصناعة المعجمية**Lexicography :

فالدراسة تبحث قواعد تكوُّن الكلمة العربية في أحد شرائحها، بما يمكّن من تعميم المقولات والإجراءات والنتائج على المعجم العربي كله: فيوضح القواعد التي ينتهجها المعجم العربي في تأليف أصوات وحداته، والقواعد التي تحكم تحقق الأصوات في صيغة صرفية دون غيرها، وتعليل ذلك صوتيا أو دلاليا.

❖ **في تعليم اللغة**Language learning:

في مجال تحليل الأخطاء، بضبط عين الفعل (أو تحديد الباب الصرفي)؛ فالخلط بين أبواب المضارع عند النطق بالفعل العربي من أكثر الأخطاء المرصودة في تعلم العربية.

❖ **في المصطلحية**Terminology :

فتعميم نتائج مثل هذه الدراسة على المعجم العربي يوضح قواعد التآلف والتنافر في تكوين الكلمة العربية، وفي ذلك كبير فائدة لمن يريد وضع قواعد لسك المصطلحات الجديدة، وتعريب المصطلحات الأجنبية، فتكون نتائج الدراسات الصوتية المعجمية هادية له في ذلك.

ثانيا: في العمل المعجمي الحاسوبي Computational lexicography:

❖ **في بناء قاعدة بيانات معجمية**Lexical database :

فبناء قاعدة بيانات معجمية يتطلب استقصاءً للكلمات والأوزان الممكنة وتلك الممتنعة، وإحصاء ذلك آليا؛ لتتميم وصف المعجم، والتوصل إلى القواعد الصوتية، والصوتية الصرفية، التي تحكم هذا المعجم.

❖ **في تعرّف الكلام** Speech recognition :

وهو مبحث مهم في العمل اللساني الحاسوبي؛ فهو يتعلق بالتعرف الآلي للكلام المنطوق وتمييزه تمهيدا لفهمه، بما يرفع نسبة الدقة في التعرف والفهم الآليين. ويكون ذلك ببناء نماذج تشمل قواعد التتابعات الممكنة صوتيا، والتتابعات غير الممكنة، مما يسهل عملية الإدراك الآلي للأصوات.

هوامش البحث

(1) وفاء كامل فايد: تراكب الأصوات في الفعل الثلاثي الصحيح -عالم الكتب- القاهرة (1991).

(2) عولجت هذه الدراسة من خلال البحوث التالية:

- " الأفعال المضعفة وأبوابها الصرفية ". المجلة العربية للعلوم الإنسانية ـ جامعة الكويت – العدد 74 ، السنة 19 - ربيع (2001).
- " الباب الصرفي للفعل المضعف وأحياز أصواته : دراسة في الأحياز الوسطية والذلقية ". بحث في الكتاب التذكاري (ثمرات الامتنان) – مكتبة الخانجي – ط 1 - القاهرة (2002).
- (الباب الصرفي وصفات الأصوات : دراسة في الفعل الثلاثي المضعف) – عالم الكتب – القاهرة 2001.
- (أثر تجاور صوتي الفعل الثلاثي المضعف فى بابه الصرفي: دراسة في حيزي الحلق والشفتين)- مؤتمر مجمع اللغة العربية بالقاهرة(2009).
- (أثر تجاور صوتي الفعل الثلاثي المضعف فى بابه الصرفي: دراسة في الأحياز الوسطية)- مؤتمر مجمع اللغة العربية بالقاهرة(2010).
- (أثر تجاور صوتي الفعل الثلاثي المضعف فى بابه الصرفي: دراسة في حيز الشفتين)- مؤتمر مجمع اللغة العربية بالقاهرة(2011).

(3) (القواعد الحاكمة لتنافر صوتي الفعل الثلاثي المضعف)- مؤتمر مجمع اللغة العربية بالقاهرة عام (2013) .

(4) (أثر الفاء الحلقي للفعل الثلاثي المضعف على الباب الصرفي لمضارعه)- مؤتمر مجمع اللغة العربية بالقاهرة عام (2014) .

(5) استقصت الباحثة (وفاء كامل) الأفعال الثلاثية الصحيحة الواردة في القاموس المحيط في عدد من البحوث: أحدها يعالج أحوال تآلف الأصوات وتنافرها في الفعل الثلاثي الصحيح، فى كتاب: (تراكب الأصوات في الفعل الثلاثي الصحيح: دراسة استقصائية في القاموس المحيط) - عالم الكتب- القاهرة (1991)، والثاني يرصد أحوال تصرف الفعل الثلاثي الصحيح على الأبواب الصرفية، وعنوانه: (مدى ارتباط الفعل الثلاثي الصحيح بالمضارع المفتوح العين – دراسة إحصائية على القاموس المحيط)، العدد 58: مجلة كلية الآداب- جامعة القاهرة : مارس (1993)، والثالث يرصد أثر صفات الأصوات على الباب الصرفي للفعل المضعف في كتاب (الباب الصرفي وصفات الأصوات: دراسة في الفعل الثلاثي المضعف)، ط1- عالم الكتب، (2001) ، والرابع بعنوان "الأفعال المضعفة وأبوابها الصرفية ": المجلة العربية للعلوم الإنسانية- جامعة الكويت- ع 74- س19، عام (2001)، والخامس بعنوان (الباب الصرفي للفعل المضعف وأحياز أصواته: دراسة في الأحياز الوسطية والذلقية)، ضمن بحوث الكتاب التذكاري (ثمرات الامتنان)- مكتبة الخانجي، ط1 – القاهرة (2002)، والسادس بعنوان: (قواعد تنافر صوتي الفعل الثلاثي المضعف)، مؤتمر مجمع اللغة العربية- ، د. 79- مارس (2013)، والسابع عنوانه (أثر أصوات الفعل الثلاثي المضعف في بنيته الصرفية: دراسة في الأصوات الشفهية)- نشر في (العربية): مجلة رابطة أساتذة اللغة العربية- المجلد 46- مطبعة جامعة جورج تاون- (2013)، والثامن بعنوان (أثر الفاء الحلقي للفعل الثلاثي المضعف على الباب الصرفي لمضارعه) مؤتمر مجمع اللغة العربية(2014).

(6) نبّه الفيروزابادي في مقدمة القاموس على أن الفعل إذا ورد بصيغة الماضي دون المضارع، أو ورد مصدره فقط، يكون الباب الصرفي فيه هو (نصر)، وإذا ذكر الماضي وبعده المضارع، دون تقييد بضبط ولا وزن، كان الفعل على باب (ضرب)، ولكن الباحثة آثرت التدقيق في الباب الصرفي للمضارع بالرجوع إلى لسان العرب؛ كي تنهض الدراسة على أساس سليم.

(7) عرفه ابن يعيش بقوله: " هو المقطع الذى ينتهى الصوت عنده ". شرح المفصل: 124/10.

(8) استخدم الخليل هذا المصطلح بكثرة في كتاب العين، ص 65/64 .

(9) عرّفه سيبويه بأنه " حرف أُشبع الاعتماد في موضعه، ومنع النفس أن يجري معه حتى ينقضي الاعتماد [عليه] ويجري الصوت": الكتاب 4 / 434. وعرفه ابن جني بأنه:" حرف أضعف الاعتماد في موضعه حتى جرى معه النفس": سر صناعة الإعراب: 60/1. وانظر: محمود السعران : علم اللغة – دار الفكر العربي – القاهرة (1992)، ص 88.

(10) عرّفه ابن جنى بأنه:" حرف أضعف الاعتماد فى موضعه حتى جرى معه النفس": سر الصناعة : 1/ 60. علم اللغة 88. وفضّل سعد مصلوح تسميته بالصوت غير المجهور، وهي تسمية أكثر دقة : دراسة السمع والكلام – عالم الكتب – الطبعة الأولى – القاهرة (2000)، ص 152.

(11)عرّفه سيبويه بأنه: " الذى يمنع الصوت أن يجرى فيه " : الكتاب 434/4، دراسة السمع والكلام: 175، علم اللغة : 153.

(12) دراسة السمع والكلام : 184.

(13) يعبر عنها سيبويه بأنها " بين الرخوة والشديدة" ، وذكر ابن جني : " والحروف التي بين الشديدة والرخوة ثمانية..، وهي الألف والعين والياء واللام والنون والراء والميم والواو " : سر صناعة الإعراب 61/1.

(14) الكتاب: 4 / 435.

(15) شرح المفصل : 128/10 : " والإطباق أن تطبق على مخرج الحرف من اللسان ما حاذاه من الحنك ". أو هو " ارتفاع مؤخر اللسان إلى أعلى قليلا فى اتجاه الطبق اللين، وتحركه إلى الخلف قليلا فى اتجاه الحائط الخلفى للحلق". وانظر : عمر (أحمد مختار) : دراسة الصوت اللغوى - عالم الكتب - القاهرة (1991): 326.

(16)علل ابن دريد تسمية الحروف المطبقة بقوله : " لأنك إذا نطقت بها أطبقت عليها حتى تمنع النفس أن يجرى معها ": مقدمة الجمهرة ص 8.

(17) استخدم سيبويه مصطلح الانفتاح: الكتاب 436/4. وهناك من يميل إلى التعبير عنه بالترقيق، في مقابل التفخيم.

(18) حدد ابن جني معنى الاستعلاء بقوله : " أن تتصعد في الحنك الأعلى ". سر صناعة الإعراب 62/1.

(19) استخدم ابن جنى مصطلح (الانخفاض) الذى يعبر عن المعنى نفسه. المرجع السابق: 62/1.

(20) الكتاب 457/4، المقتضب 192/1- 3، سر صناعة الإعراب: 47/1، شرح المفصل : 125/10.

(21) الكتاب 457/4 : " وحرفان يخالطان طرف اللسان : الضاد والشين؛ لأن الضاد استطالت لرخاوتها حتى اتصلت بمخرج اللام، والشين كذلك حتى اتصلت بمخرج الطاء "، وأيضا في شرح المفصل : 141/10.

(22) الكتاب: 448/4 " الشين تتفشى في الفم حتى تتصل بمخرج اللام "، وكذلك في شرح المفصل: 10/ 125. وفي شرح المفصل: 140/10: " الشين أشد استطالة من الضاد، وفيها تفشّ ليس في الضاد "، وانظر: أحمد مختار عمر: دراسة الصوت اللغوي - عالم الكتب، القاهرة 1991– ص 317.

(23) ذكر سيبويه أن الصوت يجري فيه لتكريره وانحرافه إلى اللام : الكتاب 435/4. وذكر ابن جني أن الراء " إذا وقفتعليه رأيت طرف اللسان يتعثر بما فيه من التكرير " : سر صناعة الإعراب 63/1.وأيضا في دراسة السمع والكلام : 181-2، برتيل مالمبرج : الصوتيات – ترجمة محمد حلمي هليّل – عين للدراسات والبحوث – القاهرة (1994)، ص 94، علم اللغة ص 171، كمال بشر: علم اللغة العام (الأصوات)-دار المعارف – القاهرة (1970)، ص 166.

(24) علم اللغة، ص 168، الصوتيات ص 91، علم اللغة العام، ص 167.

(25) علم اللغة، ص 169، علم اللغة العام (الأصوات): ص 166 – 67، دراسة السمع والكلام : 180-81، الصوتيات92 – 93.

(26) اختلف سيبويه في ترتيب الصوامت عن الخليل ، وكان ترتيب الحروف عند سيبويه كما يلى :

الهمزة والألف والهاء والعين والحاء والغين والخاء، والقاف والكاف، والجيم والشين والياء، والضاد، واللام والنون والراء، والطاء والدال والتاء، والزاى والسين والصاد، والظاء والذال والثاء، والفاء والباء والميم والواو: الكتاب 433/4. و سقط مخرج اللام من طبعة الكتاب، تحقيق (هارون). وقد اتفق ابن جني مع سيبويه في ترتيبه، واعترض على ترتيب الخليل: سر الصناعة 45/1.

(27)استخدمت الباحثة الرموز الصوتية التي اعتمدتها الرابطة الدولية للصوتيات International Phonetic Association.

(28)الكتاب: 433/4، وفي شرح المفصل: 124/10: " فمن ذلك الحلق وفيه ثلاثة مخارج، فأقصاها من أسفله إلى ما يلي الصدر مخرج الهمزة، ولذلك ثقل إخراجها لتباعدها، ثم الهاء ".

(29) الكتاب: 433/4: " من أقصى اللسان وما فوقه من الحنك الأعلى مخرج القاف. ومن أسفل من موضع القاف من اللسان قليلا، ومما يليه من الحنك (الأعلى) مخرج الكاف". والمعنى نفسه في المقتضب: 328/1 ، وسر صناعة الإعراب: 47/1، و شرح المفصل: 124/10، وهمع الهوامع : 227/2.

(30)العين: 64، الكتاب 433/4: " ومن وسط اللسان بينه وبين وسط الحنك الأعلى مخرج الجيم والشين والياء ". واتفق معه ابن جني، في سر صناعة الإعراب: 46/1. وفي المقتضب قدم مخرج الشين على مخرج الجيم: 328/1، وذكر" أن أقرب الحروف من الياء الجيم" 329/1. وفي شرح المفصل: 124/10 " الجيم والشين والياء ولها حيز واحد، وهو وسط اللسان بينه وبين وسط الحنك، وهي شجرية، والشجر: مفرج الفم، لأن مبدأها من شجر الفم .. والضاد من حيز الجيم والشين والياء".

وذكر اللسان أن ربيعة واليمن يجعلون الشين ضادا غير خالصة : مادة (م ض ط).

(31) العين: 1/ 64، وتسمى أصوات الصفير. والمقتضب 329/1، وفي شرح المفصل: 125/10: " الصاد والسين والزاي من حيز واحد، وهو ما بين الثنايا وطرف اللسان، وهى أسلية لأن مبدأها من أسلة اللسان، وهو مستدق طرف اللسان، وهي حروف الصفير".

(32) العين: 1/ 64، شرح المفصل: 125/10: " والطاء والدال والتاء من حيز واحد، هو ما بين طرف اللسان وأصول الثنايا، وهي نطعية لأن مبدأها من نطع الغار الأعلى، وهو وسطه، يظهر فيه كالتحزيز ".

(33) العين: 1/ 65، وتسمى أيضا أصوات ما بين الأسنان. شرح المفصل: 125/10: " والظاء والذال والثاء من حيز واحد، هو ما بين طرف اللسان وأصول الثنايا، وهى لثوية لأن مبدأها من اللثة ".

(34)شاع بين القدماء إطلاق اسم حروف الذلاقة على ستة أصوات هي اللام والراء والنون والفاء والباء والميم: سر صناعة الإعراب: 64، وشرح الشافية: 257/3-58. ونسب ابن يعيش إلى سيبويه إطلاق (حروف الذلاقة) على هذه الأصوات التي تجمعها عبارة (مر بنفل). ولم تعثر الباحثة في (الكتاب) على ما يشير إلى إطلاق هذه التسمية على تلك الأصوات. ويمكن أن يكون مرجع ذلك إلى أن الخليل حين تحدث في مقدمة العين عن الحروف الذلقية والشفوية، حددها وذكر سبب التسمية- وهو أن الذلاقة في المنطق إنما هي بطرف أسلة اللسان والشفتين، وهما مخرجا هذه الأحرف الستة- ولكنه عاد فقسم تلك الأصوات إلى أصوات ذليقة هي: (ر- ل- ن)، وأصوات شفوية هي: (ب - ف - م): مقدمة العين ص 57.

وقد تابع البحث هنا تقسيم الخليل للأصوات، كما ورد في مقدمة العين ص 65، مع إضافة الهمزة إليه من تقسيم سيبويه.

(35) لم يستثن من ذلك سوى الأفعال: (أهّ) و (أحّ)، و (هه) و(هعّ)، و(خعّ)، وبيانها كما يأتي:

- الفعل (أهّ) حكاية صوت؛ فقد نص القاموس المحيط ولسان العرب (أ هـ هـ) على أن: " الأهة: التحزن، وقد أهّ أهًّا وأهَّة".
- كما أورد القاموس الفعل (أحّ) بمعنى: سعل. ونص اللسان على أنه حكاية صوت، مادة (أح ح):" (أحّ) حكاية تنحنح أو توجع".
- والفعل (هه) يمكن أن يكون حكاية صوت أيضا؛ فقد ورد بالقاموس:"(هه) يهَه بالفتح ههّا وهَهَّة: لثغ واحتبس لسانه".
- وقد نص القاموس على أن الفعل (هعّ) لغة في (هاع).

- وجاء بالقاموس (خ ع ع): " خعّ الفهدُ يخِعّ: صات من حلقه إذا انبهر في عدْوه". وجاء باللسان (خ ع ع): "روي عن عمرو بن بحر أنه قال: خع الفهد يخع، قال: وهو صوت تسمعه من حلقه إذا انبهر في عدْوه. قال أبو منصور: كأنه حكاية صوته إذا انبهر، ولا أدري أهو من توليد الفهادين أو مما عرفته العرب فتكلموا به، وأنا بريء من عُهدته".

(36)القاموس: " قهقه: رجّع في ضحكه أو اشتد ضحكه، كقه فيهما، أو قهَّ: قال في ضحكه قه، فإذا كرره قيل قهقه". وفي اللسان: (ق هـ ق هـ): الجوهري: القهقهة في الضحك معروفة ... يقال: قهَّ وقهقه بمعنى، وإذا خفف قيل: قهَّ الضاحك... وإنما خفف في الحكاية. ولم يرد الفعل في تاج العروس. ومن الواضح أنه حكاية صوت.

(37)القاموس: " كه يكِه كهوها: هرِم ". وفي اللسان (ك هـ ك هـ): " كهت الناقة تكِهّ كهوها إذا هرمت ...أبو عمرو: يقال: كهَّ في وجهي أي تنفس .. والكهكهة حكاية صوت الزمر.. والكهكهة في الضحك أيضا ". ولم يرد الفعل في التاج.

(38)ورد الفعل (ضهّ) بالقاموس: " ضهَّه: شاكله وشابهه، لغة في ضاهاه". ولم يرد الفعل (ضه) في لسان العرب، ولا في تاج العروس.

(39) أشار لسان العرب (ش ع ع) إلى أن الفعل المتعدي يأتي على باب (نصر) في حين يرد اللازم على باب (ضرب)، ومعناهما واحد.

(40) ومعنى الفعل (شحّ) واحد على الأبواب الثلاثة، انظر اللسان (ش ح ح).

(41) ورد الفعل بصيغة الماضي في القاموس، ولم يرد في لسان العرب. وجاء بالتاج: " صغَّ : أهمله الجوهري وقال ابن الأعرابي: أي أكل أكلا كثيرا".

(42)ورد الفعل في القاموس. وذكر تاج العروس: " ذغ جاريته: أهمله الجوهري وصاحب اللسان، وقال أبو عمرو الشيباني: أي جامعها".

(43) ورد الفعل (لهّ) في القاموس المحيط بصيغة الماضي فحسب: " لهّ الشَّعْر: رققه وحسّنه"، ولم يرد في لسان العرب، وذكر التاج: (ل هـ هـ): " لهّ الشعْرَ والكلامَ يلِهّه لهّا: رققه وحسّنه، وهو مجاز".

(44) باستثناء الفعل (ضه) ، وقد سبقت الإشارة إلى أنه لغة في ضاهَى.

المراجع العربية والأجنبية

أحمد مختار عمـر:

- دراسة الصوت اللغوي- عالم الكتب – القاهرة (1991).

الاستراباذي:

- شرح شافية ابن الحاجب (تحقيق الزفزاف)- دار الكتب العلمية – بيروت (1982).

ابن جني:

- الخصائص – تحقيق النجار – ط 2 – دارالهدى – بيروت (ب ت).
- سر صناعة الإعراب- تحقيق هنداوي – ط2 – دار القلم – دمشق (1993).

الخليل بن أحمد الفراهيدي:

- كتاب العين– تحقيق عبد الله درويش – بغداد (1967).

ابن درستويه:

- تصحيح الفصيح وشرحه– مجلس الشئون الإسلامية- القاهرة (1419هـ).

ابن دريد :

- جمهرة اللغة – دار صادر – بيروت.

الزبيدي (محمد مرتضى الحسيني):

- تاج العروس من جواهر القاموس- ج 27- تحقيق مصطفى حجازي- سلسلة التراث العربي- دولة الكويت (1413= 1993).
- تاج العروس من جواهر القاموس - ج 36- تحقيق عبد الكريم العزباوي- سلسلة التراث العربي- دولة الكويت (1422هـ = 2001م).

السرقسطي:

- كتاب الأفعال – الهيئة العامة لشئون المطابع الأميرية– القاهرة (1992).

سيبويه:

- الكتاب ج4 – ط2– مكتبة الخانجي– القاهرة (1982).

ابن سينا :

- أسباب حدوث الحروف– نسخ وتصحيح محب الدين الخطيب – المطبعة السلفية القاهرة (1352هـ).

السيوطي:

- المزهر فى علوم اللغة وأنواعها- المكتبة العصرية- بيروت (1986).
- همع الهوامع في شرح جمع الجوامع- تحقيق أحمد شمس الدين– ط1 – دار الكتب العلمية – بيروت (1998).

علي حلمى موسى:

- دراسة إحصائية لجذور معجم الصحاح- مطبوعات جامعة الكويت- (1973).

الفيروزابادي:

- القاموس المحيط – دار الكتاب العربي – بدون تاريخ أو مكان الطبع.

ابن القطاع:

- أبنية الأسماء والأفعال والمصادر– مطبعة دار الكتب- القاهرة (1999).

ابن القوطية:

- كتاب الأفعال- الطبعة الثانية- مكتبة الخانجي- القاهرة (1993).

المبرد :

- المقتضب – المجلس الأعلى للشئون الإسلامية – القاهرة (1399).

ابن منظور:

- لسان العرب – دار المعارف – القاهرة (1981).

وفاء كامل فايد:

- تراكب الأصوات في الفعل الثلاثي الصحيح –عالم الكتب- القاهرة (1991).
- مدى ارتباط الفعل الثلاثي الصحيح بالمضارع المفتوح العين – دراسة إحصائية على القاموس المحيط)، العدد 58: مجلة كلية الآداب- جامعة القاهرة: مارس (1993).
- الباب الصرفي وصفات الأصوات- عالم الكتب- القاهرة (2001).
- الأفعال المضعفة وأبوابها الصرفية ": المجلة العربية للعلوم الإنسانية- جامعة الكويت العدد 74- س19، عام (2001).
- أثر تجاور صوتي الفعل الثلاثي المضعف فى بابه الصرفي: دراسة في حيزي الحلق والشفتين: مؤتمر مجمع اللغة العربية بالقاهرة، عام (2009).

ابن يعيش:

- شرح المفصل- عالم الكتب- بيروت، ب ت.

Handbook of the International Phonetic Association, Cambridge University Press, 13th ed. 2012.

ملخص بالتاريخ العلمي

الاسم واللقب : الدكتورة / وفاء محمد كامل أمين فايد

1. أستاذة متفرغة بقسم اللغة العربية– كلية الآداب – جامعة القاهرة.
2. أول سيدة تنتخب عضوة في مجمع اللغة العربية بالقاهرة،في 2014 (بعد ثمانين عاما من إنشائه).
3. عضو مراسل بمجمع اللغة العربية بدمشق من 2002.
4. حصلت على جائزة جامعة القاهرة التشجيعية للعلوم الإنسانية والاجتماعية عام 2004.
5. عضو مجلس إدارة الجمعية الدولية للمترجمين العرب من 2006.
6. خبيرة بمجمع اللغة العربية بالقاهرة من عام 2007 – 2014.
7. حصلت على جائزة جامعة القاهرة التقديرية للعلوم الإنسانية والتربوية عام 2013.
8. شاركت في فحص البرنامج المقدم لاعتماد "درجة الماجستير الدولية في اللسانيات الحاسوبية" ضمن برامج الدراسات العليا بكلية الآداب– جامعة الاسكندرية، 2014.

9. في إطار علم اللغة الاجتماعي تتبعت– على مدى ثلاثين عاما– ظاهرة (التغريب) في مصر. كما درست مدى انتشار هذه الظاهرة في ست من دول المشرق العربي، وقارنتها بمثيلتها في مصر.

الكتب المؤلفة والمترجمة:

1– (شرح عيون الإعراب للفزاري من إملاء علي بن فضال المجاشعي) تحقيق ودراسة، 1986.

2– (تراكب الأصوات في الفعل الثلاثي الصحيح– دراسة استقصائية في القاموس المحيط)، عالم الكتب– القاهرة 1991.

3– (اتجاهات البحث اللساني) كتاب مترجم بالاشتراك: المشروع القومي للترجمة– المجلس الأعلى للثقافة، القاهرة ، ط1: 1996، ط2 : 2000.

4– (قصيدة الرثاء بين شعراء الاتجاه المحافظ ومدرسة الديوان: دراسة أسلوبية إحصائية)– الهيئة المصرية العامة للكتاب – القاهرة 2000.

5– (الباب الصرفي وصفات الأصوات: دراسة في الفعل الثلاثي المضعف)– عالم الكتب، القاهرة 2001.

6– (بحوث في العربية المعاصرة)– عالم الكتب– القاهرة 2003.

7– (المجامع العربية وقضايا اللغة)– عالم الكتب– القاهرة 2004.

8– (معجم التعابير الاصطلاحية في العربية المعاصرة)– أبو الهول للنشر– القاهرة 2007.

9– (مدخل إلى اللغة): فيكتوريا فرومكين– ترجمة وفاء كامل– المركز القومي للترجمة – (تحت الطبع).

Arabic Speech Recognition: Challenges and Sate of the Art

Sherif Abdou

Faculty of Computers and Information Technology, Cairo University, Giza, Egypt
5 Ahmed Zewail, Orman, Gizah, Egypt

s.abdou@fci-cu.edu.eg

***Abstract*—The Arabic language has many features such as the phonology and the syntax that makes it an easy language for developing automatic speech recognition systems. Many standard techniques for acoustic and language modelling such as context dependent acoustic models and n-gram language models can be easily applied to Arabic. Some aspects of the Arabic language such as the nearly one-to-one letter-to-phone correspondence make the construction of the pronunciation lexicon even easier than in other languages. The most difficult challenges in developing speech recognition systems for Arabic are the dominance of non-diacritized text material, the several dialects, and the morphological complexity. In this article we review the efforts that have been done to handle the challenges of the Arabic language for developing Automatic Speech Recognition Systems. This includes methods for automatic generation for the diacritics of the Arabic text and word pronunciation disambiguation. Also review the used approaches for handling the limited speech and text resources of the different Arabic dialects. Finally we review the used approaches to deal with the high degree of affixation, derivation that contributes to the explosion of different word forms in Arabic. Also we will introduce the state of the art performance of Arabic Speech Recognition systems and the expectations for near future applications.**

1 Introduction

Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. The last decade has witnessed substantial advances in speech recognition technology, which when combined with the increase in computational power and storage capacity, has resulted in a variety of commercial products already on the market.

The goal of the ASR system is to find the most probable sequence of *words* $W = (w_1, w_2, \ldots.)$ belonging to a fixed vocabulary given some set of acoustic observations $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots. , \boldsymbol{x}_T)$. Following the Bayesian approach applied to ASR [1] the best estimation for the word sequence can be given by:

$$w = \arg\max_{W} P(W/O) = \arg\max_{W} \frac{p(O/W)p(W)}{p(O)} \qquad (1)$$

To generate an output the speech recognizer has basically to perform the following operations as shown in figure (1):

- Extract acoustic observations (*features*) out of the spoken utterance.
- Estimate *P(W)* - the probability of individual word sequence to happen, regardless acoustic observations. This is named the language model.
- Estimate *P(X/W)* - the likelihood that the particular set of features originates from a certain sequence of words, including both the acoustic model and the pronunciation lexicon. The latter is perhaps the only language-dependent component of an ASR system
- Find word sequence that delivers the maximum of (1). This is referred to as the search or decoder.

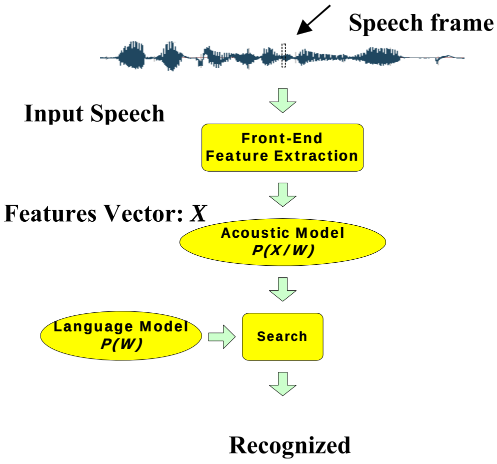


Figure 1: The ASR system main architecture.

The two terms *P(W)* and *P(X/W)*and the maximization operation constitute the basic ingredients of a speech recognition system. The goal is to determine the best word sequence given a speech input X. Actually X is not the speech input but a set of features derived from the speech. The Mel Frequency Cepstrum Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the most widely used. The acoustic and language models and the search operation will be discussed below.

A. Pronunciation Lexicon

The pronunciation lexicon is basically a list where each word in the vocabulary is mapped into a sequence (or multiple sequences) of phonemes. This allows modeling a large number of words using a fixed number of phonemes. Sometimes whole word models are used. In this case the pronunciation lexicon will be a trivial one. The pronunciation lexicon is language-dependent and for a large vocabulary (several thousand words) might require a large effort. We will discuss this for Arabic in the next section.

B. Acoustic Model

The most popular acoustic models are the so called hidden Markov models (HMM). Each phoneme (unit in general) is modeled using an HMM. An HMM [1] consists of a set of states, transitions, and output distributions as shown in figure (2).

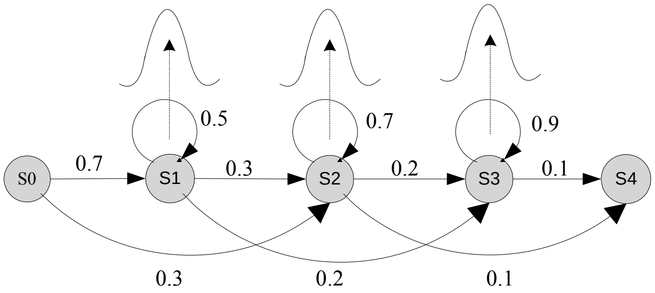


Figure (2): HMM Phone Model

The HMM states are associated with emission probability density functions. These densities are usually given by a mixture of diagonal covariance Gaussians as expressed in equation (2).

$$b_i(x) = \sum_{j=1}^{N_i} w_{ij}\mathcal{N}(x, \mu_{ij}, \sigma_{ij}) \qquad (2)$$

where *j* ranges over the number of Gaussian densities in the mixture of state S_i and expression *N (:)* is the value of the chosen component Gaussian density function for feature vector x.The parameters of the model (state transition probabilities and output distribution parameters e.g. means and variances of a Gaussian) are automatically estimated from training data. This estimation is usually referred to as model training and there exist several criteria and training algorithms that will be discussed below. Usually using only one model per phone is not accurate enough and usually several models are trained for each phone depending on its context. For example, tri-phone uses a separate model depending on the immediate left and right contexts of a phone. For example, tri-phone *A* with left context *b* and right context *n* (referred to as */b-A-n/*) has a different model than tri-phone *A* with left context *t* and right context *m* (referred to as */t-A-m/*). For a total number of phones *P* there will be P^3 tri-phones, and for *N* states/model there will be $N\ P^3$ states in total. The idea can be generalized to larger context e.g. quinphones. This typically leads to a large number of parameters. In practice context-dependent phones are clustered to reduce the number of parameters. Perhaps the most important aspect in designing a speech recognition system is finding the right number of states for the given amount of training data. Extensive research has been done to address this point. Methods vary from very simple phonetic rules to data driven clustering. Perhaps the most popular technique used is the decision tree clustering [2]. In this method both context questions and a likelihood metric are used to cluster the data for each phonetic state as shown in figure (3). The depth of the tree can be used to tradeoff accuracy versus robustness. Once the context-dependent states are clustered it remains to assign a probability distribution to each clustered state. Gaussian mixtures are the most popular choice in modern speech recognition systems. The parameters of the Gaussians are estimated to maximize the likelihood of the training data (the so-called maximum likelihood estimation). For HMMs ML estimation is achieved by the so-called forward backward or Baum-Welch algorithm.

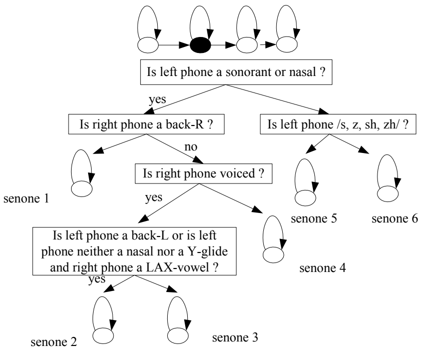

Is left phone a sonorant or nasal ?
yes
Is right phone a back-R ?
Is left phone /s, z, sh, zh/ ?
no
Is right phone voiced ?
senone 1
yes
senone 5
senone 6
Is left phone a back-L or is left phone neither a nasal nor a Y-glide and right phone a LAX-vowel ?
senone 4
yes
senone 2
senone 3

Figure (3): Decision tree for classifying the second state of K-triphone HMM

Maximum likelihood remained as the preferred training method for long time. Recently discriminative training techniques took over. It was demonstrated that they can lead to superior performance. However, this comes at the expense of a more complex training procedure [3]. There are several discriminative training criteria and perhaps this was one of the most active research area in speech recognition in the past few years. These include: maximum mutual information (MMI), minimum classification error (MCE), minimum phone error (MPE) and most recently maximum margin methods. We will not get in details of these different techniques but all share the idea of using the correct transcription and a set of competing hypotheses. They estimate the model parameters to "discriminate" the correct versus competing hypotheses. The competing hypotheses are usually obtained from a lattice which in turn requires the decoding of the training data. Model estimation is most widely done using the so-called extended Baum-Welch estimation (EBW) [4].

To summarize, acoustic model training consists of the following

- Form context dependent states and cluster them typically using the decision tree method. This step results in about several thousand states.
- Each context dependent state is represented by a Gaussian mixture model (GMM). The parameters of each GMM are estimated using the forward-backward algorithm.
- Training data is decoded and a lattice is formed for each sentence
- Statistics collected over the lattice are used to refine the model parameters by discriminative training employing variants of the so-called extended Baum-Welch estimation.

A main drawback of the Gaussian mixture model is that it is a generative model. It can provide accurate representation for the training data but does not care about the discrimination of the different classes of the data. Recently a better acoustic model was introduced that is a hybrid HMM and Deep Neural Networks (DNN). The Gaussian Mixtures Models (GMM) are replaced with Neural Networks with deep number of hidden layers as shown in figure (4).

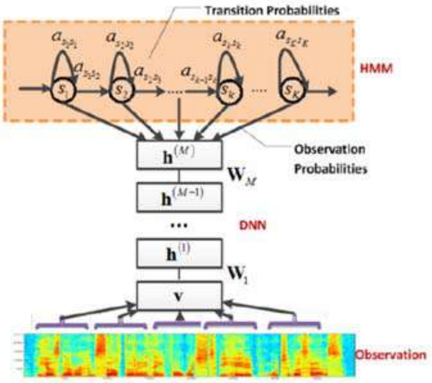


Figure (4): HMM-DNN Model

The DNN have a higher modeling capacity per parameter than GMMs and they also have a fairly efficient training procedure that combines unsupervised generative learning for feature discovery with a subsequent stage of supervised learning that fine tunes the features to optimize discrimination. The context-dependent (CD)-DNN-HMM hybrid model [5] has been successfully applied to large vocabulary speech recognition tasks and can cut word error rate by up to one third on the challenging conversational speech transcription tasks compared to the discriminatively trained conventional CD-GMM-HMM systems [6].

While the above summarizes how to train models it remains to discuss the training data. Of course using more data allows using larger and hence more accurate models leading to better performance. However, data collection and transcription is a tedious and costly process. For this reason the technique called unsupervised or better lightly supervised training is becoming very popular. First, several hundred hours of speech are used to train a model. The model together with an appropriate confidence measure can then be used to automatically transcribe thousands of hours of data. The new data can be then used to train a larger model. All the above techniques (and more) are implemented in the so-called Hidden Markov Model Toolkit (HTK) developed at Cambridge University and can be downloaded together with its source code. We strongly recommend researchers wishing to work on acoustic models to download and get acquainted with HTK [7].

C. Language Model

A language model (LM) is required in large vocabulary speech recognition for disambiguating between the large set of alternative, confusable words that might be hypothesized during the search. The LM defines the priori probability of a sequence of words. When language restrictions are well known and all the possible combinations between words can be defined, probabilities can be precisely calculated and included in finite state automata (FSA) that rules the combination of words in a sentence. Unfortunately, this scheme only applies to restricted application domains with small vocabularies. For large vocabularies and more complex configurations of sentences a simple, but effective, way to represent a sequence of *n* words is to consider it as an *n-th* order Markov chain. The LM probability of a sentence (i.e., a sequence of words $w_1, w_2, \ldots\ldots, w_n$) is given by:

$$\begin{aligned} &P(w_1)\,P(w_2\,/\,w_1)\,P(w_3\,/\,w_1,w_2)\,P(w_4\,/\,w_1,w_2,w_3)\ldots\ldots P(w_n\,/\,w_1,\ldots..w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i\,/\,w_1,\ldots.,w_{i-1}) \end{aligned} \qquad (3)$$

where in the expression such as $P(w_i \mid w_1, \ldots\ldots, w_{i-1})$, $w_1, \ldots\ldots, w_{i-1}$ is the word history for word w_i. In practice, one cannot obtain reliable probability estimates given arbitrarily long histories since that would require enormous amounts of training data. Instead, one usually approximates them in the following way:

$$P(w_i \mid w_1, w_2, \ldots\ldots.., w_{i-1}) \approx P(w_i \mid w_{i-N+1}, \ldots.., w_{i-1}) \qquad (4)$$

which is the definition of “N-grams”. On several recognition approaches, the number of predecessors considered tend to be reduced resulting in “bigrams” (for N=2) and “trigrams” (for N=3). An important feature of N-grams is that their probabilities can be directly estimated from text examples and, therefore do not need explicit linguistic rules like grammar inference systems do. Estimation of N-grams has to be carefully treated as for a vocabulary of size *V* there is as many as $(V)^N$ probabilities to be estimated in the N-gram model. Usually many word histories don’t occur with enough counts to have reliable estimate for their probabilities. Many approximation techniques were proposed to approximate these probabilities [8]. For example, in the case of bigram grammar it typically lists only the most frequently occurring bigrams, and uses a backoff mechanism to fall back on unigram probability when the desired bigram is not found. In other words, if $P(w_j|w_i)$ is sought and is not found, one falls back on $P(w_j)$. But a backoff weight is applied to account for the fact that w_j is known to be not one of the bigram successors of w_i [9]. Other higher-order backoff n-gram grammars can be defined similarly. Ideally, a good LM would ease the retrieval of the word sequence present in the speech signal by better focusing the decoding procedure, which represents another relevant step of the search procedure. In spite of the success of N-gram LMs they do not make any use of linguistic knowledge. For example, syntactic and semantic analysis, parsers and other types of linguistic structure. There is a lot of work on how to introduce such knowledge in language modeling. However, such works did not find their way in practical systems. The reason is that they often require tedious annotation, they result in complex models and are hard to introduce during decoding. However, it still interesting to see if using linguistic knowledge is capable of improving state-of-the-art systems. A widely known technique that, in a sense, introduces some syntactic or semantic knowledge to N-gram models is word classes [10]. Word classes can be determined based on human knowledge or automatic clustering. N-grams are supported by the SRILM toolkit. We strongly recommend researchers wishing to work on language models to download and get acquainted with SRILM toolkit.Classes are also supported by SRILM toolkit [11].

D. Decoding

Finding the best word (or generally unit) sequence given the speech input is referred to as the decoding or search problem. Formally, the problem reduces to finding the best state sequence in a large state space that consists of composing the pronunciation lexicon, the acoustic model and the language model. The solution can be found using the well-known Viterbi algorithm. Viterbi search is essentially a dynamic programming algorithm, consisting of traversing a network of HMM states and maintaining the best possible path score at each state in each frame. It is a time synchronous search algorithm in that it processes all states completely at time *t* before moving on to time *t* + 1. The abstract algorithm can be understood with the help of Figure (5). One dimension represents the states in the network, and the other dimension represents the time axis.

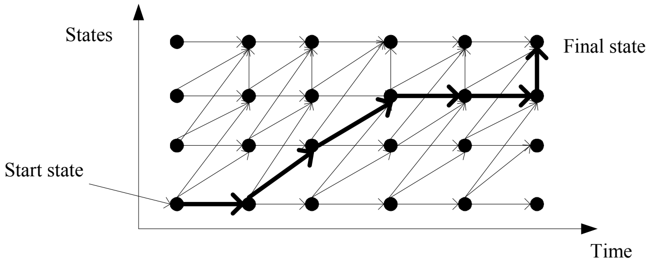


Figure (5): Viterbi Search as Dynamic Programming

Even for a moderate vocabulary, full search is prohibitive. The Viterbi beam search is a very popular and simple way to speed-up the search [12]. Using a beam is not always sufficient and there are two very popular approaches to the search problem:

- Use relatively simple acoustic and language models to generate an N-best list or a lattice. Use more detailed acoustic and/or language models to rescore the reduced search space to find the best word sequence. This is called the multi-pass approach.
- Compose the full search space and use determinaization and minimization algorithms to optimize the search space. Use a Viterbi beam search on the optimized search space to find the best word sequence. We refer to this as the single pass approach.
- A less popular approach is referred to as stack decoding that avoids visiting the whole search space [13].
- In addition to optimizing the search space calculating the Gaussian probabilities is usually time consuming especially for large vocabulary speech recognition. Techniques to accelerate the Gaussian computations are also widely used. These techniques mainly rely on using Gaussian clustering, quantization and caching.

Decoding usually requires a lot of optimization and engineering. Hence, there are usually no publicly available efficient search algorithms especially for the multi-pass approach. The HTK provides a decoder that can be used for small search problems. The single pass approach is more straightforward. The search itself is a simple beam search and can be implemented using the token passing mechanism of the HTK. The key is to optimize the network before the search. The popular finite state machine (FSM) toolkit developed by AT&T is available for download and can be used for this purpose.

E. *Model Adaptation*

Model adaptation is basically a way to modify the model parameters to better match the test utterance or alternatively to modify the input features to better match the existing models. Adaptation is one of the most active research areas in ASR and it is out of our scope to review different adaptation techniques. As far as practical systems are concerned the most popular adaptation methods are linear transformations. Namely, maximum likelihood linear regression (MLLR) in the model space, and feature space maximum likelihood linear regression (FMLLR) in the feature space. As the name suggests these transform the features or model parameters linearly in order to maximize the likelihood of the test (adaptation) data. Adaptation can be done in a supervised way where the reference transcription is given or in an unsupervised way without the reference transcription. In the latter, the system is first used to decode the input speech and the resulting output is used as a transcription for adaptation. The Maximum Likelihood Linear Regression (MLLR) [14] apply affine transforms to the means of the acoustic model as shown if Figure (6).

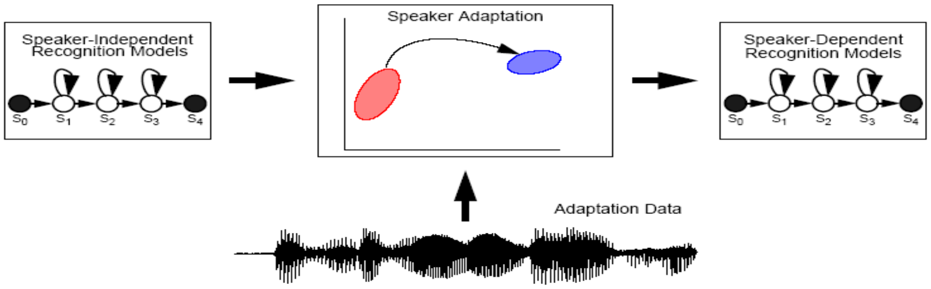


Figure 6. The adaptation of the HMM models parameters

Usually the available data from the user for making the models adaption are small and does not include samples of all the speech units. To work around such limitation a clustering technique is used to cluster the model Gaussians that can be adapted together with same adaptation matrix. A regression class tree approach [16] is used to adjust the number of regression classes to the amount of adaptation data available as shown in Figure (7).

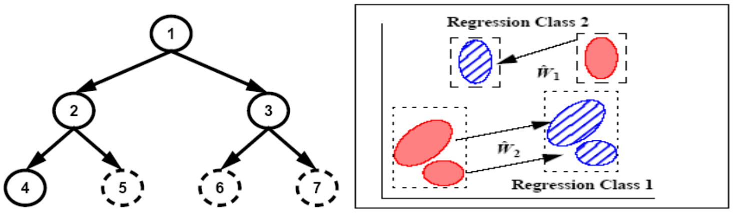


Figure 7: The regression trees for selecting the Gaussians clusters for adaptation

2 Arabic Speech Recognition

Many aspects of Arabic, such as the phonology and the syntax, do not present problems for automatic speech recognition. Standard, language-independent techniques for acoustic and pronunciation modeling, such as context-dependent phones, can easily be applied to model the acoustic-phonetic properties of Arabic. Some aspects of recognizer training are even easier than in other languages, in particular the task of constructing a pronunciation lexicon since there is a nearly one-to-one letter-to-phone correspondence. The most difficult problems in developing high-accuracy speech recognition systems for Arabic are the predominance of non-diacritized text material, the enormous dialectal variety, and the morphological complexity.

A. Restoring Diacritics

The constraint of having to use mostly non-diacritized texts as recognizer training material leads to problems for both acoustic and language modeling. First, it is difficult to train accurate acoustic models for short vowels if their identity and location in the signal are not known. Second, the absence of diacritics leads to a larger set of linguistic contexts for a given word form; language models trained on this non-diacritized material may therefore be less predictive than those trained on diacritized texts. Both of these factors may lead to a loss in recognition accuracy. Ignoring available vowel information does indeed lead to a significant increase in both language model perplexity and word error rate. Several approaches were proposed to overcome the lack of diacritized text. In the grapheme acoustic model each non-diacritized grapheme is considered an acoustic unit which is equivalent to a compound Consonant-Vowel pair. To compensate for the wide variance of these compound units in the acoustic space a larger number of mixtures are used. Although this type of model eliminates the requirement for restoring the Arabic text diacritics, the use of compound acoustic units resulted in reduction in performance compared with the phone based models.

In another approach the original diacritics of the text are restored using automatic alignment of the audio signal with a search lattice that is constructed from the reference undiacritized text script after adding all the possible diacritics for each Arabic consonant and select the best bath that has the highest score match with the audio [19]. Though that approach is simple and easy to implement it can result in large number of diacritization errors. Considering all possible diacritics combinations can add incorrect words to the search space and even some of the correct words are very rare and should be considered with low frequency. To keep the search space more focused a morphology analyzer [20] is used to produce the frequent diacritization forms for each word and only consider them in the search lattice. In another approach an automatic text diacritizer is used to restore the diacritization marks of the non-diacritized text. Such tools rely on using an advanced language model that uses the word context to predict its diacritics. From our experience in building large scale Arabic diacritized speech corpus we found that a combination between the last two approaches provides the best performance.

B. Creating Dialect Arabic Speech corpus

Whereas MSA data can readily be acquired from various media sources, there is only very limited speech corpus of dialectal Arabic available. The construction of such type of corpus is even more challenging than the MSA one. Initially the manual annotation has no standard reference, the same word can be transcribed with several ways such as “ بأشكرك، باشكرك، بشكرك“. Some transcription guidelines for Egyptian Dialectal Arabic were proposed to reduce such differences [21]. The diacritization for dialectal Arabic is more challenging than MSA since it would require a dialectal Arabic morphological analyzer to generate the different diacritization forms. Using context based diacritization would also require a robust language model for dialectal Arabic which also not currently available. Also the Dialectal Arabic diacritization using automatic alignment against the audio signal is also more harder due to the large set of Vowels. For example the Egyptian Dialectal Arabic has 15 vowels compared to only 6 vowels in MSA which raises the number of possible pronunciations for non-diacritized word greatly.

C. Large Vocabulary Decoder

Languages with morphological complexity such as Arabic are known to present serious problems for speech recognition, in particular for language modeling. A high degree of affixation, derivation etc. contributes to the explosion of different word forms, making it difficult, if not impossible, to robustly estimate language model probabilities. Rich morphology also leads to high out-of-vocabulary rates and larger search spaces during decoding, thus slowing down the recognition process. To deal with the morphological complexity of the Arabic language an effective approach for using Factored Language Models (FLM) was proposed [17]. The Arabic word can be factorized to its main morphological components, the prefix, the suffix and the stem as shown in Figure (8). Using this factorization approach the vocabulary size can be reduced with great factor. As we see in figure (8) for a dataset of size 120k the number of Arabic full form words is 14k while the number of stem units is only 6k, which is comparable with the number of stems for English data of same size. The main draw back for the factored models is the small size of the affixation units, that can be only two phones long, which make them highly confusable for the acoustic model scores.

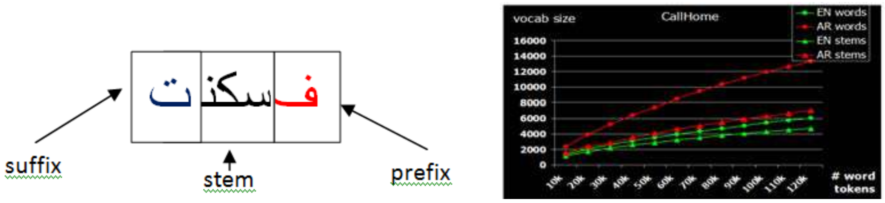


Figure (8) Left: An example of Arabic word factorization. Right: Vocabulary growth for the Arabic language

Another effective approach to deal with the large vocabulary of the Arabic language is the compilation of the whole search space in a finite state network that is optimized to the most compact size. The huge size of the search networks for Large Vocabulary Automatic Speech Recognition (LVASR) systems make it impractical or even impossible to expand the whole search network prior to decoding due to memory limitations. The other alternative approach was to expand the search network on the fly during the decoding process. But with the increase of the vocabulary size in conjunction with the usage of complex Knowledge Sources (KS) such as context dependent tri-phone models and cross word models the dynamic expansion of the search network becomes very slow and turns to be an impractical approach. With the efforts of a research team at AT&T [18] they managed to compile the search network of LVASR systems in a compact size that can fit with memory limitations and also provide a fast decoding approach. That approach relied on eliminating the redundancy in the search network that results from the approximations used in the integrated networks such as the state tying of the acoustic model units and the back-off techniques in the used language model. Let's consider a practical example of a 64k word trigram, typical of a state of- the-art LVCSR system. Among the 4 billion of possible word bigrams, only 5 to 15 million will be included in the model and, for each of these "seen" word-pair histories, the average number of trigrams will be comprised between 2 and 5. Such a LM would have about 5 to 15 million of states and 15 to 90 million of arcs, requiring between 100 and 600 MB of storage. This means a reduction by seven orders of magnitude with respect to a plain 64k trigram. Concerning cross word tri-phones, the number of distinct generalized models is typically one order of magnitude smaller than the full inventory of position-dependent contexts.

3 RESULTS

What is the current state of the art in speech recognition? This is a complex question, because a system's accuracy depends on the conditions under which it is evaluated: under sufficiently narrow conditions almost any system can attain human-like accuracy, but it's much harder to achieve good accuracy under general conditions. The conditions of evaluation - and hence the accuracy of any system - can vary along the following dimensions:

- **Vocabulary size and confusability:** As a general rule, it is easy to discriminate among a small set of words, but error rates naturally increase as the vocabulary size grows. For example, the 10 digits "zero" to "nine" can be recognized essentially perfectly , but vocabulary sizes of 200, 5000, or 100000 may have error rates of 3%, 7%, or 45%.
- **Speaker dependence vs. independence:** By definition, a speaker dependent system is intended for use by a single speaker, but a speaker independent system is intended for use by any speaker. Speaker independence is difficult to achieve because a system's parameters become tuned to the speaker(s) that it was trained on, and these parameters tend to be highly speaker-specific.

- **Task and language constraints:** Even with a fixed vocabulary, performance will vary with the nature of constraints on the word sequences that are allowed during recognition. Some constraints may be task-dependent (for example, an airline querying application may dismiss the hypothesis "The apple is red"); other constraints may be semantic (rejecting "The apple is angry"), or syntactic (rejecting "Red is apple the"). Constraints are often represented by a grammar, which ideally filters out unreasonable sentences so that the speech recognizer evaluates only plausible sentences. Grammars are usually rated by their perplexity, a number that indicates the grammar's average branching factor (i.e., the number of words that can follow any given word). The difficulty of a task is more reliably measured by its perplexity than by its vocabulary size.
- **Read vs. spontaneous speech:** Systems can be evaluated on speech that is either read from prepared scripts, or speech that is uttered spontaneously. Spontaneous speech is vastly more difficult, because it tends to be peppered with disfluencies like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter; and moreover, the vocabulary is essentially unlimited, so the system must be able to deal intelligently with unknown words (e.g., detecting and flagging their presence, and adding them to the vocabulary, which may require some interaction with the user).
- **Adverse conditions:** A system's performance can also be degraded by a range of adverse conditions. These include environmental noise (e.g., noise in a car or a factory); acoustical distortions (e.g., echoes, room acoustics); different microphones (e.g., close-speaking, omnidirectional, or telephone); limited frequency bandwidth (in telephone transmission); and altered speaking manner (shouting, whining, speaking quickly, etc.).

In order to evaluate and compare different systems under well-defined conditions, a number of standardized databases have been created with particular characteristics. Such evaluations were mostly based on the measurement of word (and sentence) error rate as the performance figure of merit of the recognition system. Furthermore, these evaluations weights were conducted systematically over carefully designed tasks with progressive degrees of difficulty, ranging from the recognition of continuous speech spoken with stylized grammatical structure (as used routinely in military tasks, e.g., the Naval Resource Management task) to transcriptions of live (off-the-air) news broadcast (e.g., NAB that involves a fairly large vocabulary over 20K words) and conversational speech. Figure (9) shows a chart that summarizes the benchmark performance of various large vocabulary continuous speech recognition tasks, as measured in formal DARPA and NIST evaluations and table (1) includes the optimum systems performance.

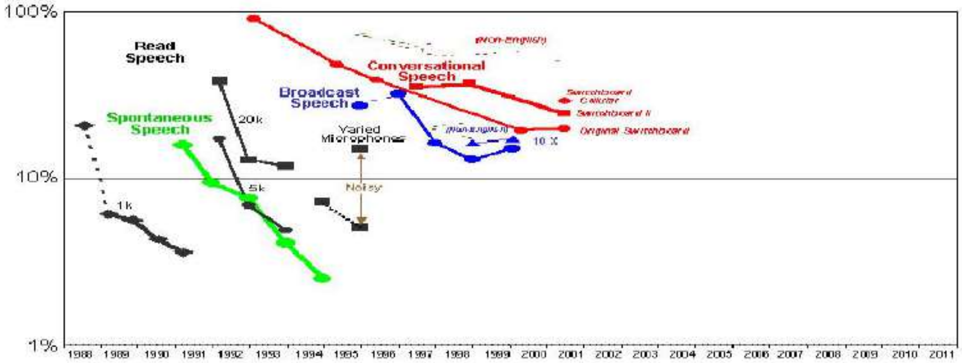


Figure 9: Speech Recognitions Systems Evaluations

In the chart, the task of "Resource Management" involves a rigidly stylized military expression with a vocabulary of nearly 1000 words. ATIS is a task that involves simple spontaneous speech conversation with an automated air travel information retrieval system; although the speech is spontaneous, its linguistic structure is rather limited in scope. WSJ refers to transcription of a set of spoken (read) paragraphs from the Wall Street Journal; the vocabulary size could be as large as 60K words. The Switchboard task is one of the most challenging ones proposed by DARPA. The speech is conversational and spontaneous, with many instances of the so-called disfluencies such as partial words, hesitation and repairs, etc. The general conclusion that can be drawn from these results is that conversational speech, which does not strictly adhere to linguistic constraints, is significantly more difficult to recognize than task-oriented speech that follows strict syntactic and semantic production rules. Also, the evaluation program showed that increasing the amount of speech data used for estimating the recognizer parameters (i.e., the size of the training set) always led to reductions of word error rate. (It is a well accepted target that in order for virtually any large vocabulary speech recognition task to become viable, the word error rate must fall below a 10% level).

TABLE I
THE OPTIMUM SYSTEMS PERFORMANCE

CORPUS	TYPE	VOCABULARY SIZE	WORD ERROR RATE
Connected Digit Strings--TI Database	Spontaneous	11 (zero-nine, oh)	0.3%
Connected Digit Strings--Mall Recordings	Spontaneous	11 (zero-nine, oh)	2.0%
Connected Digits Strings--HMIHY	Conversational	11 (zero-nine, oh)	5.0%
RM (Resource Management)	Read Speech	1000	2.0%
ATIS(Airline Travel Information System)	Spontaneous	2500	2.5%
NAB (North American Business)	Read Text	64,000	6.6%
Broadcast News	News Show	210,000	13-17%
Switchboard	Conversational Telephone	45,000	25-29%
Call Home	Conversational Telephone	28,000	40%

The size of the vocabularies that ASR systems can be handled has evolved greatly in the last decade as shown in figure (10). Now we have ASR systems that can process vocabulary in order of million words with reasonable processing time. Thanks to the approach of the Finite State Decoders (FSD) that compile the whole search space in a single network that is optimized to remove any redundancies which result in very efficient decoding.

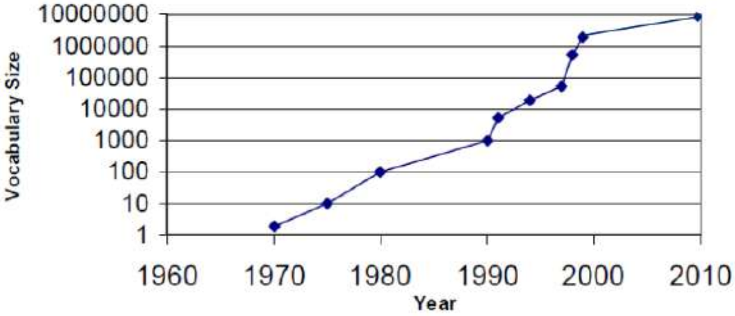


Figure 10: The vocabulary size enhancement

We believe the most challenging aspect of Arabic is the existence of many dialects. These dialects are in many instances substantially different and have very limited acoustic and language model training data. There were several attempts to perform dialect speech recognition for Egyptian, Leventaine and Iraqi but the error rate is relatively high. On the other hand MSA has sufficient resources and accordingly reasonable performance. The table below shows the performance of different systems for broadcast news transcription in the Gale project and some dialectal tasks.

Table II shows roughly state-of-the-art performance for different speech recognition tasks for Arabic. The performance is closely related to the existing resources. We can see for MSA Arabic the available resources, of vowelized training hours and Giga words of LM training text, are close to other Latin languages. So the state of art performance for MSA which is around 15% WER is very comparable with the 10% WER achieved for the similar task of Broad Cast News ASR for English. But we should keep in consideration that the complexity of the Arabic MSA ASR is much higher with vocabulary size of 560k words compared with the 210k words of the English vocabulary for the BC News ASR. The performance of dialectal Arabic, as show in the Iraqi, Egyptian and Levantine, conversational ASR is comparable with the equivalent conversational English ASR with average WER in the range 30%-40%. But we should keep in consideration that the dialectal Arabic is much challenging when compared with conversational English. The LM training data is very limited, and many required NLP tools such as morph analyzer, diacritizer and text normalizers need to be developed.

TABLE II
STATE OF ART PERFORMANCE FOR ARABIC ASR SYSTEMS

Dialect	Models	Vocabulary size	Acoustic training	LM	WER
MSA	unvowelized	589K	135hr, 1000hr (unsup)	56M 4-gram	17.0%
MSA	vowelized	589K	135hr, 1000hr (unsup)	56M 4-gram	16.9%
MSA	Vowelized + pronprobs	589K	135hr, 1000hr (unsup)	56M 4-gram	14.0%
Iraqi	unvowelized	90K	200 hr	2M 3-gram	36.0%
Egyptian	vowelized		Call home	Call home	56.1%

4 CONCLUSIONS

In this paper we reviewed the main components of ASR systems and the state of art approaches for implementing each one of them. Also we showed that the Arabic language has many features that make it an easy language for developing automatic speech recognition systems. We show that even some aspects of the Arabic language such as the nearly one-to-one letter-to-phone correspondence make the construction of the pronunciation lexicon even easier than in other languages such as English which have complex Letter to phone rules. The most difficult challenges in developing speech recognition systems for Arabic are the dominance of non-diacritized text material, the several dialects, and the morphological complexity. We show how we deal with these challenges. For the missing vowels challenge we showed that using grapheme based models would provide a close performance to the vowelized version with less cost for the data preparation. Also we showed some approaches for vowels restoring either by alignment with waves or using automatic text diacritizer. Also for the challenge of the large vocabulary we showed that it can be factored to smaller units using morphology analyzers. Also we showed that the large vocabulary can be compiled in a finite state network that can be reduced in size using finite state optimization techniques. For the challenge of limited resources form some Arabic dialectals we showed that back-of techniques with the large MSA Arabic data resources would provide some improvement in performance. Finally we introduced the state of art performance for Arabic ASR systems for either MSA and conversational dialectal speech. Also we showed that the Arabic ASR systems has very close performance it its equivalent systems in English which confirm our initial claim that Arabic is an easy language for developing ASR systems.

REFERENCES

[1] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, SA: Prentice- Hall, Inc., 1993.

[2] C. H. Lee, F. K. Soong, and K. K. Paliwal, “Automatic Speech and Speaker Recognition”, Kluwer Academic Publishers, Norwell, MA, 1996, pp 481-508.

[3] Collins, Michael. "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. "*Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

[4] Povey, Daniel, et al. "Boosted MMI for model and feature-space discriminative training." *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.

[5] Li, Jinyu, et al. "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM." *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012.

[6] Huang, Yan, et al. "A comparative analytic study on the Gaussian mixture and context dependent deep neural network hidden Markov models." *INTERSPEECH*. 2014.

[7] S. Young, et. al., the *HTKBook*, http://htk.eng.cam.ac.uk/.

[8] J. Kupiec, “ Probabilistic models of short and long distance word dependencies in running text.”, *Proceeding of ARPA Workshop on Speech and Natural Language*, pp. 290-295 Feb 1989.

[9] S. M. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer.” In IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 35, No. 6, pp. 400-401, Mar. 87.

[10] M. Jardino, “Multilingual stochastic N-gram class language models.” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, PP. 161-163, 1996.

[11] Stolcke, Andreas. "SRILM-an extensible language modeling toolkit." *Interspeech*. Vol. 2002. 2002.

[12] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm." In IEEE Transactions on Information Theory, Vol. IT¬13, Apr. 1967, pp. 260¬-269.
[13] R.L. Bahl et al "large vocabulary natural language continuous speech recognition". IEEE International Conference on Acoustics, Speech, and Signal Processing 1989, PP. 465-467.
[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Computer Speech and Language, vol. 9, no. 2, pp. 171 – 185, Apr. 1995.
[15] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol. 12, no. 2, pp. 75 – 98, Apr. 1998.
[16] C. Leggetter and P. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression," in Proc. ARPA Spoken Language Technology Workshop, Austin, TX, USA, Jan. 1995, pp. 104 – 109.
[17] J. A. Bilmes. Graphical models and automatic speech recognition. In R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, editors, *Mathematical Foundations of Speech and Language Processing*. Springer-Verlag, New York, 2003.
[18] Mohri, M. and Riley M., "Network Optimizations for Large- Vocabulary Speech Recognition", in Speech Communication Journal, Vol. 28, Nr. 1, pp. 1–12, May 1999.
[19] L. Lamel, A. Messaoudi, J. Gauvain, "Automatic Speech-to-Text Transcription in Arabic", ACM Transactions on Asian Language Information Processing (TALIP), 8(4), pp. 225-232, December, 2009.
[20]T. Buckwalter, "Arabic Morphology Analysis", The International Conference of Acoustics, Speech and Signal Processing (ICASSP), pp. 3688–3691, 2000. Istanbul
[21] N.Y. Habash, M.T. Diab, O.C. Rambow "Conventional Orthography for Dialectal Arabic (CODA) Version 0.1" 2000.

BIOGRAPHY

Dr. Sherif MahdyAbdou received his B.Sc. and M.Sc. in computer science and automatic control from University of Alexandria, Egypt in 1993 and 1997, respectively. He received a Ph.D in Electrical and Computer Engineering from University of Miami, USA in 2003. In 2003 Dr. Abdou joined BBN Technologies as a senior staff scientist in the Arabic language team of the Ears project to provide affordable reusable speech-to-text decoding for the Defence Advanced Research Projects Agency, DARPA. In 2005 Dr. Abdou joined the Faculty of Computers and Information, Cairo University where is currently Professor at the department of Information Technology. Dr. Abdou is also the research and development manager of the Research and Development International (RDI) Company where he is leading a team to develop several products for natural language processing, computer aided language learning, speech recognition, speech syntheses, optical character recognition, handwriting recognition with special focus on the technologies of the Arabic language. Dr. Abdou has more than 75 published articles and is a member of the review committee in several conferences and journals in the HLT field. Dr. abdou is the Principal Investigator and Co- Principal Investigator of several research projects in the areas of Language learning, Virtual tutors, Web monitoring and Intelligent Contact Centres.. Also Dr. Abdou is one of the holders of the patent " Systems and Methods for Quran Recitations Rules: HAFSS".

أنظمة التعرف الآلى على الكلام العربى : التحديات والنجاحات

شريف عبده
كلية الحاسبات و تكنولوجيا المعلومات-جامعة القاهرة-الجيزة-جمهورية مصر العربية

ملخص

عدد من خصائص اللغة العربية تجعلها من اللغات السهلة لبناء انظمة التعرف الآلى على الكلام. التركيب الفونولوجى المنتظم لمفردات اللغة العربية تجعل تحديد طريقة نطق الكلمة العربية من السهل أستنتاجه من الصورة المكتوبة للكلمة بخلاف كثير من اللغات اللاتينية التى تحتاج لقواعد معقدة لتحويل الصيغة المكتوبة لصورة منطوقة. ولكن يوجد ثلاثة تحديات أساسية فى اللغة العربية تعوق بناء أنظمة التعرف الآلى على الكلام وهى عدم وجود حركات التشكيل فى أغلب النصوص المكتوبة ، العدد الكبير من مفردات اللغة العربية بسبب البناء المورفولوجى المركب مما يسبب تضخم شبكة البحث ويؤدى لبطئ النظام وأخيرا المصادر المحدودة لبعض اللهجات المنطوقة مثل اللهجة المصرية والخليجية والشامية والمغربية وهى عدم توفرتسجيلات معنونة و نصوص مشكلة بالحجم الكافى لبناء انظمة التعرف الآلى على الكلام لهذه اللهجات. فى هذه الورقة البحثية تم أستعراض الحلول البحثية التى تم تطويرها للتغلب على هذه التحديات. وتختم الورقة بعرض نتائج تقييم الأداء لأنظمة التعرف الآلى على الكلام العربى مع مقارنتها بأداء الأنظمة المقابلة فى اللغات الأجنبية الأساسية.

Robust Speaker Recognition Using Adaptive Hidden Markov Models

Aya S. Mostafa[*1], Amr M. Gody[*2], Tamer M. Barakat[*3]

Department of Electrical Engineering, Faculty of Engineering, Fayuom University, Egypt
[1]ayasami89@yahoo.com
[2]amg00@fayuom.edu.eg
[3]tmb00@fayoum.edu.eg

***Abstract* - This work introduces a generalized speaker recognition system using Hidden Markov Models. The system is evaluated using samples of Corpus database with hand label files. The system is tested for speaker identification. Text dependent and text independent identification is applied to the system. Different number of speakers, male and female speakers, different time duration files are used to test the system. The identification rates reached 90% and verification reached 100%.**

1 INTRODUCTION

Speaker recognition is classified into speaker identification and speaker verification. Speaker identification is the process of determining from which speaker among a group of speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity of a tested speaker. Most of the applications in which voice is used to verify the identity of a speaker are classified as speaker verification [1].

Hidden Markov Models (HMMs) are used in general to model sequences of stochastic data. The HMM is capable of modeling temporal behavior of sequence of events like speaker produces during talking [2]. The paper is divided into three parts, introduction to introduce the HTK, then adaptation steps in an experimental way, at last the obtained results are presented and evaluated. The system is evaluated using Corpus database [4].

2 DATABASE DESCRIPTION

The Chains corpus is a novel speech corpus collected with the primary aim of facilitating research in speaker identification. The corpus features approximately 36 speakers, males and females recorded under a variety of speaking conditions, allowing comparison of the same speaker across different well-defined speech styles. Speakers read a variety of texts alone, in synchrony with a dialect-matched co-speaker, in imitation of a dialect-matched co-speaker, in a whisper, and at a fast rate. The bulk of the speakers were speakers of Eastern Hiberno-English. The corpus is being made freely available for research purposes.

There are different speaking conditions. The solo condition is used in this work because it is clear and not so fast.

Corpus speech database has two problems to be solved in order to be utilized. The first is that the database does not contain lab files. Thus, the lab files of the selected wave files from Corpus database are hand made using *SFS* software. The second one, is that all Corpus wave files are sampled at 44.100 kHz, while this work is done with 16 KHz samples wave files. The conversion from 44.100 kHz to 16 KHz is done using SNDREC software from Microsoft. Names of speakers wave files used from databases are renamed in a way to simplify the identification process.

3 SPEAKER RECOGNITION

Speaker recognition technology is closely related to speech recognition, where it means automatic speaker (talker) recognition by machine. The general area of speaker recognition includes two fundamental tasks, speaker identification and speaker verification. The speaker identification task is to classify an unlabeled voice sample as belonging to (having been spoken by) one of a set of N reference speakers (N possible outcomes), whereas the speaker verification task is to decide whether or not an unlabeled voice sample belongs to a specific reference speaker (2 possible outcomes the sample is either accepted as belonging to the reference speaker or rejected as belonging to an impostor).

The speech used for identification tasks can be either text-dependent or text independent. In a text-dependent application, the speaker is required to speak a predetermined (fixed) utterance. In contrast, text-independent speaker recognition does not rely on a specific text being spoken. The text independent speaker recognition is more difficult but also more flexible. For speaker recognition, various types of speaker models have been long studied.

The probabilistic HMM and their mathematical foundations gave rise to the speaker recognition systems using these models. HMMs have become the most popular statistical tool for the text-dependent task. The best results have been obtained using continuous density HMMs (CHMMs) for modeling speaker characteristics [5]-[6]-[7].

4 HIDDEN MARKOV TOOLKIT (HTK)

HTK is a toolkit for building Hidden Markov Models (HMMs). HMMs can be used to model any time series and the core of HTK is similarly general purpose. However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. Thus, much of the infrastructure support in HTK is dedicated to this task. As shown in Fig. 1, there are two major processing stages involved. Firstly, the HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. Secondly, unknown utterances are transcribed using the HTK recognition tools. The main body of the HTK tool kit is mostly concerned with the mechanics of these two processes [2].

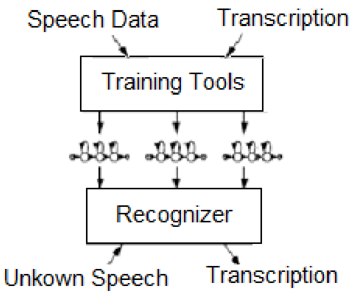


Figure1: HTK processing stages

5 EXPERIMENTAL WORK

Corpus speech [4] data base is used during experimental trails. The following sections illustrate the main activities for preparing and executing the experiments.

A. The HTK command files and configuration files

The HTK commands and configuration files used to build the speaker identification, are explained breifly hereby:

1) The Grammar, HTK uses a finite state grammar that consists of variables defined by regular expressions. A file called gram.txt is created including the following lines:

```
$speaker = frf01 | frf02|irf01 | frf04;
```

where
($speaker)
= frf01, frf02, irf01 and | frf04 are the alternative tokens to be recognized by the system

2) A word network must be created from the grammar; this can be done using HParse HTK command:

HParse gram.txt wdnet

3) Features Extraction: is to extract relevant information from the speech signal. Mel-frequency Cepstral coefficients (MFCC) is commonly used [8]-[9].
Before extracting the features, HTK will have a configuration file to configure the input and output parameters of the tool kit commands. For example of such configuration file, a file named "config_wav2mfc" is created as of the following sample.

```
# Coding parameters
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAV
SOURCERATE = Sampling Period in units of 10-7s (ex.100000)"
TARGETKIND = MFCC_0_D_A
TARGETRATE = Frame skip duration in units of 10-7s (ex. 100000)
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = Frame duration in 10-7s (ex. 250000)
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
```

```
NUMCEPS = 12
ENORMALISE = T
```

*4)*Preparing features extraction script file; list of speech signals in wav file format and the corresponding target features under test are listed into text file, named “convert.txt”, as shown in the next sample

```
Data\frf01_f01_solo.wav        Data\frf01_f01_solo.mfc
Datafrf02_f01_solo.wav         Data\frf02_f01_solo.mfc
Data\frf04_f01_solo.wav         Data\frf04_f01_solo.mfc
Data\ irf01_f01_solo.wav       Data\irf01_f01_solo.mfc
```

where Data is the directory that contains the wav files and will store the output mfcc files.

5) MFCCs are extracted from the “.wav” files using HCopy HTK command as follows:

HCopy -T 1 -C config_wav2mfc -S convert.txt

6) HMM Models Preparation;
Training and testing file lists are provided into two script files as the next example:

File Name: train.txt

```
Data\frf01_f01_solo.mfc
Data\frf02_f01_solo.mfc
Data\frf04_f01_solo.mfc
```

The same for testing.
test.txt

```
Data\frf01_f01_solo.mfc
Data\frf02_f01_solo.mfc
Data\frf04_f01_solo.mfc
```

7) In speech recognition, all acoustical events are modeled separately. In the present work, all speakers have to be modeled with a Hidden Markov Model. For each speaker a HMM will be designed. Number of states for the HMM is arranged according as the topology described in Fig. 2. There is no fixed rule for choosing the number of states, but it is found from iterations, and changing the number of states that 5 states are suitable for good results. Increasing the number of states yields slight differences in results.

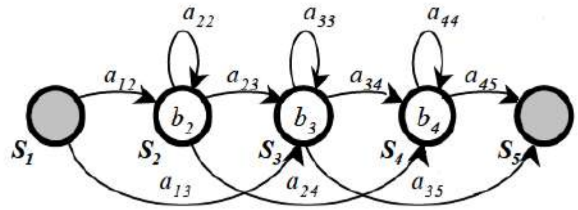


Figure 2: 5 states HMM topology

From Fig. 2, the models are actually 3 “active” states {*S2*, *S3*, *S4*}: the first and last states *S1* and *S5*, are “non emitting” states, it means that no observations are done. The observation functions *bi* are single Gaussian distributions with diagonal matrices. The transition probabilities are *aij*.
In HTK, a HMM is described in a text prototype file named proto. The proto file for the HMM illustrated in Fig.**2** is of the form:

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "proto"
<BeginHMM>
<NumStates> 5
<State> 2
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.00.0 0.0 0.0
```

```
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.01.0 1.0 1.0
<State> 3
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.00.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.01.0 1.0 1.0
<State> 4
<Mean> 39
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.00.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.00.0 0.0 0.0
<Variance> 39
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.01.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 1.0 1.01.0 1.0 1.0
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

where:
~o <VecSize> 39 <MFCC_0_D_A>
is the header of the file, giving the coefficient vector size (39 coefficients in this model are used), and the type of coefficient is chosen to be (MFCC_0_D_A).

~h "proto" <Begin HMM> (...)<End HMM>
This tag encloses the description of a HMM prototype file.
<Num States> 5
This tag gives the total number of states in the HMM, including the 2 non-emitting states 1 and 5.
<State> 2
This tag Describes the observation function of state 2. Here a single-Gaussian observation functions, with diagonal matrices is used. This function is described by a mean vector and a variance vector (the diagonal elements of the autocorrelation matrix). States 1 and 5 are not described, since they have no observation function.
<Mean> 39
0.0 0.0 (...) 0.0 (x 39)
This tag gives the mean vector (in a 39 dimension observation space) of the current observation function. Every element is arbitrary initialized to 0: the file only gives the “prototype” of the HMM (its global topology). These coefficients will be trained later.
<Variance> 39
1.0 1.0 (...) 1.0 (x 39)
This tag gives the variance vector of the current observation function. Every element is arbitrary initialized to 1.
<TransP>5
This tag gives the 5x5 transition matrix of the HMM, that is:

$$\begin{matrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{matrix}$$

where a_{ij} is the probability of transition from state i to state j. Null values indicate that the corresponding transitions are not allowed. The other values are arbitrary initialized (but each line of the matrix must sum to 1): they will be later modified, during the training process. Such a prototype has to be generated for each event to model.

In the present work, HMMs prototype is adapted to define a Hmmdefs (with headers ~h "speaker1", ~h "speaker2" and ~h "speaker3", so on) to model the speakers under observations.

Now using HCompV to initialize the models with the training data:

```
HCompV -C config_mfc -f 0.01 -m -S training.txt–Mhmm0 proto
```

- Two files are created – proto and vFloors– in the directory hmm0. These files must be edited in the following way:
- The first three lines of proto must be cut and pasted into vFloors, which is then saved as macros.
- Create a file called hmmdefs by copying and pasting the rest of the proto file once for each HMM and renaming the copies accordingly. Note that each HMM begins with ~h “model_name” and ends with<EndHMM>.

8) Model Re-estimation, now, the speakers models with global means and variances have been initialized, HERest is then used to perform Baum-Welch training. For this step two more files have to be created:"Speaker Train Models0.mlf "which contains speakers' transcriptions files.

```
#!MLF!#
"*/frf01_f01_solo.lab"
frf01
.
"*/frf02_f01_solo.lab"
frf02
.
"*/irf01_f01_solo.lab"
irf01.
"*/frf04_f01_solo.lab"
frf04
.
…...,and  so on.
```

the second file contains the tokens names .This file can be named “Tokens.txt” file,

```
frf01
frf02
irf01
```

Then re-estimate the models:

```
HERest -C config_mfc -I speakertrainmodels0.mlf –t250.0 150.0 1000.0 -S training.txt -H hmm0/macros –Hhmm0/hmmdefs -M hmm1 tokens.txt
```

Continue re-estimating the model for three or more times, each time putting the re-estimated models in a new directory: hmm1 (as above in 8.2), hmm2, hmm3, hmm4, and so on.

9) Recognition results; the dictionary file defines (in alphabetical order) speaker models by their constituent parts, i.e. by each HMM associated with them. For this speaker recognizer each model consists of only one HMM.
Create Dictionary file; call it dict– will look like this:

```
frf01 frf01
frf02 frf02
```

It is a two columns file. The first column is for the token and the next column is for the way it is constructed from the available HMM models. In our case they are the same. The system is recognizing 4 speakers. Each speaker is a token. Each speaker has a model so that the token is the same as the model.
A Hmm List file is created for naming the models. It is similar to the file tokens.txt but each model is enclosed in double quotes; it will contain the following:

```
“frf01”
“frf02”
```

Now, all the elements are in place to perform speaker recognition. Our recognized labels will be outputted to the files results_speaker.mlf for training data and testing data, when carrying out the recognition using HVite (which performs recognition using the Viterbi algorithm):

```
HVite -H hmm3/macros -H hmm3/hmmdefs -S test.txt-i results_speaker_test.mlf -w wdnet -p 0.0 -s 5.0 dict HmmList
HVite -H hmm3/macros -H hmm3/hmmdefs -S train.txt-i results_speaker_train.mlf -w wdnet -p 0.0 -s 5.0 dict HmmList
```

List the contents of the test files. The former will be similar in format to speakertrainmodels0.mlf except the models listed are at the speaker level rather than the phone level. Call it speakertestmodels0.mlf.

```
#!MLF!#
"*/frf01_f05_solo.lab"
frf01
.
"*/frf02_f06_solo.lab"
frf02
.
```

10) Results, HResultsis then used to display the HTK results analysis tables and store the output of the speakers models in an .mlf file named results_speaker:

```
HResults -I speakertestmodels0.mlf HmmListresults_speaker.mlf
```

6 EXPERIMENTALRESULTS and ANALYSIS

Two types of speaker identifications will be tested hereby: text dependent, it means speakers are saying the same utterance and text independent which means that the speakers are saying different utterances.

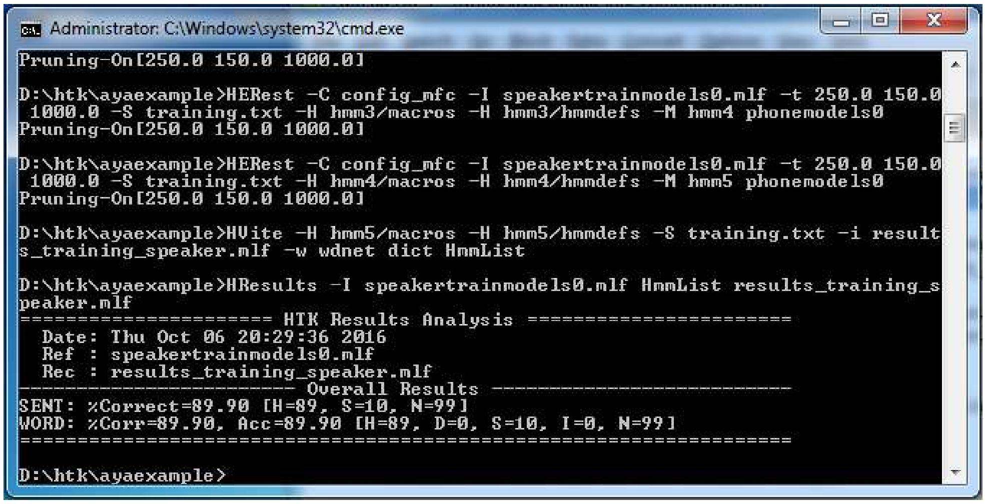
The speakers samples used from Corpus database are identified through the process by the IDs shown in Table 1 for simplicity.

TABLE I

SAMPLES NAMES AND CORRESPONDING USED IDS

Sample Name	Sample ID
solo\frf01\frf01_f01_solo	spk1
solo\frf02\frf02_f02_solo	Spk2
solo\frf04\frf04_f03_solo	Spk3
solo\irf01\irf01_f04_solo	Spk4
solo\irf02\irf02_f05_solo	Spk5
solo\irf03\irf03_f06_solo	Spk6
solo\irf04\irf04_f07_solo	Spk7
solo\irf05\irf05_f08_solo	Spk8
solo\irf06\irf06_f09_solo	Spk9
solo\irf07\irf07_f010_solo	spk10
solo\irf08_f01_solo_solo	spk11
solo\irf09_f01_solo_solo	Spk12
solo\irf10_f01_solo_solo	Spk13
solo\irm01_f01_solo_solo	Spk14
solo\irm02_f01_solo_solo	Spk15
solo\irm03_f01_solo_solo	Spk16

1- Text independent speaker identification case: 10 speakers with 10 speech different utterances each from Corpus database-solo condition. From Fig. 3, text independent identification using the HTK models for large number of speakers gives reliable results, about 90%.



```
Pruning-On[250.0 150.0 1000.0]

D:\htk\ayaexample>HERest -C config_mfc -I speakertrainmodels0.mlf -t 250.0 150.0
 1000.0 -S training.txt -H hmm3/macros -H hmm3/hmmdefs -M hmm4 phonemodels0
Pruning-On[250.0 150.0 1000.0]

D:\htk\ayaexample>HERest -C config_mfc -I speakertrainmodels0.mlf -t 250.0 150.0
 1000.0 -S training.txt -H hmm4/macros -H hmm4/hmmdefs -M hmm5 phonemodels0
Pruning-On[250.0 150.0 1000.0]

D:\htk\ayaexample>HVite -H hmm5/macros -H hmm5/hmmdefs -S training.txt -i result
s_training_speaker.mlf -w wdnet dict HmmList

D:\htk\ayaexample>HResults -I speakertrainmodels0.mlf HmmList results_training_s
peaker.mlf
====================== HTK Results Analysis =======================
  Date: Thu Oct 06 20:29:36 2016
  Ref : speakertrainmodels0.mlf
  Rec : results_training_speaker.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=89.90 [H=89, S=10, N=99]
WORD: %Corr=89.90, Acc=89.90 [H=89, D=0, S=10, I=0, N=99]
===================================================================

D:\htk\ayaexample>
```

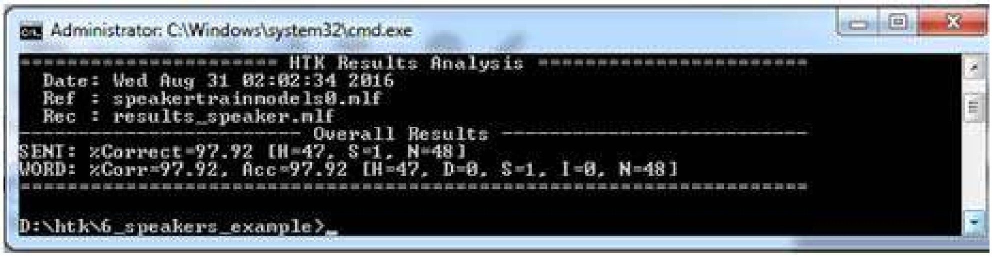
Figure 3: Text independent speaker identification for 10 speakers saying 10 different utterances, case 1

TABLE II

ANALYSIS OF RESULTS SHOWN IN FIG. 3

	spk1	Spk2	Spk3	Spk4	Spk5	Spk6	Spk7	Spk8	Spk9	spk10	%
spk1	10										10/10
Spk2		10									10/10
Spk3			9								9/9
Spk4		1		6					3		6/10
Spk5			5		4		1				4/10
Spk6						10					10/10
Spk7							10				10/10
Spk8								10			10/10
Spk9									10		10/10
Spk10										10	10/10
Results For Identification											89.9%

2- Text dependent speaker identification case: Three male speakers and 3 female speakers saying the same utterance, 8samples from each speaker. The percentage reaches 98%, see Fig. 4.



```
====================== HTK Results Analysis =======================
  Date: Wed Aug 31 02:02:34 2016
  Ref : speakertrainmodels0.mlf
  Rec : results_speaker.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=97.92 [H=47, S=1, N=48]
WORD: %Corr=97.92, Acc=97.92 [H=47, D=0, S=1, I=0, N=48]
===================================================================

D:\htk\6_speakers_example>_
```

Figure 4: Text dependent HTK speaker identification result analysis, case 2

TABLE III

ANALYSIS OF RESULTS SHOWN IN FIG. 4

	Spk11	Spk12	Spk13	Spk14	Spk15	Spk16	
Spk11	7				1		7/8
Spk12		8					8/8
Spk13			8				8/8
Spk14				8			8/8
Spk15					8		8/8
Spk16						8	8/8
Results For Identification							97.92%

From previous example cases, it is seen that HMMs could be used to produce a robust speaker identification system.

7 CONCLUSIONS AND FUTURE WORK

It is seen from the obtained results that the adaptation is succeeded in raising the HTK flexibility and robustness in the field of speaker recognition both verification and identification whether it is text dependent or text independent. The obtained results gives identification rates around 90% for difficult identification environments with large speaker numbers. The future work is opened to apply the proposed adaptation to different databases and different languages.

REFERENCES

[1] Homayoon Beigi, "*Fundamentals of Speaker Recognition*", Springer US, 2011.

[2] Steve Young, "Gunnar Evermann, and el, *The HTK Book (for HTK Version 3.4)*",Cambridge University Engineering Department.,2009.

[3] Yu, K., Mason, J. and Oglesby, J., "*Speaker Recognition Using Hidden Markov Models, Dynamic Time Warping and Vector Quantization*", IEE Proceedings – Vision, Image and Signal Processing, Vol. 142, pp.313-318 ,1995.

[4] "http://chains.ucd.ie/index.php", CHAINS, *2013*.

[5] ShwetaBansal, Atul Kumar and S. S. Agrawal, "*Speaker Adaptation on Hidden Markov Model using MFCC and Rasta-PLP and Comparative Study*", Indian Journal of Science and Technology, Vol. 9, no. 28, July 2016.

[6] M. Savic and S. Gupta, "*Variable parameter speaker verification system based on Hidden Markov Modeling*", *in Proc. of ICASSP*, pp. 281–284, 1990.

[7] B. Tseng, F. Soong, A. Rosenberg, "*Continuous probabilistic acoustic map for speaker recognition", in Proc. of ICASSP*, Vol. II, pp. 161–164, 1992.

[8] Ondřej Novotn, ."*Adaptation of speaker recognition system ",* Master's thesis, Brno University of Technology, Faculty of Information technology, 2014.

[9] Suma Swamy1, and K.V Ramakrishna, "*AN EFFICIENT SPEECH RECOGNITION SYSTEM*", Computer Science and Engineering: An International Journal (CSEIJ), Vol. 3, no. 4, August 2013.

BIOGRAPHY

Prof. Amr M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is now the head of Electrical Engineering department, Fayoum University. He is an author and a co-author of about 50 papers in national and international conference proceedings and journals. His current research areas of interest include speech processing, speech recognition and speech compression.

Assoc. Prof. Tamer M. Baraket received his BSc in communications and computers engineering from Helwan University, Cairo; Egypt in 2000. Received his MSc in Cryptography and Network security systems from Helwan University in 2004 and received his PhD in Cryptography and Network security systems from Cairo University in 2008. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt, in 2009. His main interests are: Cryptography and network security, Digital Image, and Digital Signal Processing. More specially, he is working on the design of efficient and secure cryptographic algorithms, in particular, security in the wireless sensor networks.

Eng. Aya S. Mostafa Received the BSc. in Electronics and communications from the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 2012. She is now an MSc. student in the same department. She is researching in the field of speech processing and speaker recognition. She and her colleagues got the best graduation project award in 2012. She is working now as an employee with the Beni Suef portal team in Beni Suef Governorate.

نظام متين للتعرف على الكلام باستخدام نماذج ماركوف المخفاة

آيه مصطفى، عمرو جودى، تامر بركات

قسم الهندسة الكهربية- كلية الهندسة- جامعة الفيوم- الفيوم

ملخص

يقوم هذا العمل بتقديم نظام التعرف على المتحدث با ستخدام نماذج ماركوف المخفية. وتم تقييم النظام باستخدام عينات من قاعدة بيانات لمتحدثين تسمى كوربوس، مع عمل ملفات الحواشي (اللاب) يدوياً.

وتم التطبيق على نظام التعريف المعتمد على النصوص وكذا النظام الحر الذي لا يعتمد على النصوص. وتم استخدام متحدثين مختلفين، ذكور وإناث، وفترات زمنية للصوت مختلفة، لاختبار النظام.

وقد تم الحصول على معدلات التعرف على المتحدث تقترب من 90% وأيضا معدلات تأكيد هوية المتحدث تقترب من 100%.

Enhancement Quality and Accuracy of Speech Recognition System by Using Multimodal Audio-Visual Speech signal

Eslam E. El Maghraby[*1], Amr M. Gody [*2], M. Hesham Farouk [**3]

[*] *Electrical Engineering, Faculty of Engineering, Fayoum University Egypt*

[**] *Engineering Math. & Physics Dept., Faculty of Engineering, Cairo University Egypt*

[1]eem00@fayoum.edu.eg

[2]amg00@fayoum.edu.eg

[3]mhesham@eng.cu.edu.eg

Abstract— **Most developments in speech-based automatic recognition have relied on acoustic speech as the sole input signal, disregarding its visual counterpart. However, recognition based on acoustic speech alone can be afflicted with deficiencies that prevent its use in many real-world applications, particularly under adverse conditions. The combination of auditory and visual modalities promises higher recognition accuracy and robustness than can be obtained with a single modality. Multimodal recognition is therefore acknowledged as a vital component of the next generation of spoken language systems. This paper aims to build a connected-words audio visual speech recognition system (AV-ASR) for English language that uses both acoustic and visual speech information to improve the recognition performance. Initially, Mel frequency cepstral coefficients (MFCCs) have been used to extract the audio features from the speech-files. For the visual counterpart, the Discrete Cosine Transform (DCT) Coefficients have been used to extract the visual feature from the speaker's mouth region and Principle Component Analysis (PCA) have been used for dimensionality reduction purpose. These features are then concatenated with traditional audio ones, and the resulting features are used for training hidden Markov models (HMMs) parameters using word level acoustic models. The system has been developed using hidden Markov model toolkit (HTK) that uses hidden Markov models (HMMs) for recognition. The potential of the suggested approach is demonstrated by a preliminary experiment on the GRID sentence database one of the largest databases available for audio-visual recognition system, which contains continuous English voice commands for a small vocabulary task. The experimental results show that the proposed Audio Video Speech Recognizer (AV-ASR) system exhibits higher recognition rate in comparison to an audio-only recognizer as well as it indicates robust performance. An increase of success rate by 4% for the grammar based word recognition system overall speakers is achieved for speaker independent test.**

Keywords- AV-ASR, HMM, HTK, MFCC, DCT, PCA, MATLAB, GRID.

1 INTRODUCTION

Automatic speech recognition (ASR) is currently used as an assistive tool in many fields including human computer interfaces, telephony, and robotics and has been used as an alternative method for individuals with disabilities. In spite of their effectiveness, speech recognition technologies still need more work to be employed for people with speech communication disorder especially for people who find it difficult to type with a keyboard.

In human-human communication signals from multiple channels are at work. Human communicate not only through words but also by intonation, gaze, hand and body gestures and facial expressions. Human computer interaction can benefit from modeling several modalities in analogous ways. Multimodal systems represent and manipulate information from different human communication channels at multiple levels of abstraction. So, the need to other source of information that is related to speech can introduce a novel solution compared to audio only ASR. Visual features like the movement of the lips and facial features can work as an example of such source of information. Visual features are demonstrated in many recent audio-visual ASR systems for normal speakers [1, 2].

Hearing impaired and deaf persons make extensive use of visual speech cues and some few individuals perform lip-reading to such a degree that enables almost perfect speech perception [3]. It is well known that seeing the talker's face in addition to hearing his voice can improve speech intelligibility, particularly in noisy environments [4], [5]. The main advantage of the visual signal is its complementarity to the acoustic signal [6]. Phonemes that are most difficult to perceive in the presence of noise are easier to distinguish visually and vice versa. The visual signal contains that kind of information that is acoustically most sensitive to noise [3]. Studies have also shown that visual information leads to more accurate speech perception even in noise-free environments [7]. The strong influence of visual speech cues on human speech perception is demonstrated by the McGurk effect [8] in which, for example, a person hearing an audio recording of /baba/ and seeing the synchronized video of a person saying /dada/ often resulted in perceiving /gaga/.

Automatic speech recognition (ASR) has been an active research area for several decades, but in spite of the enormous efforts, the performance of current ASR systems is far from the performance achieved by humans: error rates are often one order of magnitude a part [9]. Most state-of-the-art ASR systems make use of the acoustic signal only and ignore visual speech cues. They are therefore susceptible to acoustic noise [10], and essentially all real-world applications are subject to some kind of noise. Much research effort in ASR has therefore been directed toward systems for noisy speech environments and the robustness of speech recognition systems has been identified as one of the biggest challenges in future research [11].

The advantage of such an approach is straightforward; the weaknesses of one modality are offset by the strengths of another, resulting in higher accuracy levels. Indeed, audio-visual speech recognition (AV-ASR), in which acoustic features and visual information extracted from the speaker mouth region are jointly used, has been investigated in the literature and found to increase ASR accuracy, primarily in the presence of acoustic noise [12,13].

The above facts have motivated significant interest in automatic recognition of visual speech, formally known as automatic lip reading, or speech reading [5]. Work in this field aims at improving ASR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to audio-visual automatic speech recognition systems. Critical however to the performance of the resulting audio-visual ASR system is the choice of visual features that contain sufficient information about the uttered speech.

There are three key reasons why vision benefits human speech perception [14]: It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The latter is due to the partial or full visibility of articulators, such as the tongue, teeth, and lips. Place of articulation information can help disambiguate, for example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) from the nasal alveolar /n/ [15]. All three pairs are highly confusable on basis of acoustics alone. In addition, jaw and lower face muscle movement is correlated to the produced acoustics [16–17], and its visibility has been demonstrated to enhance human speech perception [18].

Compared to audio-only speech recognition, AV-ASR introduces new and challenging tasks, that are highlighted in the block diagram of Figure1: First, in addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker's face. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by extraction of suitable visual features. In contrast to audio-only recognizers, there are now *two* streams of features available for recognition, one for each modality. The combination of the audio and visual streams should ensure that the resulting system performance is better than the best of the two single modality recognizers, and hopefully, significantly outperform it. Both issues, namely the *visual front end design* and *audio-visual fusion*, constitute difficult problems [19], and they have generated much research work by the scientific community.

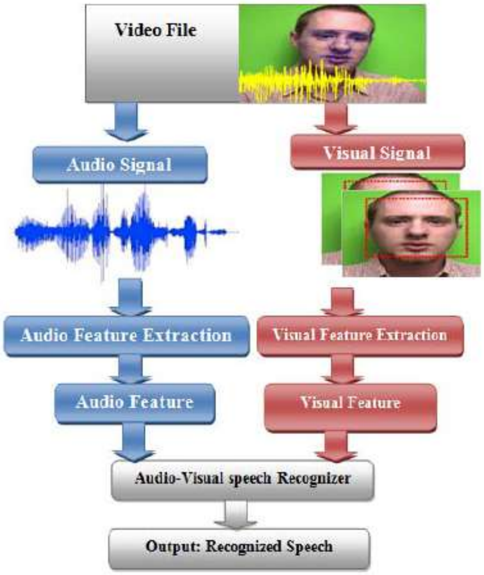


Figure 1: audio-visual speech recognition system

The accuracies obtained by the previous researches in Audio visual speech recognition system are reasonably high [19], but it is still needed to get further improvement. This paper describes a system that uses the visual features to enhance the recognition accuracy.

The proposed Audio Video Automatic Speech Recognizer (AV-ASR) system extracts solely appearance based features, and operates on full face video with no artificial face markings. As a result, both face detection and ROI extraction are required. All stages of the adopted visual front end algorithm are described below.

This paper will discuss the effect of adding visual features on the performance of speech recognition system for different visual features selection methods compared to audio only speech recognition systems and give result on English sentence corpus.

The rest of this paper is organized as follows. Section 2 discusses previous related works. Section 3 explains the block diagram of our proposed system. The experimental results are introduced in section 4. Section 5 contains a conclusion about what have been achieved through this research and future work.

2 PROPOSED SYSTEM

The proposed AV-ASR system architecture that is introduced in this paper is depicted in Figure 2. The input video that contains the speakers' spoken word is divided into audio file and its corresponding image files. There are two working threads, the audio front-end and images (visual) front-end. The audio-visual feature integration process is then performed. Finally, the Hidden Markov Model (HMM) classification is applied to classify the words to their respective classes.

A. Audio Front-End

In this subsection, preprocessing steps done on the audio files and feature extraction are described.

1) *Audio alignment with video stream*: GRID corpus is used. The audio is extracted from the composite video signal. This is accomplished by using the following command to extract mono channel audio signal from the composite signal (mpg) file, run this command line in windows command bacth file,

```
for f in *.mpg; do ffmpeg -i "$f" -ac 1 "${f%.mpg}.wav"; done
```

2) *Audio Pre-Processing:* Before extracting the ASR features, there are required pre-processes that must be applied on the speech streams.

- ***Framing***: or segmentation, means dividing the speech signal into smaller pieces to alter it as stationary with constant statistical properties. It is common in speech to use frame length window not more than 25(ms). In other words, speech signal holds its properties for small period of time typically 25ms [19].

- ***Frame Overlapping:*** Another process that is optionally used to ensure the continuity of the speech signal properties in the current frame along with the adjacent frames. The typical value for the frame overlap period is 10ms [19].

- ***Frame Scaling:*** Since Speech is a non-stationary signal where properties change quite rapidly over time. For most phonemes the properties of the speech remain invariant for a short period of time short-term which estimates of parameters and this is done by effectively cross-multiplying the signal by a window function which is zero everywhere except for the region of interest. Hamming window is applied on the current frame.

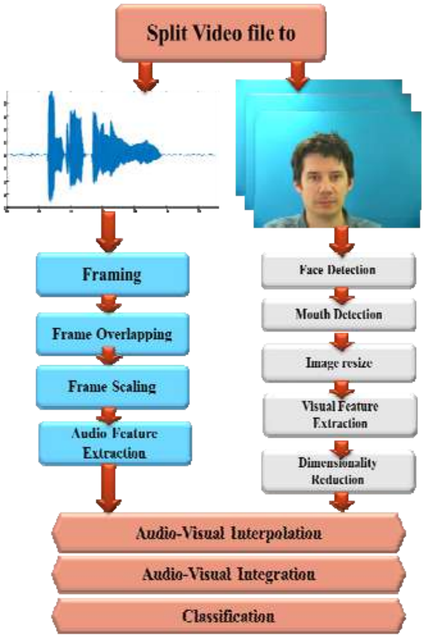


Figure 2: The block diagram of the proposed audio-visual speech recognition system.

3) *Recognition Feature Extraction from Audio signal:* Mel frequency cepstral coefficients (MFCC) is chosen in this research paper. MFCC is the most common audio features [20], MFCC is based on known variation of the human ear’s critical bandwidth with frequency. The overall process of the MFCC is illustrated in figure 3 The software system Hidden Markov Model Toolkit (HTK) [2] is used for extracting 13 MFCC features together with their 1st and 2nd derivatives producing an acoustic feature vector of length 39 elements.

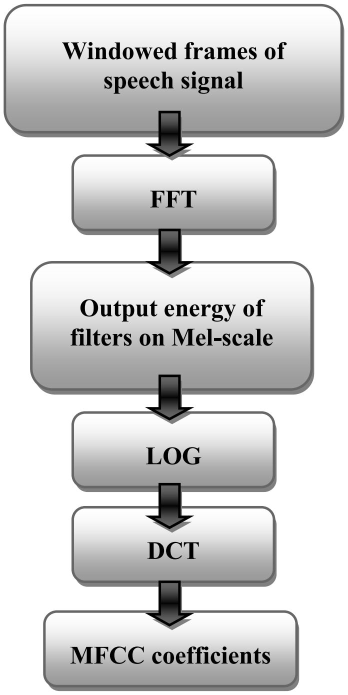

Windowed frames of speech signal
FFT
Output energy of filters on Mel-scale
LOG
DCT
MFCC coefficients

Figure 3: Steps of Calculating MFCC features.

B. Visual Front-End

The pre-processing on the images of the input video, visual feature extraction, and post processing are explained below. The mouth region within a rectangular window was detected as ROI. This was done by applying a classifier trained by the rapid and robust Viola-Jones object detection algorithm. These colored images are further transformed into gray-scale ones. By using appearance (pixel) based method, every pixel inside the detected ROI images was considered as a feature.

1) *Visual Pre-Processing*: We extract the visual features from mouth region; so, the Mouth region needs to be prepared first and this is done by some preprocessing steps which are briefly explained below:

- ***Face Detection*****:** Before we begin tracking a face, we need to first detect it. Matlab [22] is used to detect speaker's face. It has tools for object detection like the *vision.CascadeObjectDetect*or to detect the location of a face in a video frame. The cascade object detector uses the Viola-Jones detection algorithm and a trained classification model for detection. A rectangle around the face region is returned.

  The Visual Speech Recognition (VSR) system adopted the Viola-Jones detection module [23], which is much faster than any of its contemporaries via the use of an attention cascade using low feature number of detectors based on a natural extension of Haar wavelets. In this cascade, each detector fits objects to simple rectangular masks. In order to reduce the number of computations for such large number of cascades, Viola and Jones used the concept of integral image. They assumed that, for each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values above and to the left. The integral image can be computed quickly, which drastically improves the computation costs of the rectangular feature models. The attention cascade classifiers are trained on a training set as the authors have explained in [24]. As the computation progresses down the cascade, the features can get smaller and smaller, but fewer locations are tested for faces until detection is performed. The same procedure is adopted for mouth detection, except that object is different and search about mouth will be only on the lower half of the input image.

TABLE 1

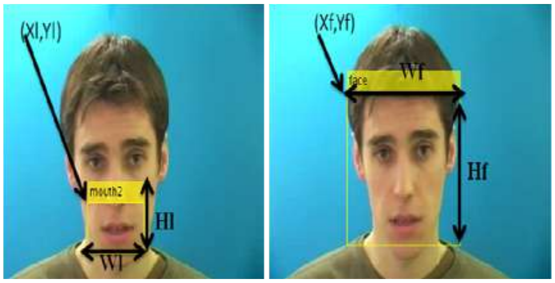
MOUTH LOCALIZATION ALGORITHM [25]

1. Grab the video frame for input.
2. Achieve the face detection and draw a box on the detected face then determined some of detection box (face) properties where:
   - the origin point
     - Xf: x-coordinate of the left border of face region
     - Yf: y-coordinate of the top border of face region
   - The width
     - Wf: the width of face region
   - The height
     - Hf: the height of face region
3. Detect the lip region is set as per the following calculations,
   - Xl=Xf+Wf/4
   - Yl=Yf+(2*Hf/3)
   - Wl=Wf/2
   - Hl=Hf/3

Where:

- Xl: x-coordinate of the left border of lip region
- Yl: y-coordinate of the top border of lip region
- Wl: the width of lip region
- Hl: the height of lip region

4. Xl,Yl,Wl and Hl are the values constituting of the lip region in the lip detection.
5. Repeat step 2, 3 and 4 for all frames.



(b) (a)

Figure 4 The mouth localizing algorithm (a) the box around face (b) the box around mouth in proportional to the face box

- **Mouth Detection:** Mouth localization algorithm based on deciding the bounding box around the mouth in a geometric way. Proportional to the bounding box around the face, the algorithm decides first the left corner point of the mouths bounding box. Then, the required size of this box can be drawn easily to the extent that encloses any possible lip movements. But the method introduced in [30] doesn't give accurate mouth detection method, so we adjust the mouth region to be calculated from the following formula:

```
Xl=fbox1(1)+((fbox1(3))*0.3);
Yl =fbox1(2)+(0.73*(fbox1(4)));
W1= (fbox1(3))*0.4;
Hl = (fbox1(4))*0.27;
mnbox=[Xl Yl W1 Hl];
```

Table 1 shows the proposed mouth localization algorithm which was explained in [25]. Figure 4 (a) and (b) show how the algorithm decides the box around the mouth in proportion to the box around the face. After getting the rectangle arround the mouth region we use imcrop Matlab function to crop the ROI around the mouth region.

- **Resize Mouth Region:** Mouth rectangle is resized to be in the form of 2^n where n is an integer. This operation is done in order not to make the calculation of DCT features be affected by the lip location in the input image. We choose the value of n to be 6 (64*64 pixels). Using Matlab *imresize* function is used to do this task.

- **Convert RGB to Grey:** The input mouth image is of RGB format. It is converted to grey format in range 0 (black) to 255 (white). We use *rgb2gray* function for implementing this process.

As simple output of each pre-processing step of visual front-end for an image from a speaker is shown in figure 5

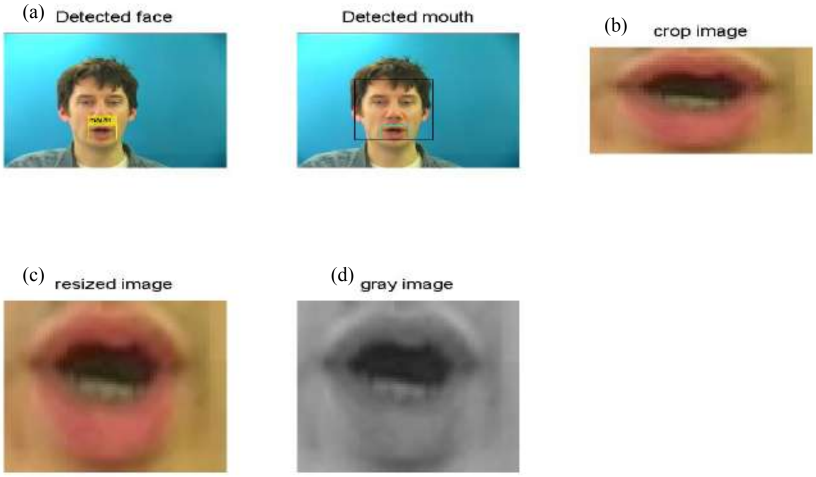


Figure 5: Visual pre-processing steps. a) Face and mouth detected regions, b) Mouth region only, c) Mouth region after resizing by 64x64, and d) Mouth region as gray scale.

C. Recognition Features Extraction for the visual signal

There are two main visual feature extraction categories that are appearance or pixel based and shape or model based. Examples of model based features are the width and the height of the speaker's lips. There is a loss of information because it depends on some information about the lips not the whole region [26]. Appearance based assumes that all mouth region pixels are informative to speech recognition [27].

Various visual features have been proposed in the literature. In general, feature extraction methods can be categorized into three kinds: 1. “pixel based” where features are employed directly from the image, 2. “lip contour based”, in which a prior template or model is used to describe the mouth area and 3. the combination of 1 and 2. Among these approaches, the one based on low level pixels is assumed to be the most efficient on [27]. As a typical method to extract pixel based features, image transforms such as Discrete Cosine Transform (DCT) [28], Principal Component Analysis (PCA) [29], Discrete Wavelet Transform (DWT) [28] and Linear Discriminant Analysis (LDA) [30] have been employed for lip-reading and have achieved high accuracy for visual-only recognition task. Among these, DCT has been shown to perform equally well or better than others.

Working at this pixel-based field faces a problem: How to reduce the high dimensional raw image data to low dimensional feature vectors without losing important information? Potamianos [29] retained the coefficients according to several sub lattices. Heckmann [32] compared 3 strategies to select the coefficients based on energy, variance and relative variance respectively and stated that the one based on energy performed best. Nefian [33] divided a 64 × 64 Region of Interest (ROI) into 64 blocks of size 8×8, and extracted the first 2x2 low frequency coefficients from each block. Projection using LDA to seek optimal classification performance in [33] can also be used for data dimensionality reduction, although this ability of LDA is limited to the number of classes. Motivated by the above studies, this paper focuses on the dimensionality reduction strategies for DCT based features for visual-only lip-reading task. In view of the excellent ability for information compression, PCA is applied to extract DCT coefficients. This combination is assumed to utilize the advantages of these two transforms. DCT is preferable to differentiate frequencies while PCA is beneficial to select the most ‘important’ components. Experimental results demonstrate that this new method does improve the speech reading performance when the final dimension is below a certain point, compared to the methods of selecting the coefficients according to specific criterion, such as ‘low frequency’.

The visual feature extraction task concentrated in the current work. Inspired by the cascade strategy by [34],

- The first stage is the image transform by using block based DCT. This step forms a 320 dimension vector by using blocked 8x8 2D DCT and then extract 5 elements from the upper left corner from each block using zigzag method as shown in figure 6.
- The second step is the dimensionality reduction procedure, using PCA form a final vector V in the final stage. Then V is used as a feature vector of the visual part of the system.
- In PCA we take the first 9 eign vectors which have highest values, the dimension reduced from 320 to 9.

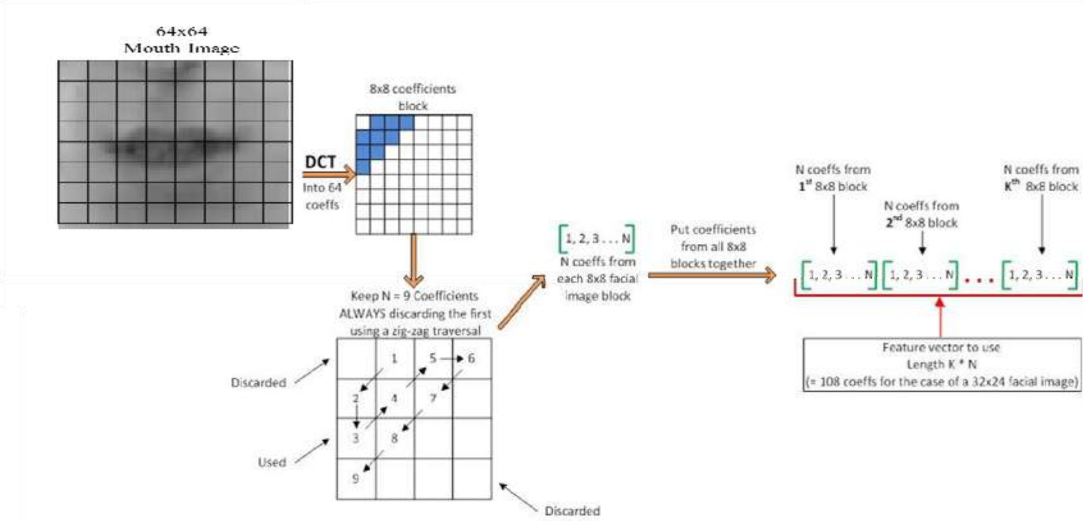

64x64
Mouth Image
DCT
Into 64 coeffs
8x8 coefficients block
Keep N = 9 Coefficients ALWAYS discarding the first using a zig-zag traversal
Discarded
Used
Discarded
1
2
3
4
5
6
7
8
9
[1, 2, 3 . . . N]
N coeffs from each 8x8 facial image block
Put coefficients from all 8x8 blocks together
N coeffs from 1st 8x8 block
N coeffs from 2nd 8x8 block
N coeffs from Kth 8x8 block
[1, 2, 3 . . . N] [1, 2, 3 . . . N] . . . [1, 2, 3 . . . N]
Feature vector to use
Length K * N
(= 108 coeffs for the case of a 32x24 facial image)

Figure 6: Extract feature vector from low frequency components from each block

D. Audio-Visual Features Integration

The features from different modalities (audio and visual features) have to be fused at some level. There are two different strategies to work with different types of features, early integration and late integration. In early integration (or what is called feature fusion), features from different sources are concatenated in one feature vector. The recognition process is applied on the combined feature vector. Late integration uses different or same classifiers for each feature type, and then the results of the classifiers are combined to get the final classification result. In this paper, early integration strategy is used by concatenating the acoustic and visual feature vectors on one vector. However, the audio and visual are with different frame rates, 44.1 KHz and 25 Hz for audio and video respectively, so linear interpolation is required first to up sample the video features rate to be with the same frame rate as audio features.

The Video features vectors are linearly distributed over the Audio features vectors to create the composite features vectors. The distribution is done by cloning the smaller set of Video features vectors in such that to build the same size array as such of the larger set Audio features vectors [18]. Then the composite features vectors are constructed by concatenating both arrays of features vectors (the cloned set video features vectors and the associated set Audio features vectors).

E. HMM Classification

Hidden Markov Model (HMM) is proven to be highly reliable classifier for speech recognition applications; most of the current successful systems for automatic speech recognition are based on Hidden Markov Models. The hidden Markov Model Toolkit by university of Cambridge (HTK) [21] is used for configuring, training and testing the HMM model are initialized using the Viterbi algorithm [35]. A total of 53 HMM models, one for each word, are trained in this paper. 44 phonemes are used to build each word's model. The proposed model uses 5-state left-to-right models with different number of Gaussian mixtures from 2 to 128 mixtures. Each state is multi Gaussian statistical model to express the observed symbols. In HTK, the conversion from single Gaussian HMMs to multiple mixture component HMMs is usually one of the final steps in building the model. The mechanism provided to do this is the HHED MU command which will increase the number of components in a mixture by a process called mixture splitting. This approach to building a multiple mixture component system is extremely flexible since it allows the number of mixture components to be repeatedly increased until the desired level of performance is achieved.

The MU command has the form: MU n itemList

where n gives the new number of mixture components required and itemList defines the actual mixture distributions to modify. This command works by repeatedly splitting the mixture with the largest mixture weight until the required number of components is obtained. The actual split is performed by copying the mixture, dividing the weights of both copies by 2, and finally perturbing the means by plus or minus 0.2 standard deviations e.g. MU 3 {*.state [2-4].mix}

It is usually increasing the number of mixtures then re-estimating, then incrementing by 1 or 2 again and re-estimating, and so on until the required numbers of components are obtained. This also allows recognition performance to be monitored to find the optimum mixture. Better start with a lesser number of mixtures and work way up. As one cannot go in the reverse direction, that is, there is no way to merge mixtures in HTK. So use single Gaussian models first then increment so as to reach a mixture of 8.

Performance analysis:

In order to analyze the system performance, HTK provides a tool HResult. It is used to compute the accuracy of the system. It compares the machine transcription of the test utterances with the corresponding reference transcription files. The performance of speech system is evaluated as:

$$\%Correct = \frac{N-D-S}{N} \times 100 = \frac{H}{N} \times 100 \qquad (1)$$

where N is the number of words in test set, D is the number of deletions, S is number of substitutions and H is the number of correct labels. %correct gives the percentage of word correctly recognized. The accuracy is computed as:

$$\%Accuracy = \frac{N-D-S-I}{N} \times 100 = \frac{H-I}{N} \times 100 \qquad (2)$$

where I is the number of insertions. The performance of speech recognition system can be evaluated by measuring the word error rate (WER) defined as:

$$\%Word\ Error\ Rate = \frac{S+I+D}{N} \times 100 = 100 - Accuracy \qquad (3)$$

3 EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to validate the effectiveness of the proposed design. Initially the performance of a baseline audio only recognition system is presented. Then, present the effect on the recognition accuracy of using visual features extracted from degraded video, and finally the result of audio visual system.

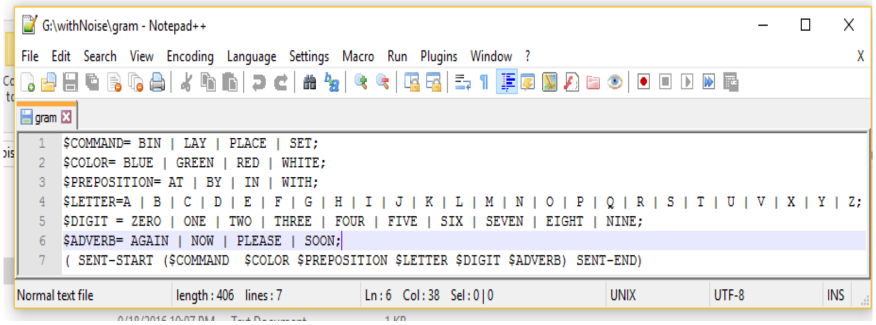
A. Data Description

To compare automatic audio visual speech recognition system performance based on the system discussed above; the GRID corpus [37] is used to perform these comparisons, which is a continuous audio-visual speech corpus for an English small vocabulary task. It contains 1000 sentences spoken by each of 34 speakers (18 male, 16 female) ages ranged from18 to 49 years. The original audio and video data were recorded under clean acoustic conditions, and the video shows only a frontal view of each subject's face. The sentences in GRID are speech commands according to a very simple grammar. Each sentence in this database contained six words including a command, color, preposition, letter, digit, and adverb. The total of 51 words within the vocabulary consist of 4 command words, 4 words representing color, 4 prepositions, 26 letters, 10 digits and 4 adverbs. Example sentences produced by a speaker in this database were "bin blue at A1 again" or "place green by D2 now". The video was recorded as a sequence of images with a frame legnth of 40ms. In the audio channel, the raw speech signal was converted into a sequence of vector parameters with a fixed 25ms frame legnth. Table 2 and figure 7 introduce the grammar file for the GRID corpus.

TABLE 2

SENTENCE STRUCTURE FOR THE GRID CORPUS [37]

Command	Color	Preposition	Letter	Digit	Adverb
BIN LAY PLACE SET	BLUE GREEN RED WHITE	AT BY IN WITH	A-Z excluding W	1-9, zero	AGAIN NOW PLEASE SOON


G:\withNoise\gram - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
gram
Normal text file
length : 406 lines : 7
Ln : 6 Col : 38 Sel : 0 | 0
UNIX
UTF-8
INS

```
$COMMAND= BIN | LAY | PLACE | SET;
$COLOR= BLUE | GREEN | RED | WHITE;
$PREPOSITION= AT | BY | IN | WITH;
$LETTER=A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | X | Y | Z;
$DIGIT = ZERO | ONE | TWO | THREE | FOUR | FIVE | SIX | SEVEN | EIGHT | NINE;
$ADVERB= AGAIN | NOW | PLEASE | SOON;
( SENT-START ($COMMAND  $COLOR $PREPOSITION $LETTER $DIGIT $ADVERB) SENT-END)
```

Figure 7: Grammar file for GRID corpus

- *The Experiment variables*

The experiment variables are listed below:

1- Number of the training and testing data used: We take 90% from the used database for training and 10% for testing
2- Audio Only Speech recognition system, Multimodal Speech recognition system.: We change the types of the feature used to check the improvement of adding the visual feature to the audio feature
3- Number of Gaussian Mixtures in HMM emitting states: We use 3 emitting HMM states and varying the numbers of the Gaussian mixtures from 2 to 128 to check if the increasing the numbers of the mixtures will increase the result or not.

First we take small amount of the database to make a small experiment so we take 100 files, 90 files for training and 10 for test.

- **Experiment** 1: speech recognition using acoustic features alone

In this experiment, we test audio only speech recognition system. HMM model with 3 emitting states and different Gaussian Mixtures in each state is used to model the recognition process. The average accuracy of using audio features only is summarized in figure 8 where the percentage correct against the experiment variables is represented, the qualifiers mono means monophone recognition, tri, means triphone recognition and Mix2 to Mix128 the number of Guassian mixtures increase from 2 to 128..

The parameters of each method used in figure 8 are explained in table 3 where A13 means Acoustic parameters with 13 MFCCs with 12 Mel cepstrum plus log energy and A39 mean the 13 elements with their delta (first order derivative) and acceleration (second order derivatives) coefficients.. Figure 8 proves that the audio only recognizer achieves high recognition rate for 39 feature vector size than using 13 elements for the feature vector, and the increase of mixtures number enhances the performance of recognition process for the two sizes of the feature vector. The optimal number of mixtures for A13 is 4 mixtures where it is 2 in A39. It gives us that increasing the feature vector size from 13 to 39 gives enhancement for the recognition rate by 6.6%.

TABLE 3

RESULT OF AUDIO ONLY SPEECH RECOGNITION SYSTEM

	Feature Type	Vector length	Best Result
A13	Audio only with MFCC_0	13	80% with Mix4
A39	Audio only with MFCC_0_D_A	39	86.67% with Mix2

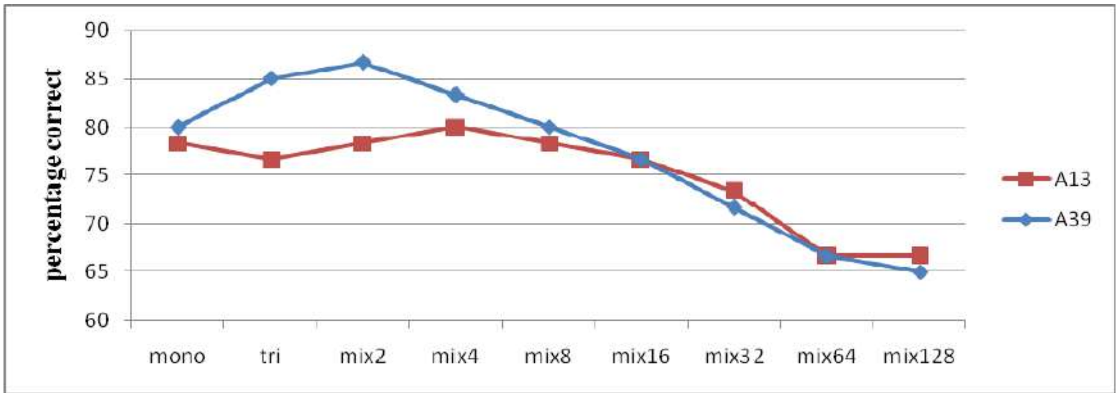


Figure 8: Percentage correct for audio only speech recognition with different mixture

- **Experiment** 2: speech recognition using Audio Visual features

The audio features and visual cues contain information related to speech production and combining these two signal streams can improve recognition accuracy; so we combined visual and acoustic features for dataset. In this experiment we check the change of the image resize effect from 64x64 and 128x128 the result is shown in table 3 which explains that resizing the image with 64x64 gives better results.

Change in the visual feature vector length from 2, 3, 5 and 6 the DCT matrix methods are applied and compared. The parameters of each method used in figure 9 are explained in table 3, where *AV45_128* means audio visual features with 45 feature vector size and 128x128 image size. The results indicate the best recognition obtained by using 64x64 image size with the DCT and PCA feature extraction. From table 4 we can see that best result is obtained when using audio visual feature with size of vector 42 and image size 64x64 with 91.67% recognition rate.

TABLE 4

RESULT OF AUDIO VISUAL SPEECH RECOGNITION

	Feature Type		Vector length	Image size	Best Result
	Audio	**Video**			
AV45_128=6+39	MFCC_0_D_A	Blocked DCT	39 audio +6 visual	128x128	76.67% At triphone and mix2
AV44_64=5+39	MFCC_0_D_A	Blocked DCT	39 audio +5 visual	64x64	80% at triphone
AV41_64=2+39	MFCC_0_D_A	Blocked DCT	39 audio +2 visual	64x64	85% At triphone and mix2
AV42_64=3+39	MFCC_0_D_A	Blocked DCT	39 audio +3 visual	64x64	91.67% at Mix2
AV45_64=6+39	MFCC_0_D_A	Blocked DCT +PCA	39 audio +6 visual	64x64	88.33% at triphone

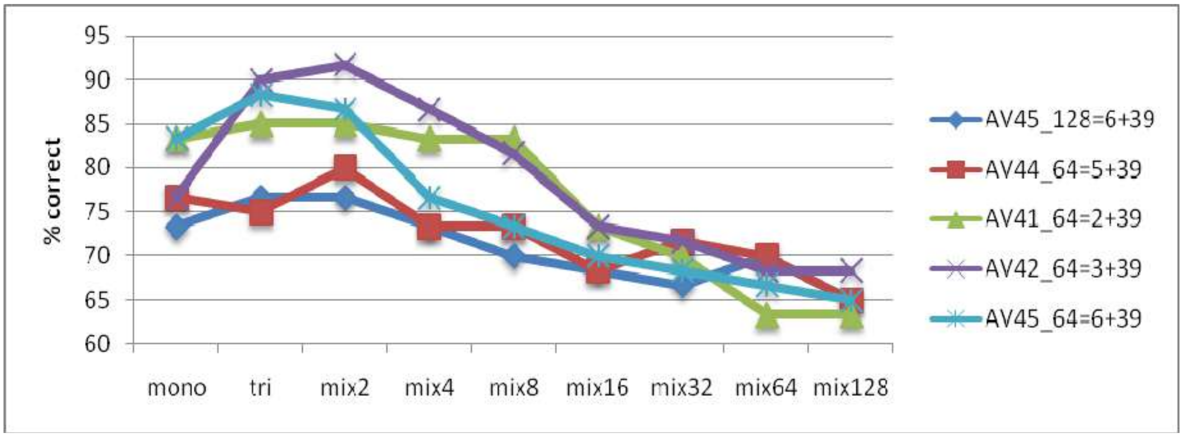


Figure 9: Percentage correct for audio visual speech recognition with different mixtures

- **Experiment 3**: speech recognition using acoustic features only Vs Audio visual feature Vs Video only speech recognition for 100 files

In this experiment, we compare the three feature types to verify the effeteness of the proposed system by obtaining the result for using audio only, video only and the combination of them in audio-visual feature. The results prove that using the audio visual system with blocked DCT visual feature gives better result with 90% recognition rate , in the second stage 88.33% by using audio-visual with blocked DCT and PCA. It means that using visual features with the audio features improve the result with 3.4%. Figure 10 explains the obtained results of getting different visual features and with different sizes for different mixtures.

TABLE 5

RESULT OF THE AUDIO ONLY VS VIDEO ONLY VS AUDIO-VISUAL SYSTEM

	Feature Type		Vector length	Image size	Best Result
	Audio	**Video**			
A39	MFCC_0_D_A	‐	39 audio only	64x64	86.67% At mix2
AV45=39+6	MFCC_0_D_A	DCT	39 audio+6 visual	64x64	90% At triphone
AV45=39+9NewPCA	MFCC_0_D_A	DCT+PCA	39 audio+9 visual	64x64	88.33% At mix2
V6	‐	DCT+PCA	6 visual only	64x64	70% At triphone

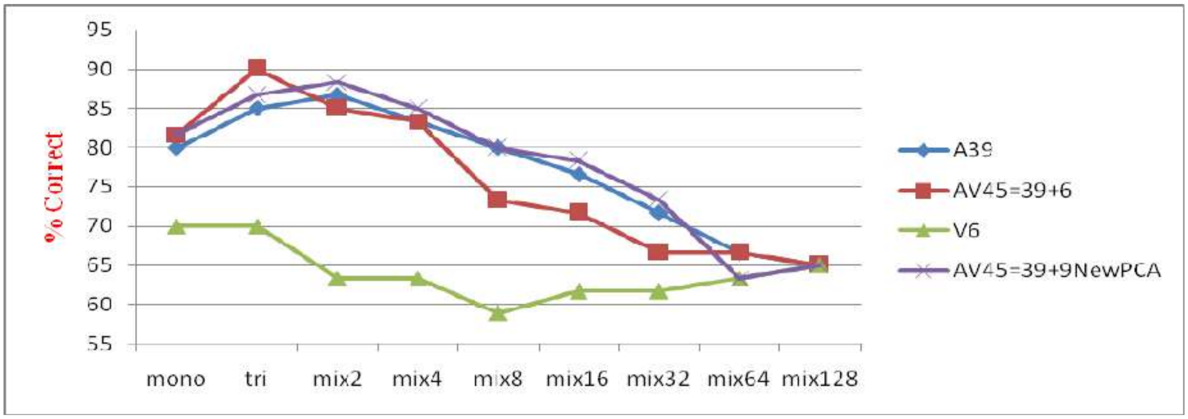


Figure 10: percent correct for Audio only Vs Video Only Vs Audio-visual

- **Experiment 4**: speech recognition using acoustic features only Vs Audio visual feature Vs Video only speech recognition for total database.
  From the results obtained we can see that using visual features which are extracted by blocked DCT and PCA with the audio features give enhancement for the performance of the recognition process and give more enhancement in case of noisy system as shown in figure 11.

TABLE 6

RESULT FOR AUDIO ONLY AND DIFFERENT AUDIO VISUAL TECHNIQUES FOR TOTAL GRID DATABASE

A 39	**MONO**	**TRI**	**MIX2**	**MIX4**	**MIX8**	**MIX16**	**MIX32**	**MIX64**	**MIX128**
WORD: %Corr	72.13	92.82	94.01	95.54	96.79	97.73	98.45	98.81	99.25
SENT: %Correct	15.99	64.11	68.79	75.77	82.04	86.81	90.96	93.6	95.51
AV 39+3zigzag									
WORD: %Corr	71.61	91.74	93.35	95.03	96.66	97.83	98.57	99.04	99.35
SENT: %Correct	15.28	59.75	65.99	73.65	81.49	87.7	91.94	94.53	96.22
AV 39+6(5DCTzigzag > 6PCA)									
WORD: %Corr	72.27	91.89	93.29	95.29	96.68	97.71	98.5	99	99.28
SENT: %Correct	16.3	60.82	66.73	75.22	81.89	87.3	91.39	94.16	95.73

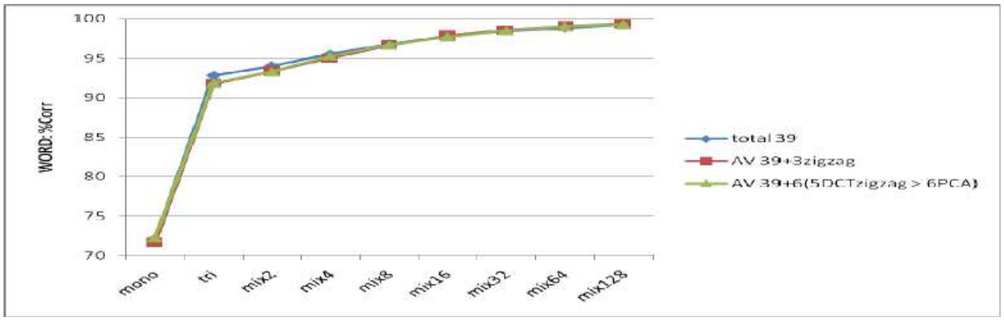


Figure 11: percent correct for Audio only Vs Video Only Vs Audio-visual for total GRID database

4 CONCLUSION

In this paper, we provided a brief overview of the basic techniques for automatic recognition of audio-visual speech, proposed in the literature over the past twenty years, with particular emphasis in the algorithms used in our speech reading system. The two main issues relevant to the design of audio-visual ASR systems are: First, the visual front end that captures visual speech information and, second, the integration (fusion) of audio and visual features into the automatic speech recognizer used. Both are challenging problems, and significant research effort has been directed towards finding appropriate solutions. This study investigates the effect of adding Discrete Cosine Transform

Coefficients DCT of mouth region as visual features which dimensionality are reduced by using Principle component analysis PCA with audio features. The proposed system is tested on the standard database, GRID sentence database. Speaker dependent and speaker independent experiments are tested and change DCT visual feature size and using PCA are applied and compared. It was found that adding the whole upper left corner region of DCT coefficients matrix with using PCA can improve the performance of AVASR. From the experiments given in this paper we find that the optimal number of audio vector size is 39 it gives enhancement for the recognition rate by 6.6%. The results indicate that the best recognition is obtained by using 64x64 image size with the blocked DCT and PCA feature extraction with best result obtained when using audio visual feature with size of vector 42 and image size 64x64 which is 91.67% recognition rate. It means that using visual feature with the audio feature improves the result with 5%. Also when testing the system under noisy environment it improves the result.

REFERENCES

[1] A.N. Mishra, Mahesh Chandra, Astik Biswas, and S.N. Sharan, "Hindi phoneme-viseme recognition from continuous speech*", International Journal of Signal and Imaging Systems Engineering (IJSISE)*, Vol. 6, No. 3, pp. 164-171, 2013.

[2] Estellers, Virginia, Thiran, and Jean-Philippe, "Multi-pose lip reading and audio-visual speech recognition", *EURASIP Journal on Advances in Signal Processing*, pp.1-23, 2012.

[3] Burnham, Douglas, et al., eds. Hearing Eye II: The Psychology Of Speech reading And Auditory-Visual Speech. Psychology Press, 2013.

[4] Massaro, Dominic W., and Jeffry A. Simpson. Speech perception by ear and eye: A paradigm for psychological inquiry. Psychology Press, 2014.

[5] Stork, David G., and Marcus E. Hennecke, eds. *Speech reading by humans and machines: models, systems, and applications*. Vol. 150. Springer Science & Business Media, 2013.

[6] Dean, David Brendan. "Synchronous HMMs for audio-visual speech processing." (2008).

[7] Chitu, Alin G., and Leon JM Rothkrantz. "Visual speech recognition." *Information Technologies and control, Sofia, Bulgaria, vol. year vii* 3 (2010): 2-9.

[8] Moore, Brian CJ. An introduction to the psychology of hearing. Brill, 2012.

[9] Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: a tutorial review." *IEEE Signal Processing Magazine* 32.6 (2015): 74-99.

[10] Gong, Yifan. "Speech recognition in noisy environments: A survey." *Speech communication* 16.3, pp.261-291,1995.

[11] Ross, Lars A., et al. "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments." Cerebral Cortex 17.5 (2007): 1147-1153.

[12] Iwano, Koji, Satoshi Tamura, and Sadaoki Furui. "Bimodal speech recognition using lip movement measured by optical-flow analysis." *International Workshop on Hands-Free Speech Communication*. 2001.

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech, "*Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326, 2003.

[14] Davis, Chris, and Jeesun Kim. "Audio-visual speech perception off the top of the head." Cognition 100.3 (2006): B21-B31.

[15] Bailly, Gerard, Pascal Perrier, and Eric Vatikiotis-Bateson. *Audiovisual speech processing*. Cambridge University Press, 2012.

[16] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*., vol. 26, pp. 23–43, 1998.

[17] J. Jiang, A. Alwan, P. A. Keating, B. Chaney, E. T. Auer Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics, " *EURASIP J. Appl. Signal Processing*, vol. 2002, pp. 1174–1188, Nov. 2002.

[18] Hennecke, Marcus E., David G. Stork, and K. Venkatesh Prasad. "Visionary speech: Looking ahead to practical speechreading systems." *Speech reading by Humans and Machines*. Springer Berlin Heidelberg, 1996. 331-349.

[19] Salama, Elham S., Reda A. El-Khoribi, and Mahmoud E. Shoman. "Audio-Visual Speech Recognition for People with Speech Disorders." *International Journal of Computer Applications* 96.2 (2014).

[20] Tiwari, Vibha. "MFCC and its applications in speaker recognition." *International Journal on Emerging Technologies* 1.1 (2010): 19-22.

[21] Steve Young, Mark Gales, Xunying Andrew Liu, Phil Woodland, et al." The HTK Book" ,Version 3.41 ,Cambridge University Engineering Department,2006 , http://www.htk.eng,cam.ac.uk

[22] http://www.mathwork.com

[23] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE, 2001.

[24] Viola, P., Jones, M. "Robust real-time object detection." *The IEEE Transactions on Computer Vision* 57(2), 137–154 (2004)

[25] Sagheer, Alaa. "Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications." Applied Mathematics & Information Sciences 9.6 (2015): 2885.

[26] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech", Proc. IEEE, Vol. 91, No. 9, pp.1306–1326, Sep, 2003.

[27] P. Scanlon and G. Potamianos, "Exploiting lower face symmetry in appearance-based automatic speech reading." *Proc. Works. Audio-Visual Speech Process. (AVSP)*, pp. 79–84, 2005.

[28] Matthews, etc. "Extraction of Visual Features for Lip reading.*" IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 2, February 2002.

[29] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. "Audio-visual speech recognition, final workshop report", *Center for Language and Speech Processing*, 2000.

[30] G. Potamianos, H.P. Graf, and E. Cosatto. "An image transform approach for HMM based automatic lip reading." *Proc. Int. Conf. Image Process*., Chicago, pp.173-177, 1998

[31] G. Chiou and J. Hwang. "Lip reading from Color Video." *IEEE Trans. on Image Processing*, 6(8) , pp.1192-1195, August (1997).

[32] P. Duchnowski, etc. "Toward movement-invariant automatic lip-reading and speech recognition."*Proc. Int. Conf. Acoust. Speech Signal Process*., Detroit, pp109-11,1995.

[33] M. Heckmann, etc. "DCT-based video features for audio-visual speech recognition." *Proc. Int. Conf. Spoken Lang. Process. Denver*, USA. September 2002. 9, pp. 1925-1928

[34] A.V. Nefian, etc. "Dynamic Bayesian networks for audio-visual speech recognition." *EURASIP Journal on Advanced Application in. Signal Processing*, Nov.2002, pp.1274-1288.

[35] G. Potamianos, etc. "A cascade image transform for speaker independent automatic speech reading." *IEEE International Conference on Multimedia and Expo*, 2000, Volume: 2, pp: 1097 -1100.

[36] Gales, Mark, and Steve Young. "The application of hidden Markov models in speech recognition." Foundations and trends in signal processing 1.3 (2008): 195-304.

[37] Cooke, Martin, et al. "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America* 120.5, pp.2421-2424, 2006.

[38] Radha, V., and C. Vimala. "A review on speech recognition challenges and approaches." doaj. org 2.1 (2012): 1-7.

BIOGRAPHY

Amr M. Gody received the B.Sc. M.Sc., and PhD. from the Faculty of Engineering, Cairo University. Egypt, in 1991, 1995 and 1999 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Fayoum University, Egypt in 1994. He is author and co-author of about 40 papers in national and international conference proceedings and journals. He is the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012, 2013 and 2014. His current research areas of interest include speech processing, speech recognition and speech compression.

Mohamed H. Farouk received the B.Sc. in Electronics Engineering from the Faculty of Engineering, Cairo University, Egypt, in 1982. He received the MSc and PhD. of Engineering Physics from the Faculty of Engineering, Cairo University, Egypt, in1988 and 1994 respectively. He joined the teaching staff of the Electrical Engineering Department, Faculty of Engineering, Cairo University, Egypt in 1984. His Current Position is full Professor, Engineering Math. & Physics Dept., Faculty of Engineering, Cairo Univ. from 2007-Till Now. He is author and co-author of about 40 papers in national and international conference proceedings and journals.

Eslam E. El Maghraby received the BSc (Honours) degree in communication and electronics from faculty of engineering, Fayoum University in 2008. She received the MSc degree in speech recognition systems from faculty of engineering, Fayoum University in 2013. She is currently a PhD student at the Faculty of Engineering-Fayoum University. She is working as Assistant Lecturer at Information system department at Faculty of Computers and Information, Fayoum University. Her research interest is in signal processing and computer networks.

TRANSLATED ABSTRACT

تحسين جودة ودقة أنظمة التعرف على الكلام باستخدام اشارة الكلام الصوتية والبصرية

***اسلام المغربى، *عمرو جودى، **محمدهشام فاروق**

**قسم الهندسة الكهربية- كلية الهندسة- جامعة الفيوم*
***قسم الرياضيات و الفيزيقا الهندسية- كلية الهندسة- جامعة القاهرة*

ملخص

بالرغم من الجهود المبذوله خلال العقود الماضيه للوصول إلي أعلي درجات التعرف علي الأصوات مازالت الانظمة التي تم الوصول إليها غير دقيقه وغير مناسبة للتطبيقات الحياتيه الحقيقية وخصوصاً تلك التي توجد في أوساط بها الكثير من الضوضاء .معظم أنظمة التعرف علي الاصوات تعتمد علي الاشارة الصوتيه كمصدر وحيد للصوت وتقوم بإهمال الجزء البصري المصاحب له. الدمج بين الاشارة الصوتيه والبصريه المصاحبة للصوت يقدم وعودا بالوصول لدرجة أعلي للتعرف علي الصوت ودقه افضل من التي يمكن الحصول عليها من خلال استخدام الاشارة الصوتيه فقط. هذا البحث يهدف إلي بناء نظام للتعرف علي الاصوات بإستخدام الاشارة الصوتية بالاضافة الي الاشارة البصريه المصاحبة للصوت لمجموعة من الجمل التي تحتوي علي كلمات منطوقة باللغه الانجليزيه. يتم استخراج خصائص الاشارة الصوتيه باستخدام خاصية MFCC واستخدام خاصية DCT لاستخراج الخصائص من الصورة المصاحبة للصوت و خاصية PCA استخدمت لغرض تقليل حجم الخصائص المستخرجه من الصورة وذلك لسهولة التعامل معها و للعمل علي سرعة أداء النظام المقترح. يتم دمج الخصائص المستخرجه من الصوت والصوره المصاحبه له لتدريب نظام التعرف علي الاصوات باستخدام HMM للتعرف علي الكلمات المنطوقة عن طريق استخدام اداة HTK. تم اختبار كفاءة النظام المقترح من خلال تطبيقة باستخدام واحدة من اكبر قواعد البيانات للصوت والصورة معا وهي قاعدة بيانات GRID. من خلال تحليل النتائج للنظام المقترح المعتمد علي الصوت والصوره معا نجد انه يقدم كفاءة اعلي ومعدل تعرف اكبر بمقدار 4% عن استخدام الصوت فقط.

MASAR: A Morphologically Annotated Gold Standard Arabic Resource

SamehAlansary

Bibliotheca Alexandrina, Alexandria, Egypt

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

sameh.alansary@bibalex.org

Abstract**—Arabic morphology poses special challenges to computational natural language processing systems. Its rich morphology and the highly complex word formation process of roots and patterns make computational approaches to Arabic very challenging. In spite of the recent progress, Arabic is still lacking the necessary tools and annotated resources, and Arabic Natural Language Processing (NLP) is still in its infancy. In this paper we present a Morphologically Annotated Gold Standard Arabic Resource (MASAR) which is planned to help in building Arabic NLP systems.The used data are selected form the International Corpus of Arabic covering different sources and genres. The analysis of this data follows the stem-based approach depending on Buckwalter Arabic Morphological Analyzer. In building MASAR, more detailed and new tags were added, more features were inserted and other features have been handled.**

1 Introduction

The importance of corpora to language and linguistics studies is parallel to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those based upon the individual's own internalized cognitive perception of language. Because language and linguistics studies cannot rely on intuition or small samples of language, they require empirical analysis of large databases of texts as in the corpus-based approach. Since corpora consist of texts, they enable linguists to contextualize their analyses of language. Hence, corpora are very well suited to more functionally based discussions of language and linguistics.

In spite of the recent progress, Arabic is still lacking the necessary tools and annotated resources, and Arabic NLP is still in its infancy. Although tagged corpora are of great importance in extracting grammatical and linguistic information and training machine learning algorithms, and although their information is useful for Natural Language Processing applications such as text indexing, information retrieval and speech processing, they are not freely available for research.

Arabic is a language of rich morphology compared to other languages especially European languages. It is based on both derivational and inflectional morphology. The richness of Arabic morphology makes the analysis process difficult to deal with[1].

Morphological analysis for text corpora is a prerequisite for many text analytics applications, which has attracted many researchers from different disciplines such as linguistics (computational and corpus linguistics), artificial intelligence, and natural language processing, to morpho-syntactically analyze text of different languages including Arabic. Recently, several researchers have investigated different approaches to morphological and syntactic analysis for Arabic text. Many systems have been developed which vary in complexity from light stemmers, root extraction systems, lemmatizers, complex morphological analyzers, part-of-speech taggers and parsers [2].

All these issues give rise to the need to have a rich morphological annotated resource for Modern Standard Arabic (MSA) since it is currently the sixth official language of the United Nations and it is also one of the most widely spoken languages in the world with estimated 422 million native speakers. In this paper, the focus is on building a Morphologically Annotated Gold Standard Arabic Resource (MASAR) which is built during the process of developing the International Corpus of Arabic (ICA) and it will be available soon for all researchers to be accessed and used.

In what follows, section 2 discusses the issues of compiling and analyzing the data morphologically, starting from data description, analysis stage tools and procedures. Section 3 reviews the description of the used tag sets and the available information (current state). Section 4 reviews one of the most prominent related works, Penn Arabic Tree Bank (PATB), and sets a comparison between it and the current resource. Finally, section 5 concludes the paper.

2 Data Compilation and Analysis

Developing MASAR began at Bibliotheca Alexandrina (BA) in 2006 during building the “International Corpus of Arabic (ICA)”, a serious effort to build a representative Arabic corpus as being used all over the Arab world to support research on Arabic. The collection of samples is limited to written Modern Standard Arabic selected from a wide range of sources designed to represent a wide cross-section of Arabic [3].

The first goal of the ICA is to morphologically analyze the collected corpus [4]. In 2007, the morphological analysis stage began and it was preferred to develop our own morphologically annotated gold standard analyzed resource to be used while analyzing the whole ICA data, since it contains more information and details than any other annotated data. It has two releases; the first one consists of about 500 thousand manually annotated words and the second one consists of about 1.5 million automatically annotated words using BASMA [5]that are verified for quality assurance. In this paper, the first release is our concern. The following is the description of the selected data of the first release and the issues that were faced during the analysis process:

A) Data Description:

In the first release of MASAR, 1,111 text documents were selected from ICA corpus from texts that were published in 2006-2007. It contains 570,137 tokens; 69,937 are punctuations, numbers, and Latin strings and 500,200 Arabic word tokens representing 81,487 word types. These texts are selected from different sources in ICA; Press, Net Articles and Books as figure 1 shows. Moreover, these selected texts covered more than one genre as Table 1 shows. In Press Source, the texts are selected from news papers, magazines and electronic press that cover different countries as figure 2 shows.

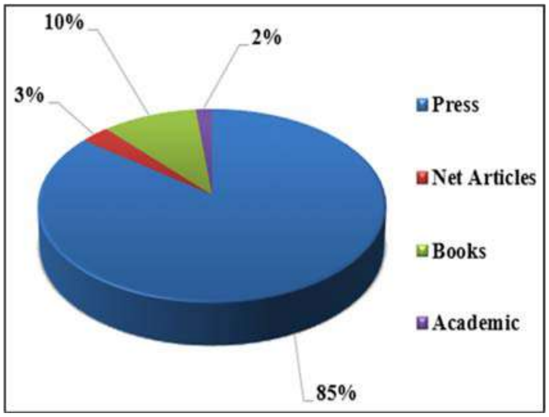


Figure 1. MASAR Texts Distribution According to Sources.

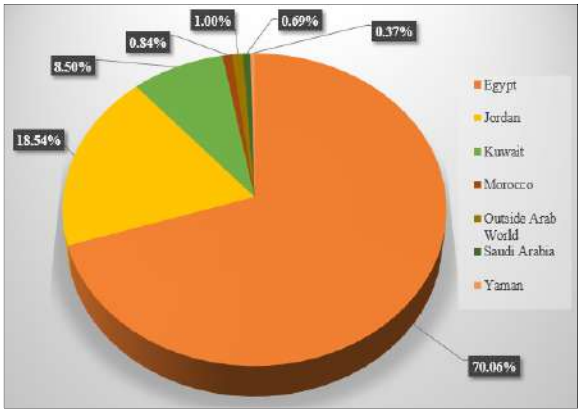
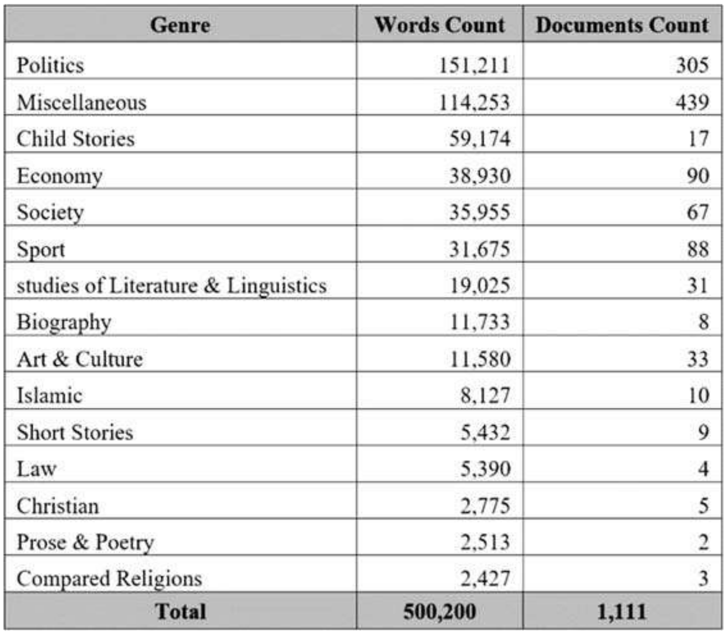


Figure 2. The Distribution of MASAR Words of the Press Source According to Countries.

TABLE 1
MASAR TEXTS DISTRIBUTION OVER THE GENRES



Genre	Words Count	Documents Count
Politics	151,211	305
Miscellaneous	114,253	439
Child Stories	59,174	17
Economy	38,930	90
Society	35,955	67
Sport	31,675	88
studies of Literature & Linguistics	19,025	31
Biography	11,733	8
Art & Culture	11,580	33
Islamic	8,127	10
Short Stories	5,432	9
Law	5,390	4
Christian	2,775	5
Prose & Poetry	2,513	2
Compared Religions	2,427	3
Total	**500,200**	**1,111**

Before the morphological analysis process, each text is preprocessed and marked up with some structural markup such as beginning and end of document, title, paragraph or question.

B) Analysis Stage: Tools and Procedures:

1. Resources and Morphological Analyzer

The stem-based approach (concatenative approach) has been adopted as the linguistic approach to analyze the ICA. The second version of Buckwalter Morphological Arabic Morphological Analyzer (BAMA 2.0) [6] has been selected since it is a well-known analyzer in the literature and has even been considered as the "most respected lexical resource of its kind" [7]. It is used in LDC Arabic POS-tagger, Penn Arabic Dependency Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms interacting with other concatenation databases. Every word form is entered separately, and the stem is used as the base form. The word is viewed as composed of a basic unit that can be combined with morphemes governed by morph tactic rules. It makes use of three lexicons: A Prefixes lexicon, a Stem lexicon, and a Suffixes lexicon [4]& [8].

Although Buckwalter has many advantages including its ability to provide a lot of information such as Lemma, Vocalization, Part of Speech (POS), Gloss, Prefix(s), Stem, Word Class, Suffix(s), Number, Gender, Definiteness and Case or Mood, it does not always provide all the information the ICA requires, and in some cases, the provided analyses would need some modification. It may give the right solution for the Arabic input word, provide more than one result that needs to be disambiguated to reach the best solution, provide many solutions, but none of them is right, segment the input words wrongly without taking the segmentation rules in consideration or provide no solutions at all. Consequently, solutions enhancement is needed in these situations. For more details about BAMA problems and how these problems have been handled see [4] &[9].

As a result of BAMA's problems, the number, gender, and definiteness features need to be modified according to their morphosyntactic properties. Some tags had been added to BAMA's lexicons, some lemmas and glossaries had been modifiedand others had been added [4], [9] & [10]. In addition, three new features had been used while developing MASAR:

- ***Name entities***: The name entity feature has been identified by adding the feature (NE) to the words that carry this feature. By adding this feature, researchers can easily identify name entities and determine their correct part of speech according to context. For example, "الولايات المتحدة الأميركية" "AlwilAyAtuAlmut~aHidapu Al>miyrokiy~apu"[1] "the United States of America" appears in the analysis as table2 shows:

TABLE 2
AN EXAMPLE OF NAME ENTITY

الولايات	NOUN	NE
المتحدة	ADJ	NE
الأميركية	ADJ	NE

- **Root:** It has been noticed that the root feature does not appear in BAMA's output, although it is found in its dicStems lexicon. Moreover, unfortunately not all of the roots that are available in this lexicon are Arabic roots and other roots are wrongly detected, so there has to be some modifications in these roots. After reviewing all roots in the dicStems lexicon, some modifications were needed in BAMA's AraMorph Perl file to show the root feature in BAMA's output.
- **Stem Pattern:** Although the stem pattern is not used in BAMA's lexicon at all, it is a very useful feature that may help the researcher in detecting the number, gender and case ending of some Arabic words automatically. The stem patterns have been detected automatically, in some cases depending on root and stem and in other cases depending on root, lemma, and stem. Then, these stem patterns have been added and mapped in the dicStems lexicon. Moreover, some modifications were needed in BAMA's AraMorph Perl file to show the stem pattern feature in BAMA's output.

Moreover, 484,595 words representing (96.88%) were provided with a suitable morphological analysis from BAMA, and 15,605 words representing (3.12%) were not provided with a suitable morphological analysis from BAMA for one of two reasons, either because BAMA does not provide any analysis, or because none of the solutions BAMA provides are suitable for the word's context.These words are manually analyzed according to their contexts, then they are added to BAMA's lexicon to be analyzed in the same manner as they would be if they have been analyzed automatically by BAMA.

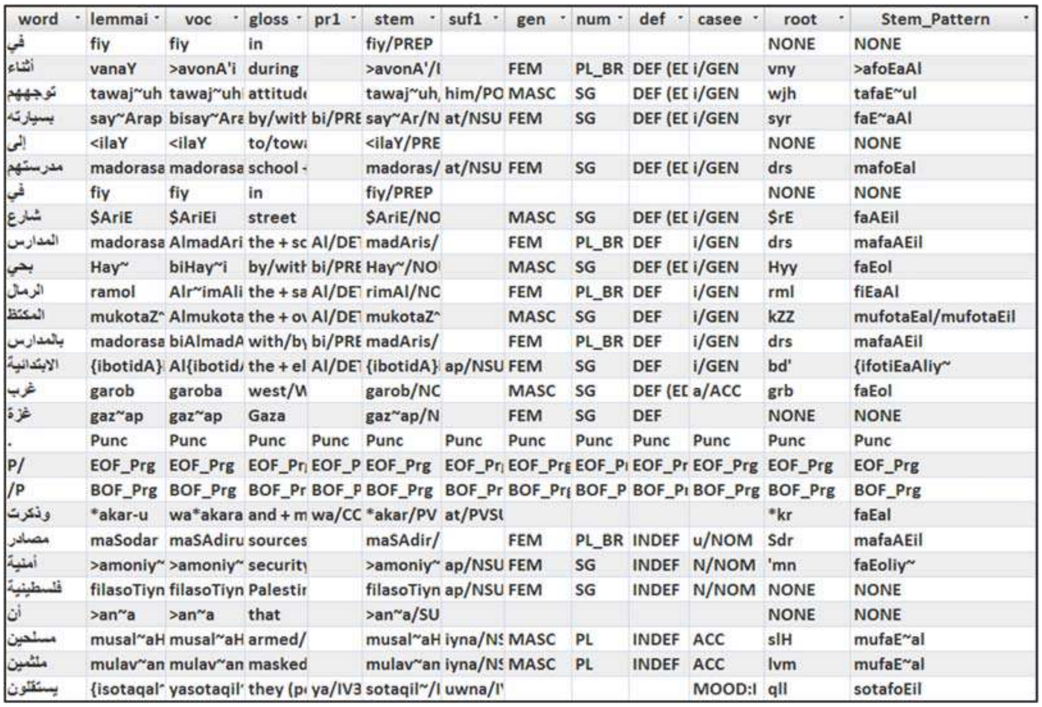
[1] The transliteration scheme follows that of BAMA: http://www.qamus.org/transliteration.htm[Accessed 1-11-2016].

2. Morphological Disambiguation Tool

In order to improve the speed and accuracy of the manual morphological annotation, an interface that allows the annotators to concentrate on the task of providing the best morphological analysis of each word according to its context has been developed. All BAMA's solutions have been parsed in a certain way that separates the clitics automatically based on each POS so that each feature is shown separately on the interface.These features are prefixes, stem, suffixes, glossary, number, gender, definiteness, lemma, case ending, root, pattern, and vocalization. In this interface, the annotator may leave a comment that the word has been disambiguated manually or that the word has a spelling mistake.

3. Morphological Disambiguation Procedure and Quality Control

The morphological annotation procedure is to use the automatic developed interface to provide a pass through the data. BAMA's handled lexicon and morphological analyzer is used to generate a candidate list of "POS tags" for each word. The POS morphological annotation task is to select the suitable POS tag from the list of alternatives provided (whether BAMA's solutions or manual words). Once the annotation process is done, the annotated files are saved in a database in a way where each feature is saved separately in order to ease the next stages of syntactic and semantic analysis processes as figure 3 shows.



word	lemmai	voc	gloss	pr1	stem	suf1	gen	num	def	casee	root	Stem_Pattern
في	fiy	fiy	in		fiy/PREP						NONE	NONE
أثناء	vanaY	>avonA'i	during		>avonA'/I		FEM	PL_BR	DEF (EI	i/GEN	vny	>afoEaAl
توجههم	tawaj~uh	tawaj~uh	attitud		tawaj~uh,	him/PO	MASC	SG	DEF (EI	i/GEN	wjh	tafaE~ul
بسيارته	say~Arap	bisay~Ara	by/with	bi/PRE	say~Ar/N	at/NSU	FEM	SG	DEF (EI	i/GEN	syr	faE~aAl
إلى	<ilaY	<ilaY	to/tow		<ilaY/PRE						NONE	NONE
مدرستهم	madorasa	madorasa	school -		madoras/	at/NSU	FEM	SG	DEF (EI	i/GEN	drs	mafoEal
في	fiy	fiy	in		fiy/PREP						NONE	NONE
شارع	$AriE	$AriEi	street		$AriE/NO		MASC	SG	DEF (EI	i/GEN	$rE	faAEil
المدارس	madorasa	AlmadAri	the + sc	Al/DET	madAris/		FEM	PL_BR	DEF	i/GEN	drs	mafaAEil
بحي	Hay~	biHay~i	by/with	bi/PRE	Hay~/NO		MASC	SG	DEF (EI	i/GEN	Hyy	faEol
الرمال	ramol	Alr~imAli	the + sa	Al/DET	rimAl/NC		FEM	PL_BR	DEF	i/GEN	rml	fiEaAl
المكتظ	mukotaZ~	Almukota	the + ov	Al/DET	mukotaZ~		MASC	SG	DEF	i/GEN	kZZ	mufotaEal/mufotaEil
بالمدارس	madorasa	biAlmadA	with/by	bi/PRE	madAris/		FEM	PL_BR	DEF	i/GEN	drs	mafaAEil
الابتدائية	{ibotidA}	Al{ibotid	the + el	Al/DET	{ibotidA}	ap/NSU	FEM	SG	DEF	i/GEN	bd'	{ifotiEaAliy~
غرب	garob	garoba	west/W		garob/NC		MASC	SG	DEF (EI	a/ACC	grb	faEol
غزة	gaz~ap	gaz~ap	Gaza		gaz~ap/N		FEM	SG	DEF		NONE	NONE
.	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc	Punc
P/	EOF_Prg	EOF_Prg	EOF_Pr	EOF_P	EOF_Prg	EOF_Pr	EOF_Prg	EOF_P	EOF_Pr	EOF_Prg	EOF_Prg	EOF_Prg
/P	BOF_Prg	BOF_Prg	BOF_Pr	BOF_P	BOF_Prg	BOF_Pr	BOF_Pr	BOF_P	BOF_Pr	BOF_Prg	BOF_Prg	BOF_Prg
وذكرت	*akar-u	wa*akara	and + m	wa/CC	*akar/PV	at/PVSU					*kr	faEal
مصادر	maSodar	maSAdiru	sources		maSAdir/		FEM	PL_BR	INDEF	u/NOM	Sdr	mafaAEil
أمنية	>amoniy~	>amoniy~	security		>amoniy~	ap/NSU	FEM	SG	INDEF	N/NOM	'mn	faEoliy~
فلسطينية	filasoTiyn	filasoTiyn	Palestir		filasoTiyn	ap/NSU	FEM	SG	INDEF	N/NOM	NONE	NONE
أن	>an~a	>an~a	that		>an~a/SU						NONE	NONE
مسلحين	musal~aH	musal~aH	armed/		musal~aH	iyna/NS	MASC	PL	INDEF	ACC	slH	mufaE~al
ملثمين	mulav~an	mulav~an	masked		mulav~an	iyna/NS	MASC	PL	INDEF	ACC	lvm	mufaE~al
يستقلون	{isotaqal~	yasotaqil~	they (p	ya/IV3	sotaqil~/I	uwna/I				MOOD:I	qll	sotafoEil

Figure 3. Sample of ICA Gold Standard Resource

The data of MASAR have been morphologically annotated by ten well-trained linguistic annotators. In order to make sure that the annotators follow the same guidelines and of almost the same level of professionalism, nine files with total of 9,153 words (and varying numbers of POS choices per word) were tagged independently by each annotator and they were compared together. Out of 9,153 words, only 449 words show some disagreement. All ten agreed on 89% of the words; the pairwise agreement is at least 94.8%.

3 DATA CURRENT STATE

After disambiguating and developing MASAR, it has been found that:

- There are about 4,100 new unique lemmas that were either modified or added.
- The root feature has been handled to appear in the output of the modified BAMA representing 3,451 unique roots.
- The pattern feature has been added to MASAR, representing 821 unique patterns.They cover all Arabic roots that are in the modified BAMA.
- There are 191 unique tags in the modified BAMA, while there were 167 unique tags in BAMA 2.0. Table 3 shows some tags that have been added to the modified BAMA:

TABLE 3
ADDED NEW TAGS IN MASAR

Tag	Example	Description
NOUN_ADV(M)	جاء الرجل **مسرعًا** jA'aAlr~ajulu**musoriEAF**	Adverb of Manner
NOUN_ADV(T)	حدث ذلك **أمس** Hadava *`lika**>amosi**	Adverb of Time
NOUN_ADV(P)	تدور الشمس **حول**نفسها taduwru Al$~amosu**Hawola**nafosihA	Adverb of Place
NOUN(VERBAL)	**تعال**نلعب **taEAla**naloEabo	Verbal noun
NOUN_PROP_ADV(T)	التقيته **الثلاثاء**الماضي {ilotaqayotuhu**Alv~ulavA'a**AlmADiya	Proper nouns that refer to adverb of time

It can be noticed that all the used tags in MASAR belong to 5 main tag set categories which are divided into 25 sub tag sets as table 4 shows:

TABLE 4
TAG SET CATEGORIES AND SUB TAG SETS

Tag Set Category	Sub Tag Sets
Verbal category	Command Verb, Imperfect Verb, Imperfect Passive, Verb Past Verb &Past Passive Verb
Nominal category	Adjective, Noun, Adverb of Manner, Adverb of Time, Adverb of Place, Verbal Noun, Proper Noun & Proper Noun (Adverb of Time)
Pronouns category	Demonstrative Pronoun, Pronoun & Relative Pronoun.
Particles category	Focus Particle, Future Particle, Interrogative Particle, Negative Particle, Particle, Verbal Particle & Exception Particle.
Conjunctions category	Conjunctions & Sub Conjunctions.
Preposition and Interjection	---

- In addition, some prefixes and suffixes have been added in modified BAMA. Table 5 shows some added tags to the prefixes and suffixes:

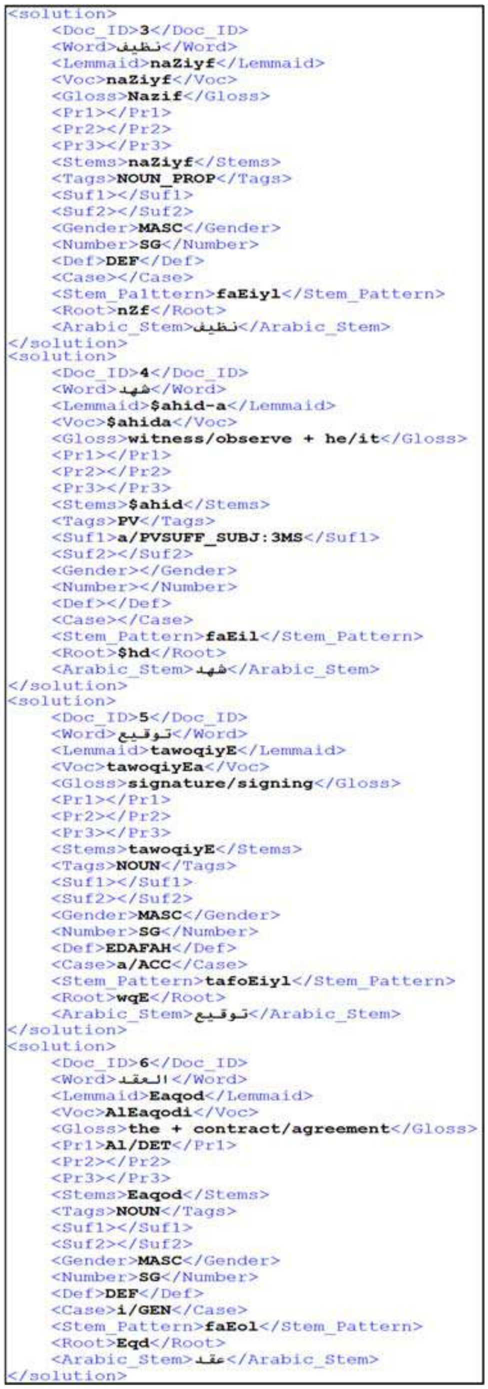
TABLE 5
SAMPLE OF ADDED PREFIXES AND SUFFIXES

Prefixes	CV_SUBJ:2FP
	CV_SUBJ:2FS
	CV_SUBJ:2MP
	CV_SUBJ:2MS
	wa/PREP
	la/PREP
	>a/INTERROG_PART
Suffixes	hAt/NSUFF
	NSUFF_SUBJ:2MS
	CVSUFF_SUBJ:2MD
	CVSUFF_SUBJ:2FP
	CVSUFF_DO:3FS
	CVSUFF_DO:3FS

- Moreover, new more detailedfeatures have been added in the number and definiteness features; the plural broken (PL_BR) and the EDAFAH features.

A) Releasing Format

In order to make MASAR easy to use, each raw text (in txt format) will be accompanied with its analyzed file that will be available in XML format as figure 4shows.



```
<solution>
    <Doc_ID>3</Doc_ID>
    <Word>نظيف</Word>
    <Lemmaid>naZiyf</Lemmaid>
    <Voc>naZiyf</Voc>
    <Gloss>Nazif</Gloss>
    <Pr1></Pr1>
    <Pr2></Pr2>
    <Pr3></Pr3>
    <Stems>naZiyf</Stems>
    <Tags>NOUN_PROP</Tags>
    <Suf1></Suf1>
    <Suf2></Suf2>
    <Gender>MASC</Gender>
    <Number>SG</Number>
    <Def>DEF</Def>
    <Case></Case>
    <Stem_Paltern>faEiyl</Stem_Pattern>
    <Root>nZf</Root>
    <Arabic_Stem>نظيف</Arabic_Stem>
</solution>
<solution>
    <Doc_ID>4</Doc_ID>
    <Word>شهد</Word>
    <Lemmaid>$ahid-a</Lemmaid>
    <Voc>$ahida</Voc>
    <Gloss>witness/observe + he/it</Gloss>
    <Pr1></Pr1>
    <Pr2></Pr2>
    <Pr3></Pr3>
    <Stems>$ahid</Stems>
    <Tags>PV</Tags>
    <Suf1>a/PVSUFF_SUBJ:3MS</Suf1>
    <Suf2></Suf2>
    <Gender></Gender>
    <Number></Number>
    <Def></Def>
    <Case></Case>
    <Stem_Pattern>faEil</Stem_Pattern>
    <Root>$hd</Root>
    <Arabic_Stem>شهد</Arabic_Stem>
</solution>
<solution>
    <Doc_ID>5</Doc_ID>
    <Word>توقيع</Word>
    <Lemmaid>tawoqiyE</Lemmaid>
    <Voc>tawoqiyEa</Voc>
    <Gloss>signature/signing</Gloss>
    <Pr1></Pr1>
    <Pr2></Pr2>
    <Pr3></Pr3>
    <Stems>tawoqiyE</Stems>
    <Tags>NOUN</Tags>
    <Suf1></Suf1>
    <Suf2></Suf2>
    <Gender>MASC</Gender>
    <Number>SG</Number>
    <Def>EDAFAH</Def>
    <Case>a/ACC</Case>
    <Stem_Pattern>tafoEiyl</Stem_Pattern>
    <Root>wqE</Root>
    <Arabic_Stem>توقيع</Arabic_Stem>
</solution>
<solution>
    <Doc_ID>6</Doc_ID>
    <Word>العقد</Word>
    <Lemmaid>Eaqod</Lemmaid>
    <Voc>AlEaqodi</Voc>
    <Gloss>the + contract/agreement</Gloss>
    <Pr1>Al/DET</Pr1>
    <Pr2></Pr2>
    <Pr3></Pr3>
    <Stems>Eaqod</Stems>
    <Tags>NOUN</Tags>
    <Suf1></Suf1>
    <Suf2></Suf2>
    <Gender>MASC</Gender>
    <Number>SG</Number>
    <Def>DEF</Def>
    <Case>i/GEN</Case>
    <Stem_Pattern>faEol</Stem_Pattern>
    <Root>Eqd</Root>
    <Arabic_Stem>عقد</Arabic_Stem>
</solution>
```

Figure 4. Analyzed Sample in MASAR

4 COMPARING ICA GOLD STANDARD RESOURCE WITH THE MOST PROMINENT RELATED WORK

Most researchers working in the field of Arabic natural language processing opt to construct their own manually collected datasets to run their experiments. Most of the time, the datasets are small and therefore their experimental findings may neither be convincing nor clear as how to scale up the results. Linguistic resources, which are required in advanced research of Arabic natural language processing, have to be built from scratch and then they should be shared with researchers in the field of Arabic NLP to expedite the development of Arabic natural language processing applications [11].

Corpora are important resources for language studies; however, Arabic lacks sufficient resources in this field. Therefore, many trials have been conducted to build Arabic corpora, but unfortunately some of them were unsuccessful trials and others were for commercial purposes.

One of these attempts that is considered one of the gold standard corpora and is being used by different researchers to develop and test their own applications; the Penn Arabic Treebank (ATB). It began in the fall of 2001; now it has completed three full releases of morphologically and syntactically annotated data: (1) Arabic Treebank: Part 1 v 2.0, roughly 166K words of written Modern Standard Arabic newswire from the Agence France Presse corpus; (2) Arabic Treebank: Part 2 v 2.0, roughly 144K words from Al-Hayat distributed by Ummah Arabic News Text, and (3) Arabic Treebank: Part 3 v 1.0, roughly 350K words of newswire text from An-Nahar press.

Penn Arabic Dependency Treebank used the output of BAMA and its main lexicon that contains over 77,800 stem entries representing about 45,000 lexical items. It used Buckwalter as the starting point for the morphological annotation and POS tagging of Arabic newswire text. For each input string, the analyzer provides a fully vocalized solution (in Buckwalter Transliteration), the word's unique identifier or lemma ID, a breakdown of the constituent morphemes (prefixes, stem, and suffixes), and their POS values and corresponding English glossaries.

Two major decisions in the actual annotation procedure regarding the morphological/part-of-speech (POS) tags have been made, both in an effort to enhance the POS tags to be more suitable for the syntactic annotation. The process of enhancing the Arabic Treebank focused primarily on making the guidelines more comprehensive, more consistent, and clearer at both the morphological and syntactic levels. However, the Part-of-Speech/morphology/gloss annotation has not yet been fully and manually revised – it is planned to revise this phase of annotation for the future releases.

There were 3781 automatic Part of Speech (POS) tag changes. These tag changes were an approximation to what the "correct" tags should be. The available corrections include correcting the core POS tag (for example, changing an active verb tag to a passive verb tag), correcting tokenization errors, and correcting case endings. The annotation tool has been modified so that Treebank annotators now have the ability to correct case endings and specific POS tags such as CONJ _ ADV or PREP _ NOUN [12].

When comparing the Penn Arabic Tree Bank with MASAR as they are both based on BAMA, a number of conclusions could be drawn:

- In Penn Arabic Tree Bank, words that do not have analysis solutions from BAMA are given a comment "NO Match", while it is analyzed in MASAR in the same manner as they would have been analyzed.
- Wrong tags and lemmas from BAMA are still found in PATB.
- MASAR makes use of more features; Name Entity, Root, and Stem Pattern.
- MASAR has more details in some features as PL_BR in number feature and EDAFAH in definiteness feature.
- BAMA's missing or wrong number and gender features have been handled in MASAR, while this is not the case in PATB.
- In disambiguating the PATB data, if the best solution of a certain word needs some modifications, the analysts could only leave a comment. For example, "Should be NOUN". However, this is not the case in MASAR where the solution can be modified and then assigned to the word.

These and other issues can be observed when comparing PATB with MASAR. Therefore, it could be claimed that MASAR may be considered as a good resource for researchers who are concerned with morphological analysis of Arabic and they can use it in developing and evaluating their systems.

5 CONCLUSIONS

This paper presented an attempt to build a morphologically annotated gold standard resource for Modern Standard Arabic. About 500,000 words are morphologically annotated using a modified version of BAMA. MASAR first release will be available soon for researchers to be used in developing and testing their application. It is expected to analyze this data syntactically and semantically in the future.

REFERENCES

[1] Gridach, M., & Chenfour, N. (November, 2011). Developing a new system for Arabic morphological analysis and generation. *In Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP'11)*, (pp. 52-57).

[2] Sawalha, M.; Atwell, E.; Abushariah, M. (February, 2013). SALMA: Standard Arabic Language Morphological Analysis. *In Communications, Signal Processing, and their Applications (ICCSPA) 1st International Conference on* (pp. 1-6). IEEE.Alansary, S. (2012). BAMAE: Buckwalter Arabic Morphological Analyzer Enhancer. *In proceedings of Arabic Language Processing Conference.* Rebate, Morocco, 2-3 May: Mohamed Vth University.

[3] Alansary, S., Nagi, M., & Adly, N. (December, 2007). Building an International Corpus of Arabic (ICA): progress of compilation stage. *In proceedings of the 7th International Conference on Language Engineering*. Cairo, Egypt, 5–6 December.

[4] Alansary, S., Nagi, M., & Adly, N. (2008). Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage. *In Proceedings of the 8th conference of The Egyptian Society Of Language Engineering (ESOLE).* Cairo, Egypt. (17-18 December).

[5] Alansary, S. (December, 2015). BASMA: BibAlex Standard Arabic Morphological Analyzer. *In the proceedings of (ESOLE).* Egypt, Cairo, 9-10 December.

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

[7] Hajic, J., Smrz, O., Buckwalter, T., & Hubert, J. (September, 2005). Feature-based tagger of approximations of functional Arabic morphology. *In Proceedings of the Workshop on Treebanks and Linguistic Therories (TLT).* Barcelona, Spain.

[8] Alansary, S., & Nagi, M. (August, 2014). The International Corpus of Arabic: Compilation, Analysis and Evaluation. *In the proceedings of EMNLP Workshop.* Doha, Qatar.

[9] Alansary, S. (2012). BAMAE: Buckwalter Arabic Morphological Analyzer Enhancer. *In proceedings of Arabic Language Processing Conference.* Rebate, Morocco, 2-3 May: Mohamed Vth University.

[10]Alansary, S. (December, 2015). BASMA: BibAlex Standard Arabic Morphological Analyzer. In the proceedings of (ESOLE). Egypt, Cairo, 9-10 December.

[11]Hammo, B., Al-Shargi, F., Yagi, S., & Obeid, N. (2013). Developing tools for Arabic corpus for researchers. In The proceedings of the second workshop on Arabic corpus linguistics (WACL-2).

[12]Maamouri, M., Bies, A., & Kulick, S. (2008). Enhanced Annotation and Parsing of the Arabic Treebank. In LREC.

BIOGRAPHY

Dr. SamehAlansary

He is professor of computational linguistics in the Department of Phonetics and Linguistics, 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.

مسار: مصدر معياري لتقييم التحليل الصرفي في اللغة العربية

سامح الأنصاري

مركز اللغويات الحاسوبية العربية – مكتبة الإسكندرية

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية

sameh.alansary@bibalex.org

ملخص

يمثل الصرف العربي تحديا خاصًا للعديد من أنظمة المعالجة الآلية للغة العربية. فالعربية غنية بالكثير من التنوعات والتعقيدات الصرفية فالجذر الواحد في اللغة يمكنه توليد العديد من الكلمات المختلفة في الوزن الصرفي. وعلى الرغم من التطور الذي تشهده العربية في بناء أنظمة متعددة لها إلا أنها ما زالت تفتقد إلى الأدوات والمصادر اللغوية وما زالت المعالجة الآلية للغة العربية في مهدها الأول. تعرض هذه الورقة مصدرًا معياريًا محللا صرفيًا - (مسار) - والذي تم بناؤه لمساعدة الباحثين في تطوير واختبار العديد من الأنظمة التي تفيد اللغة العربية. وقد بدأ العمل في بناء هذا المصدر أثناء العمل في مشروع المدونة اللغوية العربية العالمية حيث تم اختيار العينة اللغوية المراد تحليلها لبناء هذا المصدر من العديد من المصادر والفئات داخل هذه المدونة وتحليلها باستخدام أحد المحللات الصرفيةالشهيرة تيم باك ولتر. ويحتوي مسار على معلومات لغوية أكثر تفصيلا عن باك ولتر، كما يحتوي على معلومات لغوية جديدة أخرى مما يجعله مصدرا غنيا بالمعلومات الصرفية اللغوية.

Improving Alserag Arabic Diacritization Grammar through Syntactic Analysis

Sameh Alansary

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

***Abstract*— Diacritization of written text has a significant impact on Arabic NLP applications. We present an approach to Arabic automatic diacritization that integrates morphological analysis with more intensive shallow syntactic analysis which has great effects in improving the results of the case ending. The developed system (Alserag) is a rule based system. The results of the system in this phase were evaluated for accuracy against the reference using two metrics; diacritization error rate (DER) and word error rate (WER). The DER measurement was 4.42% while WER measurement was 14.75% while the previous DER measurement was 8.68% while WER measurement was 18.63%. The whole data of Alserag results have been benchmarked among other three systems; Harakat, Mishkal, Aldoaly, as well as the previous results of the Alserag system before the syntactic analysis improvements.**

1 Introduction

Diacritizing Arabic written text is crucial for many NLP tasks. Arabic diacritics are superscript and subscript diacritical marks (referred to sometimes as vocalization or vowelling), defined as the full or partial representation of short vowels, shadda (consonantal length or germination), tanween (nunation or definiteness), and hamza (the glottal stop and its support letters) [1]. Diacritization helps the reader in disambiguating the text or simply in articulating it correctly. Arabic is a language where the intended pronunciation of a written word cannot be completely determined by its standard orthographic representation; it rather depends on a set of special diacritics. The absence of these diacritics in Arabic text increases lexical and morphological ambiguity, because one written form can have several pronunciations, each pronunciation may have different meaning(s) [2,3]. The word form"ذكر" 'zakar' can have many possible pronunciations like 'zakar'(male) and 'zakara' (to mention).
However, these diacritics are generally left out in most genres of written Arabic which results in widespread ambiguities in pronunciation and (in some cases) meaning.
Although native speakers are able to disambiguate the intended meaning and pronunciation from the surrounding context with minimal difficulty, it is not the case with automatic processing of Arabic which is often hampered by the lack of diacritics. Several applications can radically benefit from automatic diacritization, such as Text-to-speech (TTS), Part-Of- Speech (POS) tagging, Word Sense Disambiguation (WSD), and Machine Translation [3].

The focus in this paper is on the improvements of the syntactic analysis of Alserag; an Arabic diacritizer. Alserag is based on 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 is based on shallow parsing and finally the morpho-phonological step.

This system has been presented before in the AISI 2016 [4], the results of the system were DER measurement that was 8.68% while WER measurement that was 18.63%. Moreover, the system was fully described in this paper. However, the system is reintroduced in this paper since intensive improvements have been made in the syntactic analysis leading to a change in the results. This paper will also present the challenges that faced the system that were overcome and the limitations of the system.

Section 2 will present the related work concerning automatic diacritization for Arabic text. Section 3 demonstrates the grammar workflow. Section 4 explains the different applied modules to fully diacritize texts. Section 5 evaluates the output and discusses the results and also is concerned with the benchmarking process. Finally, Section 6 concludes the paper.

2 THE STATE OF THE ART

Much work has been done on Arabic diacritization. The actually implemented systems can be divided into two categories [5]: Systems implemented by individuals as part of their academic activities and systems implemented by commercial organizations for realizing market applications.

One of the advantages of the first type is that they presented some good ideas as well as some formalization. The weak point about these systems was that they are mostly partial demo systems [5]. The following are examples of these systems:

[6] choose the analysis from the diacritizations proposed by the Buckwalter Arabic Morphological Analyzer (BAMA).

[7] also used (BAMA). They sought to improve automatic speech recognition (ASR) by working on diacritization. They used a language-modeling approach.

[8] follow a statistical model. The model combines different sources of information ranging from lexical, segment based, and POS features.

[9] introduce MADA system that uses BAMA, where they select the optimal full morphological tag for Arabic in context and use it to select from a list of possible analyses produced by a morphological analyzer. [10] presents an approach that integrates syntactic analysis with morphological tagging through improving the prediction of case and state features.

For the second category, the most representative commercial Arabic morphological processors are Sakhr, Xerox, and RDI [5].

Armedia is part of the Sakhr Office Tools. It is a large scale Arabic diacritizer that is achieved by native Arabic speakers. It is based on standard Arabic dictionaries. However, that has caused a problem, since any morphologically possible Arabic word that is not registered in these dictionaries will not be considered. The problem of this restriction is that even the most extensive Arabic dictionaries do not list all the used Arabic vocabulary of that time. The actual implementation of the statistical disambiguation at Sakhr is made by considering only the monograms of words in the text corpus and does not count for the correlation among neighboring words. Moreover, it is an unfree tool.

Xerox is an Arabic diacritizer, and it is the best system implemented by non-native Arabic speakers. Yet, it suffers some shortcomings, such as it is based on the standard Arabic dictionaries which means that all the morphologically possible Arabic words that are not registered in these dictionaries are not considered. Xerox system does not have a mechanism for disambiguation. Therefore, the coverage of the final system is not considered as outstanding [5].

RDI is also a large-scale Arabic diacritizer that is achieved by native Arabic speakers. It has a number of advantages over the previously mentioned systems: It is a factorizing system which allows dealing with all the possible Arabic words with no need to be tied to a fixed vocabulary. This system uses a powerful n-grams statistical disambiguation technique which means that the system considers the statistical correlation among the words and their neighbors [11, 12].

KAD is an Arabic diacritizer in which a technique that depends on two major steps is used. The first step is to create a very rich list of frequently used Arabic quad-grams, the second step is to use this list in diacritizing almost any Arabic text [13].

There are also other available systems as Mishkal Arabic diacritizer, and Harakat Arabic diacritizer; they are free Arabic diacritizers which are available online. Finally, in March 2016 Google has launched an innovative new Google Labs Arabic tool called Tashkeel, a tool that adds the missing diacritics to Arabic text. Unfortunately, the tool is not available now.

There is another system[14] that has integrated three different proposed techniques, each of which has its own strengths and weaknesses. They are lexicon retrieval, diacritized bigram and SVM statistical-based diacritizer. Firstly, lexicon retrieval technique (LR): It tries to find the diacritized word returned from an Arabic lexicon. Secondly, diacritized bigram technique (DB): When more than one solution is retrieved for an ambiguous diacritization, the bigram technique is applied. Finally, SVM-statistical technique (SVM): The main idea of this approach is to tokenize and automatically annotate tokens with the correct POS tags. Then, by searching the Arabic lexicon using a token and the corresponding POS, the correct diacritization result can be reached, even though multiple ambiguous words are retrieved from the lexicon [14].

Most of the previous approaches cited above utilize different sequence modeling techniques that use varying degrees of knowledge from shallow letter and word forms to deeper morphological information. None of the previous systems make use of syntax with the exception of [10] who have integrated syntactic analysis; however, they are not rule based.

Alserag has been developed in 2016, it is based on 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 system tokenizes a text and provides a solution for each token and restores the appropriate internal diacritics from the dictionary), the syntactic processing step that is responsible for the case ending detection is based on shallow parsing and finally the morpho-phonological step that is developed to fulfill the requirements of vowel harmony and assimilation [4].

3 WORKFLOW OF THE DEVELOPED GRAMMAR

In this section, the architecture of the system will be presented and the different processes that took place in order to convert a plain text into a fully diacritized text will be described. It is worth mentioning that it is a rule-based system [4]. Figure 1 presents the system's overall architecture, where the diacritization is achieved through 7 main phases: (i) Preprocessing which is responsible for auto-correcting the raw text and segmenting the Arabic text into sentences. (ii) Tokenization which is the process of splitting the natural language input into lexical items. (iii) Disambiguation which is a process of choosing the right internal diacritization for the word from the dictionary. (iv) Name entity recognition, and (v) Shallow syntactic parsing which is an analysis of a sentence by identifying its constituents (NPs, JPs---etc.). (vi) Case ending module which is responsible for identifying the arguments of the verb and assigning the diacritical marks accordingly. (vii) Morph-phonological module which is a series of rules that focus on the sound changes that take place in morphemes (minimal meaningful units) when they combine to form words.

There are two engines that are used in the Arabic diacritizer, the first is Interactive ANalyzer (IAN) which is used in the analysis process, the second is dEep-to-sUrface natural language GENErator engine (EUGENE) is which used is in the generation process [4].

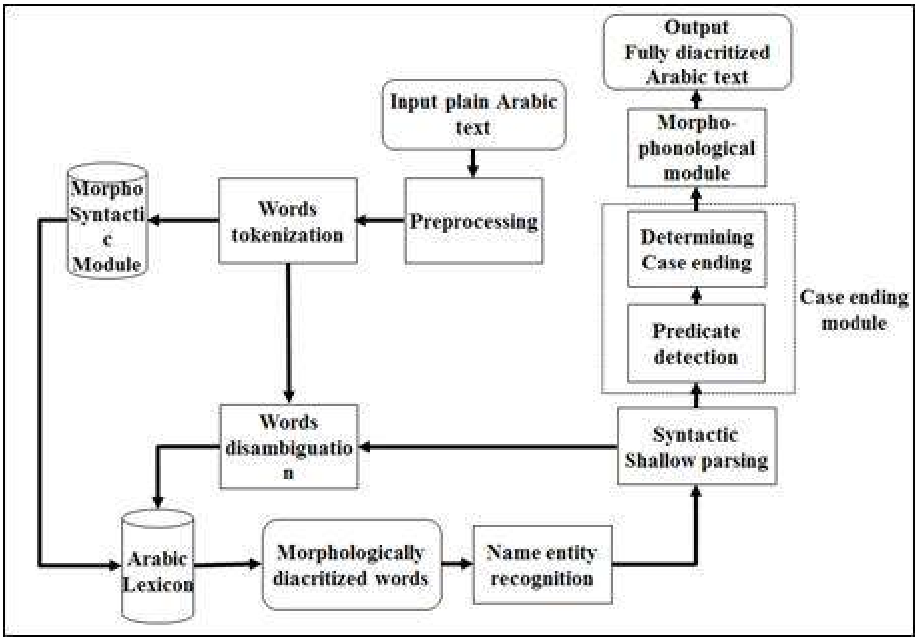


Figure 1. Architecture of ALSERAG.

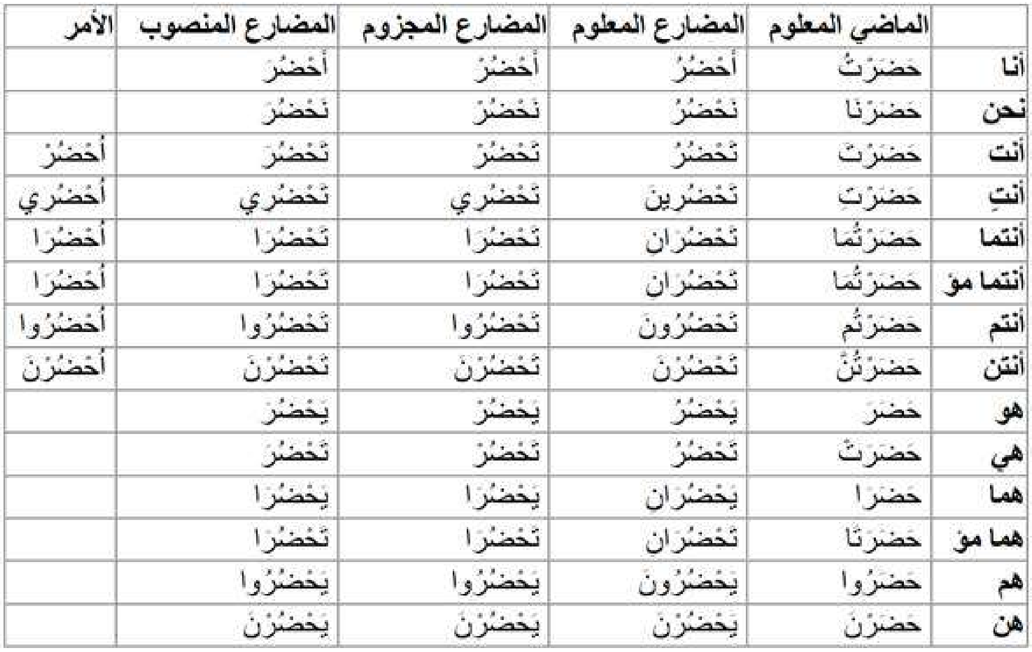
4 THE RESOURCES OF THE DEVELOPED GRAMMAR

Alserag depends on two resources; the Arabic diacritized dictionary and a set of linguistics rules. Each one will be described in details in the following subsections.

A. Dictionary

The Arabic diacritized dictionary includes Arabic natural language words in their diacritized form; each word is stored along with the corresponding linguistic features which describe it morphologically, morpho-syntactically, syntactically and semantically. For example, the Arabic word "نظر" 'look' is stored in its diacrizied form "نَظَرَ" and a list of linguistic features such as part of speech, tense, transitivity, person, gender, number, etc. are included in the dictionary. It is a word-form dictionary, for example the dictionary lists all the word forms of the verb "حضر" 'come' such as "يحضر" 'he comes, "يحضرون""they are coming' "حضرت" ,'she came' and so forth as shown in figure 2. The words in the Arabic diacritized dictionary are extracted from the Arabic dictionary in UNLarium[1]. The process of diacritization mainly depends on two resources: BAMA and Alkhalil Arabic Morphological Analyzer.

[1] http://www.unlweb.net/unlarium/



	الماضي المعلوم	المضارع المعلوم	المضارع المجزوم	المضارع المنصوب	الأمر
أنا	حَضَرْتُ	أَحْضُرُ	أَحْضُرْ	أَحْضُرَ	
نحن	حَضَرْنَا	نَحْضُرُ	نَحْضُرْ	نَحْضُرَ	
أنت	حَضَرْتَ	تَحْضُرُ	تَحْضُرْ	تَحْضُرَ	أُحْضُرْ
أنتِ	حَضَرْتِ	تَحْضُرِينَ	تَحْضُرِي	تَحْضُرِي	أُحْضُرِي
أنتما	حَضَرْتُمَا	تَحْضُرَانِ	تَحْضُرَا	تَحْضُرَا	أُحْضُرَا
أنتما مؤ	حَضَرْتُمَا	تَحْضُرَانِ	تَحْضُرَا	تَحْضُرَا	أُحْضُرَا
أنتم	حَضَرْتُم	تَحْضُرُونَ	تَحْضُرُوا	تَحْضُرُوا	أُحْضُرُوا
أنتن	حَضَرْتُنَّ	تَحْضُرْنَ	تَحْضُرْنَ	تَحْضُرْنَ	أُحْضُرْنَ
هو	حَضَرَ	يَحْضُرُ	يَحْضُرْ	يَحْضُرَ	
هي	حَضَرَتْ	تَحْضُرُ	تَحْضُرْ	تَحْضُرَ	
هما	حَضَرَا	يَحْضُرَانِ	يَحْضُرَا	يَحْضُرَا	
هما مؤ	حَضَرَتَا	تَحْضُرَانِ	تَحْضُرَا	تَحْضُرَا	
هم	حَضَرُوا	يَحْضُرُونَ	يَحْضُرُوا	يَحْضُرُوا	
هن	حَضَرْنَ	يَحْضُرْنَ	يَحْضُرْنَ	يَحْضُرْنَ	

Figure 2. The different forms of the root “حضر”.

The diacritizing process begins with Buckwalter`s analysis. Some words have only one solution, other words have more than one solution and some words couldn’t be analyzed in Buckwalter. These words are analyzed by Alkhalil which also suggests different solutions to some words. Then, these words are verified manually to select their correct diacritization. Some enhancements have to be made in the Buckwalter solutions. For example, some solutions have a missing vocalization “َ” before “ا” as in “كتاب” ‘book’, and “مدارس” ‘schools’. So, these missing vocalizations have been added manually.

The Arabic diacritized dictionary includes fully and partially diacritized entries. Partially diacritized entries are internally diacritized, but have no case ending, since their case ending depend on the context as in singular nouns such as “قَلَم” ‘pen’, singular adjectives such as “جَمِيلَة”“beautiful’, broken plural nouns such as “صُحُف” ‘newspapers’ and subjunctive and indicative present verbs where the subject of the verb is singular as in “تكتب” ‘she writes’. By default, a present tense verb is marked by a short /o/ (الضمة), in this case it is called indicative (المضارع المرفوع). However, if a present verb is preceded by certain particles, the verb will be marked by a short /a/ (الفتحة), and if the verb ends by one of the three suffixes (ين ،ان ،ون), the final (ن) will be deleted, in this case it is called subjunctive (المضارع المنصوب). On the other hand, imperative verb forms “أُكْتُبْ” ‘you write’ and past verb forms “كَتَبَتْ” ‘she wrote’ are fully diacritized, also dual and regular plural nouns in genitive, accusative and nominative cases, such as “مُدَرِّسَتَانِ” ‘two schools’, “مُعَلِّمِينَ” ‘male teachers’ are fully diacritized as well as dual and plural adjectives “نَشِيطَانِ” ‘both are active’,“مُتَمَيِّزِين” ‘they are special’, because their case endings do not depend on the context.

1) *Linguistic features of the Arabic diacritized dictionary:* the linguistic features of the Arabic diacritized dictionary is a set of linguistic information developed to describe every Arabic word. Arabic words are described on different linguistic levels: morphological information, morpho-syntactic information, syntactic information and semantic information. Each one will be described in details in the following sub-subsections:

Morphological Information: such as 1) Part of speech where the Arabic entries are classified into different classes. These classes are noun, verb, adjective, adposition, and adverb. Each class may include subclasses for example, nouns are classified into common nouns and proper nouns, so word like “معلم” ‘teacher’ is a common noun and word like “أحمد” ‘Ahmed’ is a proper noun. Verbs are also classified into subclasses, full verb, copula verb, and auxiliary verb. For example, the verb “أكل” ‘eat’ is a full verb, “بدا” ‘seem’ is a copula verb, and “كان” ‘was’ is an auxiliary verb. 2) The lexical structure feature, since Arabic words can be classified into sub-words (bound morphemes) such as “س”/sin/ the future prefix, simple words as “رسم” ‘draw’, and multiword expressions which are lexical structures made up of a sequence of two or more lexemes such as “كفر الشيخ” ‘Kafr El Sheikh’.

Morpho-syntactic Information: such as 1) **Transitivity** feature which describes the syntactic behavior of the verbs and the type of their arguments. the Arabic diacritized dictionary classifies verbs according to transitivity into two main classes, intransitive verbs and transitive verbs. The intransitive verbs are further classified into unaccusative verbs whose subject is not the agent, as in the sentence "اتسع الميدان"(The square widened) and unergative verb whose subject is the agent, as in the sentence "تكلم الرجل"(The man spoke). Transitive verbs are further classified into four types, direct transitive; a verb which takes a subject and a single direct object, such as the verb "قرأ" 'read' in "قرأ الرجل الكتاب" 'the man read the book', indirect transitive; a verb which takes a subject and a single indirect object, such as the verb "استمع" 'listen' in the sentence "استمع الطالب إلى الدرس"'the student listened to the lesson', ditransitive; a verb which takes a subject and two objects, such as the verb "منح" in the sentence "القانون يمنح المرأة حق الانتخاب"'the law grants women the right to vote'. 2) **Number** feature that is assigned to nouns to specify their number for example, the noun "هاتف" 'telephone' is singular, the noun "مهندسان" 'two engineers' is dual and the noun "مستشفيات" 'hospitals' is plural. Number is also assigned to verbs to specify the number of their subject for example, the verb "فهم" 'he understood' is assigned as singular to indicate that its subject is singular, and the verb "قالوا" 'they said' takes the plural feature to indicate that its subject is plural and so on and so forth. 3) **Tense** feature which is used in the grammatical description of verbs, referring primarily to the way the grammar marks the time at which the action denoted by the verb took place. It can be broadly classified as: past tense as in "لعب" 'played' and present tense as in "يلعب" 'plays'. 4) **Gender** feature, all Arabic words are classified into three genders; masculine such as "مكتب" 'office', feminine such as "منضدة" 'table' and common such as "عجوز" 'old man/woman'. The gender feature is assigned to nouns to describe their gender, to verbs to describe the gender of the subject of the verb and to adjectives to describe the gender of the Substantive. 5) **Diptotic noun** feature reflects that diptotic nouns are marked in nominative case with the short /o/ (الضمة), and both accusative and genitive cases with the short /a/ (الفتحة). Moreover, these nouns do not receive /ًا/ تنوين(a type of nunation). For example, the noun "مساجد" 'mosques' is diacritized "مَسَاجِدُ" in the nominative case, and "مَسَاجِدَ" in both accusative and genitive cases.6) **Shortened Nouns** feature describes shortened noun الاسم المقصور; a noun which ends with an 'alif denoting a long vowel /aa/ (ا/ى) for example, the noun "عصا" 'stick', "مستشفى" 'hospital'. 7) **Defective Nouns** feature describes defective noun الاسم المنقوص; a noun which ends with a long vowel /ee /(ي) that is original letter and belongs to the root for example, the noun "قاضي" 'judge', and "محامي" 'lawyer'. However, proper names cannot be defective nouns. 8) **Active participle** feature, the active participle اسم الفاعل is essentially related in meaning to the meaning of the verb. An active participle can be used to describe what someone is doing right now (going, leaving), and to indicate that someone/something is in a state of doing something. For example, the active participle of "دَرَسَ" is "دارِس" and the active participle of "كَتَبَ" is "كاتِب". 9) **Passive participle** feature, the passive participle اسم المفعول may express a current state of being or it may express a state of having been the result of an action that has already been performed. For example, the passive participle of "دَرَسَ" is "مدروس" Likewise, the passive participle of "كَتَبَ" is "مكتوب".

Syntactic Information: it contains subcategorization frames which determine the number and types of the necessary syntactic arguments (specifiers, complements and adjuncts) of verbs, nouns and adjectives. For example, in the sentence "أعطى التلميذ الكتاب لصديقه"'the student gave the book to his friend', the subcategorization frame of the verb "أعطى" 'give' determines that the verb has three arguments, a verb specifier (a noun phrase)" التلميذ" 'the student', a verb complement (a noun phrase) "الكتاب" 'the book', and a verb complement (a prepositional phrase) "لصديقه" 'to his friend'. In the sentence "الوصول إلى المعلومات"'access to the information', the subcategorization frame of the noun "وصول" "access" determines that the noun has a complement (a prepositional phrase) "إلى المعلومات" 'to the information'. Furthermore, in the sentence "موضوع متعلق بالبيئة""subject related to the environment" the subcategorization frame of the adjective "متعلق" 'related' determines that the adjective has a complement (a prepositional phrase) "بالبيئة" 'to the environment.

Semantic Information: the semantic classification adopted in the Arabic diacritized dictionary is the English WordNet 3.0. ontology. Each entry in this classification carries a feature that describes it semantically, for example, the noun "عالم" 'scientist' is classified semantically as 'HUM' while the noun "مصر"'Egypt' is classified semantically as 'LCT' "location".

B. The Linguistic Rules

There are three modules in order to provide fully diacritized Arabic words namely, morphological analysis module, syntactic analysis module and morph-phonological processing module.

1) *Morphological analysis module:* It is responsible for morphologically analyzing Arabic words and assigning the correct POS and the internal diacritization of words which is achieved through two processes; tokenization process and disambiguation process. Firstly, tokenization is the process of splitting the natural language input into lexical items; the tokenization process is mainly based on the dictionary, therefore the possibility of ambiguity increases with the increase in the number of entries in the dictionary. Lexical items are tokenized according to longest matched unless the possible longest match is blocked by the developed rules. Secondly, disambiguation is applied over the outcomes of the tokenization process; they are used to reject the wrong lexical choices and re-obtain the right ones. This set of rules has a specific format: (node 1) (node 2) (...)(node n)=P; Where (node 1), (node 2) and (node n) are nodes, and P is an integer expressing the possibility of occurrence.

Actually, there is a process prior to disambiguation which is the segmentation. Affixation has an important role as the first level of part of speech disambiguation, since prefixes and suffixes are the smallest processing units that rules can begin with. For example, in the verb "سيفعل" 'will do', the verb will be divided into "س" 'will' +"يفعل " 'do', because there is a rule in the segmentation module that segments the prefix "س"'will' if it is followed by a present or jussive verb by the rule in (1).

1) ([س])({^V,^I,^PRS|JUS})=0;

Another example of the segmentation process, in the phrase "أن لغتهم مغولية" 'That their language is Mongolian',"لغتهم"'their language'is segmented into "ل" 'for',"غت" (meaningless) ,"هم" 'their', such sequence is meaningless, the correct segmentation should be as follows: "لغت" 'language' with an open ت because of the morph-syntactic change, "هم" 'their' so the rule in (2) blocks the first segmentation which is wrong:

2) ([أن])(BLK)([ل])(TEMP)(SPR)(BLK)({J|N})=0;

As a result, the string "لغت" (language) becomes a Temp, as it does not exist which means that the engine cannot recognize it, therefore the morpho-syntactic rules take place which use the regular expression technique in converting the letter "ت" to "ة" to be "لغة" 'language' and hence it could be recognized and retrieved with its internal diacritics from the dictionary.

The sequence "وهناك عوائق"'and there are obstacles'would be automatically segmented as "وهنا" 'became weak' + "ك" 'as'+"عوائق "'obstacles', according to the longest match algorithm, given the fact that the dictionary includes'VER وهنا' "became weak", 'ك' "as", "و" 'and' as a conjunction (COO), "هناك" is an adverb (AAV) 'there' and "عوائق" noun (NOU) 'obstacles', the rule in (3)will reject the verb "وهنا" 'became weak' and the 'ك' "as":

3) (SHEAD)([وهنا],V)([ك],P)=0;

As mentioned before disambiguation is concerned with preventing the wrong automatic lexical choices and obtaining the right internally diacritized words. Some linguistic indicators can help in solving the lexical ambiguity which are semantic, morphological and adjacency indicators as mentioned in the publication of [4].

Semantic indicators: In the sequence "وفي الصباح"'in the morning, it would be automatically segmented as "وفي" 'be honest/ honest', "ال" 'the' and "صباح"'morning', according to the longest match algorithm, given the fact that the dictionary includes"وفي" 'be honest' VER (verb) 'honest' an adjective, "ال" 'the' is an article (ART), "و" 'and' is a conjunction (COO), "في" 'in' is a preposition (PRE) and the noun "صباح" 'morning'. If the word "صباح" 'morning' exists in dictionary with the semantic class TIM (time), the rule (4) will reject the verb "وفي" 'be honest' and the adjective 'honest'.

4) (SHEAD)([وفي],{V|J})(BLK)(ART)({TIM|LCT})=0;

Morphological indicators: In the phrase "لرفع", the noun "رَفْع" 'lifting' is chosen instead of the verb "رَفَعَ" 'to lift', since it is preceded by the preposition 'prefix' "ل" (to) by rule in (5).

5) (P)(V)=0;

Agreement in Arabic plays a crucial role in disambiguating the morphological solutions of words, adjectives agree with nouns in definiteness, gender, number, and case, for example, "الطالب الجديد" 'The new student'. There is a general Arabic rule that states that all plural nouns which don't refer to human beings are considered to be grammatically feminine singular. For example, in the phrase "مقالات كثيرة"many articles', the plural noun is modified by a feminine singular adjective as the rule in (6).

6) (N,^MCL,^PLR,^ NANM)(BLK)(J,^SNG,^FEM)=0;

The subject should also agree with the verb in number and gender. In the phrase "وهم يكتبون" 'and they write', the automatic choice will be the noun "وَهْم" 'illusion' according to the longest match algorithm, but because it is followed by a plural verb "يكتبون" 'they Write', this tokenization will be rejected and will be retokenized as "وَ" (and) + "هُمْ" (they) by the rule in (7) since the subject and the verb should agree in number and gender.

7) (SHEAD) ([وهم] ,%x) (BLK) (V, ^NUM=%x)=0;

Nevertheless, Number and Gender qualifiers represented in Subject-Verb-Agreement which is concerned with making the verb agree with subject follow some conditions. For example, if the verb follows the subject, the agreement will be in both number and gender; however, if the verb precedes the subject, they should agree in gender only. For example, in the sentence “درس المدرسون كتابَهم العربيّ” ‘The teachers studied their Arabic books’, the verb will always be singular and will agree with the subject only in gender.

Adjacency indicators:Functional word qualifiers can also be used in solving disambiguation problems, some particles are used with certain parts of speech[4]. For example, in the sequence “بعد عودتها” ‘after her arrival’ the preposition “بعد” ‘after’ should be followed by a noun, so rule (8) uses the regular expression technique in order to reject the segmentation of “عودتها” into “عودت” ‘accustomed’ Verb + “ها” ‘her’ or any other verb that ends with “ت” if it follows certain prepositions:

8) (#FINAL, {[من] , ^R | [مع] | [عن] | [في] | [حوالي] | [على] | [قبل] | [حول] | [مثل] | [إلى] | P , ^[لكي] | [عبر] | [بعد] | [نحو] | [دون] | P }, %01) (BLK , %02) (V , [/.+ت/] , %03) (SPR , %04) = 0;

Sometimes, specific sequences can be disambiguated depending on specific features and sometimes specific words. For example, the sequence “سوى زوجته” ‘except his wife’, “زوجته” (his wife) can be segmented as “زوجت” ‘to marry’ (verb) + “ه” ‘him’ which is rejected because it is preceded by the particle “سوى” ‘except’ by the rule (9).

9) (PTC,{[سوى]|[غير]},#FINAL)(BLK)(^J,^N,^ART,^[أن],^TEMP)=0;سوى زوجته

The developed rules can deal with coordinated elements. For example, “لغدر الكلاب أو الأشرار” ‘for the betrayal of the dogs and the villains’, the two coordinated elements should belong to the same part of speech. “الكلاب” ‘the dogs’ is segmented into “ال” ‘the’ (Article) + “كلاب” ‘dogs’ (Noun); however, “الأشرار” (the villains) is segmented into “ال” (the) (Article) + “أشرار” ‘villains’’ (Adjective) which is not linguistically accepted, so the selected adjective is rejected by rule (10) and the noun will be chosen:

10) (P)(N)(BLK)(ART)(N)(BLK)(COO)(BLK)(ART)(J)=0;

Moreover, the collocation can be used in solving many disambiguation cases. Collocation is essentially a lexical relation [15]. It is co-occurrence of specific words together with specific features. This system adopts the bigram language model (the occurrence of 2 words). The collocation rule is designed to express either the high possibility of occurrence or the low possibility of co-occurrence. An example of low possibility of occurrence in rule (11), “سوق المال” ‘money to market/market’ “سوق” should be disambiguated as the noun “سوق” ‘Market’ not as the verb “سَوَّقَ” ‘to market’:

11) ([سوق],V)(BLK)(ART)(LEMMA=مَال)=0;

In high possibility of occurrence, first the bigram is determined and their frequency of occurrence together is increased. For example, in a phrase such as “جرس المدرسة” ‘The school bell’, where each lexical item has different diacritization possibilities because of the large amount of entries in the dictionary, both words can be wrongly selected, “جرس” is selected as the verb “جَرَّسَ”‘disgrace’ and “مدرسة” is selected as the noun “مُدَرِّسَة” ‘female teacher’ which are both wrongly disambiguated. So, this case can be handled by the rule in (12):

12) ([جرس],N)(BLK)(LEMMA=مَدْرَسَة,N)=255;

In certain cases, the collocation depends on a lexical combinatory preference imposed by the language usage within a particular community and not based on grammatical rules. The repeated use of combined lexemes over time becomes so frequent that, eventually, the speakers of the community automatically associate a lexeme A with lexeme B. For example, the sequence “فرض حظر” ‘imposed restrictions’, the rule in (13) states that “حظر”‘restriction’ is always preceded by “فرض”’ imposed’as a noun not as a verb.

13) ([فرض],N)(BLK)([حظر],^N)=0;

2) *Syntactic analysis module:* After segmenting the input successfully and choosing the correct internal diacritization of each word, shallow parsing of the input takes place. Shallow parsing is considered necessary for case ending assignment [4]. It starts by grouping words under the different phrasal categories. Accuracy in grouping these phrasal categories helps in increasing the efficiency of assigning the different case endings. Phrasal grouping is necessary for identifying the sentence components and linking them by a predicate. In this module the phrases are built; those phrases are noun phrases (NPs), adjectival phrases (JPs), prepositional phrases (PPs) and finally adverbial phrases (Aps) in certain cases. The adopted linguistic theory in this module is the X-bar theory [4]. In this paper, the focus is on improving the results of Alserag system by improving the parser. It is noteworthy to mention that this module has borne the most amount of work in this phase as a result the case ending has been enhanced to score an error rate %14.75and by comparing this percentage with the

previous percentage; before the syntactic improvements, it becomes clear that great enhancements have been made.

This module starts by composing the adjectival phrases first if they exist. Then the smallest NP in the sentence is built and then it is linked with the preposition; if there is one, to build a PP, with an adverb to build an AP, or with another noun to build a bigger NP. Those phrases are the constituents' boundaries. If the input contains a verb, it is considered as the core of the sentence, since it is the verb that communicates the three most important elements of any message - the what, who and when. The verb will determine the case ending of the sentence`s elements. However, in nominal sentences the process of case ending assignment is different. Moreover, the sentences that begin with enna and her sisters (إن وأخواتها) or kana and her sisters (كان وأخواتها) differ in their case endings from other verbal sentences. In what follows there will be an explanation of how the grammar has handled the different challenges of case ending assignment.

Building constituents' boundaries: Syntactic constituents are groups of words linked on the basis of their relationship with other words in the sentence; two or more words in a sentence can serve and function as one single unit. A set of rules has been developed in order to group the related words together to form different phrasal categories.

In order to have a comprehensive idea about building the different constituents' boundaries, let`s consider the following example in sentence (1). Sentence (2) represents the output after applying the morphological analysis module to the sentence in (1); each word is assigned with the right internal diacritization. Rule (14a) starts by projecting all the adjectives in the sentence to the intermediate constituent J-bar (JB). So, by rule (14a) both the adjective "دولية" 'international' and "أهم" 'most important' are projected to be the intermediate constituent J-bar (JB). Then, the adjective "دولية"'international' will be projected to the maximal projection JP as it is not followed by a complement by rule (14b).

Sentence 1. بطولة كأس العالم لكرة القدم هي أهم مسابقة كرة قدم دولية تقيمها الفيفا
Sentence 2. بُطُولَة كَأْس اَلْعَالَم لِكُرَة اَلْقَدَم هِيَ أَهَمّ مُسَابَقَة كُرَة قَدَم دَوْلِيَّة تُقِيمهَا اَلْفِيفَا

14)

a. (J,^JB,%j):=(%j,JB);
b. (JB,%j,^JP)({^ART|STAIL},%z):=(JP, -JB,%j)(%z);
c. (NOU,^NB,%x):=(%x,NB);
d. (ART,%z)(NB,^np,%n)({^ART,^NB|STAIL},%y):=((%z,)(%n,np),POS=%n,SEM=%n,NP,DEF=def,GEN=%n,NUM=%n,LEMMA=%n,ARS=%n)(%y);
e. ({^ART|SHEAD},%z)(NB,^np,^def,^indef,%n)(NP,def,%y)({^COO|COO,SKIP|STAIL},%c):=(%z)((%n, np)(%y,casatt=GEN),NB,DEF=def,GEN=%n,NUM=%n,LEMMA=%n,SEM=%n,ARS=%n,synf=edafa)(%c);
f. ({SHEAD|^ART},%s)(NB,^np,%n)(JP,GEN=%n,^def,%adjc):=(%s)((%n,np)(np,%adjc),NB,SEM=%n,GEN=%n,NUM=%n,DEF=def,LEMMA=%n,ARS=%n);
g. ({^ART,^DEM|SHEAD},%z)(NB,def,^NP,%n)({^NB,^SPR,^ART,^DEM|STAIL},%y):=(%z)(NP,%n,NB)(%y);
h. (JB,{SUP|CMP},^jp,%j)(NP,%y)({^COO|COO,SKIP|STAIL},%c):=((%j,jp)(%y,casatt=GEN),JB, GEN=%j,DEF=%j,ACAS=%j,FRA=%j,DEG=%j,LEMMA=%j)(%c);
i. (%x,P,^pp)(%n,NP)({STAIL|PUT,^BLK},%s):=((%x,pp)(%n,casatt=GEN),PP)(%s);

Then, the heads of the noun phrases "بطولة" 'championship', "كأس" 'cup', "كرة" 'ball', "قدم" 'foot', "مسابقة' 'contest' and "فيفا" 'FIFA' will be projected to the intermediate constituent N-bar (NB) by rule (14c). Then, rule (14d) starts to build the different NPs in the sentence by combining the definite article "ال" 'the' with the following noun "قدم" 'foot' to project a noun phrase (NP), the same rule will be applied recursively to link both "ال" 'the' with "فيفا" 'FIFA 'and "ال" 'the' with "عالم" 'world'. The composed NPs are automatically assigned with the features of their heads such as gender, number, animacy and semantic class that are necessary to describe those NPs. Rules are able to build a bigger NBs. For example, rule (14e) combines the indefinite noun "كرة" 'ball' with the definite NP "القدم" 'foot' and assigns the NP "القدم" 'foot' with the genitive case. Rule (14e) will be applied recursively to link "كأس" 'cup' with "العالم" 'world' and "بطولة" 'championship' with the definite NP "كأس العالم" 'the world cup' and finally "مسابقة" 'contest' with "كرة قدم" 'football'. Since that the adjectival phrase "دولية" 'international' agrees with the preceding NB in gender and definiteness, they will be linked to form a bigger NB by rule (14f). Then, if there is any other complements or modifiers that are still not linked with their heads, rule (14g) will project all the NBs to NPs.

Some other phrases require NPs as their complement, so after building all the NPs, it is their turn to be built. The superlative adjective "أهم" 'most important' requires a complement to complete its meaning. So, rule (14h) will combine the JB "أهم" 'most important' with the NP "مسابقة كرة قدم دولية" 'international football contest' to build a bigger JB and assign the NP with the genitive case, then it will be projected to the maximal projection JP by rule (14b). Next, prepositional phrases (PPs) will be built. The composed NP "كرة القدم" 'football' will be combined with the preceding preposition "لِ" 'for' to form the prepositional phrase (PP) by rule in (14i) and also this rule will assign the genitive case to the NP "كرة القدم" 'football'. Figure 3 shows the composed phrases of the sentence (1).

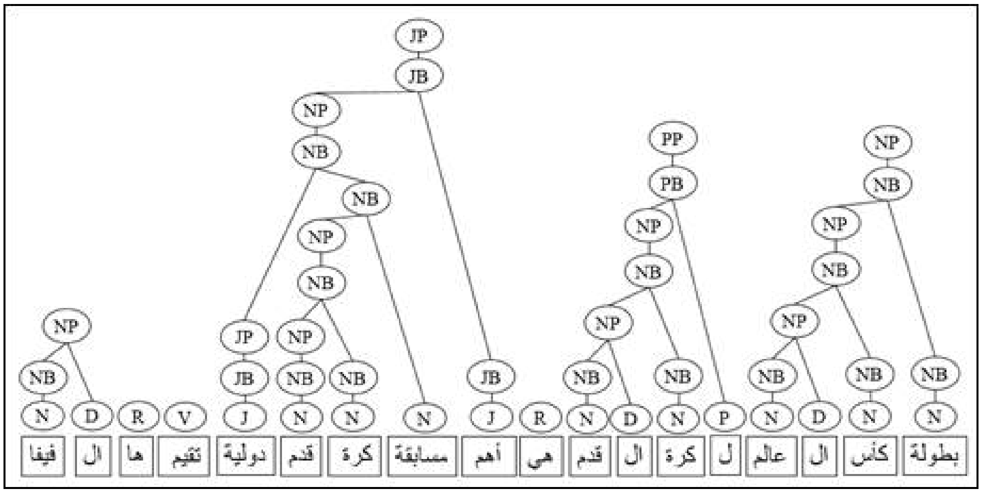


Figure 3: shows the different composed phrases in the sentence (1).

Nominal sentences diacritization: Nominative case is automatically assigned to noun phrases in the beginning of sentences, because it is considered as the topic "مبتدأ" 'mobtadaa'. So, in sentence (1) the phrase "بطولة كأس العالم" 'world cup championship' will be assigned with the nominative case (NOM) by rule (15a). In Arabic, there are different types of comments "خبر" 'khabar'; single, the noun phrase, or the prepositional phrase comment. In sentence (1), the nominal phrase "هي أهم مسابقة كرة قدم دولية" 'it is the most important international football contest' is the comment of the sentence. It consists of a noun phrase and an adjectival phrase. The noun phrase "هي" 'it' is also considered as the topic so, it is assigned with (NOM) case, and the adjectival phrase "أهم مسابقة كرة قدم دولية" 'the most important international football contest' is considered as the command so, it is assigned with (NOM) case, by rule (15b).

15)

a. (SHEAD,%s)(NP,^casatt,%n)({^COO|STAIL},%c):=(%s)(%n,topic,casatt=NOM)(%c);
b. ({SHEAD|^V|PSV},%s)(PPR,NP,%np)({NP|JP},^casatt,%w):=(%s)(%np)(%w,command,casatt=NOM);

Since Arabic is a free word-order language, it permits the advancement and delaying of some constituents of the sentence [4]. Delayed topic is a frequently occurring phenomenon in Arabic nominal sentences, such as the sentence " في الدار محمد" 'Mohammad in the house'. Delayed topic can be detected by rules in the case of the prepositional phrase comment"خبر شبه الجملة" by the rule in (16).

16) (SHEAD,%x)(%c,{PP|ADV_NP})(NP,^CAS,%y):=(%x)(%c)(%y,mobtadaa,CAS=NOM);

Rule in (16) states that if a prepositional phrase or an adverbial noun phrase comes in the beginning of the sentence and is followed by a noun phrase, this noun phrase is assigned with nominal case (NOM). However, these nominal phrases case is changed if Anna and her sisters precede it. For example, "إن محمد في البيت"'that Mohamed is in the house', the NP "محمد" 'Mohammad' became accusative. Similarly, the noun phrase or single comment of Kanna and her sisters "الخبر المفرد أو خبر الجملة الاسمية" is changed from the default case of the comment; nominative case, to the accusative case as in "كانت البنت جميلة" 'the girl was pretty'.

Verbs and their Arguments diacritization: Verb is the predicate of the sentence [4]. In sentence (1), The verb "تقيم"' 'hold' is a transitive verb that requires two arguments, one functions as a subject and the other as an object. After identifying the phrasal constructions in the sentence as in figure 3, grammar rules have been developed to assign the function and the case ending of the composed verb arguments by rule in (17). The rule states that, if a verb is preceded by a noun phrase and is followed by a pronoun and a noun phrase, the preceding noun phrase will be considered as the object of the verb if it agrees with pronoun "ها" and there is gender agreement between the verb and the following noun phrase. The second NP will be considered as the subject. Once the functions of the arguments have been determined, the case ending will be assigned to each noun phrase; the nominative case (NOM) will be assigned to the subject and the accusative case (ACC) will be assigned to the object if there is no other diacritic sign assigned earlier. The final diacritized sentences is shown in sentence(3).

17) (NP,^casatt,^subj,%n1)(V,TSTD,^TST2,%v)(SPR,GEN=%n1,NUM=%n1,%n)(NP,GEN=%v,^obj,%n2):=(obj,%n1)(%v)(%n,casatt=ACC)(subj,casatt=NOM,%n2);

Sentence 3. بُطُولَةُ كَأْسِ اَلْعَالَمِ لِكُرَةِ اَلْقَدَمِ هِيَ أَهَمُّ مُسَابَقَةِ كُرَةِ قَدَمٍ دَوْلِيَّةٍ تُقِيمُهَا اَلْفِيفَا

In fact, determining constituents' boundaries isn't an easy task, many challenges have been faced. If there is a mistake in building the constituents, it will cause problems in assigning case endings and nunation.

In what follows the focus will be on how constituents' boundaries affect case endings assignment and nunation, as well as the challenges that were faced in determining the correct boundaries.

Challenges in nunation: Before describing the challenges, the criteria that determine the assigning of nunation should be described. Since Arabic does not have indefinite article, indefinite nouns and adjectives are marked for indefiniteness with nunation. So, in order to correctly assign the nunation the noun phrase should be correctly built first and then its definiteness should be determined. In sentence in (3), rules succeeded in determining the right boundaries and the definiteness of each noun phrase. So, only "قدم"'foot' is assigned with nunation, because it is the only indefinite noun phrase, while the other noun phrases are definite by edafa or by the definite article "ال" 'the'. In sentences in(4) there are three examples that have the same syntactic pattern which is "N COO+N ART+N"; however, they function differently which poses a challenge.

Sentence 4.
(a) بتفوق وسيادة الجنس الأبيض
(b) ما بين ساعة ونصف الساعة
(c) على مختلف فئات وطبقات المجتمع

Both sentences in (4a) and (4c) should be linked in the same way. In sentence in (4a) the NB "تفوق" 'superiority' and the NB "سيادة" 'predominance' should be linked with the coordinator "و" 'and', in (4c) the NB "فئات" 'categories' and the NB "طبقات" 'classes' should also be linked with the coordinator "و" 'and' to form bigger NBs. Then the generated NBs should be linked with the NPs "الجنس الأبيض" 'the white race' in (4a) and "المجتمع" 'the society' in sentence (4c) to form yet bigger NBs. Finally, rules will project them to the maximal projection NPs and assign them with the feature definite "def". While in sentence (4b), first the NB "ساعة" 'hour' should be projected to the maximal projection NP and be assigned with the feature indefinite "indef", then it should be linked with the other NP "نصف الساعة" 'half an hour' by the coordinator "و" 'and' to form a bigger NP. The challenge is how rules could determine which constituents should be linked together and which NB should be projected directly to the maximal projection as in (4b).

The boundaries will be correctly determined for sentence (4b) and the nunation will be assigned correctly as in sentence (5b). The boundaries are also correctly determined for sentence (4a),since the two coordinated; NBs"تفوق" 'superiority' and "سيادة" 'predominance', have the same semantic feature which is state "STA". So, the boundaries will be correctly determined as in sentence (5a). But, in sentence (4c) the two coordinated NBs; "فئات" 'categories' and "طبقات" 'classes', have different semantic features. Therefore, the boundaries will be wrongly determined and the nunation will be assigned wrongly as in (5c); it still a limitation in the current version of Alserag.

Sentence 5.
(a) بِتَفَوُّقِ وَسِيَادَةِ اَلْجِنْسِ اَلْأَبْيَضِ
(b) مَا بَيْنَ سَاعَةٍ وَنِصْفِ اَلسَّاعَةِ
(c) عَلَى مُخْتَلَفِ فِئَاتٍ وَطَبَقَاتِ اَلْمُجْتَمَعِ

Challenges in verb argument diacritization: Determining the verb`s arguments is based on some criteria such as the number of constituents that follow the verb, verb`s transitivity, and the semantic features of the following noun phrases. So, any mistake in one of these factors will cause wrong case ending assignment. In sentences (6) there are two sentences that have the same syntactic pattern "V ART+N". However, the NP in (6a) should be assigned with NOM case because it is the subject of the verb, while in sentence (6b) the NP should be assigned with ACC case because itis the object of the verb. The developed rules were able to differentiate between them by using the semantic feature of the NPs as a clue. In (6a) the NP "الولد"'the boy' is assigned with the semantic feature human (HUM) which implies that this NP is the doer of the verb. In sentence (6b) the NP "التفاحة" 'the apple'is assigned with the semantic feature food (FOO) which implies that NP is the object that have been eaten. So, it will be correctly diacritized as shown in sentences(7).

Sentence 6.
(a) أكل الولد
(b) أكل التفاحة

Sentence 7.
(a) أَكَلَ اَلْوَلَدُ
(b) أَكَلَ اَلتُّفَّاحَةَ

Although the system was able to overcome many challenges, it still has some limitations. The system has been improved in this phase and still there is more potential for further improvements. One of the most difficult problems that faces a parser is structural ambiguity, sinceit leads toproblems in determining the boundaries between constituents. Despite these limitations, the system is proved to be promising when tested and demonstrated.

3) *Morpho-phonological process:* Morphological and phonological processes are tightly interrelated in spoken production. During processing, morphological processes must combine the phonological content of individual morphemes to produce a phonological representation that is suitable for driving phonological processing. Further, morpheme assembly frequently causes changes in a word's phonological well-formedness that must be addressed by the phonology [16]. Many morpho-phonological alternations occur in Arabic due to the concatenative nature of Arabic morphology, the interaction between morphological and phonological processes is usual [4]. There are different cases where morpho-phonological change is necessary, such as vowel harmony and assimilation necessity. Vowel harmony takes place in the diacritization process (i.e. phonological) [4].

5 RESULTS AND EVALUATION

The corpus has been selected from a Morphologically Annotated Gold Standard Arabic Resource (MASAR). The selected corpus size is 400,000 Modern Standard Arabic words; they are divided into 300,000 words as tuning data and 100,000 words as testing data. The selected texts are from different sources; Newspapers, Net Articles and Books representing the following genres; politics: 148,211, miscellaneous: 100,253, child stories: 57,174, economy: 34,930, society: 32,955 and sports: 26,477 as referred in [4].

However, the results that were evaluated automatically for accuracy against the reference, which is fully diacritized texts by Arabic linguists,were only 100,000words,using the following 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, the evaluation system calculates internal diacritics and case ending separately. Alserag results were compared with the output of other three known diacritization systems; Harakat, Mishkal, and Aldoaly as they are the only available systems. The outputs of these three systems were evaluated using the same data. Figure 4 shows the evaluation results of the 100.000 words. Figures 5 and 6 show the normal distribution curve of DER and WER respectively. Table 1 shows benchmarking of the 100.000 words after the syntactic improvements among the other three systems; Harakat, Mishkal, and Aldoaly as well as the result of Alserag system (Alserag1) before the improvement in the previous phase.

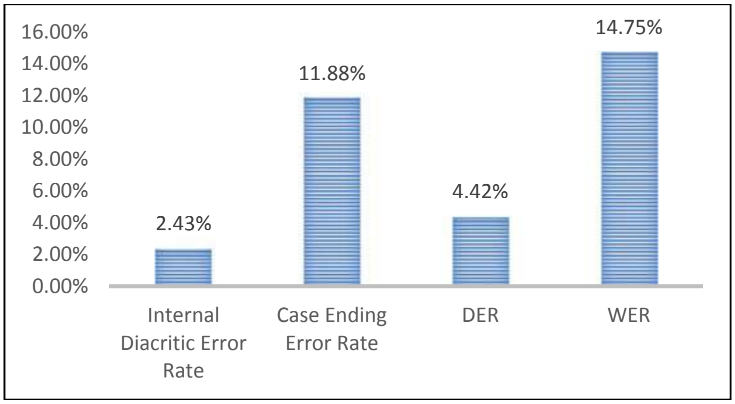


Figure 4: Evaluation of the whole data of Alserag

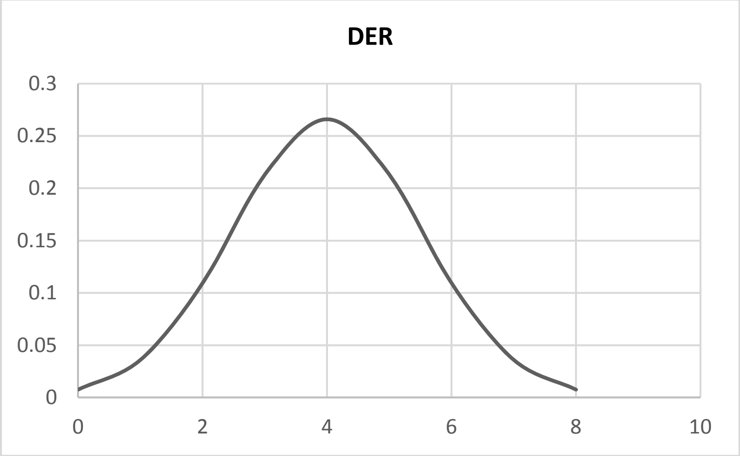


Figure 5: Normal distribution curve of DER

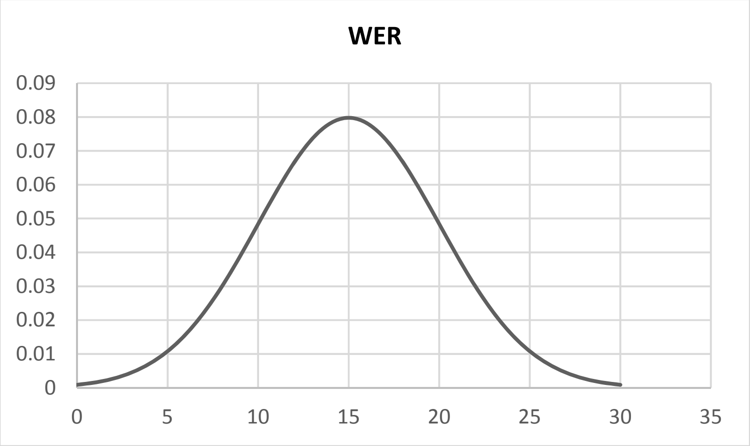


Figure 6: Normal distribution curve of WER

Table 2: Benchmarking of the whole data of Alserag among the other three systems and the previous phase (Alserag1).

	Alserag2	Alserag1	Harakat	Mishkal	Farahidy
Internal Diacritic Error Rate	%2.43	%5.77	%17.16	%22.15	%75.85
Case Ending Diacritic Error Rate	%11.88	%14.77	%16.89	%33.09	%90.27
Diacritic Error Rate (DER)	%4.42	%8.68	%17.11	%24.44	%78.87
Word Error Rate (WER)	%14.75	%18.63	%43.24	%66.28	%97.87

According to the results obtained by the benchmarking process, our system scored the least error rate followed by Harakat and Mishkal and finally Aldoaly which scored over 80% error rate.

6 CONCLUSIONS&FUTURE DIRECTIONS

The history and the definition of automatic diacritization for Arabic texts systems were presented. The techniques that are used and the different approaches in this field were thoroughly described. Also a historical and linguistic background has been given, survey of all the existing diacritization systems has been made. It is planned to perform the evaluation and benchmarking of Alserag by using the dataset of LDC (Arabic Treebank) which was used by more robust systems such as Sakhr, RDI and Microsoft system in the evaluation process so that we can compare our results with the published results of such systems.The paper presents an automatic diacritization system Alserag that is developed based on the rule- based approach, which is considered as our contribution to the automatic diacritization field. All of the other available systems that were mentioned are statistical based. The results of the system were evaluated against the reference. The DER was 2.43% while WER measurement was 14.75%.

REFERENCES

[1] M. Maamouri, A.Bies, and S. Kulick, *"Diacritization: A Challenge to Arabic Treebank Annotation and Parsing"*. Linguistic Data Consortium, University of Pennsylvania, USA, 2006.

[2] H.Bouamor,W.Zaghouani, M.Diab,O. Obeid, K.Oflazer,M. Ghoneim, and A. Hawwari. *"A Pilot Study on Arabic Multi-Genre Corpus Diacritization Annotation"*, Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 80–88, Beijing, China, c2014 Association for Computational Linguistics, 2015.

[3] A. EL-Desoky, M. Fayz, D. Samir,*"A smart Dictionary for the Arabic Full-Form Words"*, (IJSCE). ISSN: 2231-2307, Volume-2, Issue-5, 2012.

[4] S. Alansary, *"Alserag: An Automatic Diacritization System for Arabic"*. The 2nd International Conference on Advanced Intelligent Systems and Informatics (AISI'16), Cairo, Egypt, 2016.

[5] M. Al Badrashiny,*"Automatic Diacritizer for Arabic Text"*. A Thesis Submitted to the Faculty of Engineering, Cairo University in Partial Fulfillment of the Requirements for the Degree of master of science in electronics & electrical communication, 2009.

[6] D. Vergyri, and K. Kirchhoff,*"Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition"*, COLING Workshop on Arabic-script Based Languages, Geneva, Switzerland, 2004.

[7] S. Ananthakrishnan, S. Narayanan, and S. Bangalore,*"Automatic diacritization of Arabic transcripts for asr"*. In Proceedings of ICON-05, Kanpur, India, 2005.

[8] I. Zitouni, J.S. Sorensen, and R. Sarikaya,*"Maximum Entropy Based Restoration of Arabic Diacritics"*, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL); Workshop on Computational Approaches to Semitic Languages; Sydney-Australia, 2006.

[9] N. Habash, and O. Rambow,*"Arabic Diacritization through Full Morphological Tagging"*. Proceedings of the 8th Meeting of the North American Chapter of the Association for Computational Linguistics (ACL); Human Language Technologies Conference (HLT-NAACL), 2007.

[10] S.Anas, S. Khalifa and N. Habash,*"Improving Arabic Diacritization through Syntactic Analysis"*, In Proceedings of EMNLP, Lisbon, 2015.

[11] M. Attia, *"A Large-Scale Computational Processor of the Arabic Morphology, and Applications"*, M.Sc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

[12] M. Attia,*"Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications"*, PhD thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Sept. 2005.

[13] M. Alghamdi, and Z. Muzafar,*"KACST Arabic Diacritizer"*, The First International Symposium on Computers and Arabic Language, 25-28, 2007.

[14] K.Shaalan, H.M. Abo Bakr, and I. Ziedan,*"A hybrid approach for building Arabic diacritizer"*. In Semitic '09: Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, 2009.

[15] K. H. Nofal,*"Collocations in English and Arabic: A comparative study"*. *English Language and Literature Studies, 2*(3), 75-93, 2012.

[16] A.M. Cohen-Goldberg, J. Cholin, M. Miozzo and B. Rapp,*"The interface between morphology and phonology: Exploring a morpho-phonological deficit in spoken production"*, 2014.

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

تحسين نتائج المشكل العربي السراج خلال التحليل النحوي

سامح الأنصاري

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

ملخص

يعد التشكيل الآلي للنص المكتوب ذا تأثير كبير في نظم المعالجة الآلية للغة العربية. وتقوم هذه الورقة على تقديم نهج محسن لأحد أنظمة التشكيل الآلي للنصوص في اللغة العربية (السراج) والقائم على نظام قاعدي. ويقوم النظام المحسن على تضافر التحليل الصرفي مع مزيد من التحليل النحوي السطحي والذي كان له أثر كبير في تحسن نتائج العلامة الإعرابية. وقد تم تقييم هذا النظام في هذه المرحلة للوقوف على دقته وإلى أي مدى تم تحسن نتائجه. ومن أجل عملية التقييم تم استخدام مرجع لغوي بالاعتماد على مقياسين أساسيين هما معدل الخطأ على مستوى تشكيل الحروف (DER) والذي سجّل حاليا 4.42% في حين أنه كان 8.68%، ومعدل الخطأ على مستوى تشكيل الكلمات (WER) والذي سجّل حاليا 14.75% في حين أنه كان 18.63% مما يظهر تحسنا ملموسا في النتائج بعد إضافة المزيد من قواعد التحليل النحوي. وقد تم مقارنة نتائج هذا النظام بنسختيه مع ثلاثة أنظمة أخرى هي حركات ومشكال والدؤلي.

Arabic Document Pre-processing and Layout Analysis

Hassanin M. Al-Barhamtoshy
IT Dept., Faculty of Computing and Information Technology, King Abdulaziz University (KAU)
Jeddah, Saudi Arabia
hassanin@kau.edu.sa

***Abstract*- One of the greatest significant challenging tasks in image and document recognition is pre-processing and document layout analysis. The pre-processing stage includes: binarization, enhancement, noise removing, and skew/deskewing methods. The layout analysis includes developing structures detail to simplify syntactic and semantic processing over the documents. Therefore, the layout analysis includes two mean modules: page segmentation and documents category. The document categories are early printed, calligraphy document, historical document, newspaper, professional letter, etc.**

This paper provides a survey of the state of the art in the field of document analysis automation, taking into consideration the main approaches established in Arabic goal applications, and provides real practical references for implementing an Arabic document analysis system. Consequently, this paper presents an outline of basic image analysis processes used in document analysis and focus on the Arabic document pre-processing and layout tasks.

***Keywords* Image analysis, layout analysis, text region, segmentation, classification, Arabic OCR.**

1 INTRODUCTION

In the most recent centuries, large volumes of historical, early printed and historical/handwritten documents existing in public and private libraries, institutions, archives, galleries and other organizations have been digitized to create them accessible to the universal public. The automation transcript of these types of documents is difficult due to *image analysis*, *optical character recognition* (OCR) and *natural language processing* (NLP). The developments in these research areas make it possible in recent years to resolve this problem.

The main challenging tasks related to design and implement such solutions are: preprocessing and layout analysis. Therefore, image processing and document recognition for such documents, should include interactive transcription for ground truth data set and customized interfaces design. Most of the research works are related to text segmentation, recognition and text retrieving through image analysis tasks.

To achieve this layout analysis, the level of data existing in a document should be augmented to the desired analyzed elements. Consequently, an image should be transformed into an analyzed electronic form which is appropriate for the presentation structure. Firstly, an image is described by an object data of different types:

- Graphic information where the whole image is represented as a sequence of orthogonal pixel runs [1]
- Segmented information to describe texture regions
- Layout data description to represent the arrangement of objects (their geometry)
- Symbolic data representing that multiple glyph images

Many algorithms such as: *binarization*, *document enhancement*, *page orientation*, *skew and slant detection* are essential algorithms that are used in document and image pre-processing and layout analysis phases. In this section, we investigate and overview of these algorithms and related attributes. Detailed description of image processing and document analysis can be found in books, [1, 2].

Typical modules of a document understanding architecture is illustrated in Fig. 1. The related internal processes depend on the application approaches. Some modules do not need to OCR results for logical tagging.

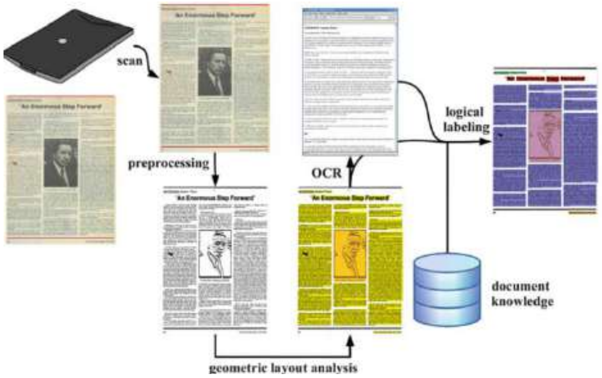


Figure 1: Typical modules of a document understanding system [2]

To understand the structure of the layout analysis, a general model represents the geometric structures of such image is generated. The layout structure is an assembly of components such that each component reflects the different interpreted analysis of this component. More details about Arabic documents analysis have been described in [3]. Each type of these components is represented as a graph using components itself as vertices and relation between component objects as edges. This flexible representation can be a directed graph or a tree structure graph. Besides, the multiple components of objects representation, in the layout, a logical labeling algorithm is used to arrange them as valid components to represent the analyzed image. Also, Fig. 1 shows the case of a document segmented into segments and classified with text regions. This segment/region is fed to an optical character recognition system. The layout segments are linked with the recognized text by logical labeling mechanism and language model stored in a document knowledge.

The organization of this paper is as the following. Section 2 presents basic image analysis algorithms used in document analysis. Section 3 illustrates page layout decomposition section with emphasis on text regions tasks, especially for binarization techniques and segmentation processes, taken into consideration all types of Arabic documents.Testing and results with be illustrated in Section 4. Finally, Section 5 concludes the paper and future work.

2 PREPROCESSING OF DOCUMENTS

As a result, a significant amount of research has been devoted to the ***binarization*** process. The binarization is used to classify all document pixels' as text or non-text regions. In general, binarization process involves threshold selection value T for each analyzed document. If *D(i, j)* represents the original gray scale document, then the subsequent binary document is defined as the following:

$$B(\,i\,,j\,)= \left\{ \begin{matrix} 1\; if\; D(i,j) \leq T \\ 0\; if\; (D(i,j) > T \end{matrix} \right\} \ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

where T represents optimal or global thresholding value [2], other research naming as adaptive thresholding [4]. There exist other strategies aiming to overcome thresholding selection from combination techniques and also from training samples.

In the second important algorithm, "***document and image enhancement***", this task improves the readability of text zones and certifies minimized image storage space. Accordingly, three categories of image artifacts can be found in the documents dataset [2, 4]. These artifacts include: (1) Not clear text regions due to low contrast and illumination of background intensity, (2) Shining or shadow effects- due to the transparency of the paper that interferes with the document on the other side, (3) Damage character or noisy background, and (4) Frames/Borders or part neighboring page. A histogram equalization is used for noise reducing and text readability increasing by Leung et al. [5]. Tonazzini [6] has proposed fast procedures projection-based to handle shining and shadow effects. Damaged characters or noisy background can be corrected by restoration algorithm [7]. Black borders or frames removing can be performed by cleaning algorithm of Shafait et al. [8]. Recently, image ***page orientation*** was presented by achieving of"*Run Length Smoothing Algorithm" (RLSA)* and black/white transition change in the two directional (vertical BW_v and horizontal BW_h) [2]. Therefore, the page orientation is identified by:

$$Page\ Orientation(i\,,j)= \left\{ \begin{matrix} \text{Landscape}\; if\; BWv \geq BWh \\ Portrait\; if\; BWv < BWh \end{matrix} \right\} \ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

Caprari [8] proposed an algorithm to exploit *text page orientation*. Other methods use local analysis and projection histograms to find the ratio between textual squares and non-textual squares for vertical and horizontal projection profiles [4].

Skewing of document is affecting the OCR subsequent phases (especially segmentation and recognition phases). Many of skew detection techniques are described: projection profiles [2,4], Hough transform using Hough space [9] and connected components [10], nearest neighboring clustering of connected components [11,12], and cross-correlation [2, 4] based on vertical deviations measurement among image pixels. Accordingly, the document image rotates at a range of angles calculated by the previous techniques. Our implemented technique is based on [13], we first segment the image, then detect the skewing angle form the segmented objects, based on page borders. If the page border is not included, therefore the page is segmented into lines using a histogram technique, obtain the skew angles for each line by using curve fitting, and finally the average skew angle calculated and rotate the whole page by this skew angle. The proposed technique is given by the following pseudo code algorithm:

1. Read the image from the stored dataset
2. Convert the image into a binary image
3. Find the connected components of the binary image and its related neighbors
4. Compute the centroid for each connected component and each related neighbor
5. For each centroid do: (a) Compute the angle between this centroid and its nearest neighbors, (b) Accumulate the angle in the histogram array
6. Determine the max value in the histogram array
7. Return by the skew angle = the max value in the histogram array

3 Page Layout Analysis and Decomposition

Any Arabic document has a hierarchical layout as shown in Fig.2. The page layout analysis includes: (1) document categories, (2) text /non-text classification, and (3) segmentation. First, the categories of Arabic documents include early printed documents, new book documents, and calligraphy documents. The document layout analysis classifies the entire document into textual and non-textual regions, in case of decomposing the document in smaller regions. Each of them is analyzed and recognized later. Given a segmented region, the objective of segmentation is to achieve a decomposition of the document image into smaller regions or pieces (lines and words segments).

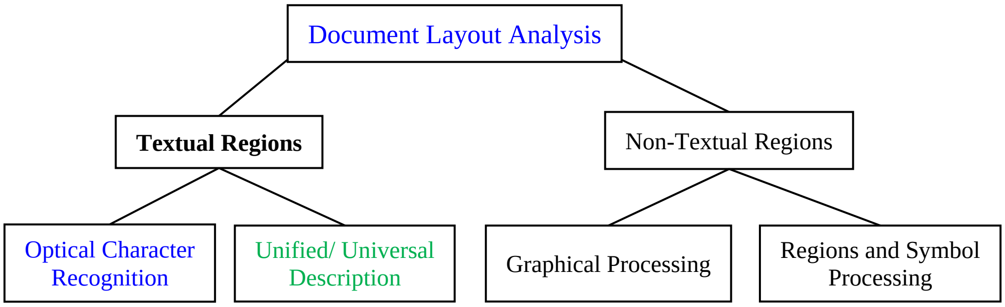
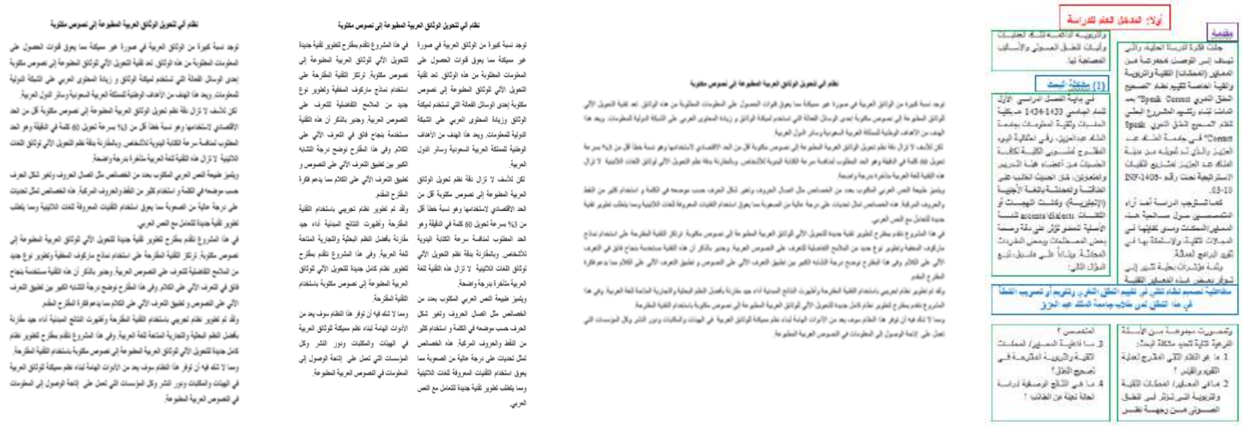

Document Layout Analysis
Textual Regions
Non-Textual Regions
Optical Character Recognition
Unified/ Universal Description
Graphical Processing
Regions and Symbol Processing

Figure 2: Document Images Layout Decompositions

Text region components are text blocks. It consists of textual lines, and a textual line is composed from words, and a word is composed from characters. Therefore, we need page segmentation to extract such text components from the input images.

A.Text-Region Components Classes

The text region level can be classified into (1) Rectangular, (2) Manhattan, (3) Overlapping, and (4) Curved layouts. The most common layout is called "*Rectangular*", whose borders are perpendicular or parallel to the sides of the page. Fig. 3 illustrates examples of this layout. Other techniques can be found in [3].

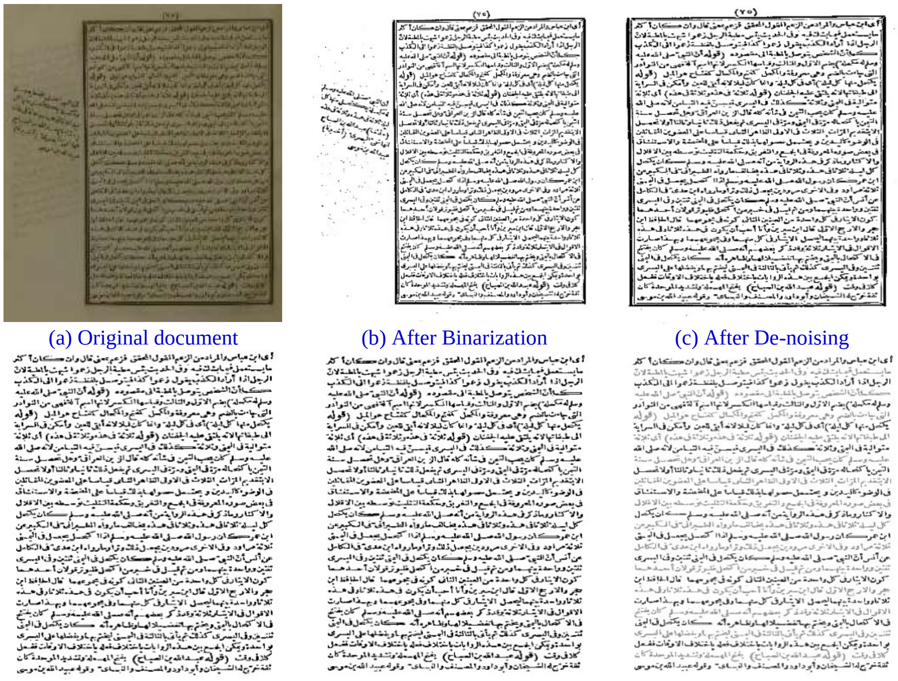


(a) Portrait 1 column (b) Portrait 2 columns (c) Landscape document (d) Textual segments

Figure 3:Three examples of Arabic printed Text-Region Components

B. Analysis of Calligraphy Pages

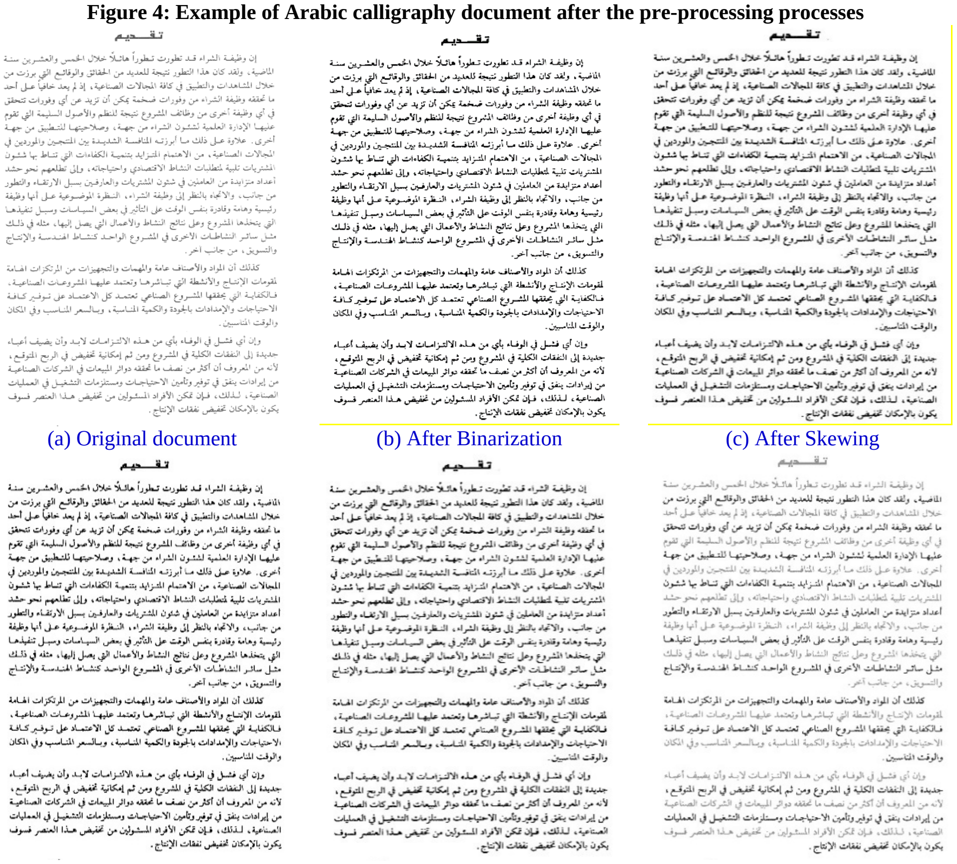
In order to analyze the calligraphy pages (good Arabic handwritten documents), let us briefly explain a process anticipated by Zheng et al.[14]. Such method employments connected components as elements of processing. Therefore, the goal here is to categorize connected components into either early printed or calligraphy/good handwriting. To complete this objective for noisy or not clear documents, Zheng et al. [14] describe their classification task into either machine printed, handwriting, or noise. Accordingly, the Zheng et al. task is improved to be used in the three categories of the Arabic document types (early printed, books, and Calligraphy). The used extracted features are like those in page **segmentation** (e.g. Gabor filters). The next step includes "features extraction" from each connected component, with classification achievement. In other way, "Fisher classifier" is applied after feature selection by principal component analysis. Later the classification task is error prone due to the limited amount of information from each connected component, the language model and extra processing are necessary to filter the results by considering the appropriate and contextual information. Fig. 4, 5 and 6show examples of the Arabic documents after binarization, noise removing, frame removing, skewing and de-skewing, and segmentation processes.



(a) Original document (b) After Binarization (c) After De-noising

(d) After De-Framing (f) After De-skewing (g) AfterSegmentation

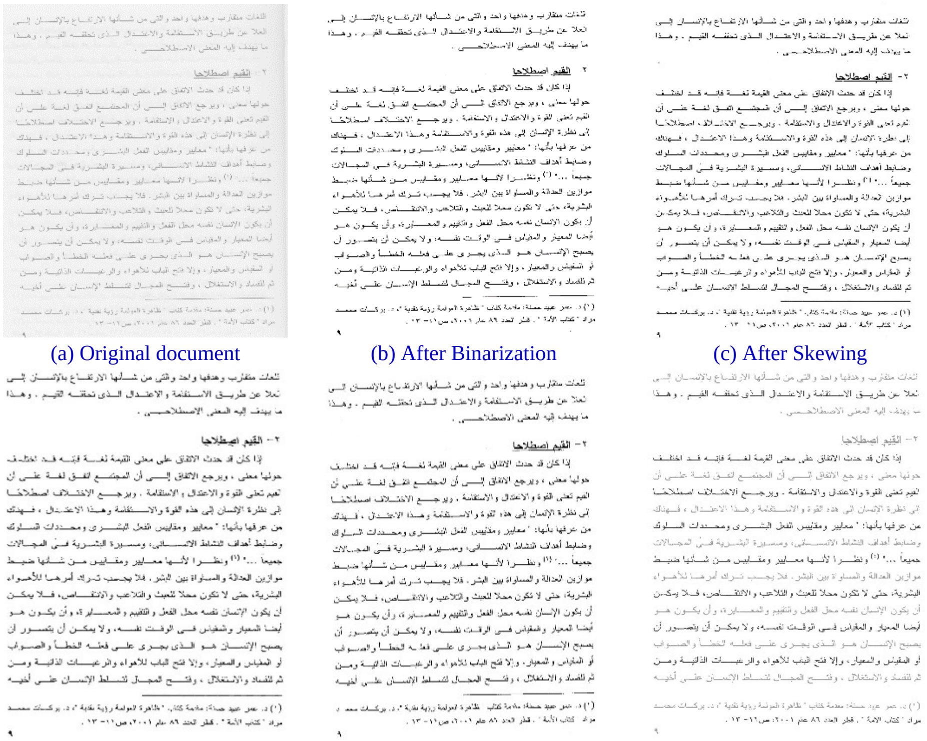
Figure 4: Example of Arabic calligraphy document after the pre-processing processes



(a) Original document (b) After Binarization (c) After Skewing

(d) After De-noising (f) After De-Framing (g) AfterSegmentation

Figure 5: Example of Arabic document after the pre-processing processes



(a) Original document (b) After Binarization (c) After Skewing

(d) After De-noising (f) After De-framing (g) AfterSegmentation

Figure 6:Another Example of Arabic document after the pre-processing processes

4 TESTING AND RESULTS

The previous mentioned discussion was implemented and tested on Arabic document images with three categories of domains. Fig. 5 and 6 show samples of images (Arabic printed documents from standard books) that have been tested. The comparative test has been evaluated between the five tasks (binarization, skewing/de-skewing, noise removing/ De-noising, frame removing/ De-framing, and segmentation) processes. In segmentation process two colors are used to differentiate between lines' segments. Therefore, two aspects of measurements are used: speed and accuracy.

(A) Speed

The proposed Arabic OCR have been implemented in windows environment with visual C++ language. The first test has been done on laptop with Intel i7-3667U CPU, 2.0 GHz with 8 GB of memory and 256 GB hard drive. Table 2 listed the computational time using 5 images from each category. The proposed Arabic OCR system is tested according to the universal Meta data description of the dataset for training.

TABLE I
DETAILED OF PRE-PROCESSING TASKS FOR THE PROPOSED ARABIC OCR (SECONDS)

Image #	Categories														
	Early Printed					New Books					Calligraphy				
	Binarization	Skewing/De-skewing	Noise removing/De-noising	Frame removal/De-framing	Segmentation	Binarization	Skewing/De-skewing	Noise removing/De-noising	Frame removal/De-framing	Segmentation	Binarization	Skewing/De-skewing	Noise removing/De-noising	Frame removal/De-framing	Segmentation
Doc. 1	3.0	1.2	.05	0.3	6.4	2.4	0.6	.04	0.3	4.5	3.3	1.3	.06	0.3	7.2
Doc. 2	3.2	1.1	.06	0.2	7.0	2.8	0.7	.03	0.2	5.0	3.4	1.2	.07	0.3	7.5
Doc. 3	3.5	1.4	.07	0.4	8.0	3.0	0.8	.04	0.4	5.5	3.7	1.4	.08	0.5	8.4
Doc. 4	3.4	1.3	.06	0.3	7.1	2.7	0.8	.05	0.3	6.0	3.6	1.5	.07	0.5	7.6
Doc. 5	3.3	1.5	.06	0.4	6.5	2.6	0.6	.04	0.3	4.0	3.5	1.6	.07	0.4	6.7

Average	**3.3**	**1.3**	**.06**	**0.3**	**7**	**2.7**	**0.7**	**.04**	**0.3**	**5**	**3.5**	**1.4**	**.07**	**0.4**	**7.5**

(B) Accuracy

Comparing the speed of different processes (binarization, skewing and de-skewing, frame removing, noise removing, and segmentation) of the Arabic documents for different categories.

TABLE II
SPEED PROCESSING (SECOND)

Image Type	**Processes speed (Seconds)**				
	Binarization	**Skewing/ De-skewing**	**Noise removing**	**Frame removing**	**Segmentation**
Early Printed	3.3	1.3	0.06	0.3	7
New Books	2.7	0.7	0.04	0.3	5
Calligraphy	3.5	1.4	0.07	0.4	7.5
Average	3.16	1.11	0.056	0.33	6.5

5 CONCLUSION

This paper has described several approaches of pre-processing and layout analysis modules of the Arabic document analysis, focusing mainly on *binarization*, *document enhancement*, *page orientation*, *and skew* algorithms. Accordingly, many algorithms are mentioned such as projection profiles, Manhattan, and non-Manhattan, connected components, etc. The layout analysis undertakings are tested taken into consideration:1-page segmentation, and 2-document categorization in the Arabic images types (early printed, new books, and calligraphy documents). We have seen that connected components with the language model classification by means of texture features is a capable way to complete the task.

Due to faster processors of high-performance computing, and big storage with bigger data sets and experimentation will finally bring us to improve the accuracy of the proposed Arabic OCR system. Therefore, a deep-believe neural networks becomes more compatible with the mature of recognition tasks.

ACKNOWLEDGEMENT

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology -the Kingdom of Saudi Arabia– award number (11-INF-1997-03). The author thanks Science and Technology Unit, King Abdulaziz University for technical support. Also, the author would like to thank RDI team, especially, Eng. Shaimaa Samir for her developing and testing results.

References

[1]. Doermann D., Tombre K., (2014) Handbook of Document Image Processing and Recognition, Vol. (1), Springer Reference.

[2]. Dengel A. and Shafait F. (2014) Analysis of the Logical Layout of Documents, Editors: D. Doermann, K. Tombre, *Handbook of Document Image Processing and Recognition*, (Ch. 7: pp 177-222).

[3]. Al-Barhamtoshy H. (2016) Towards Large Scale Image Similarity Discovery Model, 2nd International Conference on Advanced Technologies for Signal and Image Processing ATSIP'2016, March 21-24, Monastir Tunisia, http://ieeexplore.ieee.org/stamp/ stamp.jsp?tp=&arnumber=7523047

[4]. Gatos B. G. (2014) Imaging Techniques in Document Analysis Processes, Ch. 4: Handbook of Document Image Processing and Recognition, Vol. (1), pp. 73-131.

[5]. Leung CC, Chan KS, Chan HM, Tsui WK (2005) A new approach for image enhancement applied to low-contrast–low-illumination IC and document images. Pattern Recognit Lett 26:769–778.

[6]. Tonazzini A (2010) Color space transformations for analysis and enhancement of ancient degraded manuscripts. Pattern Recognition Image Analysis 20(3):404–417.

[7]. Drira F, LeBourgeois F, Emptoz H (2011) A new PDE-based approach for singularity preserving regularization: application to degraded characters restoration. Int J Doc Anal Recognit. doi:10.1007/s10032-011-0165-5.

[8]. Shafait F, van Beusekom J, Keysers D, Breuel TM (2008) Document cleanup using page frame detection. Int J Doc Anal Recognit 11:81–96.

[9]. Li S, Shen Q, Sun J (2007) Skew detection using wavelet decomposition and projection profile analysis. Pattern Recognit Lett 28:555–562.

[10]. Amin A, Fischer S (2000). A document skew detection method using the Hough transform. Pattern Anal Appl 3(3):243–253.

[11]. Cao Y, Li H (2003) Skew detection and correction in document images based on straight-line fitting. Pattern Recognit Lett 24(12):1871—1879.

[12]. Lu Y, Tan CL (2003). A nearest-neighbor chain based approach to skew estimation in document images. Pattern Recognit Lett 24:2315–2323.

[13]. Liu, Hong, et al. "Skew detection for complex document images using robust borderlines in both text and non-text regions." Pattern Recognition Letters 29.13 (2008): 1893-1900.

[14]. Zheng Y, Li H, Doermann D (2004) Machine printed text and handwriting identification in noisy document images. IEEE Trans PAMI 26(3):337–353.

[15]. Al-Barhamtoshy H., Khemakhem M., Eassa F., Fattouh A., Al-Ghamdi A., Jambi K., (2016) Universal Metadata Repository for Document Analysis and Recognition, the 13th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2016, (Accepted).

BIOGRAPHY

Hassanin Al-Barhamtoshy received his B.Sc. degree from Electronics and Communication Engineering Department, ***Cairo University***, and the M.Sc. degree in Systems and Computers Engineering from ***Azhar university***, Cairo. In 1992, he received the Ph.D. degree in Systems and Computers Engineering from **Azhar University**, he was an Assistant Professor in the Systems and Computer Engineering dept. During 1996-1997 he was an Assistant Professor in Computer Science at **KAU University**, Jeddah, Saudi Arabia. During 1998-2002 he was an Associate Professor in Computer Science at **KAU University**. He is currently Professor in the Information Technology dept., at Faculty of Computing and Information Technology (2003-now).

تحليل شكل الوثيقة العام والمعالجة المبدئية للوثيقة العربية

أ.د. حسنين محمد البرهمتوشي

قسم تقنية المعلومات – كلية الحاسبات وتقنية والمعلومات – جامعة الملك عبد العزيز

المستخلص. تعد عمليتي المعالجة المبدئية وتحليل البناء العام للوثيقة أحد التحديات المهمة في التعرف على الصور والمستندات. مرحلة "ما قبل المعالجة (المعالجة المبدئية)" تشمل: binarization، وتعزيز وإزالةالضوضاء، واكتشاف الانحراف وتصحيحه. وتشمل عملية "تحليل البناء العام" التعرف على تفاصيل مكونات المستند وتبسيط تفاصيل التركيب النحوي والدلالي للوثيقة، وبالتالي يتضمن هذا التحليل نموذجي: تقسيم الصفحة(مكونات نصية، مكونات رسومية) وتحديد نوع المستند. وحدد نوع المستند بفئات قد تكون من النوع المطبوعة مبكراًearly printed، وثيقةالخط المحسن calligraphy، وثيقةتاريخيةhistorical، صحيفة، بريدإلكتروني مهني، الخ

تقدم هذه الورقة مسحا علمياً لفنون أتمتة تحليل الوثائق العربية، وتأخذ بعين الاعتبار النهج الرئيسية في الكتابة العربية، ويقدم إشارات عملية حقيقية لتنفيذ نظام تحليل الوثائق العربية. ونتيجة لذلك، تقدم هذه الورقة عرضا لعمليات تحليل الصور الأساسية المستخدمة في تحليل الوثائق والتركيز على الوثيقة العربية في التجهيز الأولي (المعالجة المبدئية) للوثيقة ومهام تحليل الشكل العام.

كلمات مفتاحية: تحليل الصور، التحليل والتخطيط، منطقة النص، تجزئة، تقسيم، والتعرف على العربية المكتوبة أو المطبوعة.

Software and Hardware Implementation for Documents Classification using Self-Organizing Maps (SOM)

Abdelfattah ELSharkawi[1], Ali Rashed[2], Hosam Eldin Fawzan[3]

[12]Department of Systems and Computer Engineering, Al-Azhar University, Egypt.

[1]sharkawi_eg@yahoo.com

[2]a_m_rashed@hotmail.com

Department of Electrical and Computer Engineering, Faculty of Engineering Science, Sinai University, Egypt.

[3]Hos.9876@yahoo.com

***Abstract*: Self-Organizing Maps (SOM) is one of the most popular unsupervised neural networks and used for clustering unseen patterns and visualization of data. Clustering and classification plays a large part in today's world. SOM in this paper will be used for classification instead of clustering documents. Reuters 21578 Corpus from the news agency Reuters International and Movies Reviews from Amazon datasets are trained and tested to measure the accuracy of Classification. The experiments were carried out on different dataset sizes to determine their influence on Classification results. The testing results approve that Self-Organizing Maps as it is well suited for clustering also suited for classification of document collections and introduces good results. The hardware implementation of SOM using Field Programmable Gate Array (FPGA) aims to accelerate the classification process.**

1 INTRODUCTION

There has been a massive increase in use of electronic documents in the recent past due to the increase of World Wide Web. Therefore, organizations tend to store most of their data in digital format [1]. Text categorization is one major research area of text mining. Text categorization is the process of grouping documents in a supervised manner based on the pre-defined labels.

The Self-Organizing Map (SOM) [2], [3] is a neural network based clustering algorithm highly recognized for its visualization capabilities. Therefore, it has been shown to be one of the best text clustering and visualization algorithms [4]. SOM [5] is an unsupervised neural network model used effectively to map high-dimensional data to a low dimensional space (usually two dimensional). The low-dimensional space (also called output space) consists of a grid of neurons connected with each other; according to a specific structure (can be hexagonal, rectangular, etc.). This structure allows the topology preservation of input data (i.e., similar input patterns are expected to be mapped to neighboring neurons in the output grid) [6]. By this way, SOM manages to achieve dimensionality reduction, abstraction, clustering and visualization of the input data and in classification as in this research. This is the reason that it has been applied successfully to many different domains and datasets like financial data [7], speech recognition [8], image classification [9], document clustering [10], [11].

Categorization of a text corpus in which each article is attributed with a set of categories; this called a classical supervised classification task. Most supervised classification methods learn parameters from a training set of labeled instances, and use the learned model to score test instances. The Self-Organizing Map (SOM) is in contrast an unsupervised technique, clustering similar training instances together, without knowledge of their categories. The resulting maps display visually identifiable, but non-delimited, clusters [12]. SOM uses continuous Vector Space Model (VSM) that maps words and documents into a low dimensional space [13] called Term Document (TD) matrix as feature extraction. TD matrix evaluates the rare terms with low weight which (in some cases) is considered as more informative than defining frequent terms. In practice, a weighting scheme that better captures the importance of a word in the document than VSM is TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is one of the feature factorization methods widely used in text mining that can reflect the importance of terms in documents, and hence it is used as the first process in text mining to extract the features of documents in a dataset.

Field-programmable gate arrays (FPGAs) are generic, programmable digital devices that can perform complex logical operations. FPGAs can replace thousands or millions of logic gates in multilevel structures. Their high density of logic gates and routing resources, and their fast reconfiguration speed give them the advantage of being extremely powerful for many applications. FPGAs are widely used because of their rich resources, configurable abilities and low development risk, making them increasingly popular. FPGAs have been available since the 1980s, but have only recently started to become popular as computation accelerators.

On the other hand, Kohonen's self- organizing map (SOM) represents one of the most machine learning techniques used in clustering and information retrieval. There are many challenges facing SOM parameters that govern the clustering process and hence achieve the expected results. Among these parameters are the initializations with random weights, the scheme of the neighborhood shrinking function, the map size, and the definition of the learning rate [14]. This paper has used two global datasets for testing the accuracy of classification. The first one is Reuter-21578 "ApteMod. The "ApteMod" is a collection of 10,788 documents partitioned into a training set of 7769 documents and a test set of 3019 documents. 100,250 and 500 documents for each one of four categories (earn acquisition, crude, and trade). The second is 2000 movies reviews from Amazon (1000 positive and 1000 negative). The training is run for 500 for each category and then 1000 for each one

2 SELF-ORGANIZING MAPS (SOM)

Kohonen's self-organizing maps (SOM) are abstract mathematical models used for clustering of data [16]. The SOM consists of a topological grid of neurons typically arranged in one or two dimension lattice [17]. The SOM Learning algorithm steps are:

1. Select an input vector X (t) = (x1 (t), x2 (t), ..., xn (t))
2. Find winning node by calculating the Euclidian distance
   $d_s = \min\|X(t) - W(t)\|$, where Wk (t)= (wk (t), wk (t), ..., wk (t)).
3. Adjust weights as follows:
   $$W(t+1) = W(t) + \eta(t)*(X(t) - W(t)), \text{ where } 0< \eta(t) <1. \quad (1)$$

where the learning rate function is:

$$\eta\ (t, k, s) = A1 * \frac{1}{e^{\left(\frac{t}{A2}\right)}} * e^{\left(\frac{-d(k,s)}{2\sigma 2}\right)} \quad (2)$$

where d (k, s) is the Euclidian distance between the node k and the winning node s in the two-dimensional grid, while A1 and A2 will be defined latter in Eq.(3) and Eq.(4). In the formula, the first Gaussian function controls the weight update speed and the second Gaussian function defines the neighborhood shrinkage function in SOM. The standard deviation σ decreases monotonically with time [18].

ηstart: 0 < ηstart < 1, is the starting value (value at time t = 0) for η for the winning node s. Note that the time t goes from 0 to (C-1).

ηend: 0 < ηend < ηstart, is the final value (value at time t = (C-1)) for η for the winning nodes [19].

From Eq. (5) it is clear that at time t=0, η (0, k, s) = A1. Hence:

$$A1=\eta start \quad (3)$$

$$A2=(C-1)/\ln(\eta start/\eta end) \quad (4)$$

3 THE NEW MODEL

The suggested model comprises a collection of processes; namely choosing the training dataset, preprocessing, feature extraction, training using SOM and hence calculates the average weight of each cluster in the SOM map for testing as in Fig. (1). TF-IDF has used in feature extraction stage which is known as the best weighting scheme data mining for training documents. TF-IDF represents the documents that train by SOM and clustering them in a map. After finishing the training process then calculate the average weight of all documents that belong to each cluster. This process a sign only one weight vector for each cluster which influences in determining the accuracy of classifying the tested document correctly by measuring the cosine similarity between the weight vectors and tested document vector.

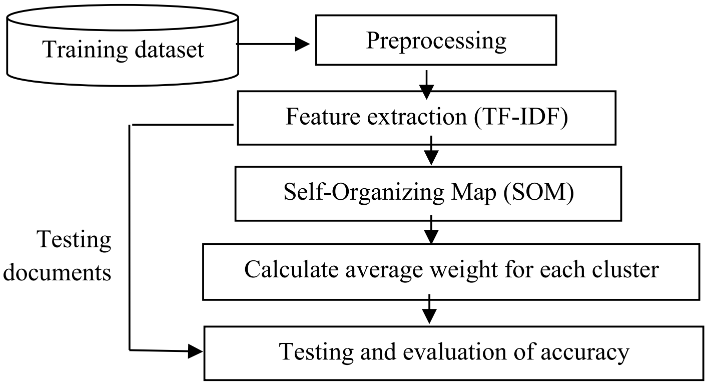

Training dataset
Preprocessing
Feature extraction (TF-IDF)
Self-Organizing Map (SOM)
Testing documents
Calculate average weight for each cluster
Testing and evaluation of accuracy

Fig 1: The main algorithm for clustering and searching of text documents

3.1. PREPROCESSING

Data mining techniques aims at having the data in a structured form and hence can easily obtain the knowledge. We have used software tools called Weka (Waikato Environment for Knowledge Analysis) to implement those processes which includes: tokenization (breaking up a sequence of strings into pieces such as keywords, functional phrases, symbols and other elements called tokens), remove stop words (e.g. the, am, is, are etc.), stemming using porter stemming algorithm [22] (which means returning each word to its original form) and generate counting terms. All these steps are summarized in Fig. (2).

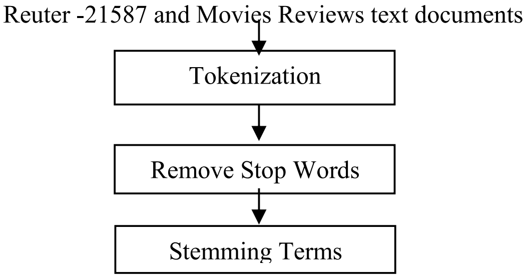


Fig 2: Steps of preprocessing

3.2. FEATURE EXTRACTION (TFIDF)

TF-IDF algorithm calculates an index for measuring the importance of a term to a document in a corpus. It is used for calculating the frequency of terms of a given word in a given collection of documents and calls it Term Frequency (TF) as shown in equation (4). It also calculates the Inverse Document Frequency (IDF) as in equation (5). The term count or the number of times the term appears in document indicates the importance of that term in this document [7]. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [7]. The TF-IDF is the product of term count (TF) and Inverse Document Frequency (IDF). The term frequency TF of term t_i in document d_j is calculated by equation (1).

$$\mathrm{TFij} = \frac{\mathrm{Nij}}{\Sigma \mathrm{Nij}} \qquad (5)$$

where N_{ij} , is the number of occurrences of the considered term t_i in document d_j.

$$\mathrm{IDFi} = \frac{\log|\mathrm{Dtotal}\,|}{|d: ti\ \varepsilon d|} \qquad (6)$$

where |Dtotal | is total number of documents in the corpus and d : t_i ε d is number of documents where the term t_i appears. The TF-IDF for each term t can be defined as in Eq. (6) [7].

$$\text{(TF-IDF) weight} = TF_{ij} * IDF_i \qquad (7)$$

3.3. SOM CLUSTERING ALGORITHM FOR DOCUMENT CLUSTERING

1. Each node's weights are initialized randomly
2. Select a random vector from a feature TF-IDF and presented in the lattice
3. Calculate the BMU.
4. Adjust the weights of the winning node and the weights of its neighboring node in the grid.
5. Adjust the learning rate L(t) as explained in Eq. (4).
6. Repeat step 2 N iterations.

3.4. CALCULATE AVERAGE WEIGHT FOR EACH CLUSTER

The map size for training document is 18*18 which represent 324 vectors. We should measure the cosine similarity between each of 324 weight vectors and each test document to determine to which class the test document belongs. This introduces 324 angels and these angels are convergent. But when calculating one average weight for each cluster by selecting the files that belong to each group and calculate the average weights for them will introduce 4 weight vector and the angels between them and the test documents becomes not convergent. This will influence the accuracy of tested documents and achieving best results using the Reuters and Movies Reviews datasets.

3.5. TESTING OF SOM PERFORMANCE

The evaluation of a classification task is defined by three measures namely; Precision, Recall and F measure. The precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class) as Eq. (8)[23].

$$Precision = \frac{TP}{TP + FP} \qquad (8)$$

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been) as Eq. (9) [23].

$$Recall = \frac{TP}{TP + FN} \qquad (9)$$

The accuracy of two measures (Precision +Recall) can be combined using F score as in the following Eq. (10).

$$Accurecy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

The accuracy (F) values are within the interval [0, 1] and larger values indicate higher classification quality.

3.6. IMPLEMENTATION OF SOM BY FPGA

The most critical aspect of any hardware design is the selection and design of the architecture that provides the most efficient and effective implementation [20]. SOM has implemented in a digital version of the SOM on FPGA by replacing the Euclidean distance computations to avoid the expense of hardware multiplication [21].

The implementation of the proposed model for SOM onto hardware results in large designs, which consumes substantial hardware internal resources (slices, registers and look-up table (LUT) units), limiting the scale of network implementation. FPGA vendors provide tools that allow the designer to build embedded systems on efficiently on FPGAs. This hardware design is implemented using Xilinx Software Development Kit (SDK) and Xilinx Embedded Development Kit (EDK) as in Fig. 3.

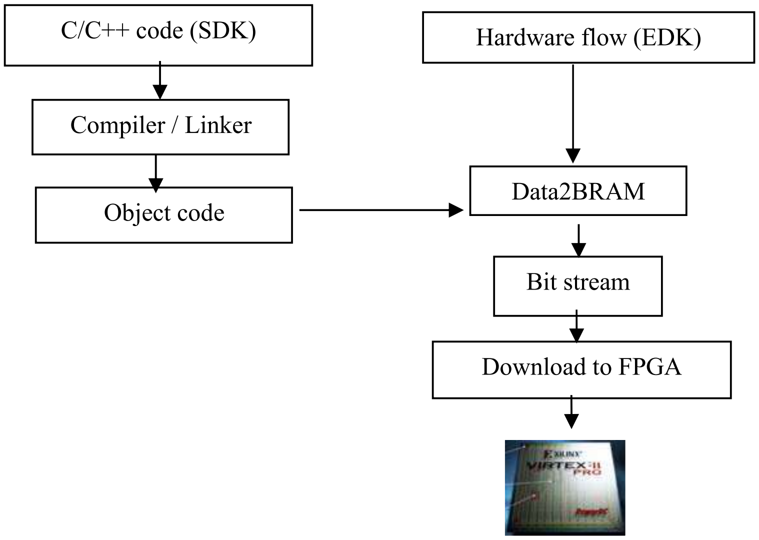


Fig.3: EDK and SDK model to run SOM on FPGA

Xilinx Embedded Development Kit (EDK) which allows the designer to build MicroBlaze embedded processor systems in Xilinx FPGAs. Software Development Kit (SDK) is a collection of software used for developing applications for a specific device or operating system. The tool provides a C/C++ compiler for that processor and an IDE based on Eclipse framework. In this design we will build a processor system based on MicroBlaze using the EDK and run this system on Nexy3 FPGA board. The Nexys 3 Spartan-6 FPGA, that has used to implement the SOM shows in Fig. 4.

Fig 4: Nexys 3 Spartan-6 FPGA Trainer Board

The FPGA contains several memory blocks up to 128 KB. This is different from the on board memories. The 64KB of these block memories has connected to the MicroBlaze as a local processor memory.

4 IMPLEMENTATION AND DISCUSSION OF RESULTS

SOM learns to classify data without supervision. With this approach an inputs vector is presented to the network through the training process for 800 iterations. Each 100 iterations the testing process is run to determine the documents that belong to each cluster and then calculate the average weight vector of these documents belong to represent each cluster. Then measure the cosine similarity between the average weight of each cluster and the tested document. The accuracy of classification documents is measured each 100 iterations using (Precision, recall and F-measure) for tested documents.

4.1 CLASSIFICATION RESULTS

The training process has implemented for two different dataset router dataset and movies review. At first the router dataset trained by using 400 documents (100 for Earning, Acquisition, Trade and Crude class) are used for training process. The testing process is done by using 120 documents (30 for each class). Table 1 makes comparative results between SOM [24] and the new model.

TABLE 1: PRECISION (P), RECALL(R) and F MEASURE FOR 4 CLASS

	SumithMatharage [23]			**Suggested model**		
Category name	P	R	F	P	R	F
Earning	0.82	0.84	0.83	0.93	0.82	0.87
Acquisition	0.75	0.80	0.77	0.78	0.86	0.82
Trade	0.78	0.82	0.80	0.83	0.96	0.89
Crude	0.72	0.73	0.72	0.96	0.80	0.87

From the four classes appear in Table (1), we choose the first two classes to increase the space for each class during training documents. 2000 documents (1000 for earning class and 1000 for Acquisition class) have trained using SOM. The testing process evaluated using 1000 documents (500 for each class). The following table makes comparative results between Sumith Matharage [23] and the suggested model.

TABLE 2. PRECISION (P), RECALL(R) AND F -MEASURE FOR EARNING AND ACQUISISION CLASSES

		Sumith Matharage [23]			**Suggested model**		
dataset	Category name	P	R	F	P	R	F
1000	Earning	0.82	0.84	0.83	0.98	0.97	0.98
	Acquisition	0.75	0.80	0.77	0.97	0.98	0.98
2000	Earning	0.82	0.84	0.83	0.97	0.98	0.97
	Acquisition	0.75	0.80	0.77	0.98	0.97	0.97

The results in Table (2) show that, the accuracy of Precision, Recall and F-measure has increased for the two classes compared with the two classes result in Table (1). This means that reducing the number of clusters for training by SOM from 4 to 2 has increased the accuracy of classification because each of the two classes occupies more space in SOM map instead of training 4 classes. This influence on calculating the average weight of each of the two class and reduce the distance between clusters.

The second dataset is the movies reviews consist of 2000 documents (1000 reviews positive and 1000 reviews negative). The SOM trains these documents as two classes (the first has label 1 and second has label 2). The results in table 5 show the comparative between research in[23] and the proposed model using the same dataset for classification. These results show that, the classification accuracy of two classes (positive and negative documents) has been increased comparative with [23]. The accuracy percent has increased F-totally by 6% approximately.

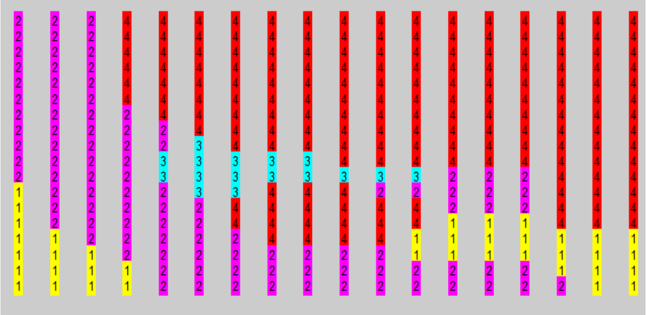
TABLE 3: PRESCISION, RECALL AND F-MEASURE FOR MOVIES REVIEWS DATASET

	B. Ohana and B. Tierney [24]		**A. Hamouda [25]**		**Suggested model**	
Category name	P	R	P	R	P	R
Positive	55.55%	80.35%	62.88%	83%	52.8%	92.8%
Negative	64.5%	35.7%	75%	51%	88.0%	66.3%
F-measure	58.03%		67.00%		72.8%	

The resulting maps display visually each cluster of documents separated. This is because the SOM has a neighborhood function affect the node to the neighbor weight. Each cluster of documents has a specific label putted when the weight of any node inside the grid is closer to that cluster. In this way, the SOM algorithm makes underlying similarities in high-dimensional space visible in lower dimensions. Through a two-step methodology, the labels of a training corpus associate with areas of the map; areas that in turn can be used to classify previously unseen documents.

The network is created from a 2D lattice of 'nodes', each of which is fully connected to the input layer. Fig.3 shows a very small Kohonen's network of 18 * 18 nodes connected to the input layer representing a two dimensional vector. All neurons in the output layer are well connected to adjacent neurons by a neighborhood relation depicting the structure of the map. Generally, the output layer can be arranged in rectangular lattice.

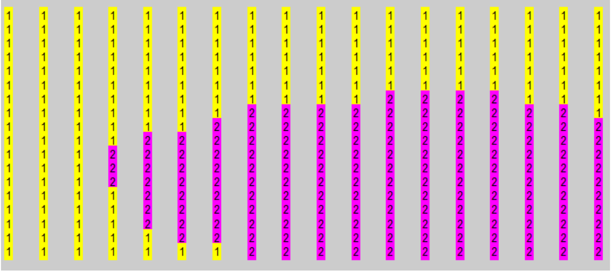
The training process produces 18*18 map each 100 cycles until reaches to the end of training cycles. Each one of the 4 classes has a different label (1,2,3 and 4) as view in Fig (5).



2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4
2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4
2	2	2	2	3	3	3	3	3	3	3	3	2	2	2	4	4	4
1	2	2	2	2	3	3	4	4	4	2	2	2	2	2	4	4	4
1	2	2	2	2	2	4	4	4	4	4	4	2	2	2	4	4	4
1	2	2	2	2	2	4	4	4	4	4	4	1	1	1	4	4	4
1	1	2	2	2	2	2	4	4	4	4	1	1	1	1	1	1	1
1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	1	1	1
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1

Fig 5: Clustering of four classes by 18*18 Kohonen map

When using the SOM to train two classes the first has a label 1 and the second has label 2. One of the maps will view as shown in Fig (6).



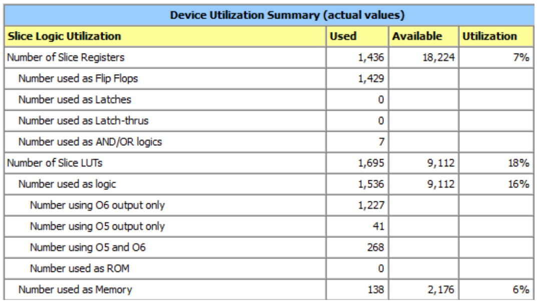
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	1	1	1
1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2

Fig 6: Clustering of two classes by 18*18 Kohonen map

After many experiments, the size 10*10 looks to be appropriate for arbitrary queries. Of course, increasing the size of the map will result in longer processing times, since many more weight vectors will need to be considered.

4.2 FPGA RESULTS

One of the results of implement the suggested model on FPGA is determine the Device utilization summary as in Fig. (7).



Device Utilization Summary (actual values)			
Slice Logic Utilization	**Used**	**Available**	**Utilization**
Number of Slice Registers	1,436	18,224	7%
Number used as Flip Flops	1,429		
Number used as Latches	0		
Number used as Latch-thrus	0		
Number used as AND/OR logics	7		
Number of Slice LUTs	1,695	9,112	18%
Number used as logic	1,536	9,112	16%
Number using O6 output only	1,227		
Number using O5 output only	41		
Number using O5 and O6	268		
Number used as ROM	0		
Number used as Memory	138	2,176	6%

Fig 7: Device Utilization Summary

There is a comparative between [21] and this study that views the resources available and the usage percentage. This comparative will views in Table (4).

TABLE 4: IMPLEMENTATION RESULTS FOR THE SOM

	SOM by ref[21]			**The proposed SOM**		
Resources Name	Available	Used	Per.(%)	Available	Used	Per.(%)
Flip flops	135,168	4,095	3%	18,224	1,436	7%
4 input LUTs	135,168	18,387	13%	9,112	1,536	18%
Bounded IOBs	768	147	19%	232	18	7%
Occupied Slices	67,584	11,468	16%	2,278	752	33%

where LUT is basically a table that determines what the output is for any given input(s). Bounded IOB (Input/output Buffer) determines the number of pins of FPGA were used on the device and Occupied Slices refer to the basic building block components in the FPGA. However, each slice contains a number of LUT's, flip-flops, and carry logic elements which make up the logic of your design before mapping

5 CONCLUSIONS

Self-organizing map is good in clustering and visualization data mining such as documents. The method of calculating the average weight vector to each cluster has enhanced the accuracy of classification of documents using SOM map. Implementation of the SOM for classification using Reuters-21578 dataset has increased the accuracy percentage between 4%-15%. When reducing the number of classes to 2 and increasing the numbers of documents for training and testing the accuracy percentage again, the accuracy has increased to become between 15%-25% approximately for the same two classes. When using movies reviews dataset, the accuracy has increased by 6%. The hardware implementation of the suggested model has consumed little component in FPGA which lead researchers to use another FPGA beard with low price to implement the SOM. SOM has a proved that it's good in clustering as well as in classification of documents.

REFERENCES

[1] A. Haug, "The implementation of enterprise content management systems in smes", Journal of Enterprise Information Management, vol. 25, no. 4, pp. 349–372, 2012.

[2] T. Kohonen, "Self-organized formation of topologically correct feature maps," Biological Cybernetics, vol. 43, pp. 59–69, 1982.

[3] T. Kohonen, "Essentials of the self-organizing map," Neural Networks, vol. 37, pp. 52–65, 2013.

[4] D. Isa, V. Kallimani, and L. Lee, "Using the self organizing map for clustering of text documents," Expert Systems with Applications, vol. 36, no. 5, pp. 9584–9591, 2009.

[5] Kohonen, T. "Self-organizing Maps", vol. 30 of Springer Series in Information Sciences. Springer Berlin. (2001).

[6] Kohonen, T. Automatic formation of topological maps of patterns in a self-organizing system. (1981).

[7] Deboeck, G. and Kohonen, T. (2013). Visual explorations in finance: with self-organizing maps. Springer Science & Business Media.

[8] Kohonen, T., "The neural phonetic typewriter. Computer", 21(3):11–22 (1988).

[9] S. Lu, "Pattern classification using self-organizing feature maps". In 1990 IJCNN International Joint Conference on, pp. 471–480. (1990).

[10] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen" WEBSOM for textual data mining. Artificial Intelligence Review", 13(5-6):345–364 (1999).

[11] G. Spanakis, G. Siolas, and Stafylopatis," ADoSO: a document self-organizer. Journal of Intelligent Information Systems", 39(3):577–610 (2012).

[12] Lars Bungum and Bj and orn Gamback," Self-Organizing Maps for classification of a Multi-Labelled Corpus", 2012.

[13] Kai-Wei Chang, Wen-tau Yih, Christopher Meek," Multi-Relational Latent Semantic Analysis", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1602–1612, Seattle, Washington, USA, 18-21 October 2013.

[14] J. Baltic, "Investigation on Learning Parameters of Self-Organizing Maps", Modern Computing, No. 2, 45-55, Vol. 2 (2014).

[15] D. Chandrakala, S. Sumathi, R. Bharath Raj,V. Kabilan, "Enhanced Emergent Trend Detection System Using PSO Based High Dimension Growing Self Organizing Map", European Journal of Scientific Research, India, 2011

[16] Abdelfattah Elsharkawi , Ali Rashed , Hosam Eldin Fawzan, "Comparative Study of clustering phenomena of 2-D SOM Against 3-D SOM", Journal of Al Azhar University, Engineering sector, ISSN:1110-6409, Egypt, 2014.

[17] Dumidu Wijayasekara, "Visual, Linguistic Data Mining Using Self Organizing, Compute". Sci. Dept., Univ. of Idaho, Idaho Falls, ID, USA, June 2012.

[18] Z. Mohd Zin, M. Khalid, E. Mesbahi and R. Yusof, "Data clustering and topology Preservation using 3D visualization of self organization maps", Proceedings of the World Congress on Engineering, 2, 2012.

[19] Mohamed Salah Hamdi, Ahmed Bin Mohammed, "SOMSE: A semantic map based meta-search engine for the purpose of text information customization", Applied Soft Computing, Vol.11, pp. 3870-3876, 2011.

[20] C. Chang, M. Shibu and R. Xiao Self Organizing Feature Map for Color Quantization on FPGA FPGA implementations of neural networks -Springer, 2006.

[21] Kofi Appiah Andrew Hunter Hongying Meng and Shigang Yue Mervyn Hobden Nigel Priestley Peter Hobden and Cy Pettit, "A Binary Self-Organizing Map and its FPGA Implementation", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009.

[22] Porter, M.F. An algorithm for suffix stripping. Program: electronic library and information systems, Vol. 40 Issue: 3, pp.211–218, (2006).

[23] Sumith Matharage, Damminda Alahakoon, "Growing Self Organising Map Based Exploratory Analysis of Text Data", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol. 8, No. 4, 2014

[24] B. Ohana and B. Tierney, "Sentiment Classification of Reviews Using SentiWordNet, 9th. IT&T Conference", Dublin Institute of Technology, Dublin, Ireland, 22nd. -23rd. October, 2009

[25] A. Hamouda, M. Marei and M. Rohaim, Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis, Journal of Advances in Information Technology, Volume 2, No. 4, November 2011.

Bibliography

Dr. Abdelfattah El-Sharkawi

Associate Professor , Software Engineering , Al –Azhar University, Cairo, Egypt.
Associate Professor, Software Engineering at Systems and Computer Engineering, Faculty of Engineering –Al-Azhar University, Cairo, Egypt
Ph.D. degree in Systems and Computer Engineering (1993), Faculty of Engineering, Al-Azhar University. M. Sc. degree in Systems and Computer Engineering (1986) Faculty of Engineering, Al-Azhar University. B.Sc. in Systems and Computer Engineering (1981) Faculty of Engineering, Al-Azhar University

Prof. Dr Ali Mahmoud Rashed

Ph.D. degree in electronics and communications engineering (1982) Faculty of Engineering, Ain Shams University. M. Sc. degree in electrical engineering (1977), Faculty of Engineering, Al Azhar University. B.Sc. of Electronics and Communications Engineering (1968) Military Technical College. Grade: Very Good. Supervise more than 50 Ph.D. and M.Sc. thesis and 100 B.Sc. projects. A member in the researches reviewers committee for Al Azhar Engineering research Journal. A member in the researches reviewers committee for Teacher Higher Staff Position – Al Azhar University. Holding a technical consultant position in ETCP (Egyptian Technical Colleges Project), Including Courses Revision, Quality Assurances and Human Resource managements since 2006 -2015.

Hosam Eldin Fawzan received the B.Sc. from the faculty of Engineering, Al Azhar University, Egypt, in 2003. M. Sc. degree in system and Computer Engineering - Faculty of Engineering - Al Azhar University Egypt, in 2010. I'm joined the Electrical and Computer Engineering Department, Sinai University, Egypt in 2008.

تطبيق برمجى وعتادى لتصنيف الوثائق باستخدام خرائط التنظيم الذاتى (SOM)

عبد الفتاح الشرقاوى[1] *، على راشد*[2] ، حسام الدين فوزان**[3]

*قسم هندسة النظم والحاسبات،جامعة الأزهر

القاهرة- جمهورية مصر العربية

sharkawi_eg@yahoo.com[1]
a_m_rashed@hotmail.com[2]
**قسم الهندسة الكهربائية والكمبيوتر،كليه العلوم الهندسيه ،جامعة سيناء

شمال سيناء – جمهورية مصر العربية

Hos.9876@yahoo.com[3]

ملخص

تعتبر خرائط التنظيم الذاتي(SOM) واحدة من أكثر الشبكات العصبية غير الخاضعة للرقابة شعبية وتستخدم لتجميع أنماط البيانات الغير مرئية و عرض البيانات. إن تجميع وتصنيف الوثائق يلعب دورا كبيرا في عالم اليوم. فى هذا البحث سوف يتم استخدام ال SOM فى تصنيف البيانات بدلا من تجميعها. إن كلا من قاعدة البيانات Reuters21578 من وكاله الانباء العالميه رويترز و Movies Reviews من موقع امازون قد استخداما فى التدريب والاختبار لقياس دقة التصنيف. وأجريت التجارب على أحجام بيانات مختلفة لتحديد تأثيرها على نتائج التصنيف. النتائج اثبتت أن ال SOM كما هى جيدة فى التجميع كذالك ايضا فى التنصنيف. ان تطبيق العتاد لل SOM باستخدام ال FPGA يهدف الى تسريع عمليات التصنيف.

Semantic Approach for Classification of Web Documents

Passent Elkafrawy[1], Dina ElDemerdash[2]

Faculty of Science, Menofia University, Egypt

[1]basant.elkafrawi@science.menofia.edu.eg

***Abstract*—We present a semantic approach method for classification of web documents. The proposed approach required only a domain ontology and a set of user predefined categories. Currently, most approaches to text classification represent document as (bag of words) and training the large set of documents to train the classifier. Our approach doesn't require a training set of documents. In our method we use DBpedia ontology as the main classifier, representing documents as (bag of concepts). We extract the terms from the document, extract their resources from DBpedia Spotlight, use *S*parqle query to determine class ontology and map them to their concepts then we determine the best category.**

1 INTRODUCTION

The amount of information in the World Wide Web has been overloaded heavily and rapidly in current era. At the same time organizing and managing information for knowledge extract is crucial problem. Web content consists of text documents and multimedia documents. Web Document Classification process plays an important role in organizing and managing data in the World Wide Web for better knowledge understand.

Web document classification is the process of classifying documents into predefined categories. Classification is one of web mining techniques, to solve the problem of information overload. Moreover, web mining has been improved by utilize semantic techniques. Semantic techniques provide deeper understanding of information by machine. Traditional web document classification methods represent document as (bag of words). Traditional supervised text classification methods use machine learning to perform the task. Most of them learn classification definitions and create the classifier from a set of training documents pre-classified into a number of fixed categories. Such methods, including Support Vector Machines [1], Naïve Bayes [1], decision trees [1], and Latent Semantic Analysis [2] are effective, but they require a set of pre-classified documents to train the classifier. There are major problems in (use context analysis to the words of document i.e. (bag of words)) First, count word occurrences and not consider meaning. Second, in order to train classifier, collect large number of documents must be collected. Third, the meaning of web content is not machining accessible due to lack of semantics.Fourth, it's simply difficult to distinguish the meaning between two sentences.

In this paper, we proposed ontology as classifier. The novel in our method doesn't require a training set of documents. Using general encyclopedic knowledge–based ontology such as DBpedia ontology.The DBpedia Ontology organizes the knowledge on Wikipedia in 320 classes which form a substitution hierarchy and is described by 1,650 different properties. It features labels and abstracts for 3.64 million things inupto97 different languages of which 1.83 million are classified in a consistent ontology, including 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases. Additionally, there are 6,300,000linkstoexternal web pages, 2,724,000 links to images, 740,000 Wikipedia categories and 690,000 geographic coordinates for places.

2 RELATED WORK

Bin Shi, Liying Fang, Jianzhuo Yan, Pu Wang, and ChenDong(2009)[3]proposed a uniform representation for the content, which include concepts and relations, of semantic documents based on WordNet. Use WordNet (ontology) to mapping relations between concepts. Use SVM to classification semantic web document. This method only considerstwo semantic relations in WordNet. Proposed a method to get only the path between two concepts in the WordNet.

Bai Rujiang Shandong and Liao Junhua (2009)[4] proposed a system that uses ontologies and Natural Language Processing techniques to index texts. Traditional BOW matrix is replaced by “Bag of Concepts” (BOC). Presented a new ontology-based methodology for automated classification of documents. To improve text classification, they enrich documents with related concepts, and perform explicit disambiguation to determine the proper meaning of each polysemous concept expressed in document. They use three ontology WordNet, open Cyc and SUMO to find concept of key words and compare concept in different ontology they not consider meaning between all concept in document and relation between it.

Jun Fang, Lei Guo and Yue Niu (2010)[5]proposes a novel ontology-based documents classification method by using ontology reasoning and similarity measure. They solve drawbacks of current ontology-based documents classification methods of classifier training divide concept to high concept and low concept by ontology reasoning and similarity measure but not consider relation between concepts.

Shikha Agarwal, Arachana Singhal and Punam Pedi (2012)[6] used weighted concept frequency–inverse document frequency (cf-idf) with background knowledge of domain ontology,for classification of RSS feed news items. They have shown that a rich and comprehensive ontology can be successfully used as text classifier.They consider the important concept only by (cf-idf) not consider the meaning and relation between terms but using ontology.

Chaaminda Manjula Wijewickrema (2014)[7] improved the classification accuracy of an automatic text classification system by using ontology.He proposed a solution to reduce the number of misclassification due to vocabulary ambiguities of the language used. Use ontology to represent the relationship among the concepts. But he used the first four highest frequency terms are chosen to decide the subject of the test document.although the ontology is using to increase the accuracy of automatic classification, the final decision still has to be made manually.

Henrihs Gorskis1 and Arkady Borisov2 (2015)[8] examined the feasibility of using rules and concepts discovered during the classification tree building process in the C4.5 algorithms, in a completely automated way, for the purposes of building an ontology from data. By building the ontology directly from continuous data, concepts and relations can be discovered without specific knowledge about the domain. The main novelty of this approach is the creation of concepts, which reflect unique and important data value intervals or spans for every attribute. Ontology building approach have some drawbacks. The number of intervals found by the C4.5 algorithms can be large and not intuitively understandable to a human user. The reasonability of the found value intervals can only be evaluated by a domain expert. Maybe the complexity of the span hierarchy is only a perceived one, maybe to an expert the value spans make sense and he will be able to give them appropriate names.

3 SEMANTIC CLASSIFICATION FOR WEB DOCUMENTS

We propose to build semantic classification system that classified web documents by semantic approach based on the DBpedia Ontology. Hence, the next sections describe the necessary steps to build such systems

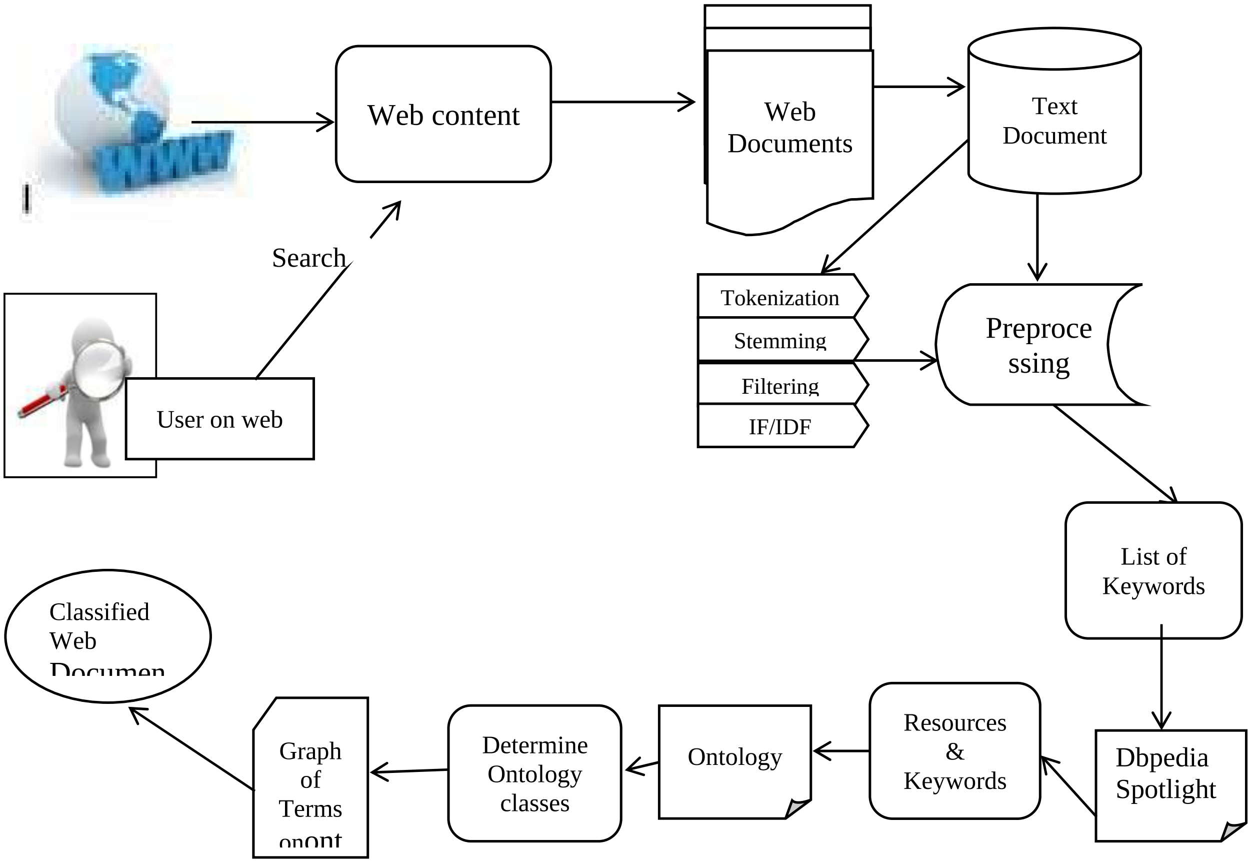


Figure 1: Classificatin Pipeline

A. First step: Text document preprocessing

Preprocessing method plays a very important role in text mining techniques and applications. It is the first step in the text mining process.Preprocessing steps such as Tokenization, Stopwords removable,stemming and TF/IDF algorithms for the text documents [9](figure 1)

Tokenization: is the process of breaking a text into words, phrases, symbols.

Stop words removable: removing the unimportant words from documents content by using a list of stop words. Stop-words are words that from non-linguistic view do not carry information such as (*a, an, the, this, that, I, you, she, he, again, almost, before, after*).

Stemming: Removes the affixes in the words and produces the root word known as the stem
TF/IDF: Term Frequency–Inverse Document Frequency (TF/IDF) is a numerical statistic which reveals that a word is how important to a document in a collection. The value of TF/IDF increases proportionally to the number of times a word appears in the document [10].
After Text document preprocessing we get the important words in document and the number of times a word appears in the document.

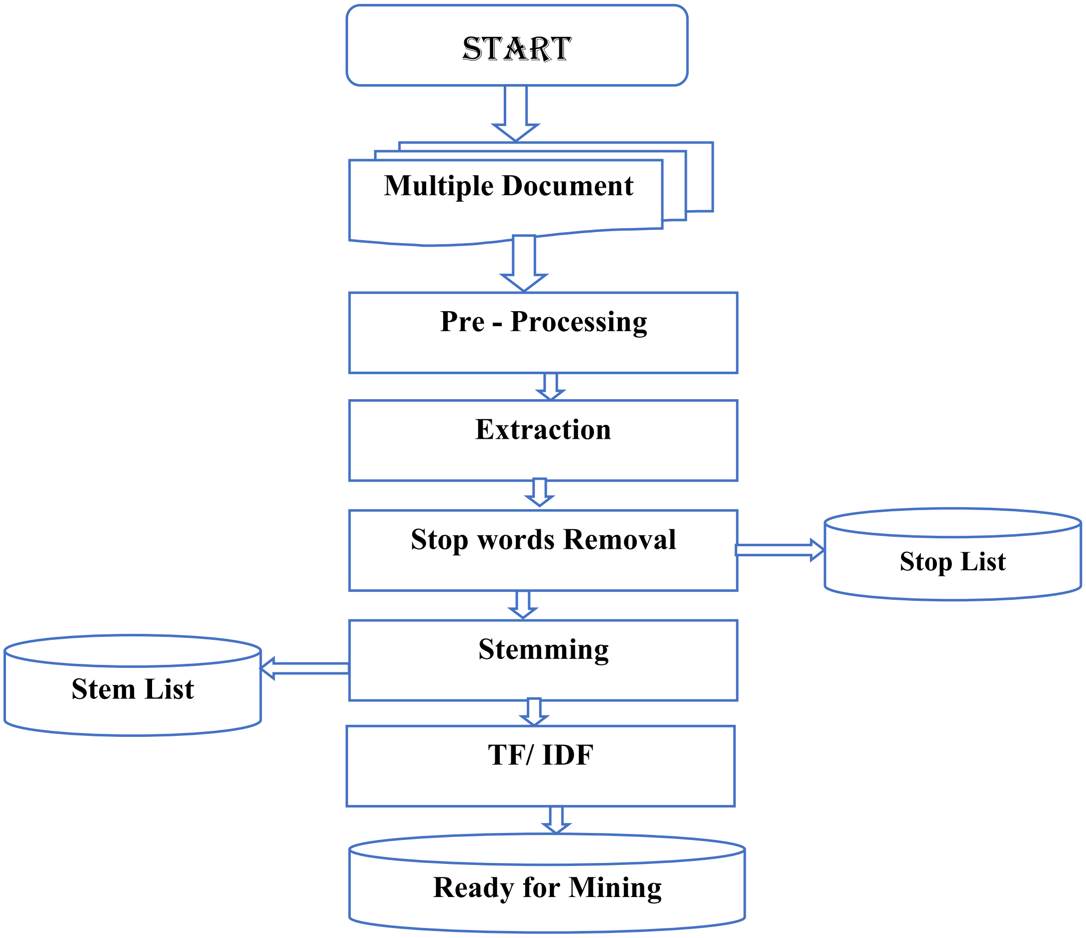


Figure 2:Text mining pre-processing steps

B. Second step: keywords and resource Extraction

We use DBpedia ontology to classification documents in semantic approach. The DBpedia ontology has been created for the purpose of classifying this extracted data. It's a cross-domain ontology based on info box templates in Wikipedia articles. The ontology currently covers 359 classes which form a consumption hierarchy and are described by 1,775 different properties.

The DBpedia 3.8 knowledge base describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases.

*1) Resource Extraction:*Our system use DBpedia spotlight [11] to extract resources from the text document. DBpedia spotlight is a tool designed for automatically annotating mentions of DBpedia resources in text [12]

2) *Keyword extraction:* Extract keywords from the question, the system chooses any word in the document satisfying one of the following seven conditions [13]:

1. All non-stop words in quotations.
2. Words recognized as proper nouns by DBpedia Spotlight.
3. Complex nominal and its adjective modifiers.
4. All other complex nominal

5. All nouns and their adjectival modifiers
6. All other nouns
7. All verbs

C. Determine ontology class

After we Extract resources and keywords determine ontology class by using Sparqle query. Retrieve the DBpedia ontology classes and properties we have to build a SPARQL query with the resource itself. The result of the query (Sparqle query) is an RDF file which holds ontology classes and properties and other information belonging to that resource. we compute similarity between the keywords and the ontology classes and properties. Thus, similar classes are selected.

D. Determine categories

After the last step we know keywords, resources and ontology classes. we show the relationships between all to Determine the best categories to these document.

4 CONCLUSIONS

We presented a novel approach to web document classification, depend on DBpedia ontology our method we use ontology as classifier to do not make training classifier. We extract terms from document and extract resources to determine classes and mapping them to their concepts then we determine the best category. Our method depends on meaning of terms and sentence.

In future work, we plan to improve our method by making modification to ontology if term not found.

BIOGRAPHY

Passent elKafrawy, Associate Professor, Faculty Science, Menofia University
Dr. Passent M ElKafrawy is an Associate Professor since 2013, she got her PhD from the University of Connecticut in United states on 2006 in Computer Science and Engineering. In the field of computational geometry as a branch of Artificial Intelligence. Then she taught in Eastern State University of Connecticut for one year. In 2007 she worked as a Teacher in Faculty of Science, Menoufia University, Mathematics and computer science department since that time till now.

Dina El Demrdash, computer Instructor, Information center, Menofia University
Graduated from Faculty of Science, Menoufia University, Mathematics and computer science department2009. Certified computer trainer from UNESCO and ICDL ARABIA since 2011
then, since 2012 she works as a Specialist Systems Analysis and Design, Menoufia University until now.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, pp. 1-47, 2002.
[2] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis". Discourse processes, vol. 25, pp. 259-284, 1998.
[3] Bin Shi, Liying Fang, Jianzhuo Yan, Pu Wang, and Chen Dong. "Classification of Semantic Documents based on WordNet". International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, 2009.
[4] Bai Rujiang Shandong and Liao Junhua "Improving Documents Classification with Semantic Features". Second International Symposium on Electronic Commerce and Security, 2009.
[5] Jun Fang ,Lei Guo and Yue Niu. "Documents Classification by Using Ontology Reasoning and Similarity Measure". Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)
[6] Shikha Agarwal, Arachana Singhal and Punam Pedi "Classification of RSS feed News Items using ontology". 978-1-4673-5119-5/12. 2012 IEEE.
[7] Chaaminda Manjula Wijewickrema "Impact of an ontology for automatic text classification". Annals of Library and Information Studies Vol.61, December 2014, pp.263-272.
[8] Henrihs Gorskis1 and Arkady Borisov2 "Ontology Building Using Classification Rules and Discovered Concepts" Information Technology and Management.
[9] Vishal Gupta , Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" Journal of Emerging technologies in web intelligence, vol 1, no 1 August 2009.

[10] Jacob Perkins "Python Text Processing with NLTK 2.0 Cookbook". 2010.
[11] Tartir, Samir; McKnight, Bobby and Arpinar, I. Budak (2009) “SemanticQA: web-based ontology driven question answering”. In Proceedings of the 2009 ACM symposium on Applied Computing (SAC '09).
[12] Pablo N. Mendes1, Max Jakob1, Andrés García-Silva2, Christian Bizer1 (2011). "DBpedia Spotlight: Shedding Light on the Web of Documents".
[13] Guo, Qinglin and Zhang, Ming “Question answering system based on ontology and semantic web”. In Proceedings of the 3rd international conference on Rough sets and knowledge technology (RSKT'08), 2008.

المنهج الدلالي لتصنيف وثائق الويب

ونحن نقدم وسيلة لنهج الدلالي لتصنيف الوثائق على شبكة الإنترنت.النهج المقترح يتطلب فقط انطولوجيلنطاق محدد ومجموعة من فئات المستخدمين معرفة مسبقا.حاليا، معظم النهج لتصنيف النص تمثل الوثيقة باعتبارها (حقيبة من الكلمات) وتدريب مجموعة كبيرة من الوثائق لتدريب المصنف.لا يتطلب نهجنا مجموعة من الوثائق للتدريب . في طريقة استخدمنا الأنطولوجيا DBpediaكمصنف، تمثل الوثيقة باعتبارها (حقيبةمن المفاهيم).نحن استخراج الشروط من وثيقة، واستخراج مواردها من DBpedia الضوء، وتستخدم الاستعلام SPARQL لتحديد الفئة الأنطولوجيا ثم تربت لمفاهيمهم ثم نحدد الفئة الأفضل.

Building Language-Topic Based Index for Multi-lingual Information Retrieval

Ebtsam Sayed[*1], Mostafa Aref[**2], Samir Elmougy[***3]

[*1]*Computer Science, Faculty of Computers and Information, Minia University*
Minia, Egypt

[1]ebtsamabd@gmail.com

[**2]*Computer Science, Faculty of Computers and Information, Ain Shams University*
Cairo, Egypt

[2]mostafa.aref@cis.asu.edu.eg

[***3]*Computer Science, Faculty of Computers and Information, Mansoura University*
Mansoura, Egypt

[3]mougy@mans.edu.eg

***Abstract*—Multi-Language Information Retrieval (MLIR) is a subfield of information retrieval for retrieving information from a multi-lingual collection. There is a need to search for information not only within the native language of the user, but also within the other languages. However, the user can't benefit from retrieving miss understandable information due to its language. So, Machine Translation field presents a solution by translating text from one language into another. Multi-lingual Search System enables users to search the web contents written in different languages. However, there are some difficulties that users may face to get relevant information, especially in multi-language such as language selection, query formulation and reformulation. Users can manually select the source language for search, but unfortunately the user can get bad results. The reason is that the user may select a language with poor available online contents. For example, if a user want to search using his native language (e.g. Arabic) for a specific topic (e.g. Computer Science, Technology, medicine), he/she will get bad results because of Arabic poor online contents about that topics. In this paper, we aim to enable users to translate the original query into a suitable language based on the online available contents to retrieve more relevant information to the query without their interventions. We propose a building language index based on the topic that can be used by a web search system to select automatically the suitable language/s for search, based on the available online contents with for each language about the user's query topic.**

Keywords: Multi-Language Information Retrieval (MLIR), Language-Topic Based Index, Available Contents, Contents Web Mining.

1 INTRODUCTION

Actually, English is the dominant language in the web, however more than half of web contents are written in the other languages (non-English). Users of search engines want to get relevant results to their queries/needs. Users often know only one language/native that can be used for good query formulation. The problem is, the online most of the related contents may be available in other language. **What they can do?** If they write the query in their second language, it is expected to get bad results, due to they aren't professional in that language. The existing Multi-Language Information Retrieval (MLIR) systems ask the user to identify the source language (language of query) and the target language/s (the language/s of webpages/resources he want to search in) before starting the search process. For example, user can write Arabic query (about computer networks) and limit his search to English web pages/resources. There are online contents available in Arabic but the most/best online contents are written in English.

In this paper, we suggest a different system that uses another way of selecting the target language/s. It doesn't ask the user to enter the target language. We assume that, if the user is an academic researcher, he can decide which language he should use for his search. However, not all users of search engines have experience/efficient background to decide the suitable language for search. So, our paper presents a solution through providing automatic selection of the most suitable language/s for search based on available contents.

To achieve our goal of developing a web system with automatic language selection, **we propose building an index that includes a set of languages and set of topics where the intersection is the language rank.** Table I illustrates a part of the expected index of topic-language based, where each language has score that measured by the available contents about a specific topic. We can use the proposed index to develop a web search system that selects automatically the suitable language for search. Our idea for ranking languages is based on the available online contents with that language related to the topic of user's query. For example, if a user searches for "software engineering", the suitable language for his search results is in English, because software engineering is a computer science field and the most of its online contents are written in English. Here, we are mainly interesting in *how to select the more suitable language to retrieve the more relevant information with more degree of quality than any other possible language.*

This paper is organized as follows. We discuss the process of building language-topic based index and how to select the languages and topics to be included in the index in Section2. Section 3 illustrates how we can collect the dataset for the different topics & different languages**.** Section 4 draws our conclusion and future directions.

TABLE I

PART OF THE EXPECTED LANGUAGE- TOPIC INDEX

Domain	**Language Ranks**				
	English	Chinese	German	French	Japanese
Medical	3	1.5	2.5	2	1
Technology	2.5	1	3	1.5	2
Agriculture	2.5	3	2	1	1.5

2 BUILDING LANGUAGE-TOPIC BASED INDEX

Up to our knowledge, no previous work that we can refer to ranking languages based on available contents for different topics/domains. So, we collect information from different resources about the available online contents for different topics with different languages. There are some important issues should be considered for illustration before starting in building that index such as languages, topics that should be included in the index, and how dataset for mining is collected, identifying the required information to be extracted from webpages, and the technique for extraction.

A. SELECTING LANGUAGES

We can select some of the top languages (e.g. English, Russian, German, Japanese, Spanish, French) based on Fig. 1, that shows the Percentages of websites using various content languages[1]**.** Figure 1 shows the top languages used for the internet contents /websites. Of course, English is the most used language in the web, although other languages like Chinese and Arabic are increased through the last few years.

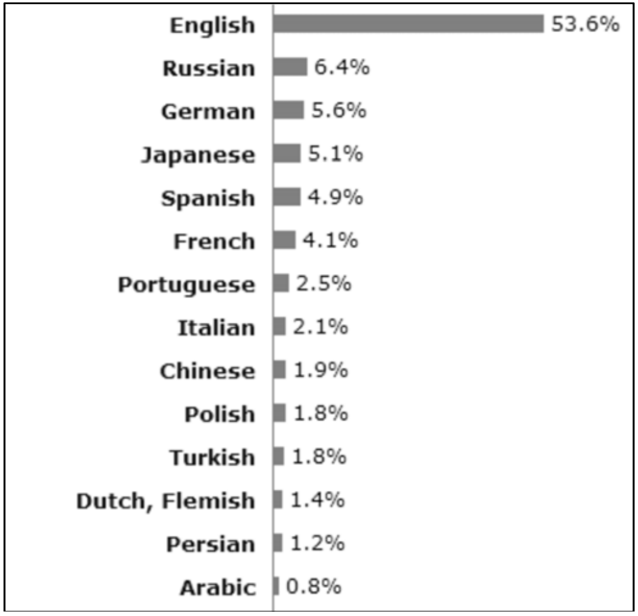


Figure 1: Top used languages in the web

B. SELECTING TOPICS

Search engines usually use topic/web directories to narrow searches. Topic directories build a hierarchical structure of web pages as taxonomy or ontology according to their contents. We can use taxonomy of **Wikipedia** (300,000 category), Google directory (**directory.google.com**), Open Directory Project or any web directories to identify the topics of our index. We can select the topics that it is expected to affect positively with Multi-Lingual IR, such as medical, ecommerce, industry, technology and tourism domain. The Open Directory data file is available as RDF format through its web site. The file includes a list of Open Directory Project categories and the external URLs relevant set for each category. Liu and et al. [2] presented a heuristic-based approach for topical crawling through using link-context and implements DOM tree for anchor text detection. They used link-context necessity to correctly guide topical crawling. They used Open Directory categories to identify topics for their crawling experiments.

3 COLLECTING REQUIRED RESOURCES

Web mining applies data mining methods to mine Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining has three categories depending on the type of data i.e. Web Content Mining, Web Structure Mining and Web Usage Mining. **Web structure mining** extracts useful knowledge from hyperlinks that draw the Web structure [3]. For example, we can detect communities of users with common interests. **Web usage/log mining** mines the behavior of website users. It discovers user access patterns from Web usage logs, which record every user's clicks [4]. **Web content mining** extracts information that is related to the website page contents. It mines useful information from Web page contents e.g. customer reviews and forum postings. For example, we can automatically cluster Web pages based on their topics [5].

In our work, we interest in Web content mining that can be divided into two subcategories, *webpage content mining* and *search result mining*. After identifying the used languages and the topics, we need to identify the available contents of a language about specific topic. In this paper, we present two approaches. The first method is **webpage contents mining**, through collecting dataset by crawling the internet webpages and the second method is **search result mining** based on using search engine search results given pre-classified queries. Assume that we have a set of languages (L) and a set of topics (T) and a language rank (R), and we need to compute Rank $R_i [T_{i,} L_i]$ to identify which language L_i is suitable for search about a specific topic T_i.

A. Webpages Contents Mining

For mining webpages contents, we need to obtain a huge amount of webpages at first. Web crawling is the process that can be used to collect the webpages. Web crawler begins with a list of URLs to visit (called the seeds). While the crawler visits seeds URLs, it extracts all the hyperlinks in the page and adds them to the list of URLs (the crawl frontier) to visit. Crawlers can be one of the following categories: Focused/Topical Crawler, Collaborative Web Crawler, Incremental Crawler, Parallel Crawler, Distributed Crawler and Mobile Crawler. Fig. 2 shows the basic crawler operations that include the following steps:

- Start with known “seed” URLs (list of starting URLs).
- Fetch and parse them
  - Extract URLs they point to
  - Add the extracted URLs to a queue
- Fetch each URL on the queue and repeat (Stop criterion can be anything).

The focused or topical crawlers aim to download only those pages that are about a specific topic. they also identify which URLs to scan and in what order to parse based on previous downloaded web pages. Focused crawlers consider that pages about a topic tend to have links to other pages on the same topic. There are many previous works that study topical crawling e.g. Ali [6] who proposed an approach for focused crawling. Also, De-Assis *et al.* [7] presented a focused crawling approach that depend on content-related information and genre information presented in web pages to guide the crawling process. Arya and Vadlamudi [8] designed a topical crawling algorithm based on an ontology to access hidden web content.

In our work, we need to crawl a huge amount of web pages about a specific topic for all selected languages. So, the topical crawler is the suitable type for our case to limit crawling to only webpages that are related to a topic. While topical crawling suffers from some challenges such as context of links consideration during the crawling process. The crawling process can described be follows giving a language L_i = {English, Russian, German, Japanese, Spanish, French, Arabic}

1- For each topic T_i, collect a set of relevant documents.
2- Train the classifier using these documents/examples.
3- Use **web crawler** to crawl all web pages belongs to a Topic T_j, which it classified as relevant to topic T_j.
4- For each topic T_i,
5- Remove repeated webpages by identifying similar webpages.
6- For topic T_j, count webpages of a language L_i.
7- Select the language L_i with large number of relevant webpages to a topic T_j as the suitable language for search.

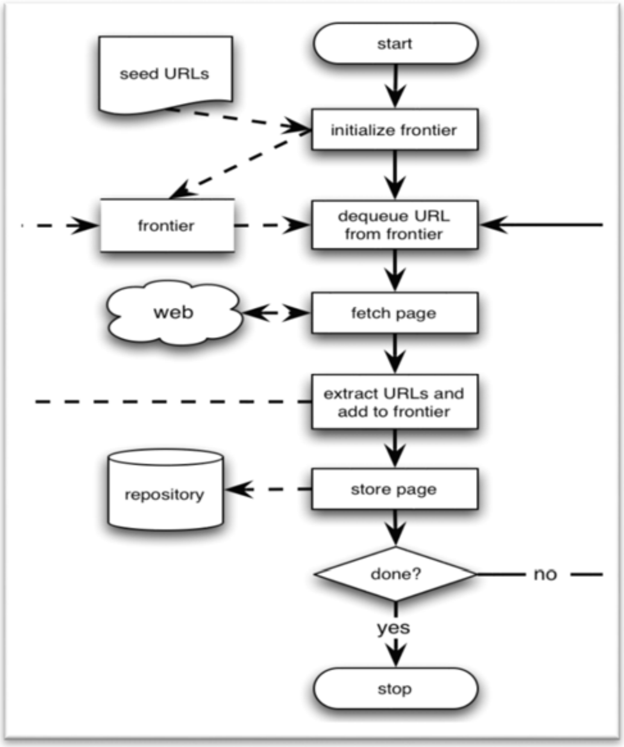


Figure 2: Basic crawler operation

B. Search Results Mining

Instead of web crawling webpages, we can use a search engine results given pre-classified queries. Search results are retrieved in seconds without any need for any effort. There are many datasets available for research purpose including queries and their tagged classes. These classified queries can be given to search engine (Google) and easily get the results.

For each language L_i where [L_i = {English, Russian, German, Japanese, Spanish, French, Arabic}]

1- For each topic $T_{i,}$ Collect queries that are classified into predefined classes.
2- Translate the same queries to other languages (e.g. from English to other languages)
3- Use a search engine to retrieve all web pages belongs to topic T_j, given queries dataset which is classified into Topics T_i.
4- Remove repeated retrieved results by identifying similar webpages.
5- For all queries, compute the average of retrieved results for each topic T_i
6- Select the language with large number of results for that topic T_j as the suitable language for search.

C. Comparison between the two techniques

Our work is application of web content mining, where we need to identify the amount of the available contents for each top language about the selected topics. There are two directions, including webpage content mining and search results mining. To decide which approach is the best, we study each one, we have to identify what information actually needed to be extracted from webpages. For each language, we need to get the number of available webpages, description, and keywords for all selected topics, the number of webpages to count the amount of the available contents and the other information (i.e. description, keywords) to identify/remove the repeated webpages. Table II shows the comparison between the two methods.

4 Conclusion and Future Work

In this paper, we studied two approaches used for collecting the required information for building language index to be used by a web search system for selecting automatically the most suitable language/s for search. We concluded that that the second approach (search results mining) is more suitable to use because it satisfies what we need for building the index. Also it is more easily and has fewer requirements than the crawling approach. Also, in this paper, we could identify the suitable one for building the index. Our next steps for future include perform experiments for computing language scores to produce the index.

Table II

Webpages Contents Vs. search results mining

	Webpages Contents mining	Search Results Mining
Produced Information	- Webpage contents and other details, e.g. description, keywords	- For each query, We get the number of results, webpage 's description, keywords, URL name, the number of links to other pages, etc. in few seconds.
Technique Requirements	- Large RAM and storage for saving web pages. - Building index for web pages. - Web crawler need for classifier to classify web pages and compute their relevance to the specific topic before downloading. - Building crawler from scratch or using the crawling tools.	- For each topic, after collecting the search results of all queries in one file. It is possible that one query have the same results of another. - We have to write our own program to remove repeated web pages to get the exact number of web pages for each topic. - We need to get a dataset that contains classified queries into set of topics. There are a lot of dataset available online. - Mapping the classes of queries into selected topics.
Difficulties and Challenges	- We need to large fraction of the web, how many hours we need for crawling e.g. 10,000 pages for each topic? - Topical Crawling isn't accurate as the crawler may retrieve irrelevant pages to the topic, it depend on bag of words (need for adding the context). - Classification increase the time of crawling. - How to identify Seed URLs to be start point of crawler for each topic & for each language? - It can be easy to identify Seed URLs for English for Arabic, but it difficult to other languages. - Building crawler from scratch is not simple task and need for time. - Tools for general crawling are available, but topical crawlers are not.	- How we can translate the queries to other languages? Manually or Google translate / create our translation method? - We have to use good translation methods to get accurate search results. - For query translation, we think that there is no ambiguity because queries are classified into topics, so translation can be more accurate.

Acknowledgment

My sincere and heartfelt thanks to Dr. Shahinda Sarhan for her insightful comments and guidance in my research idea. My special appreciation to my parents, my husband Refaat for their support and to my children Nooran and Albraa for their love and patience.

BIOGRAPHY

Ebtsam Sayed is an Teaching Assistant of Computer Science at Minia University. She received B.Sc in Computer Science in 2007 from Mina University and the M.Sc. degree in Computer Science in 2011from Cairo University. She is currently a PhD student at Computers and Information faculty, Mansoura University. Her research interests are Web mining, Multilingual information retrieval, and machine learning.

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.

Samir Elmougy received the B.Sc in Statistics and Computer Science in 1993 and the M.Sc. degree in Computer Science in 1996, both from Mansoura University, Egypt. He received the PhD degree in computer science from College of Engineering, Oregon State University, USA, in 2005. He is working as the Chair of the Department of Computer Science, Faculty of Computers and Information, Mansoura University since Dec. 2014. From 2008 to 2014, he had been with King Saud University, Riyadh, Saudi Arabia as an assistant professor at the Dept. of Computer Science. His current research interests are algorithms, error correcting codes, computer networks, software engineering, and Natural Language Processing for Arabic languages.

REFERENCES

[1] W3Techs.com website: https://w3techs.com/technologies/overview/content_language/all, (Last accessed W3Techs.com, 9 September 2016).

[2] L. Liu, T. Peng and W. Zuo, "*Topical Web Crawling for Domain-Specific Resource Discovery Enhanced by Selectively using Link-Context*", The International Arab Journal of Information Technology, Vol. 12, No. 2,pp. 201-203, March 2015.

[3] Hussein, Mohamed-K, and Mohamed-H, Mousa. *"An Effective Web Mining Algorithm using Link Analysis."*, International Journal of Computer Science and Information Technologies (IJCSIT) , pp. 1-3, 2 010.

[4] S. Kavita, G. Shrivastava and V. Kumar. *"Web mining: Today and tomorrow."* Electronics Computer Technology (ICECT), 2011 3rd International Conference Volume 1, 2011.

[5] S.Balan, and P.Ponmuthuramalingam,*"Astudy of Various Techniques of Web Content Mining Research Issues and Tools"*, International Journal of Innovative Research and Studies, Vol 2, Issues 5, May 2013.

[6] H. Ali, *Self Ranking and Evaluation Approach for Focused Crawler Based on Multi-Agent System, the International Arab Journal of Information Technology*, vol. 5, no. 2, pp. 183-191, 2008.

[7] T. De-Assis, F. Laender, A. Goncalves, and A. Da Silva., *"A Genre-Aware Approach to Focused Crawling, World Wide Web-interest and Web Information Systems"*, vol. 12, no. 3, pp. 285-319, 2009.

[8] V. Arya and R. Vadlamudi,*" An Ontology Based Topical Crawling Algorithm for Accessing Deep Web Content"*, in Proceedings of the 3rd International Conference on Computer and Communication Technology, Pradesh, India, pp. 1-6, 2012.

بناء فهرس للغات والموضوعات لإسترجاع البيانات متعددة اللغات

ابتسام سيد[1]*, مصطفى عارف[2]**, سمير الموجى[3]***

كلية الحاسبات والمعلومات- جامعة المنيا-المنيا- مصر*

[1]ebtsamabd@gmail.com

كلية الحاسبات والمعلومات-جامعة عين شمس-القاهرة-مصر**

[2]mostafa.aref@cis.asu.edu.eg

كلية الحاسبات والمعلومات-جامعة المنصورة-المنصورة-مصر***

[3]mougy@mans.edu.eg

الملخص:

ان عملية استرجاع البيانات متعدد اللغات ماهي الا عملية بحث في مجموعة من البيانات المكتوبة بلغات متعددة. ومن المؤكد انه ليس هناك فائدة ان يسترجع المستخدم البيانات بلغة لا يفهمها. ولذلك جاءت الترجمة الآلية بهدف ترجمة النصوص من لغة معينة لأخري. وقد اصبح الآن هناك حاجة الي البحث باللغات الأخرى وليس فقط البحث بلغة المستخدم الأم. ويمكن تحقيق ذلك باستخدام نظام بحثى متعدد اللغات الذي يدعم عملية البحث في محتويات الويب متعددة اللغات. ولكن هناك بعض الصعوبات التي يواجها المستخدمون لهذه الأنظمة متعددة اللغات لاسترجاع المعلومات المرتبطة بالبحث وذلك مثل اختيار لغة البحث وكتابة كلمات البحث والاستعلام واعادة صياغتها. ومن الممكن ان يحدد المستخدم اللغة المستخدمة للبحث ولكنه من المحتمل حصوله على نتائج سيئة لهذا البحث. ويرجع السبب الي عدم توافر معلومات على الويب باللغة التي قام بتحديدها للبحث. مثال: اذا قام أحد المستخدمين العرب بالبحث عن معلومات متعلقة بمجال علوم الحاسب أو التكنولوجيا أو الطب, فانه سوف يحصل علي نتائج بحث سيئة, وذلك لعدم توفر المعلومات الكافية باللغة العربية عن تلك الموضوعات التي قام بالبحث عنها. ولذلك فان هدفنا في هذا البحث هو عمل نظام بحثى متعدد اللغات بحيث يقوم هذا النظام بتحديد اللغة المناسبة للبحث بناءاً على كمية المعلومات المتوفرة بتلك اللغة بدون تدخل من المستخدم. ولبناء هذا النظام البحثي نحتاج الي عمل فهرس للغات والموضوعات يحدد فيه ماهي اللغة المناسبة للبحث لكل موضوع. ويتم تحديد هل اللغة مناسبة للبحث في موضوع بعينه وذلك على حسب كمية المعلومات المتوفرة عن هذا الموضوع بتلك اللغة.

Ambiguity Detection and Resolving in Natural Language Requirements

Somaia Osama[*1], Safia Abbas[*2], Mostafa Aref[*3]

[*]*Computer Science Department, Faculty of Computer and Information Science, Ain Shams University*
Cairo, Egypt

[1]somaia.osama.r@gmail.com

[2]safia_abbas@yahoo.com

[3]aref_99@yahoo.com

***Abstract*--Natural language is the most used representation for describing software requirements specification on computer based systems in industry. On one hand, natural language is flexible, universal, and wide spread. On the other hand, natural language requirements are recognized generally as being ambiguous. Ambiguity occurs when a sentence can be interpreted differently by different readers. Ambiguity in natural language requirements has long been recognized as a challenge in requirements engineering. In this paper we focus on the application of Natural Language Processing (NLP) technique for addressing ambiguity in natural language requirements. We describe an automated approach for detecting and resolving ambiguities in order to avoid misinterpretations. The paper also provides a case study on real world software requirements specification to show that the use of approach in automated ambiguity detection and resolving, where we have assessed the efficiency of the approach.**

***Key words*: Natural language processing (NLP), ambiguity, Requirement engineering, Software Requirement Specification, ambiguity detection, ambiguity resolving.**

1 INTRODUCTION

Requirements engineering (RE) is the activity that involves the functions associated with the extraction, modeling, analysis, verification and specification of the user's requirements [1]. The RE activity often starts with the vaguely defined requirements [2] and results finally into a Software Requirements Specification (SRS) document. The SRS is a part of the contract and it must define the user and the system requirements obviously, accurately and unambiguously. An SRS that has inconspicuous, incomplete, unmanaged, unspecified, inaccurate or ambiguous requirement definition may eventually lead to cost and time overruns [3, 4, and 5]. An important research problem in Requirements Engineering is resolving ambiguity. An ambiguity is “a statement having more than one meaning”. An ambiguity can be lexical*, syntactic, semantic, pragmatic, vagueness, generality and language error* ambiguity [6]. Although the fact that the requirements specified in natural language tend to inappropriate interpretations, the requirements are most often specified in natural language. So, it is necessary to develop the approaches that deal with resolving the ambiguities from the user requirement specifications. Manually resolving ambiguity from software requirements is a tedious, time-consuming, error-prone, and therefore expensive process [6]. Therefore, an automated and semi-automated approach to resolve ambiguities from the requirements statement is needed. There exist various approaches, starting from manual glossaries approach to automatic ontology based approach to reduce ambiguity from the Software Requirement Specification. In addition, there are a number of diverse tools such as, QuaARS [7], RESI [8], WSD [9], SREE [10, 11], ARM [12], NAI [13, 14], and NL2OCL [15], SR-Elicitor [16] developed to detect and resolve ambiguities.

2 AMBIGUITY

“An important term, phrase, or sentence essential to an understanding of system behavior has either been left undefined or defined in a way that can cause confusion and misunderstanding.” [17]. Ambiguous requirements lead to confusion, wasted effort and time and rework. Ambiguity is the possibility to interpret a phrase/word in several ways. It is one of the problems that occur in natural language texts. An empirical study by Kamsties et al [6] depicts that "Ambiguities are misinterpreted more often than other types of defects". An ambiguity has two sources: incomplete information and communication mistakes. Some errors can be resolved without domain knowledge like grammatical error though some error needs domain knowledge like the lack of detail that wants user. The Ambiguity Handbook [6] presents different types of ambiguities, categorized as Lexical, Syntactic, Semantic, Pragmatic, Vagueness, Generality and Language Error as shown in table I.

TABLE I. TYPES OF AMBIGUITY

Type of Ambiguity	Subtype	Description with example
Lexical Ambiguity	Homonymy Ambiguity	Two different words have the same written and phonetic representation, but unrelated meanings and different etymologies. E.g.: The airport shall be a major hub for Departures from Australia to Asia. "major" (important/an army officer of high rank/ specialize in a particular subject at a college)
	Polysemy Ambiguity	A word has several related meanings but one etymology.
Syntactic Ambiguity	Analytical Ambiguity	The role of the constituents within a phrase or sentence is ambiguous. E.g.:The software will follow the applicable regulatory and utility technical requirements in its speculated calculations and selection process.(can refer regulatory technical requirements and utility technical requirements or regulatory requirements and utility technical requirements)
	Attachment Ambiguity	A particular syntactic constituent of a sentence, such as a prepositional phrase or a relative clause, can be legally attached to two parts of a sentence. Or a phrase can be placed in different positions in the parse tree.
	Elliptical Ambiguity	When it is not certain whether or not a sentence contains an ellipsis.
	Coordination Ambiguity	More constituents joined by coordinative conjunctions (and, or). E.g.: The system shall print a login session report to every Manager and Database Administrator. (can refer The system shall print a login session report to very Manager and every Database Administrator or The system shall print a login session report to every person who is both a Manager and a Database Administrator.
Semantic Ambiguity	Scope Ambiguity	A sentence has more than one way of reading it within its context although it contains no lexical or structural ambiguity.
Pragmatic Ambiguity	Referential Ambiguity	An anaphor can take its reference from more than one element, each playing the role of the antecedent. E.g.: If the ATM accepts the card, the user enters the PIN. If not, the card is rejected.
	Deictic Ambiguity	Pronouns, time and place adverbs, such as now and here, and other grammatical features, such as tense, have more than one reference point in the context. The context includes a person in a conversation, a particular location, a particular instance of time, or an expression in a previous or following sentence.
Vagueness		If it is not clear how to measure whether the requirement is fulfilled or not. E.g.: The System shall be easy as possible.

3 THE PROPOSED ARCHITECTURE OF OUR TOOL

We will develop an automated system to detect and resolve ambiguities from full text documents. The system architecture is shown in Figure 1. The initial input is a complete requirement text. The output is unambiguous requirement texts.

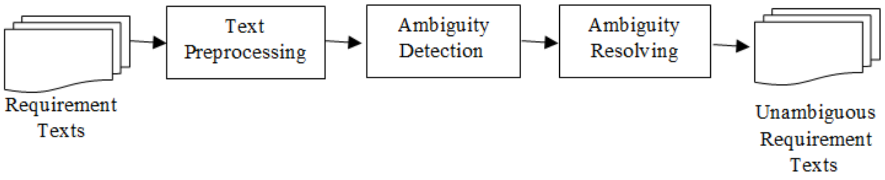


Figure 1: System Architecture

The system consists of three major functional process modules

(a) **The Text Preprocessing Module**

The input requirements document is split into separate sentences using an established sentence boundary detector. The individual sentences are then passed to Tokenizer, the Tagger which identifies the individual words' part of speech, and marks phrase boundaries and the finally syntactic parser.

(b) **The Ambiguity Detection Module**

This module would apply a set of ambiguity measures to a RS in order to identify potentially ambiguous sentences in the requirement specification. The main goals for the tool for identifying and measuring ambiguities in natural language requirement specifications are: to identify which sentences in a natural language requirement specification are ambiguous and, for each ambiguous sentence, identify the ambiguity word in the sentence.

(c) **The Ambiguity Resolving Module**

Finally, this section focuses in removing the ambiguity. For each ambiguous sentence, remove the ambiguity from the sentence automatically as the final step using resolve rules, and thus improve the natural language requirement specification.

3.1 The Text Preprocessing Module

The Text preprocessing module consists of four stages as shown in figure 2.

- **Sentence splitter:** the sentence splitter separates each sentence from the input string and returns a list of strings.
- **Tokenizer:** the tokenizer takes each sentence as an input and splits them into tokens such as numbers, words and punctuation.
- **Parts of speech (POS tagger):** It is used to perform the process of marking up the words in a text as corresponding to a particular part of speech.
- **Syntactic parser:** sequences of words are transformed into structures that indicate how the sentence's units relate to each other. This step helps us in identifying the main parts in a given sentence such as object, subject, verb…etc.

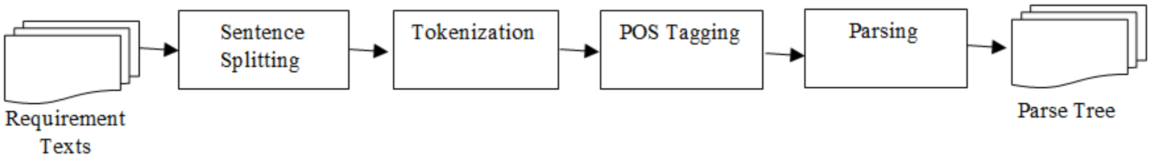


Figure 2: Text Preprocessing Module

E.g.:**"The system provides maximum output."**

After POS Tagging

"**"The/DT system/NN provides/VBZ maximum/JJ output/NN ./."** "

The text is syntactically analyzed and a parse tree is produced for further semantic analysis. Figure 3 shows the generated parse tree of the above example. Parse tree is the output of the text preprocessing module.

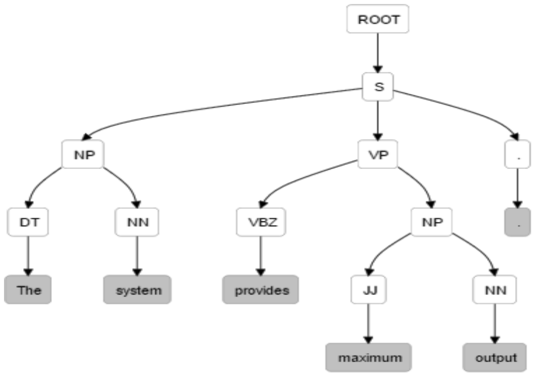


Figure 3: Parse Tree

3.2 The Ambiguity Detection Module

This module could apply a several ambiguity measures to a requirement specification to recognize possibly ambiguous sentences in the requirement specification. The core goals for this tool for detecting and measuring ambiguities in natural language requirement specification are: to detect which sentences in a natural language requirement specification are ambiguous and, for each ambiguous sentence, identify the ambiguity word in the sentence. The Ambiguity Detection Module architecture is shown in Figure 4.

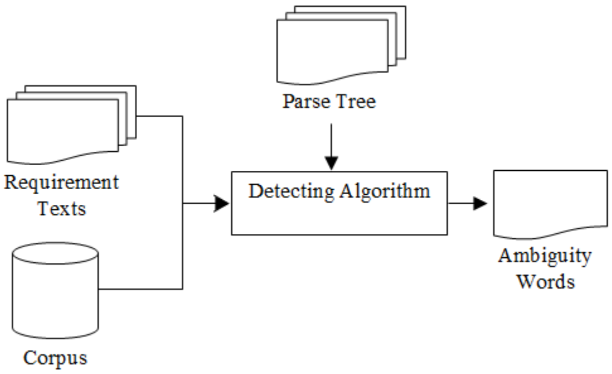


Figure 4: Ambiguity Detection Module

Corpus is the main element of ambiguity detection. Ambiguous words that result in misinterpreted requirements are analyzed and stored into the corpus. The major aim of this process is to check and validate whether the data which is a part of Software Requirements Specification document is ambiguous or not.

i. Identify Referential Ambiguity

The Referential corpus contains the possible ambiguity indicators: I, it, its, itself, he, she, her, hers, herself, him, himself, his, me, mine, most, my, myself, that, their, theirs, them, themselves, these, they, you, your, yours, yourself, and yourselves, anyone, anybody, anything, everyone, everybody, everything, nobody, none, no one, nothing, our, ours, ourselves, someone, somebody, something, this, those, us, we, what, whatever, which, whichever, who, whoever, whom, whomever, whose, and whosever.

ii. Identify Coordination Ambiguity

The Coordination corpus contains the possible ambiguity indicators: and, and/or, or, but, unless, if then, if and only if, and also.

iii. identify Scope Ambiguity

The Scope corpus contains the possible ambiguity indicators: a, all, any, few, little, several, many, much, each, not, and some.

iv. Identify Vague

The *Vague* corpus contains the possible ambiguity indicators: /, <>, ( ), [ ], { }, ;, ?, !, adaptability, additionally, adequate, aggregate, also, ancillary, arbitrary, appropriate, as appropriate, available, as far as, at last, as few as possible, as little as possible, as many as possible, as much as possible, as required, as well as, bad, both, but, but also, but not limited to, capable of, capable to, capability of, capability, common, correctly, consistent, contemporary, convenient, credible, custom, customary, default, definable, easily, easy, effective, efficient, episodic, equitable, equitably, eventually, exist, exists, expeditiously, fast, fair, fairly, finally, frequently, full, general, generic, good, high-level, impartially, infrequently, insignificant, intermediate, interactive, in terms of, less, lightweight, logical, low-level, maximum, minimum, more, mutually-agreed, mutually-exclusive, mutually-inclusive, near, necessary, neutral, not only, only, on the fly, particular, physical, powerful, practical, prompt, provided, quickly, random, recent, regardless of, relevant, respective, robust, routine, sufficiently, sequential, significant, simple, specific, strong, there, there is, transient, transparent, timely, undefinable, understandable, unless, unnecessary, useful, various, and varying [10].

Algorithm for Ambiguity Detection

Ambiguity Detection works on following algorithm. This algorithm is used to classify the ambiguities as Lexical, Syntactic or Syntax ambiguity. The steps of the algorithm are shown in algorithm 1:-

Step-1: Read corpus of ambiguous words from a text file, and store it in data store named as "s".
Step-2: Read the software requirement specification document line by line.
Step-3: For each line, match all words against the corpus. If word/words are matched then store the sentence in another data store named as "d". Continue this step for each line of SRS, till the end of software requirement specification document is reached.
Step-4: Classifies the sentences into Lexical, Syntactic or Syntax ambiguities, depending upon the types of ambiguous words/phrases.
Step-5: Calculate the percentage of ambiguities.

Algorithm 1 Ambiguity Detection Algorithm

3.3 The Ambiguity Resolving Module

Finally, this section focuses in resolving the ambiguity. For each ambiguous sentence, resolve the ambiguity from the sentence automatically as the final step using resolve rules, and thus improve the NL RS. The Ambiguity Resolving Module architecture is shown in Figure 5.

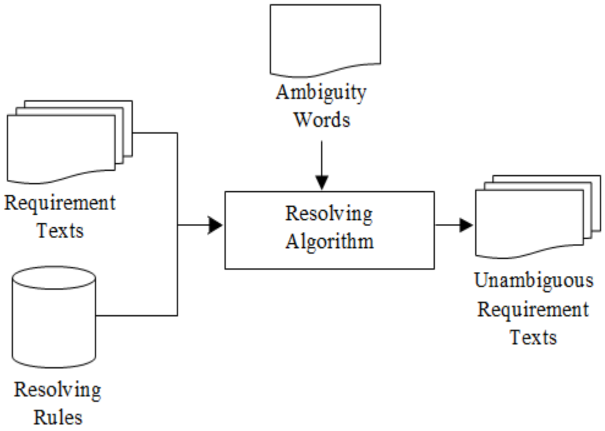


Figure 5 The Ambiguity Resolving Module

In this part we describe a resolving ambiguity approach using disambiguation rules for ambiguous word. Our approach resolves ambiguities by common rules; TABLE II shows some rules for some individual ambiguous words.

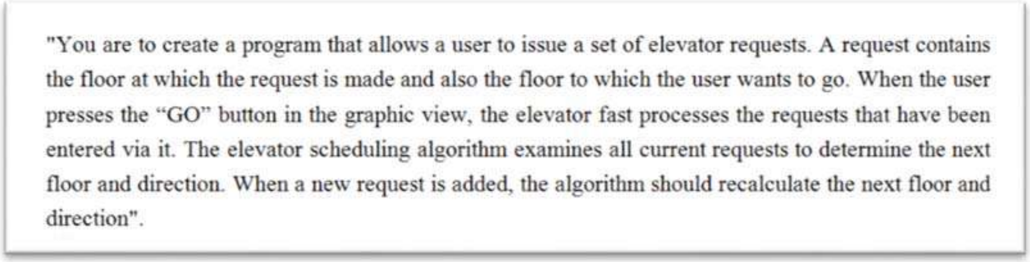
TABLE II RESOLVING RULES

Rule	Example	Ambiguity Type
Rule 1: when sentence containing vague adjective such as **prompt, fast, routine**, replace with specific time.	E: The system should give prompt respond to all user inputs. E.1: The system shall give within 1 second respond to each user input. E: The System is responsible for routine processing of the data. E.1: The System is responsible for daily processing of the data.	Vague
Rule 2: when sentence containing **both**, split it to two sentences.	E: The system should print reports for both users and clients. E.1: The system should print inventory report for users. E.2: The system should print inventory reports for clients.	Vague
Rule 3: when sentence containing **not only, but also, as well as,** split it to two sentences.	E: A reward system must be established not only for the individuals, but also for organizations and teams of employees. E.1: A reward system must be established for the each individual. E.2: A reward system must be established for each organization. E.3: A reward system must be established for each team of employees. E: The system shall process data received from users and clients as well as to produce a standard report on it. E.1: The system shall process data received from each user and each client and to produce a standard report on it. E.1.1: The system shall process data received from each user and each client. E.1.2: The system shall produce a standard report on it.	Vague
Rule 4: when sentence containing **eventually, at last, finally,** replace with specific time.	E: When a client makes a request, the server must eventually receive data. E.1: When a client makes a request, the server must receive data no later than 24 hours. E: The System shall finally be able to receive data from mirrored sites. E.1: The System shall be able to receive data from mirrored sites no later than 24 hours after completion of processing.	Vague
Rule 5: when sentence containing **maximum or minimum**, replace with specific number.	E: The system shall return minimum results to the user. E.1: The system shall return at least 1 search result to the user.	Vague
Rule 6: when sentence containing **as much as possible, as many as possible, as little as possible, or as few as possible,** replace with specific unit.	E: Simulated output should accommodate as many events as possible. E.1: Simulated output shall accommodate at least Second Level Event.	Vague
Rule 7: when sentence containing **unless**, replace with **if not**.	E: Unless the user has the administrator's authorization, the user will not be able to access the database. E.1: If the database user does not have the administrator's authorization, the database user shall not be able to access the database. E: The system will display registration alert unless the user has registered. E.1: The system shall display registration alert if the authorized user has not registered.	Vague
Rule 8: when sentence containing **But**, split it to two sentences.	E: The Cask Loader software shall provide not only cask loading tracking support, but optimization of heat loading. E.1: The Cask Loader software shall provide cask loading tracking support. E.2: The Cask Loader software shall provide optimization of heat loading.	Coordination Ambiguity
Rule 9: when sentence containing **and/or**, split it to two sentences.	E: An authorized user shall have the ability to edit and/or void a log entry. E.1: An authorized user shall have the ability to edit a log entry. E.2: An authorized user shall have the ability to void a log entry. E: Avoid stop or start message. E.1: Avoid stop message. E.2: Avoid start message.	Coordination Ambiguity
Rule 10: when sentence containing **all, any, some, many, few, or several** replace with **each**.	E: The operator log will record all warning messages prompted by the system. E.1: The operator log will record each warning message prompted by the system.	Scope Ambiguity

4 Case Study

Elevator Case Study

To show how the presented modules work, its tasks are going to be applied on a real example "Elevator case study". Our input is software requirement specification in natural language as shown in figure 6.



"You are to create a program that allows a user to issue a set of elevator requests. A request contains the floor at which the request is made and also the floor to which the user wants to go. When the user presses the “GO” button in the graphic view, the elevator fast processes the requests that have been entered via it. The elevator scheduling algorithm examines all current requests to determine the next floor and direction. When a new request is added, the algorithm should recalculate the next floor and direction".

Figure 6: Software Requirement Specifications

The input requirements document is split into separate sentences. The individual sentences are then passed to Tokenizer, the Tagger which identifies the individual words’ part of speech, and marks phrase boundaries and the finally syntactic parser. The output of Text Preprocessing Module is shown in figure 7, figure 8, figure 9, figure 10 and figure 11.

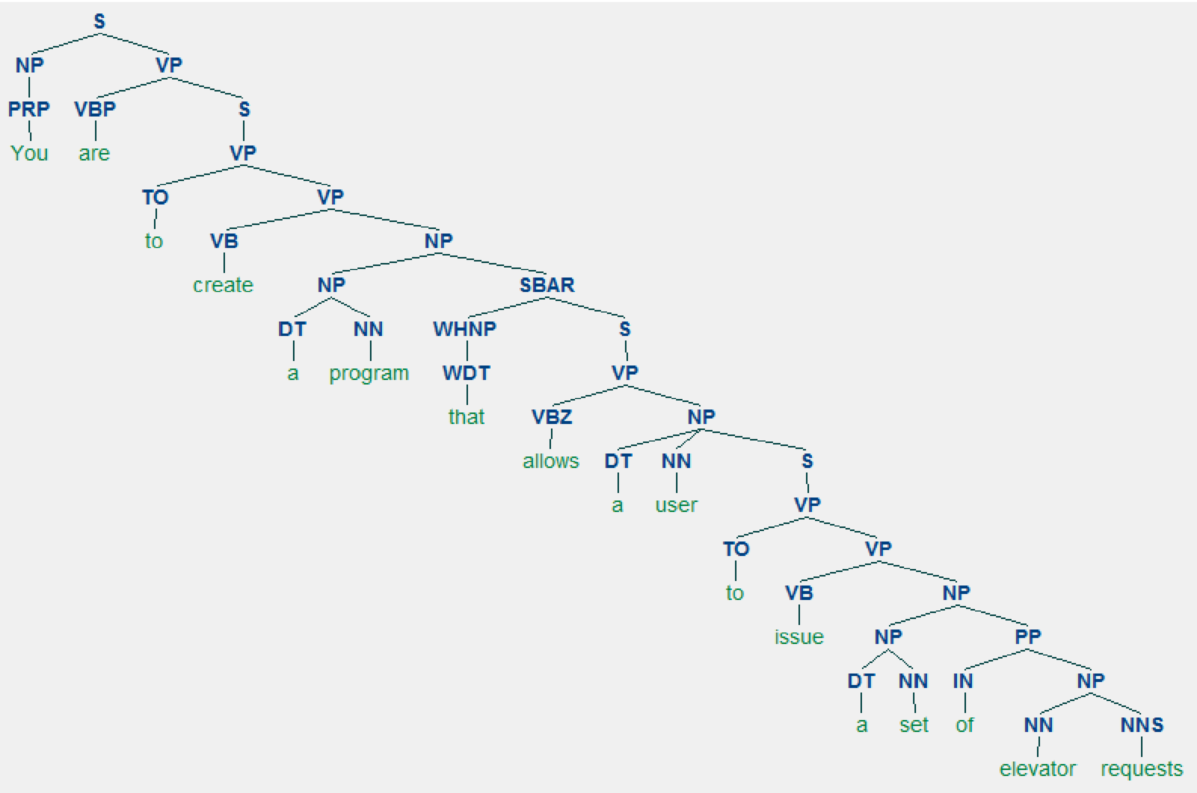


Figure 7: Parsing

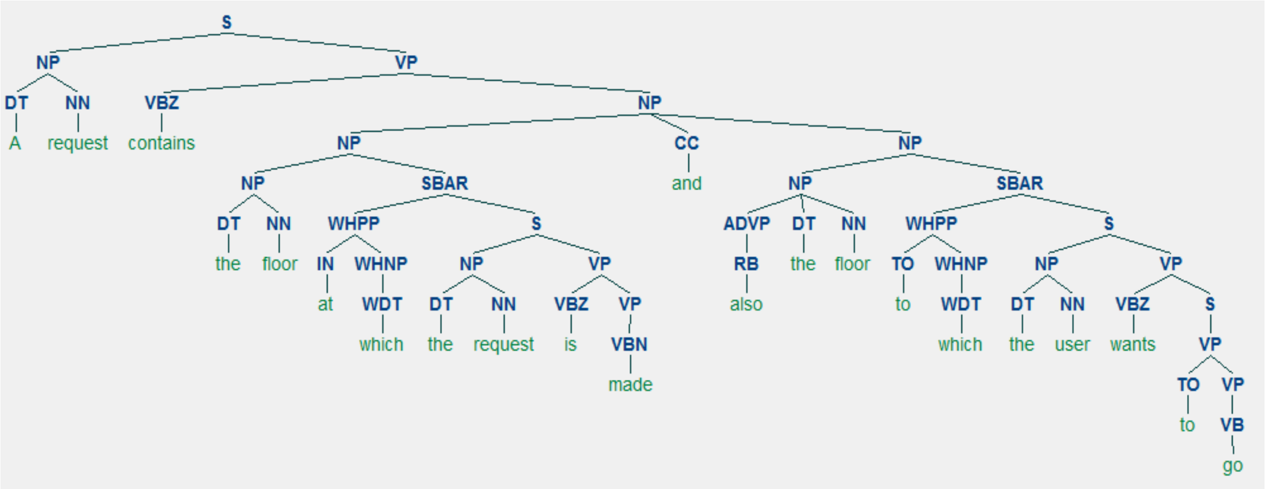

S
NP VP
DT NN VBZ NP
A request contains
NP CC NP
NP SBAR and NP SBAR
DT NN WHPP S ADVP DT NN WHPP S
the floor IN WHNP NP VP RB the floor TO WHNP NP VP
at WDT DT NN VBZ VP also to WDT DT NN VBZ S
which the request is VBN which the user wants VP
made TO VP
to VB
go

Figure 8: Parsing

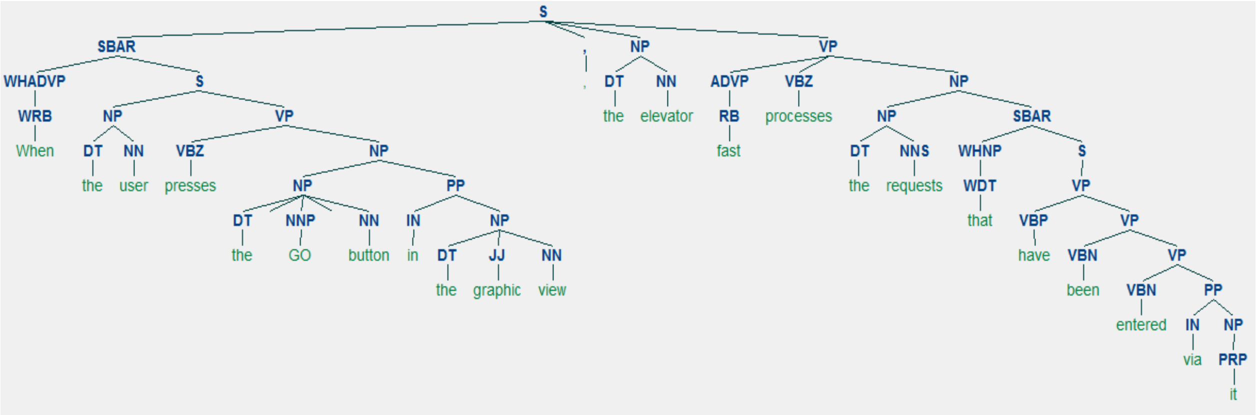

S
SBAR , NP VP
WHADVP S DT NN ADVP VBZ NP
WRB NP VP the elevator RB processes NP SBAR
When DT NN VBZ NP fast DT NNS WHNP S
the user presses NP PP the requests WDT VP
DT NNP NN IN NP that VBP VP
the GO button in DT JJ NN have VBN VP
the graphic view been VBN PP
entered IN NP
via PRP
it

Figure 9: Parsing

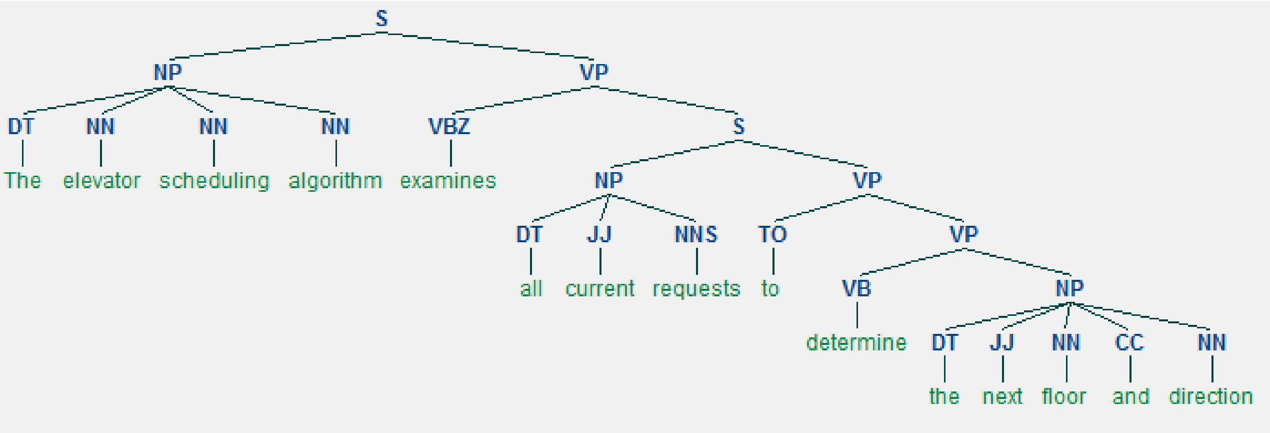

S
NP VP
DT NN NN NN VBZ S
The elevator scheduling algorithm examines NP VP
DT JJ NNS TO VP
all current requests to VB NP
determine DT JJ NN CC NN
the next floor and direction

Figure 10: Parsing

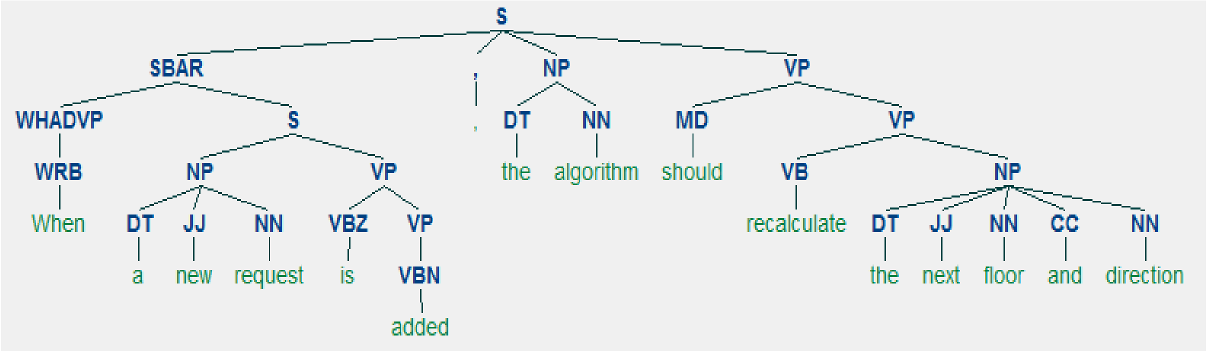
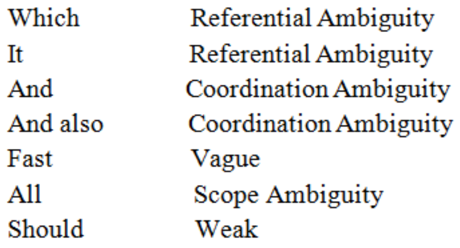


Figure 11: Parsing

In Ambiguity Detection Module detect which sentences in a natural language requirement specification are ambiguous and, for each ambiguous sentence, identify the ambiguity word in the sentence. The output of Ambiguity Detection Module is shown in figure 12.

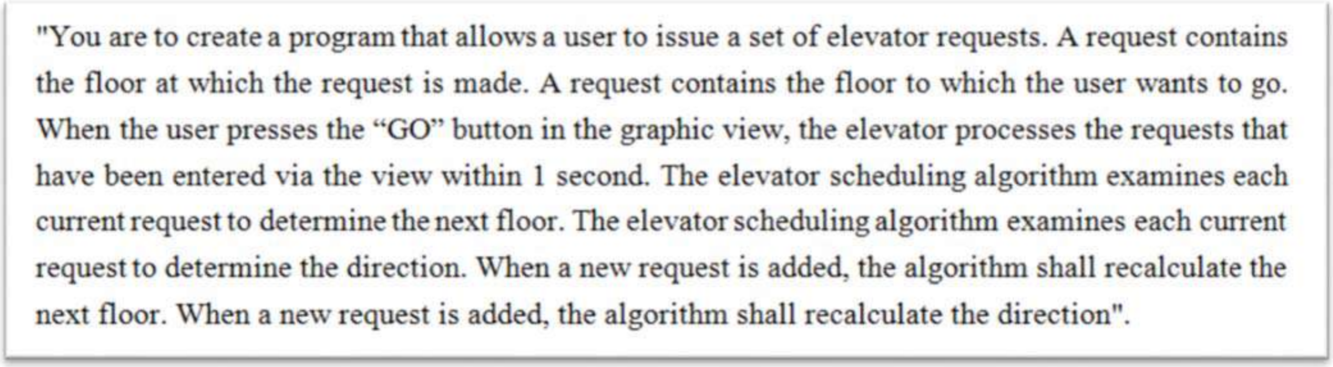


Which	Referential Ambiguity
It	Referential Ambiguity
And	Coordination Ambiguity
And also	Coordination Ambiguity
Fast	Vague
All	Scope Ambiguity
Should	Weak

Figure 12: Ambiguity Detection Module Output

In Ambiguity Resolve Module resolve the ambiguity. Resolve the ambiguity word "And" according to rule 9. Resolve the ambiguity word "Should" according to rule. Resolve the ambiguity word "And also" according to rule 3. Resolve the ambiguity word "It" according to rule. Resolve the ambiguity word "Fast" according to rule 1.and Resolve the ambiguity word "All" according to rule 10.

The output of Ambiguity Detection Module is Unambiguous requirement specification as shown in figure 13.



"You are to create a program that allows a user to issue a set of elevator requests. A request contains the floor at which the request is made. A request contains the floor to which the user wants to go. When the user presses the "GO" button in the graphic view, the elevator processes the requests that have been entered via the view within 1 second. The elevator scheduling algorithm examines each current request to determine the next floor. The elevator scheduling algorithm examines each current request to determine the direction. When a new request is added, the algorithm shall recalculate the next floor. When a new request is added, the algorithm shall recalculate the direction".

Figure 13: Unambiguous Software Requirement Specifications

4 Conclusions

One of the most essential stages of software development is requirement gathering. Rest of the project depends on this step i.e. how requirements are understood, collected and described. If requirements are not correctly understood, or software requirements specification is not correctly designed, then the result will be ambiguous software requirements specification document. Ambiguities in software requirements specification presents conflicts in the software project, as different interpretations can be stated by team members while understanding requirements, which finally affect the quality of system to be develop. One way to resolve this problem is to detect and resolve ambiguities early, in the requirement analysis stage. So our tool is designed that finds ambiguities in software requirements specification document and resolve it. The future work, our tool will extract the object oriented information from software specification requirements such as classes, instances and their respective attributes, operations, associations, aggregations, and generalizations to enhance the text analysis process to generate UML diagrams like use-case, activity diagram, collaboration diagram and sequence diagram.

REFERENCES

[1] Sommerville, I. and Sawyer, P. 1997."Requirements Engineering A good practice guide". Chichester: John Wiley & Sons Ltd.

[2] Nuseibeh, B., & Easterbrook, S. 2000, May. "Requirements engineering: a roadmap". *In Proceedings of the Conference on the Future of Software Engineering* (pp. 35-46).

[3] Belev, G. C. 1989, January. "Guidelines for specification development". In Reliability and Maintainability Symposium, 1989. Proceedings., Annual (pp. 15-21). IEEE.

[4] Christel, M. G., & Kang, K. C. 1992. Issues in requirements elicitation (No. CMU/SEI-92-TR-12). CARNEGIE-MELLON UNIV., PITTSBURGH, PA SOFTWARE ENGINEERING INST.

[5] Donald G. Firesmith. 2007. Common Requirements Problems, Their Negative Consequences, and Industry Best Practices to Help Solve Them. In Journal of Object Technology, vol. 6, no. 1, January-February 2007, pp. 17-33

[6] Berry, D.M., Kamsties, E., Krieger, M.M.: "From contract drafting to software specification: Linguistic sources of ambiguity," http://se.uwaterloo.ca/˜dberry/handbook/ambiguityHandbook.pdf, 2003.

[7] Fabbrini, F., M. Fusani, S. Gnesi, and G. Lami. 2001. The Linguistic Approach to the Natural Language Requirements Quality: Benefit of the use of an Automatic Tool. SEW'01 proceeding of the 26th annual NASA Goddard Software En gineering Workshop, IEEE Computer Society Washington, DC, USA, 97.

[8] Sven Körner and Torben Brumm. 2009. RESI-A natural language specification improver. IEEE International Conference on Semantic Computing (ICSC).

[9] Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics - Special issue on word sense disambiguation, Volume 24, Issue 1, 2-40.

[10] Sri Fatimah Tjong. 2008. Avoiding ambiguity in requirements specifications. Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy.

[11] Tjong, Sri Fatimah, and Daniel M. Berry. 2013. The Design of SREE—A Prototype Potential Ambiguity Finder for Requirements Specifications and Lessons Learned. Requirements Engineering: Foundation for Software Quality. Springer Berlin Heidelberg, 2013, pp. 80-95.

[12] Willis, Alistair, Francis Chantree, and Anne De Roeck. 2008. Automatic Identification of Nocuous Ambiguity. Research on Language & Computation, 6 (3-4), 1-23.

[13] Hui Yang, Alistair Willis, Anne De Roeck, Bashar Nuseibeh. 2010. Automatic Detection of Nocuous Coordination Ambiguities in Natural Language Requirements. Proceedings of the IEEE/ACM international conference on Automated software engineering, 53- 62. ISBN: 978-1-4503-0116-9. DOI=10.1145/1858996.1859007.

[14] Hui Yang, Anne de Roeck ,Vincenzo Gervasi, Alistair Willis Bashar Nuseibeh. 2011. Analyzing anaphoric ambiguity in natural language requirements. Requirements Engineering - Special Issue on Best Papers of RE'10: Requirements Engineering in a Multifaceted World, Volume 16, Issue 3, 163- 189. DOI=10.1007/s00766-011-0119 y.

[15] Imran Sarwar Bajwa. 2012. Resolving Syntactic Ambiguities in Natural Language Specification of Constraints. CICLing'12 Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing, Volume 1, pp. 178-187.

[16] Basili, Victor R., Scott Green, Oliver Laitenberger, Filippo Lanubile, Forrest Shull, Sivert Sorumgard. 1995. The Empirical Investigation of Perspective-Based Reading. Technical report the empirical investigation of perspective based reading.

[17] Gracia, Jorge; Lopez, Vanessa; d‟Aquin, Mathieu; Sabou, Marta; Motta, Enrico and Mena, Eduardo, "Solving semantic ambiguity to improve semantic web based ontology matching," in The 2nd International Workshop on Ontology Matching, Busan, South Korea, 2007.

BIOGRAPHY

Somaia Osama: She graduated from faculty of Computer Science in 2009 at Akhabr El Yom Academy, Cairo, Egypt. She started working as a teaching assistant in the Computer Science department at Akhabr El Yom Academy since Sept 2009 till now. Then she got a diploma in software architect from Information Technology Institute, Smart Village, Egypt in 2011.

Dr. Safia Abbas: She received his Ph.D. (2010) in Computer science from Nigata University, Japan, her M.Sc. (2003) and B.Sc.(1998) in computer science from Ain Shams University, Egypt. Her research interests include data mining argumentation, intelligent computing, and artificial intelligent. She has published around 15 papers in refereed journals and conference proceedings in these areas which DBLP and Springer indexing. She was honored for the international publication from the Ain Shams University president.

Mostafa Aref is a professor of Computer Science and Vice Dean for Graduate studies and Research, 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.

كشف الغموض وحله في مواصفات المتطلبات باللغة الطبيعية

سمية اسامة، صفية عباس، مصطفى عارف
قسم علوم الحاسب، كلية الحاسبات والمعلومات، جامعة عين شمس

ملخص

اللغة الطبيعية هي التمثيل الأكثر استخداما لوصف مواصفات المتطلبات على الأنظمة المستندة إلى الكمبيوتر في الصناعة.من جهة اللغة الطبيعية مرنة وعالميه و واسعة الانتشار. من ناحية أخرى يتم التعرف على مواصفات المتطلبات باللغة الطبيعية عموما بأنها غامضة.يحدث التباس عندما يمكن أن تفسرمواصفات المتطلباتبطريقة مختلفة من جانب مختلف القراء. منذ فترة طويلة تم الاعترافبالغموض في مواصفات المتطلبات باللغة الطبيعية باعتبارها تحديا في مجال المتطلبات الهندسية. في هذا البحث نحن نركز على تقنية معالجة اللغات الطبيعية (NLP) لمعالجة الغموض في مواصفات المتطلبات باللغة الطبيعية. وقد قمنا بوصفالنهج الآلي للكشف عن الغموض وحله من أجل تجنب التفسيرات الخاطئة. ويقدم البحث أيضا دراسة حالة لمواصفاتمتطلبات البرنامج عن العالم الحقيقي لإظهار كيفية استخدام هذا النهج في الكشف الآلي للغموض وحله حيث قمنا بتقييم كفاءة هذا النهج.

A Combined DES and Elliptic Curve Cryptography Cryptosystem to Secure Audio Data

Mohamed Ahmed Seifeldin*1, Abdellatif Ahmed Elkouny**2, Salwa Hussein Elramly*3

**Electronics and Communication Department, Faculty of Engineering, Ain Shams University*
Abbasia, Egypt

1mseif34@gmail.com

3salwahelramly@gmail.com

***Computer Science Department, Faculty of Engineering, Ahram Canadian University*
6thOctober, Egypt

2aelkouny@gmail.com

***Abstract*—Nowadays digital communications are considered everything in our daily dealings. Audio telecommunications are one of the principal forms of digital data communications that play a vital role in our life. Often, the communicated audio data via the involved networks have a level of privacy and secrecy. Audio data's confidentiality must be attained to fulfill the objectives of privacy and secrecy. In this paper, we introduce a new cryptosystem to achieve the task of preserving the audio data's confidentiality thoroughly. The proposed cryptosystem is based on both the Data Encryption Standard (DES) and the Elliptic Curve Cryptography (ECC). DES algorithm was a prevalent technique that was withdrawn in 2000 according to its weakness against the brute-force attack. In this paper, we revive DES using the art of the ECC, which is the youngest member of the public-key family. We eliminate DES's weakness and bring it as a very robust cryptosystem. We apply our proposed technique to secure the audio data via any network. We conduct comprehensive statistical analyses to assess our technique. The obtained analyses' results are very motivating and promising.**

1 INTRODUCTION

Information security is the concerned discipline to secure the communicated digital data via networks against all the prospective security threats. These security threats are the potential security attacks which always try to utilize and exploit the communicated digital data either actively or passively [1]. There are many security mechanisms that were introduced to fulfill the requirements of many of security services and to provide a kind of fortification to the communicated data against a variety of security attacks [1]. Any security mechanism is implemented by applying a reversible algorithm for both modifying the transmitted data and revealing it on the reception. Cryptography is the security mechanism that is required to provide the needed data confidentiality service to conceal the communicated data from any attacker [2]. One of the most known cryptographic systems (cryptosystems) is the Data Encryption Standard (DES). DES was proposed in 1977 and was the first approved and announced algorithm by the US government [3]. DES is categorized as a symmetric-key cryptosystem, where the secret key of the encryption and decryption processes is the same [3]. DES's structure depended on the **FEISTEL** structure [4], which is known as the *product cipher* structure. From its first emergence as a standard, DES algorithm was a controversial topic according to some obscured criteria of its structure [5]. Although all the demonstrated disputes about DES's internal structure, DES shows high level of immunity against both differential and linear attacks [5]. The brute-force attack was the only way that brought DES as insecure algorithm. In essence, the brute-force attack is the repetitive trials of every possible key in the key space of any algorithm until one trial hits the correct secret key. In 1999, DES was withdrawn by National Institute of Standards and Technology (NIST) and approved as an insecure cryptosystem [6]. Nevertheless, DES still appears in many cryptosystems. Triple-DES is the successor of DES, which involves applying DES three times in a row to enlarge the key space of the classic DES three times its size. In addition, the **FEISTEL** structure, which is the cornerstone of the DES's structure, influences many modern cryptosystems. Accordingly, DES, as a structure, is not ended and still around.

Elliptic Curve Cryptography (ECC) is a public-key cryptosystem that depend on the mathematical model of the Elliptic Curve (EC) equation [7]. ECC was first introduced in 1980s as the newest member of the public-key family after the key factorization problem members like **RSA**[8] [9]. ECC shows an impressive performance and a high immunity against the brute-force attack. ECC requires more difficult arithmetic operation than any other public key model. NIST have proposed **10** securely certified ECs to be embodied in any type of applications, either hardware or software [10].

This paper aims to enhance the security of classical DES by overcoming its vulnerability against the brute-force attack. We also use the ECC to eliminate the disadvantage of secret key generation and distribution which is accompanied with any symmetric-key cryptosystem. As an application to our proposed technique, we simulate a communication session between two entities via network to send and receive an audio data file securely. We conduct comprehensive statistical analyses to assess our proposal thoroughly. The results are very motivating and promising.

2 DATA ENCRYPTION STANDARD

DES is categorized as a symmetric-key (*block cipher*) algorithm, i.e. the same secret key is used in both encryption and decryption processes. DES deals with the digital data as blocks of bits, each is 64-bit long block [11]. DES comprises **16** identical steps, which are called *rounds*. The secret key of the DES algorithm is 56-bit key. DES transforms the input plaintext to the output ciphertext using this secret key [11]. Each round of the 16 rounds of DES is fed with a 48-bit different subkey. All the 16 subkeys are generated from the main secret key. The process responsible for producing all the subkeys is called the *key schedule process*. DES's decryption process is identical to the encryption one except that the 16 subkeys are applied in reverse order. Fig. 1 shows the internal structure of DES, which is a **FEISTEL** structure [11].

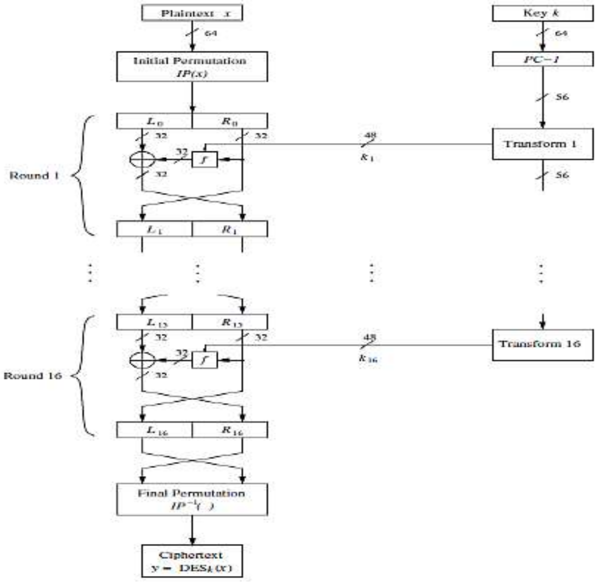


Figure 1: DES's inside structure [11]

A. *The involved operations in each DES's round*

DES algorithm consists of three stages, the initial permutation (IP), the rounds' operation, and the final permutation (IP^{-1}). Fig. 1 depicts all of these stages. The IP permutes (shuffles) the input 64-bit plaintext in a determined manner [11]. After the IP stage, the data's block is split to 32-bit halves, R_0 (right half) and L_0 (left half). In the rounds' stage, each round does the same task as its previous round. For the first round, $R_0$and the round's subkey k_i are fed to the $\boldsymbol{f}$ block. Also, R_0 is passed directly to the next round's left half as in Fig. 1. The output of the $\boldsymbol{f}$ block is then XORed with L_0 and the result is passed to the next round's right half. For the second round, the same operation holds. The operation of any of DES's rounds can be expressed mathematically as:

$$L_i = R_{i-1} \tag{1}$$

$$R_i = L_{i-1} \oplus f(R_{i-1}, k_i) \tag{2}$$

After the 16th round, the output halves are swapped and then are fed to the final stage, IP^{-1}. IP^{-1} operation is just the direct inverse of the IP operation. After the IP^{-1} operation, the output ciphertext of the DES algorithm is obtained [11].

B. *DES's key schedule process*

The key schedule process of DES has the responsibility to generate and deliver the concerned subkeys to their respective rounds. The key schedule process incorporates both permutation and rotation processes. Fig. 2 Shows the internal structure of the DES's key schedule process. The input secret key (64 bits) is entered to the permutation choise-1 (**PC-1**) block, which permutes the secret key and discards the extra added 8 bits (they are used as parity bits to check the main 56-bit key) [11]. After that, the output of PC-1 is split to two halves. For each round, the input two halves are left rotated with a certain number of bits correspond to the concerned round. After each rotation process, the output goes through the permutation choise-2 (**PC-2**) block. PC-2 block permutes the input bits and outputs 48-bit key to be entered to the respective rounds' operation.

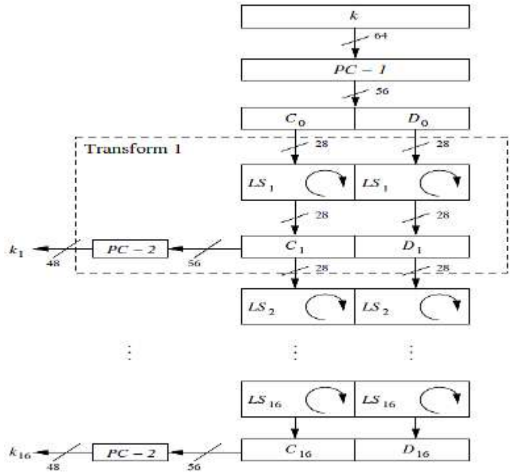


Figure 2: The internal structure of DES's key schedule process [11]

3 ELLIPTIC CURVE CRYPTOGRAPHY (ECC)

ECC is a public-key cryptosystem that was introduced in 1980s by **Victor Miller** and **Neal Kobiltz** [8] [9]. ECC is based on the mathematical model of the Elliptic Curve (EC) over the real number. For the cryptographic use, the EC is defined over a finite group or the finite set$\mathbb{Z}\boldsymbol{p}$ [7]. The EC is defined by all points that fulfill (3) accompanied with the imaginary point of infinity$\boldsymbol{O}$along the y-axis [7]:

$$y^2 = x^3 + a.x + b \qquad mod\ p \tag{3}$$

where a and b are constants, p is the prime order of the finite set $\mathbb{Z}\boldsymbol{p}$and it must be greater than 3[7] [12]. There is a condition for any EC to be utilized cryptographically that it must not have any vertices or singularities [7] [12]. This is achieved if (4) is satisfied.

$$4.a^3 + 27.b^2 \neq 0 \qquad mod\ p \tag{4}$$

If the previous conditions are valid, then the EC can be employed in a cryptosystem [7] [12].

To construct a cryptosystem over a finite set $\mathbb{Z}\boldsymbol{p}$, a group operation must be defined over the concerned set [12]. Point addition is the defined group operation over the EC. Point addition over the EC is a unique operation and unlike the ordinary algebraic addition. If there are a point $P = (x_1, y_1)$ and a point $Q = (x_2, y_2)$ on the curve, then the result of adding $P + Q$ is the point $R = P + Q = (x_3, y_3)$ and R is also on the curve [7]. By adding the point this way, the finite group is constructed [12].

A. The discrete logarithm problem (DLP) over the ECs

Any cryptographic algorithm must have a one-way intractable arithmetic problem, which is easy from one side and very impossible from the other one. The arithmetic problem over the EC is categorized as DLP [7]. This problem is called Elliptic Curve Discrete Logarithm Problem (**ECDLP**) [7] [12].

For any EC, there are domain parameters that are publicly known to everyone even the attacker. One of these parameters is the base point $\boldsymbol{P}$ that can generate all the points on the curve. The group cardinality $\boldsymbol{\#E}$ is the total number of the points on the curve. If there is a point $\boldsymbol{T}$ that equals a defined number of addition of $\boldsymbol{P}$ to itself, say $\boldsymbol{x}$ times of addition [7]. The ECDLP from the attacker perspective is to find the integer $\boldsymbol{x}$ (where $1 \leq x < \#E$) given $\boldsymbol{P}$, $\boldsymbol{T}$, and the prime $\boldsymbol{p}$ using the equation:

$$x = log_P T \qquad mod\ p \tag{5}$$

The ECDLP is very difficult to solve and it will be considerably intractable if $\boldsymbol{\#E}$ is large enough.

NIST have proposed standard ECs to work with in any application [10]. The NIST's curves are cryptographically tested and accredited. In this paper we employ the P-192 curve in our approach to enhance the DES security. The domain parameters of the P-192 curve are shown in Table I, each item of these parameters is 192-bit long. The implementation of the P-192 curve requires two levels of arithmetic, the EC's operations and the finite field operations [13] [14].

TABLE I
P-192 DOMAIN PARAMETERS

Parameter	***Value***
p	6277101735386680763835789423207666416083908700390324961279
a	6277101735386680763835789423207666416083908700390324961276
b	2455155546008943817740293915197451784769108058161191238065
P (the base point)	x_P= 602046282375688656758213480587526111916698976636884684818
	y_P= 174050332293622031404857552280219410364023488927386650641
#E (group cardinality)	6277101735386680763835789423176059013767194773182842284081

B. The employment of the ECC in our approach

In this paper, we implement the ECC by utilizing the well-known algorithm for sharing the secret key, Diffie-Hellman Key Exchange (DHKE) algorithm [7]. When using the ECs with the DHKE algorithm, the resulted algorithm is known as Elliptic Curve Diffie-Hellman Key Exchange (ECDHKE) algorithm. The operation of the algorithm is depicted in Fig. 3. The two involved entities in the communication session choose an integer ***A***, and ***B*** respectively, each is less than ***#E***. Then each participant applies the point multiplication process [7] to generate their public key. After that, each entity shares its public key with the other one. Then, another point multiplication process is applied on the received public key to generate the *joint secret* ***T_{AB}***, which can be utilized as the secret key of any cryptosystem [7]. Using this way, we have generated and distributed the secret key for DES securely. In addition, we have eliminated the disadvantage of sharing the secret key for a symmetric-key cryptosystem, like DES, via trusted secure channel. The used communication channel is assumed to be authenticated before the ECDHKE application to avoid the masquerade attacks [15].

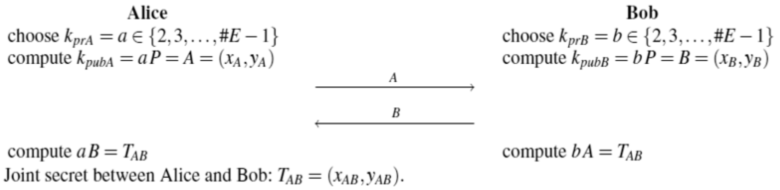

Alice
choose $k_{prA} = a \in \{2,3,\ldots,\#E-1\}$
compute $k_{pubA} = aP = A = (x_A, y_A)$
Bob
choose $k_{prB} = b \in \{2,3,\ldots,\#E-1\}$
compute $k_{pubB} = bP = B = (x_B, y_B)$
A
B
compute $aB = T_{AB}$
compute $bA = T_{AB}$
Joint secret between Alice and Bob: $T_{AB} = (x_{AB}, y_{AB})$.

Figure 3: ECDHKE algorithm operation [7]

4 THE PROPOSED APPROACH TO SECURE AUDIO DATA

From the previously described DES's key schedule process, each round requires 48-bit subkey. Since, there are 16 rounds, so that DES's operation requires 16 48-bit subkeys, which is a total of 768 bits. In our approach, we replace the ordinary DES's key schedule process with the ECC by using ECDHKE algorithm with P-192 as follows:

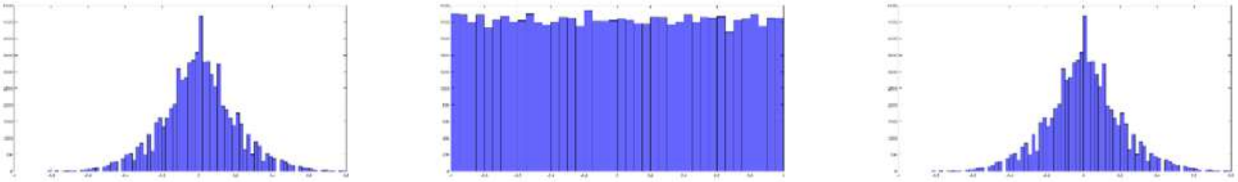
a. In the initialization of any communication session between two entities, each entity gets the joint secret ***T_{AB}*** as it is previously explained.
b. ***T_{AB}*** is just a point on the EC, which incorporates ***x*** and ***y*** coordinates of 192 bits each.
c. Each entity calculates the point ***2T_{AB}*** that is the double of ***T_{AB}***, the result obtained must be the same for each entity.
d. Now each entity has two points with ***x*** and ***y*** coordinates of 192 bits each, so that a total of 768 bits.
e. After that each entity concatenates the binary representation of these coordinates and then reshapes it to ***16x48*** matrix.
f. The obtained matrix represent the required subkeys for the DES's operation, where each row of the matrix acts as a round's subkey.

After the initialization, the two entities are ready to share data securely. We apply our implementation on sending and receiving audio data securely. All the implementations are done using the MATLAB® and the SIMULINK®software. We conduct the experimental analyses to assess our proposed approach thoroughly.

5 THE STATISTICAL ANALYSES AND RESULTS

A. Histogram analysis

Histogram analysis is used to show how the encrypted data appear as unintelligible and obscure data to the attacker. The histogram plot of the encrypted data must be almost uniform to ensure the infeasibility of the statistical attack. The histogram plots of the original, encrypted, and decrypted audio data are shown in Fig. 4, respectively. From Fig. 4, it is very obvious that there is no loss of data between the original and decrypted data. In addition, the encrypted data are fairly uniform and so the statistical attacks are infeasible.

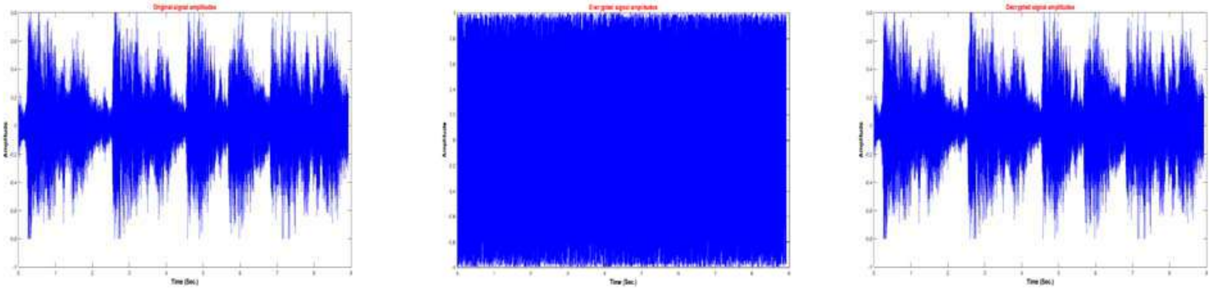


a: The original data histogram **b: The Encrypted data histogram** **c: The decrypted data histogram**

Figure 4: The histogram analysis

B. Amplitude analysis

The amplitude analysis is the plot of the amplitudes of the audio data samples versus time for the original, decrypted and encrypted audio data. For a good cryptosystem, the encrypted amplitudes must be almost uniform, so that it represents nothing than noise [16]. Fig. 5 shows the amplitudes of the original, encrypted, and decrypted data respectively. It is clear that the encrypted data resembles noise, thus the attacker cannot reconstruct or guess the original data from it. In addition, the original and the decrypted data are very identical, i.e. there is not data loss.

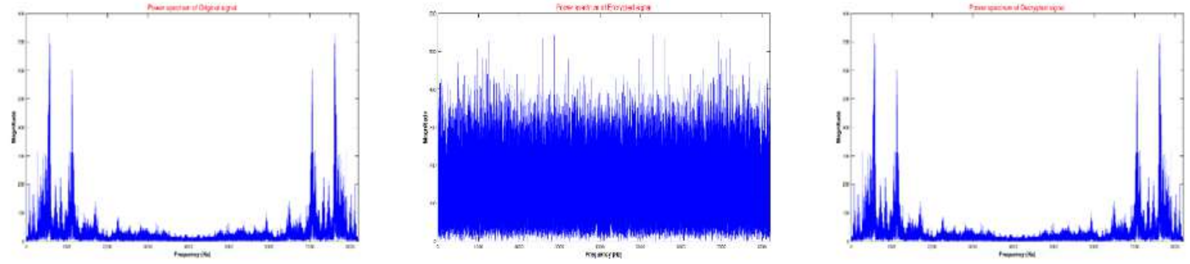


a: The original data amplitude **b: The encrypted data amplitude** **c: The decrypted data amplitude**

Figure 5: The amplitude analysis

C. The power spectrum analysis

The power spectrum analysis is an observation of the audio data magnitudes in the frequency domain. From the attacker's perspective, it could be attainable to construct the audio data if the frequency domain's data magnitudes are intelligible [17]. Fig. 6 depicts the frequency domain's data magnitudes of the original, encrypted, and decrypted data respectively. It is clear that the encrypted data's magnitudes resemble noise, thus the attacker cannot reconstruct or predict the original data from it. In addition, the original and the decrypted data are very identical, i.e. the decryption process reconstructs the original data identically.



a: The original data magnitudes **b: The encrypted data magnitudes** **c: The decrypted data magnitudes**

Figure 6: The power spectrum analysis

6 SIGNAL TO NOISE RATIO (SNR) ANALYSIS

This analysis is used as a measurement to detect how much the original data are embedded in the encrypted data. The smaller the value of the SNR, the more different the encrypted data with respect to the original data [18]. The SNR is calculated according to the equation:

$$SNR = 10.\,log_{10}\frac{\sum_{i=1}^{L} x_i^2}{\sum_{i=1}^{L}[x_i - y_i]^2} \quad (dB) \tag{6}$$

where x and y are the original and the encrypted data samples' vectors respectively. L is the length of the samples vector for either the original or the encrypted data. The smaller the value of the SNR, the more different the encrypted and the original data. The desired value of the SNR is in the negative region. The calculated SNR of our case here is **-9.86 dB**. Accordingly, the original data are undistinguishable from the encrypted data.

7 KEY SPACE ANALYSIS

This analysis is the most important analysis, since it deals with weak point of the DES. DES's key space is actually 56 bits. Our proposed algorithm's key space is 768 bits. According to the mutual independence of the generated 768 bits from each other, the actual key space of our approach is$2^{768} \cong 15.53\times10^{230}$, which is huge enough in comparison to the classical DES's key (2^{56}) [19]. Correspondingly, the brute-force attack is highly infeasible and our proposed cryptosystem is very robust against it.

8 CONCLUSIONS

In this paper, we have proposed a new cryptosystem that is based on merging of the classical DES and the art of the ECC. We have applied our approach on securing audio data communication between two entities by generating and distributing the secret key then sending encrypted audio data and decrypting it on the reception. We have conducted experimental analyses to judge our algorithm. The results show that the encrypted data have a good histogram and time series amplitudes. In addition, the power spectrum present that the frequency attacks are unachievable. The SNR analysis shows that the encrypted audio data is almost totally different from the original audio data. The key space of our new cryptosystem is substantially very wide comparing with its DES's counterpart and so that the brute-force attack is infeasible. As a conclusion, we have provided a new technique that can bring the classical DES to live again. The new DES is very robust against many kinds of attacks like the statistical attacks, the frequency attacks, and the brute-force attack.

BIOGRAPHY

Eng. Mohamed Ahmed Seifeldin has graduated from the department of Electrical engineering, the Faculty of engineering, Alexandria University. He prepared his Master's degree at Ain-Shams University. He is very interested in the field of *Information Security*.

TRANSLATED ABSTRACT

نظام تأمين بيانات الصوت يجمع بين نظامى تشفير البيانات و نظام تشفير منحنى القطع الناقص

ملخص

تعتبر الاتصالات الرقمية في الوقت الحاضر كل شيء في التعاملات اليومية. الاتصالات الصوتية هي واحدة من الأشكال الرئيسية لإتصالات البيانات الرقمية التي تلعب دورا حيويا في حياتنا. في كثير من الأحيان، البيانات الصوتية التي ترسل عبر الشبكات المعنية لها مستوى عالي من الخصوصية والسرية. يجب أن تتحقق سرية البيانات الصوتية لتفي بغرض تحقيق أهداف الخصوصية والسرية. في هذه الورقة العلمية، تم تقديم نظام تشفير جديد لتحقيق مهمة الحفاظ على سرية البيانات الصوتية بدقة. يستند نظام التشفير المقترح على كل من مقياس تشفير البيانات (**DES**) ونظام تشفير منحنى القطع الناقص (**ECC**). كان **DES**الخوارزمية السائدة والتي تم سحبها في عام 2000 وفقا لضعفها ضد هجوم القوة الغاشمة. في هذه الورقة العلمية، سوف يتم إحياء **DES** باستخدام فن **ECC**، وهو أصغر عضو في عائلة المفتاح العام. ز ذلك للقضاءعلي ضعف النظام**DES**، وجعله نظام تشفيري قوي للغاية. تم تطبيق الأسلوب المقترح على تأمين البيانات الصوتية عبر أي شبكة. تم إجراء تحليلات إحصائية شاملة لتقييم الأسلوب المقترح. نتائج التحليلات التي تم الحصول عليها محفزة جدا وواعدة.

REFERENCES

[1]I. Rec, "X. 800 security architecture for open systems interconnection for CCITT applications,"*ITU-T CCITT Recommendation*, 1991.

[2] M. Bellare and P. Rogaway, "Introduction to modern cryptography," *UCSD CSE*, vol. 207,p. 10, May11 2005.

[3] Data Encryption Standard *et al.*, "Federal information processing standards publication 46," *US Departmentof Commerce National Bureau of Standards*, 1977.

[4]H. Feistel, "Cryptography and computer privacy," *Scientific American*, vol. 228, pp. 15–23, 1973.

[5] W. Stallings, *Cryptography and Network Security*, 5th ed. New York, USA: PrenticeHall, 2011, pp. 77–96.

[6]J. Gilmore, "Cracking DES: Secrets of encryption research, wiretap politics & chip design,"1998.

[7] C. Paar and J. Pelzl, *Understanding Cryptography*. Germany: Springer-Verlag BerlinHeidelberg, 2010, pp. 239–255.

[8] N. Koblitz, "Elliptic curve cryptosystems," *Mathematics of computation*, vol. 48, no. 177,pp. 203–209, 1987.

[9] V. S. Miller, "Use of elliptic curves in cryptography," in *Advances in Cryptology—CRYPTO'85 Proceedings*. Springer, 1985, pp. 417–426.

[10] P. FIPS, "186-2. Digital Signature Standard (DSS)," *National Institute of Standards andTechnology (NIST)*, 2000.

[11] C. Paar and J. Pelzl, *Understanding Cryptography*. Germany: Springer-Verlag BerlinHeidelberg, 2010, pp. 55–78.

[12] W. Stallings, *Cryptography and Network Security*, 5th ed. New York, USA: PrenticeHall, 2011, pp. 308–320.

[13] D. Hankerson, A. Menezes, and S. Vanstone, *Guide to Elliptic Curve Cryptography*. NY10010, USA: Springer-Verlag New York, 2003.
[14] M. Brown, D. Hankerson, J. López, and A. Menezes, "Software Implementation of theNIST Elliptic Curves Over Prime Fields," in *Proc. Topics in Cryptology — CT-RSA*,Nagoya, Japan, Apr. 2001, pp. 250–265.
[15] W. Stallings, *Cryptography and Network Security*, 5th ed. New York, USA: PrenticeHall, 2011, p. 305.
[16] A. A. Tamimi and A. M. Abdalla, "An Audio Shuffle-Encryption Algorithm," in *The World Congress on Engineering and Computer Science*, 2014.
[17] S. Sharma, H. Sharma, and L. Kumar, "Power Spectrum Encryption and Decryption of anAudio File," *International Journal of Research in Computer Science*, vol. 1, 2013.
[18] E. Mosa, N. W. Messiha, O. Zahran, and F. E. abd El-Samie, "Chaotic encryption of speechsignals," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 285–296, 2011.
[19] M. A. S. Eldeen, A. A. Elkouny, and S. H. Elramly, "DES algorithm security fortificationusing elliptic curve cryptography," in *10th International Conference on ComputerEngineering & Systems (ICCES)*. IEEE, 2015, pp. 335–340.

A Speech Cryptosystem Based On Chaotic Modulation Technique

Mahmoud F. Abd Elzaher[*1], Mohamed Shalaby[**2], Yasser Kamal[**3], Salwa El Ramly[*4]

[*]*Department of Electronics and Electrical Communications, Ain Shams University*
Cairo, Egypt

[1]8273@eng.asu.edu.eg

[4]Salwa_elramly@eng.asu.edu.eg

[**]*Department of Computer Science, Arab Academy for Science, Technology*
and Maritime Transport, Cairo, Egypt

[2]myousef73@hotmail.com

[3]hockm1983@gmail.com

***Abstract*—In this paper, an encryption approach for Speech communication based on direct chaotic modulation (non-autonomous modulation) is presented, in which speech signal (S_m) is injected into one variable of the master system (using Lorenz system) without changing the value of any control parameter. This approach is based on the change of chaotic signal by injecting Speech samples into one variable in chaotic system and hence generating a new chaotic signal. The Speech signal is then extracted from the chaotic signal in the receiver side. Furthermore, a high dimension chaotic system is used, which increases the security of the encrypted signal. Non-autonomous modulation technique is suitable for securing real-time applications. A comparative study of approach and Speech masking technique is also presented. Experimental results show that modifying chaotic approaches increases the security of the encryption system.**

Keywords: Encryption, Speech encryption, Chaotic Modulation, Non-autonomous modulation, Lorenz system.

1 INTRODUCTION

The Speech communication is in close relation with daily life, such as education, commerce, politics, e-learning and news telecasting. With the advancement of modern telecommunication and multimedia technologies, Modern Speech communication systems demand a huge amount of information to be exchanged across Social Networks and the Internet every day so the need for encryption and security has increased. The conventional cryptographic techniques may be efficient for the text data; however, they are unsuitable to the bulk data capacity. One of the techniques that provides fast and highly secure encryption methods is chaos-based techniques.

Continuous cryptographic systems have been developed which use the synchronization between the transmitter and receiver to retrieve data transmitted through an insecure medium. The first generation of these systems is masking. A Speech masking technique based on Lorenz System is presented in [1, 2] which uses Lorenz equation to generate Chaotic Signals, these signals are used as a base carrier signal on which the information signal is modulated at the transmitter side. The information signal is then recovered at the receiver side. The method of masking has been shown to be insecure as there are various cryptanalysis methods [3] that make it possible to estimate the sender dynamics and decoding of the message signal.

The second generation is the parameter and non-autonomous modulation techniques. Non-autonomous techniques were developed to overcome the chaotic parameter modulation break, which includes the return map, and adaptive observer [4]. Non-autonomous modulation is considered to be more secure than parameter modulation.

The main goal of this paper is proposing a Speech encryption system that provides users with a high degree of confidence and key sensitivity, and preserving a good quality of the reconstructed speech signal by chaotic maps. In section 2, we discuss Chaos-based cryptography systems are discussed. In Section 3, a speech masking technique based on Lorenz System is presented. In Section 4, the proposed encryption approach is presented. The results of applying our proposed approach are shown in Section 5. Finally, our work is concluded in section 6.

2 CHAOTIC SYSTEM

Chaos theory was originally developed by mathematicians and physicists. The theory deals with the behaviors of nonlinear dynamic systems. Chaos theory has desirable features, such as deterministic, nonlinear, irregular, long-term prediction, and sensitivity to initial conditions. Therefore, and based on chaos theory features, the security research community adopts chaos theory in modern cryptography. A function that possesses a kind of chaotic behavior is defined as a chaotic function or map. In the following subsections, we discuss one types of chaotic systems (which we used to implement our proposed system), namely, Lorenz system.

Lorenz system can be described with three dimensions as shown in equations (1, 2 and 3).

$$\dot{X}(t) = \sigma(Y(t) - X(t)) \quad (1)$$
$$\dot{Y}(t) = rX(t) - X(t)Z(t) - Y(t) \quad (2)$$
$$\dot{Z}(t) = X(t)Y(t) - pY(t) \quad (3)$$

where $\dot{X}(t), \dot{Y}(t), \dot{Z}(t)$ are the Lorenz chaotic variables, $X(0), Y(0), Z(0)$ are initial conditions, and σ, r and p are positive constants with $r > 24.74$. Figure 1 shows a 3D figure of Lorenz chaotic system.

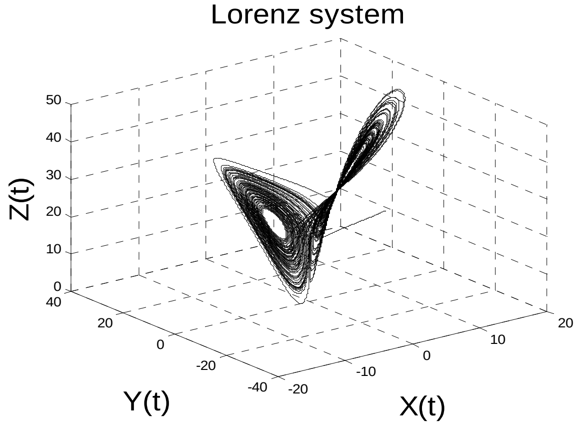


Figure 1. The 3D figure of Lorenz chaotic system

3 A SPEECH MASKING TECHNIQUE BASED ON LORENZ SYSTEM

The block diagram of the designed chaotic masking scheme is shown in Figure 2. The speech signal S_n is added to the Lorenz chaotic generator signalX_m which also acts as a driving signal for synchronization as will be explained later (Pecora-Carroll Synchronization). The speech signal is precisely recovered at the receiver by the subtraction of the receiver's regenerated drive signal from the received signal [1, 2].

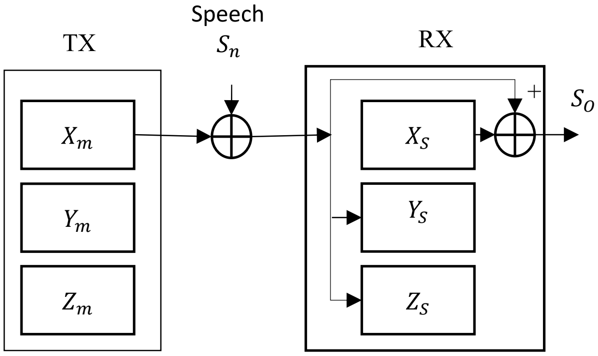


Figure 2. Chaotic masking and recovery information based on Lorenz system.

Here, we implement the master subsystem using equations (4, 5, and 6) related to Lorenz equations (1, 2, and 3). We note that S_n is the input Speech sample.

$$\dot{X}_m = F(X_M, Y_M, Z_M) + S_n \quad (4)$$
$$\dot{Y}_m = G(X_M, Y_M, Z_M) \quad (5)$$
$$\dot{Z}_m = W(X_M, Y_M, Z_M) \quad (6)$$

The slave subsystem uses Lorenz equations (7, 8, 9, and 10) to decrypt the encrypted signal.

$$\dot{X}_s = F(X_s, Y_s, Z_s) \quad (7)$$
$$\dot{Y}_s = G(X_M, Y_s, Z_s) \quad (8)$$
$$\dot{Z}_s = W(X_M, Y_s, Z_s) \quad (9)$$
$$S_O = X_M - \dot{X}_s \quad (10)$$

4 THE PROPOSED CRYPTOSYSTEM

The proposed cryptosystem is shown in Figure 3. The samples of Speech signal S_n are injected into the chaotic generator (master system) which also acts as a driving signal for synchronization. The Speech signal is precisely recovered at the receiver side (slave system) by the subtraction of the receiver's regenerated drive signal from the received signal [4]. We implement our proposed system using Lorenz system.

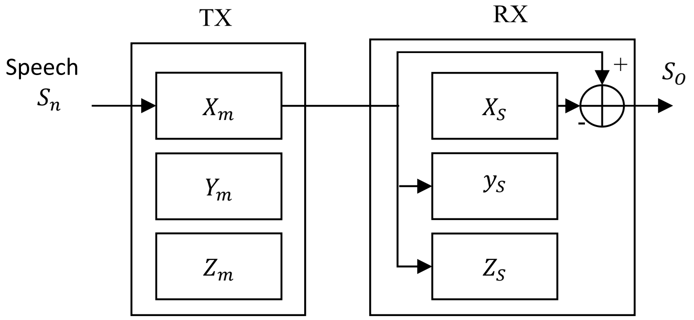


Figure 3. The proposed Cryptosystem

Here, the master subsystem is implemented using equations (11, 12, and 13) related to Lorenz equations (1, 2, and 3). It is to be noted that $S_n$is the input Speech sample, it is clear that S_n is now a parameter of function F.

$$\dot{X}_m = F((X_M + S_n), Y_M, Z_M) \quad (11)$$
$$\dot{Y}_m = G(X_M, Y_M, Z_M) \quad (12)$$
$$\dot{Z}_m = W(X_M, Y_M, Z_M) \quad (13)$$

The slave subsystem uses Lorenz equations (14, 15, 16, and 17) to decrypt the encrypted signal.

$$\dot{X}_s = F(X_s, Y_s, Z_s) \quad (14)$$
$$\dot{Y}_s = G(X_M, Y_s, Z_s) \quad (15)$$
$$\dot{Z}_s = W(X_M, Y_s, Z_s) \quad (16)$$
$$S_O = X_M - \dot{X}_s \quad (17)$$

- *Pecora-Carroll (PC) Synchronization*

In order to receive the Speech signal sample successfully, chaotic signals on both Transmitter (Master) and Receiver (Slave) must be synchronized, one of the efficient synchronization schemes that can be used is Pecora-Carroll (PC) Synchronization [5]. In this scheme, a driving signal is sent from the chaotic generator at the transmitter, to the chaotic generator at the receiver. At the receiver, state error vectors which describe the difference between the encryption and decryption state variables are constructed (equations 18, 19, 20). Figure 4 shows the block diagram of the mechanism of PC synchronization of Lorenz map.

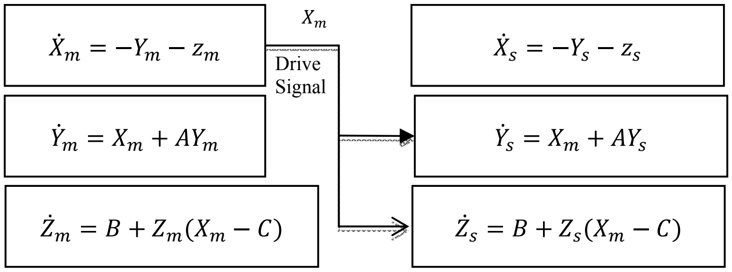


Figure 4. PC synchronization of Lorenz system

State error vectors (synchronization error) which describe the difference between the master and slave state variables are constructed. Equations (24, 25 and 26) show that the variables e_x, e_y, and e_z represent the synchronization error of $X,\ Y$, and Z, respectively, in our proposed system using Lorenz equations. Figure 5 shows this synchronization error.

$$e_x = X_M - X_s \quad (18)$$
$$e_y = Y_M - Y_s \quad (19)$$
$$e_z = Z_M - Z_s \quad (20)$$

It has been shown that with the aid of the driving signal these states errors can be reduced to zero after a certain amount of time as shown in Figure 5.

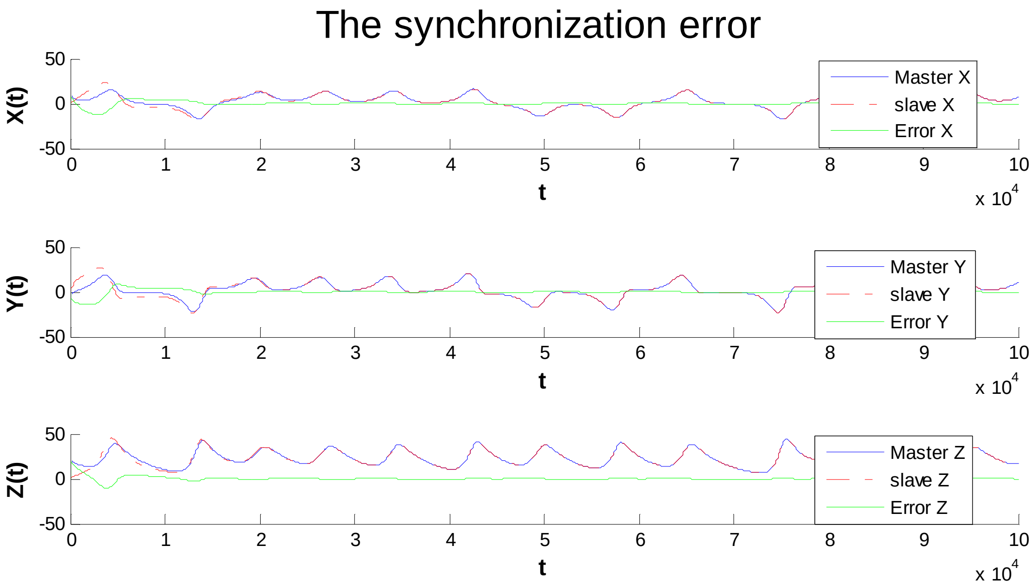
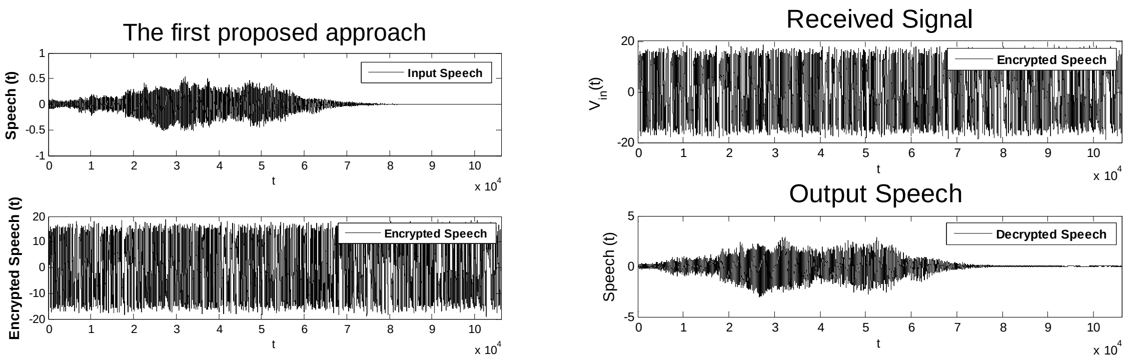


Figure 5. The synchronization error of $(X, YandZ)$ in the proposed system using Lorenz system.

5 Experimental Results And Analysis

In section 4, we presented our proposed system, which is implemented using Lorenz system, and the Speech samples, which are embedded to the chaotic signal to generate a new chaotic signal. The Speech signal is then extracted from the chaotic signal at the receiver side. Figure 6(a) shows the waveform of the original signal and the waveform of the encrypted signal for the proposed approach. Figure 6(b) shows the waveform of the received signal and the waveform of the decrypted signal for the proposed approach. Figure 7 shows the autocorrelation of the proposed approach transmitted signal. The autocorrelation function used to measure randomness, an ideal random sequence should be uncorrelated.



(A) **(B)**

Figure 7. First proposed approach (encrypted signal - decrypted signal)

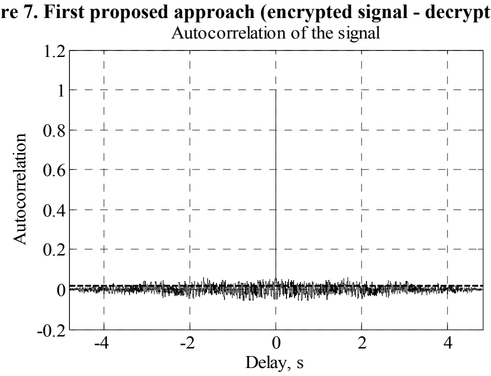


Figure 7. Proposed approach (autocorrelation of the transmitted signal)

Figure 8 shows that, unlike masking $X(t)$using Lorenz map to the original signal, which makes slight change to the original signal, embedding voice samples to $X(t)$using Lorenz map (direct modulation) makes significant changes to the original signal.

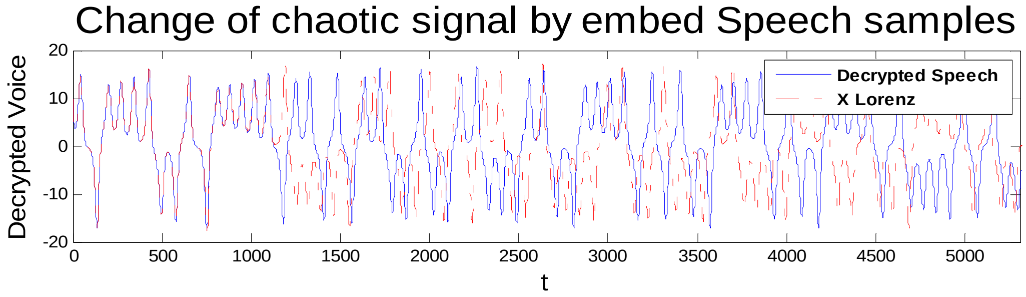


Figure 8. The effect of injecting Speech samples to Lorenz system

Figure 9 shows the effect of masking Speech samples to $X(t)$of Lorenz map.

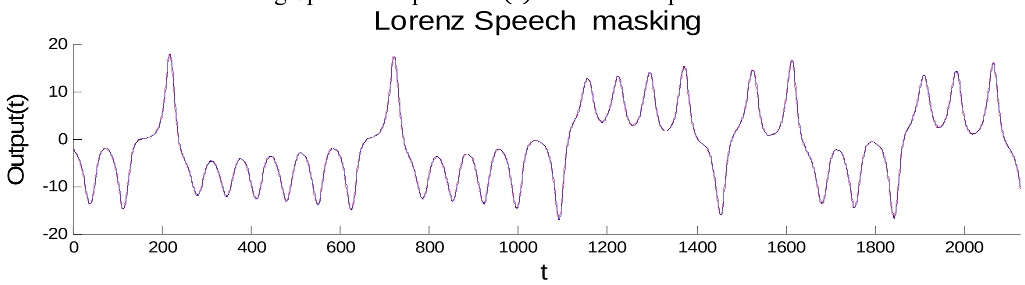


Figure 9. The effect of masking Speech samples to Lorenz system

In the following subsections, we analyze the results of applying the proposed approach according to different perspectives. In section (A) we present a comparative study between our proposed Non-autonomous approach and its chaotic masking counterpart because the results are not similar. In sections (B) and (C) there is no significant difference between the results of our proposed Non-autonomous approach and its chaotic masking counterpart, therefore we limit our self to show the results of our proposed Non-autonomous approach.

A. Statistical Analyses

To statistically analyze our results, four different measures [6] are used, Signal-to-Noise-Ratio (SNR), Segmental signal-to-Noise-Ratio (SNRseg), Log-Likelihood Ratio (LLR), and Correlation Coefficient Analysis (CCA). Tables 1, and 2 show the average result of these measures and chaotic masking based on Lorenz system.

TABLE 1

STATISTICAL ANALYSES OF SNR, SNRSEG, LLR, AND CCA FOR ENCRYPTED SIGNAL

Approach	**SNR**	**SNRseg**	**LLR**	**CCA**
Proposed approach	-38.55 dB	-38.91 dB	0.89	0.0345
Masking using Lorenz map	-35.51 dB	-35.84 dB	0.80	0.012

TABLE 2

STATISTICAL ANALYSES OF SNR, SNRSEG, LLR, AND CCA FOR DECRYPTED SIGNAL

App	**SNR**	**SNRseg**	**LLR**	**CCA**
Proposed approach	5.01dB	4.99 dB	0.213	0.82
Masking using Lorenz map	2.75 dB	2.50 dB	0.1	0.8519

B. Spectrogram Analyses

A spectrogram is a powerful tool that divides the Speech sample into multiple "blocks" (in the time domain) then plotting the Fast Fourier Transform (FFT) of each block and displaying all of them in the same graph [7, 8]. Figure 10 shows the spectrogram of the original signal frequency versus time and the spectrogram of the encrypted signal frequency versus time (proposed approach).

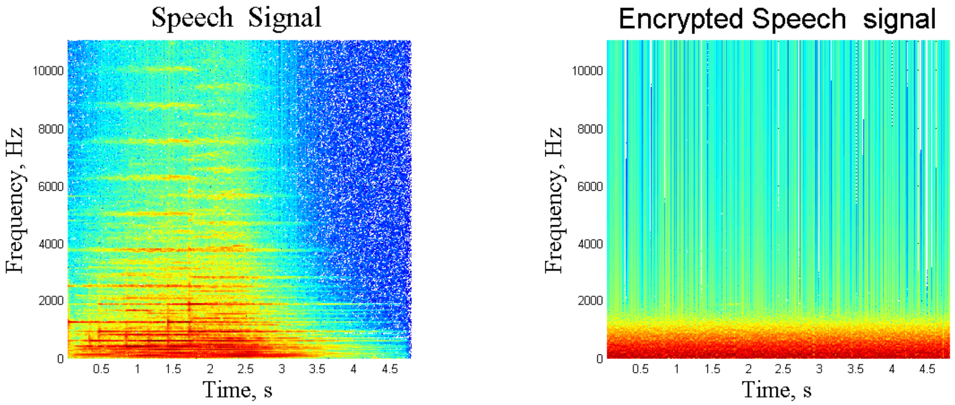


Figure 10. Proposed approach spectrogram (Original signal - encrypted signal)

C. Histogram analysis

Distributions of data values in a system comprise the histogram. Histogram analysis can be made by examining data distributions in many different fields [9]. In encryption practices, if the distributions of numbers that represent encrypted data are close, this means that encryption is performing well. The closer the encrypted data distributions are, the higher their encryption levels. Figure 11 shows the distribution versus sample value (for proposed approach).

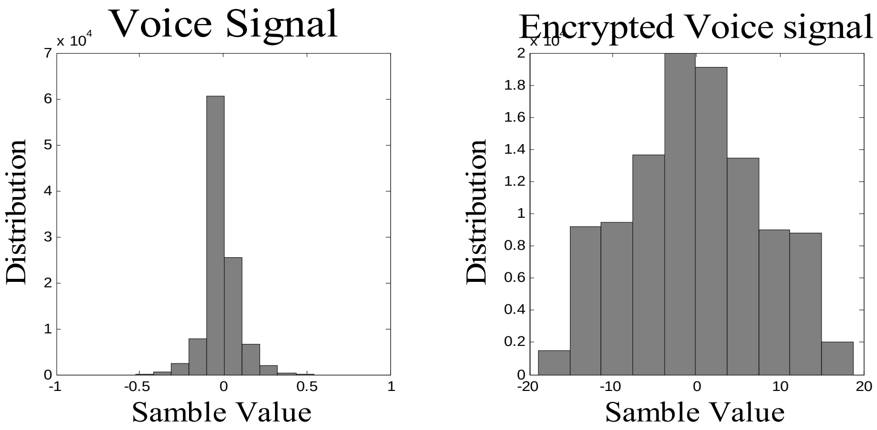


Figure 11. First proposed approach Histogram (Original signal - encrypted signal).

D. Key Sensitivity and Key Space

Key sensitivity analysis is the most important criteria of the performance analysis of the encryption system. A good encryption algorithm should be sensitive to the initial condition and key value. Lyapunov exponent (LE) [4] can be used to evaluate the chaotic system sensitivity to the initial condition. The larger value of Lyapunov exponent value the chaotic system has the more sensitivity of this system to the initial condition. Figure 12 shows dynamics of Lyapunov of Lorenz system.

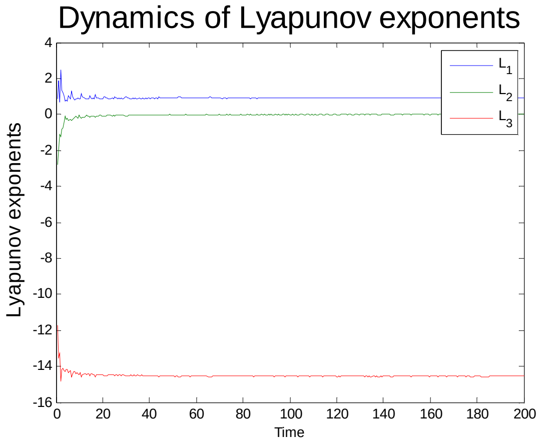


Figure 12. Dynamics of Lyapunov exponents of Lorenz map.

It is well demonstrated that the Lorenz system has two positive Lyapunov exponents and a small negative Lyapunov exponent [10], that is, the leading positive LE l_1 of the Lorenz system is equal to 0.9051 and the second positive LE l_2 has a value of $8.12*10^{-5}$, the negative Lyapunov exponent positive LE l_3 of the Lorenz system equals to -14.5718. The Lyapunov exponents provide a good indication of how chaotic the Lorenz systems are. Hence, this explains why the system is very sensitive to initial conditions and more unpredictable than other systems. In our approaches a small change in parameters leads to different results during the decryption, the data cannot be decrypted without knowing all parameters because the decryption does not happen in the correct order. The size of the key space defines the total number of different keys that are used for the encryption / decryption algorithm. It should be large enough to resist the attack. The key space depends on the initial conditions and the control parameters of chaotic map. In Lorenz map, we use three initial conditions and three control parameters.

6 CONCLUSIONS

We proposed a new chaotic-based crypto system; this system depends on the change of chaotic signal by injecting Speech samples into one variable of the master system to generate a new chaotic signal. The dynamics and decoding of this new chaotic signal is very hard to be estimated, and hence, the proposed system overcomes the disadvantages of Chaotic Masking and parameter modulation techniques. Non-autonomous modulation approaches were used to implement the proposed system using Lorenz map. Although Non-autonomous approaches give similar results of spectrogram, histogram, key sensitivity and key space analysis compared with their chaotic masking counterpart, experimental results show that Non-autonomous approaches give better performance than their chaotic masking counterpart when they analyzed against Signal-to-Noise-Ratio, Segmental signal-to-Noise-Ratio, Log-Likelihood Ratio, and Correlation Coefficient Analysis. The proposed Non-autonomous modulation approach is sensitive to the initial conditions and control parameters, which means it is difficult to decrypt the encrypted signal correctly if there is a very small change between encryption and decryption keys.

REFERENCES

[1] Rahul Ekhande, Sanjay Deshmukh, "*Chaotic Signal for Signal Masking in Digital Communications*", IOSR Journal of Engineering ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 04, Issue 02, V5. PP. 29-33 (February 2014).

[2] Hikmat N, Saad S.,"*Design of Efficient Noise Reduction Scheme for Secure Speech Masked by Chaotic Signals*", Journal of American Science 2015. PP.49-55 (Nov. 2015).

[3] KEVIN M. SHORT, "*Steps toward unmasking secure communications*", International Journal of Bifurcation and Chaos, vol. 4, pp. 959-977, (1994).

[4] M. Haroun, T. A. Gulliver, "*Real-Time Image Encryption Using a 3D Discrete Dual Chaotic Cipher*", International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering Vol.: 9, No: 3, (2015).

[5] Rahul Ek., Sanjay De."*Chaotic Synchronization in Digital Communication*", International Journal of Engineering Research, Volume 3, Issue No.7, pp. 458-461 (July 2014).

[6] Hala B., Sundus I. Mahdi, "*Modify Speech Cryptosystem Based on Shuffling Overlapping Blocks Technique*", International Journal of Emerging Trends & Technology in Computer Science Volume 4, Issue 2, (2015).

[7] E. Mosa, O. Zahran, "*Chaotic encryption of speech signals*", International Journal of Speech Technology, Volume 14, Issue 4, pp. 285-296 (December 2011).

[8] M. Ashtiyani, P. Moradi Birgani, S. Karimi Madahi, "*Speech Signal Encryption Using Chaotic Symmetric Cryptography*", Journal of Basic and Applied Scientific Research, Vol. 2, No. 2, pp. 1678-1684, (2012).

[9] Osama Faragallah, Elsayed Elshamy, Sayed El-Rabaie, "*Speech Encryption Based on Arnold Cat Cap and Double Random Phase Encoding*", International Journal of Speech Technology. No.14, pp. 14-24, (2013).

[10] Dennis Luke, Guoyuan, "*Message Signal Encryption Based on Qi Hyper-Chaos System*", e-Technologies and Networks for Development Volume 171 of the series Communications in Computer and Information Science pp. 145-155(2011).

BIOGRAPHY

Mahmoud Fawzy, BSc. in Electrical Engineering, Alexandria University. Currently works with the Air Force Defence. Fields of interest: electrical engineering and communication systems.

نظام تأمين للكلام مبنى على تقنية التعديل الفوضوى المباشر

محمود فوزى*، محمد شلبى**، ياسر كمال**، سلوى الرملى*

**كلية الهندسة، جامعة عين شمس*
***كلية الهندسة و العلوم والتكنولوجيا و النقل البحرى*

ملخص

نقدم تقنية لتشفيرنظم التخاطب عبر وسائل الإتصال مبنيا على التعديل الفوضوى المباشر حيث يتم تضمين أو حقن إشارات المحادثة الصوتية داخل متغير واحد من النظام الفوضوى الرئيسي (بإستخدام نظام لورينز) دون تغيير قيم الخصائص المسيطرة على النظام الفوضوى. ويستند هذا النهج على تغيير الإشارة المولدة من النظام الفوضوى عن طريق حقن الإشارات لمتغير واحد في نظام الفوضوى وينتج عن ذالك توليد إشارة فوضويه جديدة. ثم يتم إستخراج إشارات المحادثة الصوتية من الإشارة الفوضوية بواسطة النظام الفوضوى المستقبل. وعلاوة على ذلك، يتم إستخدام نظام الفوضوى متعدد الابعاد ليزيد من تأمين الإشارة المشفرة. تقنية التعديل الفوضوى المباشره مناسبة لتأمين التطبيقات في الوقت الحقيقي. ونقدم أيضا دراسة مقارنة بين هذه التقنية و تقنية الإخفاء الفوضوى لإشارات التخاطب. وقد أظهرت النتائج أن التعديل الفوضوى المباشر يزيد من أمن نظام التشفير.

A Proposed Arabic Text to Sign Language Translator

A. S. Elons*[1], A.Ali[2], M. F. Tolba*[3]

* *Scientific Computing Department- Faculty of Computers and Information Sciences- Ain Shams University-Cairo-Egypt*

[1]ahmed.new80@hotmail.com
[2]ahmed4a@hotmail.com
[3]fahmytolba@gmail.com

***Abstract*- This research aims to improve accessibility for deaf persons to understand words those written in Arabic language. The objective is designing and developing a 3D avatar based intelligent tool (Text-to-Sign Translator) that generates signs and play them. Automated sign to text and vice versa translation systems are vital in a world that shows a continuously increasing orientation toward removing barriers faced by physically challenged. These translation systems can impact the community communication between the hearing normal and the hearing impaired (HI) individuals.**

Key words: ***Arabic Sign Language (ArSL),Arabic Sign Language Database, Hearing Impaired (HI) .***

1 Introduction

The term 'deaf and dumb' is unfavorable to be used since HI persons suffer in their hearing abilities not their mental level Schwartz [1]. For many Hearing Impaired (HI) persons, sign language is the main media for communication. The main issue is that few number of normal hearing persons ever learn to sign. Another problem is that many HI people are not able to understand a spoken language. These communication issues increase the barrier between HI people and community. Despite of the public common sense, sign Language is not a unified universal language. Where people speak a different phonetic language, there is also a different Sign Language. Besides the locality nature issue of sign languages, Arabic Sign Language (ArSL) is not unified for all Arabian countries. ArSL differs in each Arab region or/and country with many dialects, this difference causes the difficulty of communicating among HI persons themselves in different Arabian countries. A need appeared to unify Arabic sign language in all Arabian countries. Lately, standard ArSL dictionary is accredited and published to Arabian HI community [2, 3]. (CAMSA) Council of Arab Ministers of Social Affairs made a decision of developing a unified Arab sign language dictionary and publish it to all countries, in an attempt to help [4]. This dictionary is mostly used in education and in common communication such as sign language interpreters in TV and media. In Egypt, the estimated number of HI citizens according to the last study done by "Central Agency for Public Mobilization and Statistics" in 2015 around 4.5 millions. Breaking the barriers between HI individuals and Normal-Hearing individuals can greatly reduce the problems of frustration, hate and ignorance for HI persons; it can permit better education, health, etc… for them.

ArSL like other sign languages relies on three main factors these are used to represent the manual features: hand shape, hand location and orientation. In addition to the non-manual features which are related to head, face, eyes, eyebrows, shoulders and facial expression like puffed checks and mouth pattern movements. ArSL is limited to represent nouns, adjectives and verbs, while prepositions and adverbs are represented in the context of articulation by specifying locations, orientations and movement. Signs forming and sequencing in the articulation, are done depending on the Arabic sign language grammar and rules. The goal of this work is to build cartoon 3D avatar-based, real-time, efficient, fully translation system, which translates input Arabic text to the visual Arabic Sign Language. Our review has concluded that very few research attempts world-wide succeeded to develop practical and public efficient products. None of these attempts succeeded to develop a practical commercial product to ArSL used by public. In this paper, section 2 shows related work and section 3 shows the state-of-art. Section 4 proposes the developed system and section 5 draws the experimental results and discusses the future work.

2 RELATED WORK

It has been reported in 2003 that 80% of deaf people lack education or are undereducated, are illiterate or semi-literate, the World Federation of Deaf confirmed [5]. Moreover, the average number of HI high school graduate is incapable to exceed the 4th grade level. Through the use of artificial intelligence ,researchers are trying to develop hardware and software solutions that will affect positively the way deaf persons communicate and learn. With technology waves strikes, few research projects tried to develop a bi-way translation system from and to sign language, few of these addressed ArSL. Furthermore the translational system which translates the written or spoken language to sign language mainly depend on recording videos for the sign and be saved in a video database, it just retrieves the saved gesture which corresponds to the input word. The translation system can be considered as a dictionary-retrieval for the text without any semantic post processing (meaningful translation). The main challenge in sign to text and vice versa translation is to develop a computational application that can be released over internet and that combines two important features: efficiency and ease of use.

- Efficiency mainly addresses the quick response time with low bandwidth connection.
- Ease of use concerns with the fullness of user interaction.

Toward building an automated text to sign translators, different approaches were exploited. There are 3 main approaches, researchers addressed:

- Writing and Drawing symbols based approach.
- Recorded Videos based Approach.
- A 3D Avatar Animation based approach.

Table 1 illustrates a comparison among the 3approaches.

TABLE 1
THE MAIN APPROACHES

	Writing and Drawing Approach	**Recorded Video Approach**	**3D Avatar Approach**
Explanation	- drawing is the first transcription of signs. -Examples of these transcription systems appeared such as HamNoSys [6] and SignWriting [7] although it is very hard to encode sign Language, Fig 1.	-The video based systems consist in constructing a video sequence of frames.	-Smooth transitions between signs is very hard to accomplish. -Research into synthesizing is still immature. -Existing systems employ avatars to synthesize sign language in real time.
Historical Background	-In 1984, The first version of HamNoSys was Released [6]. -In 1974, SignWriting is proposed by Valerie Sutton for the Center of Sutton Movement Writing [8]. - SignWriting system is now used as a handwriting edition of sign language and taught to HI students in the world [9]	The Personal Communicator [10], LSF Lexiquel	Grieve-Smith, [11]; Krňoul et al. [12]). Using a tool called eSIGNeditor, Kennaway et al. [13] developed during the eSIGN project.
Disadvantages	-It's not an efficient approach such that it's not a fully automated approach. -Needs another layer for signs synthesizing	-Production of video with high sign quality is resource intensive. -changing, deleting or insertion of frames is very difficult. - Errors on sign recording require re-recording the scene. -2D video cause that the 3rd dimension information can be lost. -High resolution requires high data storage space. - changing the signer requires re-recording the signs.	-Requires a pre-developed Annotation module (transcription). -Visualizing sign language in 3D animations is very sophisticated process
Advantages	Extendible	-Easy to develop. -Very human like appearance.	-Very efficient retrieval. -Transitions between signs are manageable. -Extendible and adaptable.
Comments	-It can be considered as an intermediate transcription layer between sign language and spoken ones rather than an approach.	-It's very suitable for small dataset and word by word translation	-It’s very suitable for E-documents and sentences translation.

The prior research trials in translating Arabic text to ArSL are very rare and disorganized, most of these trials worked only on translating in word-based mode and did not consider the semantics of the translated sentence .To tackle this problem, we aim in the proposed system to extend the previous research in this field by appending a higher layer of semantics during translating Arabic text to ArSL.

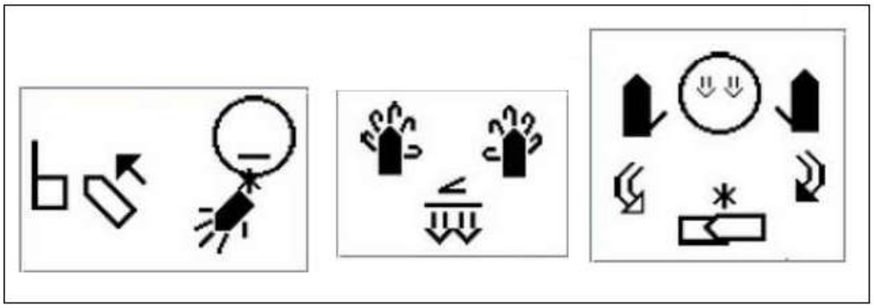


Figure 1: A sample of SignWriting representation

3 STATE OF ART

New trends for this challenge include:

- Sign Language Synthesizing and Semantic Web.
- Sign Language Synthesizing and On-Air TV.
- Sign Language Synthesizing and Mobile.

3.1 Sign Language Synthesizing and Semantic Web.

Previous research trials these exploit semantic web technologies to improve the accuracy of text-to-SL translation are very limited. ATLAS [14] is a research project for automated translation from Italian text to Italian Sign Language. The translation system communicates with the HI user via 3D avatar cartoon signer: the system input is a text written in Italian language and translates it into an intermediate designated representation of a sign language sentence called ATLAS Written Italian Sign Language (AWLIS). ATLAS [15] Linguistic analysis is composed of three main steps:
1) Italian source sentence Syntactic analysis
2) Semantic interpretation
3) target LIS sentence Generation.
The project succeeded to translate only hand gestures but ignored facial expressions and lips movements. In (ArSL), Ameera et al[16] customized Ontology to translate Arabic text to SignWriting ArSL. They developed a translation system for a specified field: Religion words.

3.2 Sign Language Synthesizing on Air TV

Sign language translation is not available on almost every current transmitted content on television. However, the great jump from analogical to digital transmitted television can offer some extra features to traditional television. These additional features can provide a new level of content's accessibility for HI individuals, Fig 2.The importance and viability of sign language on the digital television system was studied on its reference model [17], and achieved impact level of 2.5 and relevance of 3 points, in a convergence scenario. Digital television systems permit the usage of digital applications, turning the TV into a more efficient and usable communication device. With digital applications, usability issues related to the sign language interpreter window can be improved.

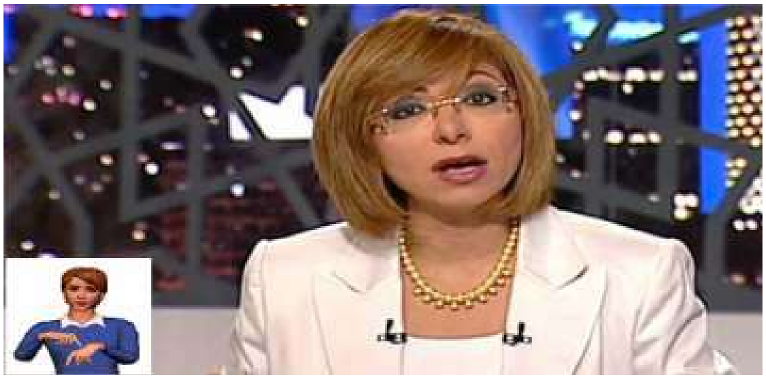


Figure 2: A sample of TV and sign language synthesizing

3.3 Sign Language Synthesizing and Mobile

Most of sign language applications on mobile are developed to establish a communication between two HI persons. University of Washington and Cornell University worked together in a project called MobileASL[18] which is a video compression project aims to make wireless cell phone communication through sign language a reality in the U.S, Fig 3.

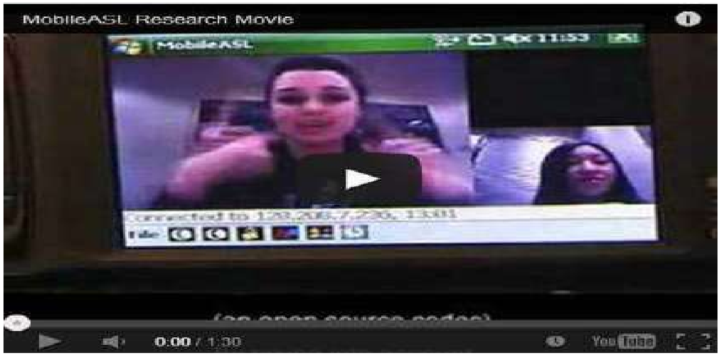


Figure 3: A screenshot for MobileASL application

One famous application in Play Store is Mimix[19], it uses a technology that translates spoken and written words into sign language with a 3D character. It uses ASL (American Sign Language)based Dictionary as signed English, Fig 4. Although the importance of mobile-based Sign languages applications, they are very limited due to storage and response time issues.

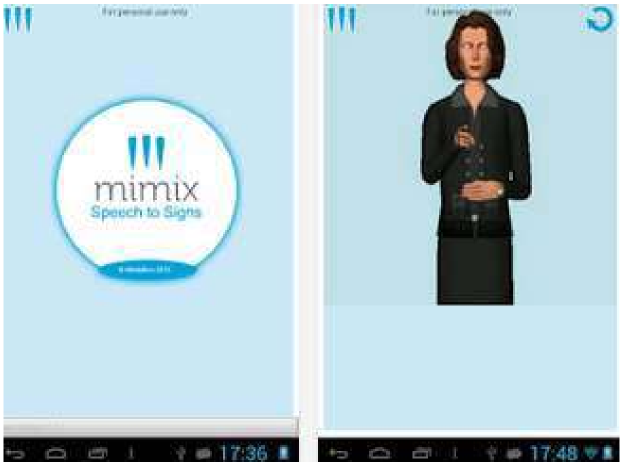


Figure 4: A screenshot for Mimix application

4 PROPOSED MODEL

The proposed prototype enables an input text to be converted to ArSL animating. The system is composed of sub-modules, some of them already existand need to be customized (NLP Processing Module) and others will be developed from scratch (ArSL Transcription Module and ArSL Synthesizer module). Figure 5 illustrates the proposed system architecture.

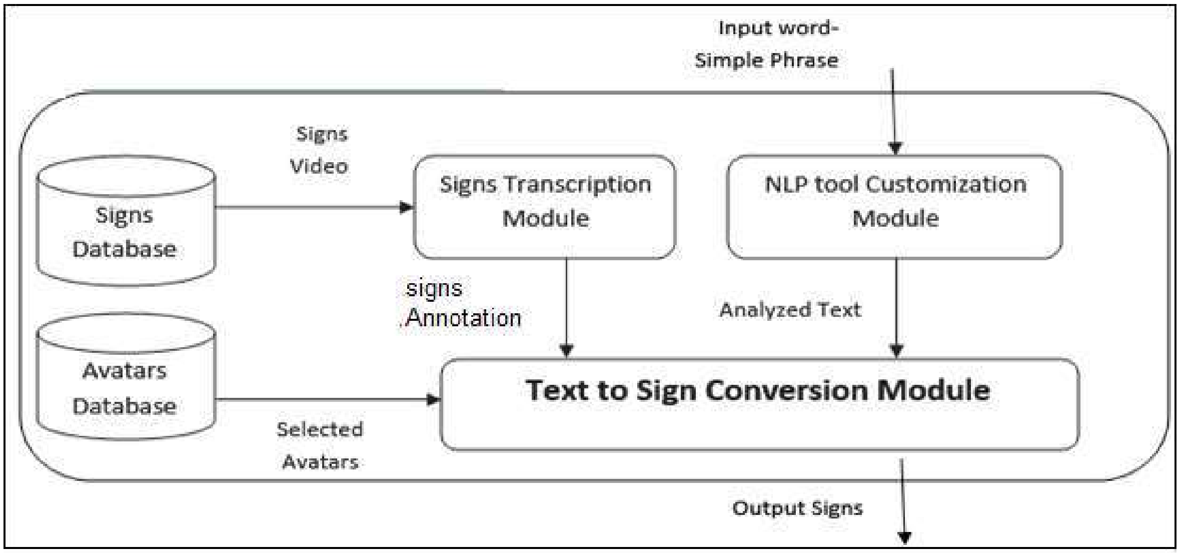


Figure 5: Proposed system architecture

The proposed system sub-modules each are:

- **NLP Processing Module**

This includes a customization for ready-made tool that is enabling of analyzing the input Arabic text, dividing the input phrase and performing lexical, syntactical and semantic analysis.

- **ArSL Transcription Module**

In ArSL, a sign can be mapped to a corresponding word and also to entire concepts and complex phrases .Finger spelling is used to spell out proper names and technical terms. It's hard to ignore the fact that most signs can be viewed as a frame sequence of hand shapes, location and movement. The transcription module is responsible of generating a computer standard description of a sign, describing the following:

- Hand shape.
- Facial Expressions.
- Body Movements.
- Eyes Movements.
- Lips movements.

- **Sign Synthesizer module**

This module employs a chosen cartoon avatar to perform the signs description generated by sign transcription module. The challenge in building this module is, making the transition between signs is smooth and real like appearance.

4.1 Proposed Sign Language Transcription

One of the noticeable differences between spoken and sign languages is the absence of a formally adopted writing system for sign languages. There have been some prior trials to originate a writing systems for sign languages, many of these are based on the seminal work of Stokoe [20] and describe the hand shape, location and articulated movement of a sign. These systems include the Hamburg Notation System HamNoSys, Hanke [21] and SignWriting, Sutton [22]. Despite the development of these approaches, they currently fall short of being computationally efficient for Sign Language Translation. Based on this, we propose customizing the "Annotation system" proposed by Pizzuto [23]. It enjoys both required properties: it is computationally tractable and it is sufficient to represent all signs necessary details. A study done by the project team shows that relatively small set of 158 hand shapes (one hand and two hands) generates the majority of signs in ArSL.Figure6, illustrates the signs transcriber module which include sub-transcription modules for hand shape, facial expressions, eyes and body movement.

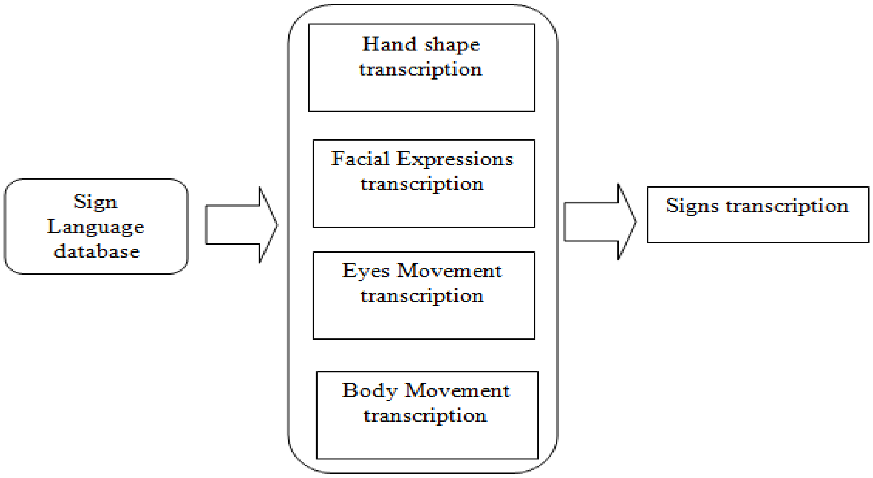

Sign Language database
Hand shape transcription
Facial Expressions transcription
Eyes Movement transcription
Body Movement transcription
Signs transcription

Figure 6: Proposed Sign Transcription Module

The hand shape transcriber is responsible for building the hand shape data which represents most of the geometric information contained in signs. This data is stored in the hand shape database for use by the sign transcriber. This approach is an appropriate one since it permits users of the system to generate a hand shape once and reuse it in different signs making the sign construction process simpler. To create the hand shapes, the transcriber employs a geometric model of the human hand.

4.2 Proposed Synthesizing Module

Recently, interest in sign language automated interpretation applications has increased substantially with increasing rate. Such computer applications include sign language dictionaries, teaching tools and various machine translation systems from text to sign language and vice versa. All these applications need a signing avatar that can perform sign language for HI persons. There are main differences between normal avatars and signing avatars; normal avatars such as those used in movie and gaming industries, usually require gross motor movement and computationally expensive collision avoidance with the environment. While signing avatars require extremely fine motor movements and minimal collision avoidance modules. Moreover, a signing avatar is required to perform “non-manual” signs which comprises facial movements, such as lifting of the eyebrows or a slight puffing of the cheeks. These non-manual signs are linguistically meaningful and essential components of any sign language. All signing avatars must tackle three distinct issues:

- The avatar appearance.
- The signs animations.
- The notational interface between the sign description in the dictionary and the commands for the animation of the sign.

Realistic human avatars are time-consuming to construct, and as in movies and games, even the best avatars appear somewhat ‘plastic’ to observers. In addition, users who use signing avatars in assistive technology tools for learning sign language, often have difficulty to pick up subtle facial expressions in human-like avatars. In order to build a generic signing avatar, we had to design a system that would work on any ‘reasonable’ avatar definition. Usually, the transitions between the signs should be smooth, a simple and straight forward approach for the smoothness is to have the beginning and ending of every sign performed in a constant posture. Although this solution offers smooth continuous transitions, starting and finishing each sign in the same posture is very unnatural. Another computationally expensive approach is to define transitions between every single pair of possible motions. This solution is very naïve and not adaptable for any modifications or additions for new signs [24]. This approach succeeds in avoiding returning to a required `neutral' pose, but it does not necessarily guarantee natural and rational transitions. In this work, PaT-Nets (Parallel Transition Networks) [25-27] is proposed to solve the motion blending problems. A PaT-Net is a simultaneously executing finite state automaton (FSA) in which the nodes are associated with actions and connections between the nodes are associated with transition conditions.

5 RESULTS AND DISCUSSION

this work has been applied for 20 different Arabic words and 10 different sentences. The remarkable point is the used time to generate the hand transcription files decreases with each new sign since the frames start to be repeated. The system gives the facility to control the avatar speed via considering and ignoring some transcription frames based on frame rate. This work can be extended in 3 different ways:

- Enabling the user input through voice instead of text, this can be done by adding a speech recognition layer before signs Synthesizing.
- Generating transcription files using a special human movements sensors such as Microsoft Kinect, Leap motion, etc… these can capture human body (including hands).
- Extending the signs database to cover the ArSL dictionary.

6 CONCLUSIONS

The In this paper, we described the Arabic text to sign language translator which meets the requirements for an Arabic sign language users. A main objective of this project is to distribute this tool to wide range of Hearing Impaired users (i.e. instructors, teachers, students and researchers) though encourage its wide use by different communities. The proposed system emphasizes the importance of 3D avatar approach for displaying the signs. The proposed system contains a complete transcription component for signs definition in animation files describe each sign components in each movement frame. This system could be a step toward a complete multi-lingual translation system among different sign languages.

BIOGRAPHY

Prof. Dr. Mohamed Fahmy Tolba is a Professor of Scientific Computing, FCSIS (1996-Present). Dr. Tolba has more than 150 publications in the fields of AI, Image Processing, Pattern Recognition, OCR, Scientific Computing, Simulation and Modeling. Also Dr. Tolba has supervised more than 50 M.Sc. and 25 Ph.D. degrees in Ain Shams University and other Egyptian Universities.

Dr. Ahmed Samir is a Lecturer at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. His research interests: Image Processing and, pattern recognition and AI. He worked in Arabic Sign Language Recognition field from 2004 till now.**(the photo is not clear}**

Dr. Ahmed Ali is a Lecturer at Scientific computing department, Ain Shams University, Cairo, Egypt His research interests: in graphics rendering and real time systems, high performance computing, gamification, modeling and simulation, neural network, and remote sensing.

REFERENCES

[1] Schwartz, B. L. "Sources of information in meta memory Judgments of learning and feelings of knowing" .Psychonomic Bulletin and Review, 1, 357 – 375, 1994.

[2] M.F. Tolba, A.S. Elons ―Recent developments in sign language recognition systems‖ In proceeding of: Computer Engineering & Systems (ICCES), 8th International Conference on Computer Engineering and Systems. 26-28 Nov. 2013. IEEE Publisher.

[3] Feras, F., Eman, and Mohamed. O, "Automatic isolated-word Arabic sign language recognition system based on time delay neural networks: New improvements", Journal of Theoretical and Applied Information Technology.Vol. 57 No.1, 2013.

[4] A. Samir Elons,M. Aboul-Ela and M.F Tolba." 3D Object Recognition Using Multiple 2D Views for Arabic Sign Language" has been published in the "Journal of Experimental & Theoretical Artificial Intelligence" Volume 25, Issue 1, 2013.

[5] World Federation of the Deaf (WFD), Position Paper regarding the United Nations Convention on the Rights of People with Disabilities, Ad Hoc Committee on a Comprehensive and Integral International Convention on the Protection and Promotion of the Rights and Dignity of Persons with Disabilities , 24 June 2003.

[6] Hanke, T.HamNoSys—representing sign language data in language resources and language processing contexts. Paper presented at the Workshop on the Representation and Processing of Sign Languages on the occasion of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.

[7]Gleaves, R., & Sutton, V. SignWriter. Paper presented at the Workshop on the Representation and Processing of Sign Languages on the occasion of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.

[8] OfficialSignWritingsite,http://www.signwriting.org/, last seen Sep-2014.

[9] Guido Gybels, Deaf and Hard of Hearing Users and Web Accessibility, white paper of The Royal National Institute for Deaf People, October 2004.

[10] http://commtechlab.msu.edu/. Last seen May-2016

[11] Grieve-Smith, A.: English to American Sign Language Machine Translation of Weather Reports. In: Nordquist, D . (ed): Proceedings of the Second High Desert Student Conference in Linguistics. High Desert Linguistics Society, Albuquerque, 1999.

[12] Krňoul, Z., Kanis, J.,Železný, M. and Müller, L. Czech text-to-sign speech synthesizer. In A . Popescu-Belis, S. Renals and H. Bourlarded)Machine Learning for Multimodal Interaction), 4thInternationalWorkshop,Berlin,Springer, pp. 180–191, 2008.

[13] Kennaway, J.R., Glauert, J.R.W. and Zwitserlood,Providing signed content on the Internet by synthesized animation. ACM Transactions on Computer-Human Interaction 14, pp. 1–29, 2007.

[14] L. Lesmo, A. Mazzei and D. Radicioni, ―Linguistic Processing in the Atlas Project,‖ SLTAT 2011, Berlin, 2011.

[15] L. Lesmo, A. Mazzei and D. P. Radicioni, ―Linguistic Descriptions in Ontology-Based Machine Translation,‖ In: B. Kokinov, A. Karmiloff-Smith and N. J. Nersessian, Eds., European Perspectives on Cognitive Science, 2011.

[16] Ameera. Almasoud and H. Al-Khalifa, "SemSignWriting: A Proposed Semantic System for Arabic Text-to-SignWriting Translation," Journal of Software Engineering and Applications, Vol. 5 No. 8, pp. 604-612. doi:10.4236/jsea.2012.58069, 2012.

[17] R.F. de Brito and A.T.C. Pereira, ―A model to support sign language content development for digital television, IEEE International Workshop on Multimedia Signal Processing, Rio De Janeiro, Brazil pp. 1-6, 2009.

[18] Anna Cavender , Richard E. Ladner , Eve A. Riskin, MobileASL:: intelligibility of sign language video as constrained by mobile phone technology, Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility, Portland, Oregon, USA doi: 10.1145/1168987.1169001, 2006.

[19] http://speechtosign.com/ last visited: Sep-2015.

[20] Stokoe, W. Sign language structure: An outline of the visual communication systems of the American Deaf. Studies in linguistics, occasional papers 8. Silver Spring, MD:Linstok Press, 1960.

[21] H., Hanke, Prillwitz, S., Leven, R., Zienert, T. andHenning, J. HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide. Hamburg: Signum, 1989.

[22] Sutton-Spence, R. &Woll, B. The linguistics of British Sign Language: An introduction. Cambridge, England: Cambridge University Press. 1999.

[23] E. Pizzuto, P. Pietrandrea. The Notation of Signed Texts: Open Questions and Indications for Further Research. Journal of Sign Language and Linguistics 4(1/2) 29-45, 2001.

[24] Torresani, L., Hackney, P., &Bregler, C. Learning motion style synthesis from perceptual observations. Adv. in Neural Inf. Proc. Sys. (pp. 1393--1400), 2007.

[25] Chautard E, Thierry-Mieg N, Ricard-Blum S, Interaction networks: from protein functions to drug discovery. A review.PathologieBiologie 57: 324.333–, 2009.

[26] Garg A, Xenarios I, Mendoza L, DeMicheliG An efficient method for dynamic analysis of gene regulatory networks and in silico gene perturbation experiments; Springer. 62–76, 2007.

[27] Zheng D, Yang G, Li X, Wang Z, Liu F, et al. An Efficient Algorithm for Computing Attractors of Synchronous And Asynchronous Boolean Networks.PloS one 8: e60593, 2013

مقترح نظام ترجمة من النص العربي للغة الاشارة العربية

أحمد سمير - قسم الحسابات العلمية - كلية الحاسبات و المعلومات - جامعة عين شمس
أحمد علي عبد المجيد- قسم الحسابات العلمية - كلية الحاسبات و المعلومات - جامعة عين شمس
محمد فهمي طلبة - قسم الحسابات العلمية - كلية الحاسبات و المعلومات- جامعة عين شمس

خلاصة:

يهدف هذا البحث الي اتاحة اللغة العربية المكتوبة في شكل لغة اشارات لكي يتمكن ذوي الاعاقة السمعية من فهمها. ان النظام المقترح يعتمد علي تصميم و تنفيذ نظام ترجمة من النص المكتوب لاشارة يتم تأديتها من خلال شخصية كرتونية. ان النظام المقترح يكسر حواجز التواصل من خلال تسهيل اندماج ذوي الاعاقة السمعية في المجتمع.

Data Preparation and Handling for Written Quran Script Verification

Mohsen A. Rashwan *1, Ali Ramadan *2, Hazem M. Safwat *3, Salah Ashraf *4, Hazem Mamdouh *5

*Department of Electronics & Electrical Communication Engineering
Faculty of Engineering, Cairo University
Giza, Egypt
1mrashwan@rdi-eg.com
2eng.aliramadan.2016@gmail.com
3hazem.safwat93@eng-st.cu.edu.eg
4salahashraf@rocketmail.com
5hazem.mamdouh.fekry@hotmail.com

Abstract— **In this paper, a system is proposed to prepare a digital or a scanned Quran version for a verification process. The system handles the skew errors in the scanned image, text extraction from ornamentation, a successful line segmentation for Arabic scripts, verse pattern detection for different versions, and powerful diacritics classifier. The proposed system has been tested on different paper Quran versions and gives very promising results.**

***Keywords*: Quran, Surah, Islamic Image Processing, Diacritic Classifier, Tilt Handling, Text Extraction, Line Segmentation, Verse Detection.**

1 INTRODUCTION

Quran is the most preserved book in the mankind history. The steadiness of Quran contents is considered one of the most important tasks to Muslims, thus they must ensure that any newly published version of Quran has no difference than version of 1400 years old. In the current technology era, the verification method does not change for a while as it is accomplished by reviewing committees by checking each holy book character and diacritic, such a process could take at least 3 months and till now this process could be done only on paper version, many Muslims depend on digital applications even though in their religious doctrine due to the prevalence of smart phones, tablets and personal computers. These electronics versions have no ruler as anyone can publish whatever he wants, thus the traditional method of verification cannot be efficient in Quranic paper versions and cannot be applicable to Quranic digital versions.

The verification process can be done in a semi-automated way using image processing and machine learning techniques since each Quranic version has its own ornamentation, frame shape and verse patterns which differ from another version. Pre-processing phase is needed to start any verification process and this phase comprises of five main tasks which pose the entry image of a qur'anic page to be able to be verified later. Previous work in this field is not quite similar but there are many authors could help in the preprocessing phase with indirect ways as shown in the next chapter, the main tasks of pre-processing phase will be explained in the methodology section 3, the results will be shown in section 4, and conclusion and future work will be discussed in section 5.

2 PREVIOUS WORK

To our knowledge, there is no previous contribution in processing images of the holy Quran, but there are some resources help indirectly to build the system. Reference [1] shows the tilt back algorithm using Hough transform, while other authors in [2]suggest handling Arabic scripts tilt error using the detection of the Alif positions and its orientation to tilt back the script. Text detection can be done using textural based model as discussed in [3]. Reference [4]could help in text line segmentation process through many algorithms such as projection based methods and smearing methods.

3 METHODOLOGY

Various Quranic versions lead to variation in the number of pages and number of lines, thus neither pages can be reference nor lines. The reference which never changes from one version to another is the verse, so the objective of this process is to reach the verse level in the same line without dividing into multiple lines. Verification process could be accomplished by verifying text itself and its diacritics, the verse is needed to be split up into two categories, the first is the text which is called main Body and the second is the diacritics and recital symbols which can be called secondaries.

The process took place in five main steps as shown in Figure 1. Tilt correction is responsible for rectifying the page if its text is angled and set it back to be in a horizontal way, leading to extracting text of the page only which makes the next task is to separate each line alone and by separating each verse alone then concatenating its lines in single one, the separation process can be done.

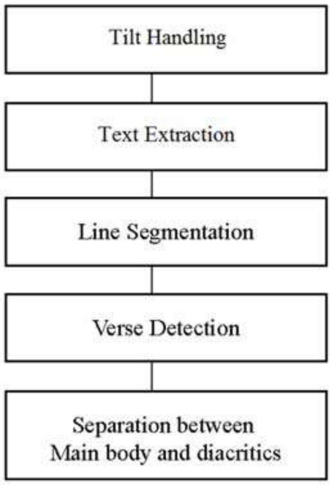


Figure 1 Main tasks of pre-processing

A. Tilt Handling

Tilt error could result from scanning or printing faults and this error could cause a failure of the whole system, so it should be handled. The simplest way to tilt back the script is using some morphological operations like erosion and dilation to get the frame complete as one component and as the frame has a nearly fixed size that allowed to search on it then calculate its orientation angle, the principle of computation the orientation angle according to [5][6]is to determine the inertia ellipse of a shape from the normalized centered moments resulting the matrix of inertia and by extracting Eigen vectors and Eigen values of the matrix the orientation angle can be obtained according to Figure 2 and then tilt the original page back with the same calculated angle.

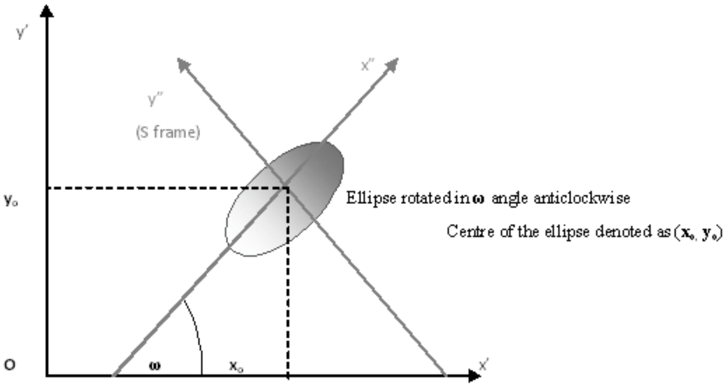
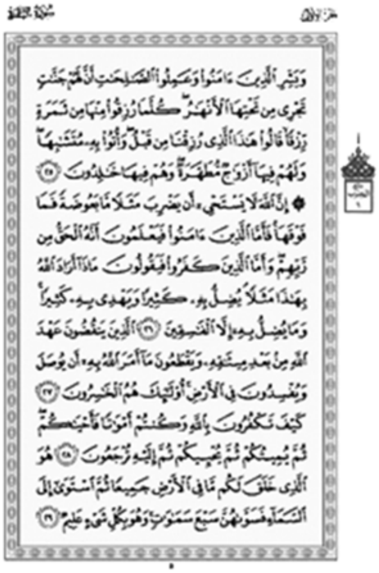

y'
x''
y''
(S frame)
y₀
Ellipse rotated in ω angle anticlockwise
Centre of the ellipse denoted as (x₀, y₀)
O
ω
x₀
x'

Figure 2 Computation of orientation angle

B. Text Extraction

After tilting back, the page if it is angled, frame and unimportant components should be removed from the given page. According to Figure 3, any qur'anic page contains not only the text but also ornamentation and indicator symbols like surah name, Hezb symbol, page number, …etc.

Thus the achievement of text extraction is to keep only text and remove other sources of destruction which negatively affect the verification process. The frame bounds the text block which seems to be the biggest connected component over the whole page so it can be easily removed according to [7] but when applying it on different versions, some drawbacks as frame could be separated in many parts due to faults of printing or binarization.



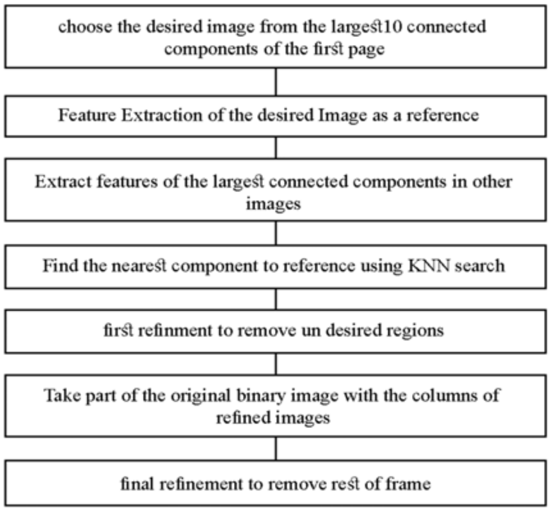
سورة البقرة الجزء الأول

وبشر الذين ءامنوا وعملوا الصالحات أن لهم جنات
تجري من تحتها الأنهار كلما رزقوا منها من ثمرة
رزقا قالوا هذا الذي رزقنا من قبل وأتوا به متشابها
ولهم فيها أزواج مطهرة وهم فيها خالدون ٢٥
۞ إن الله لا يستحيي أن يضرب مثلا ما بعوضة فما
فوقها فأما الذين ءامنوا فيعلمون أنه الحق من
ربهم وأما الذين كفروا فيقولون ماذا أراد الله
بهذا مثلا يضل به كثيرا ويهدي به كثيرا
وما يضل به إلا الفاسقين ٢٦ الذين ينقضون عهد
الله من بعد ميثاقه ويقطعون ما أمر الله به أن يوصل
ويفسدون في الأرض أولئك هم الخاسرون ٢٧
كيف تكفرون بالله وكنتم أمواتا فأحياكم
ثم يميتكم ثم يحييكم ثم إليه ترجعون ٢٨ هو
الذي خلق لكم ما في الأرض جميعا ثم استوى إلى
السماء فسواهن سبع سماوات وهو بكل شيء عليم ٢٩

٥

Figure 3 Quranic page Sample

Algorithm 1: Text Extraction

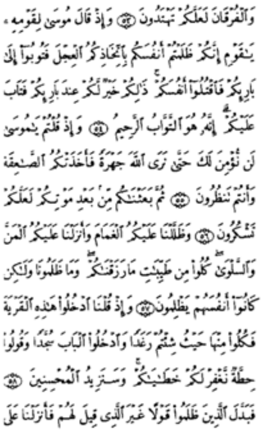


choose the desired image from the largest10 connected components of the first page
Feature Extraction of the desired Image as a reference
Extract features of the largest connected components in other images
Find the nearest component to reference using KNN search
first refinment to remove un desired regions
Take part of the original binary image with the columns of refined images
final refinement to remove rest of frame

Algorithm 1 has been applied starting with showing the user some of the largest connected components of the first page of the test Quranic version and let him choose the component that contains the text only. Extracting some features from the selected component which are perimeter and size in rows and columns as it was noticed that these features are very close to pages of different versions. After comparing the saved features with each page features using the nearest neighbor then estimating the position of the fit component in the whole page then crop it to be a text block only without any another component of frame or symbols as shown in Figure 4.

C. Line Segmentation

The role of this process is to automate the line segmentation which influences the accuracy of character and word recognition, for that reason this process considers the main core of preprocessing stage. The nature of Quranic text with its diacritics makes the segmentation too difficult to be done in a simple way like getting the projection profile approach to detect the separation between lines and the baseline of each line in the script as mentioned in [8]. Smearing methods of dilation operation in the horizontal direction could be used to fill the gap between the words to force the line to be one connected component. Reference [9]proposed a horizontal dynamic mask used to perform the segmentation of Arabic text lines. Initially the smearing method gives a promise result on one of the qur'anic version but when tested on another version it fails in some cases which is not acceptable in the verification process.



وَٱلْفُرْقَانَ لَعَلَّكُمْ تَهْتَدُونَ ٥٣ وَإِذْ قَالَ مُوسَىٰ لِقَوْمِهِۦ
يَٰقَوْمِ إِنَّكُمْ ظَلَمْتُمْ أَنفُسَكُم بِٱتِّخَاذِكُمُ ٱلْعِجْلَ فَتُوبُوٓاْ إِلَىٰ
بَارِئِكُمْ فَٱقْتُلُوٓاْ أَنفُسَكُمْ ذَٰلِكُمْ خَيْرٌ لَّكُمْ عِندَ بَارِئِكُمْ فَتَابَ
عَلَيْكُمْ إِنَّهُۥ هُوَ ٱلتَّوَّابُ ٱلرَّحِيمُ ٥٤ وَإِذْ قُلْتُمْ يَٰمُوسَىٰ
لَن نُّؤْمِنَ لَكَ حَتَّىٰ نَرَى ٱللَّهَ جَهْرَةً فَأَخَذَتْكُمُ ٱلصَّٰعِقَةُ
وَأَنتُمْ تَنظُرُونَ ٥٥ ثُمَّ بَعَثْنَٰكُم مِّنۢ بَعْدِ مَوْتِكُمْ لَعَلَّكُمْ
تَشْكُرُونَ ٥٦ وَظَلَّلْنَا عَلَيْكُمُ ٱلْغَمَامَ وَأَنزَلْنَا عَلَيْكُمُ ٱلْمَنَّ
وَٱلسَّلْوَىٰ كُلُواْ مِن طَيِّبَٰتِ مَا رَزَقْنَٰكُمْ وَمَا ظَلَمُونَا وَلَٰكِن
كَانُوٓاْ أَنفُسَهُمْ يَظْلِمُونَ ٥٧ وَإِذْ قُلْنَا ٱدْخُلُواْ هَٰذِهِ ٱلْقَرْيَةَ
فَكُلُواْ مِنْهَا حَيْثُ شِئْتُمْ رَغَدًا وَٱدْخُلُواْ ٱلْبَابَ سُجَّدًا وَقُولُواْ
حِطَّةٌ نَّغْفِرْ لَكُمْ خَطَٰيَٰكُمْ وَسَنَزِيدُ ٱلْمُحْسِنِينَ ٥٨
فَبَدَّلَ ٱلَّذِينَ ظَلَمُواْ قَوْلًا غَيْرَ ٱلَّذِى قِيلَ لَهُمْ فَأَنزَلْنَا عَلَى

Figure 4 Result of text extraction

Some challenges which affect the algorithm exist in the ornament and shapes rather than the text itself as shown in Figure 5. Because of these challenges algorithm has been consisted of three main steps as illustrated in Algorithm 2.

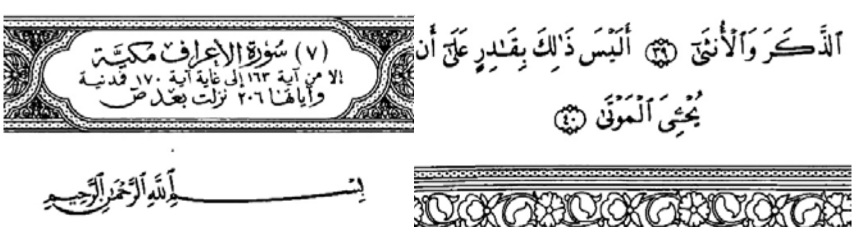
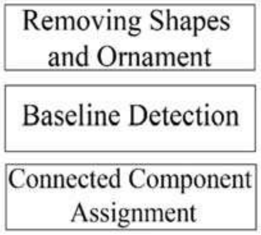

(٧) سُورَةُ الأعراف مكيّة
إلا من آية ١٦٣ إلى غاية آية ١٧٠ فمدنية
وآياتها ٢٠٦ نزلت بعد ص
ٱلذَّكَرَ وَٱلْأُنثَىٰٓ ٣٩ أَلَيْسَ ذَٰلِكَ بِقَٰدِرٍ عَلَىٰٓ أَن
يُحْـِۧيَ ٱلْمَوْتَىٰ ٤٠
بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Figure 5 Samples of the faced challenges

Removing shapes and ornament can be done by filling operation to make the detection of the non-text component is easy by removing the biggest component compared by pixels. The text script remains in the image, thus detection of baseline should be done and as said before that diacritics of two lines may overlap so projection profile cannot be used directly, so by removing components smaller than certain size resulted from testing various versions, then using horizontal projection profile computed for each strip to get the position of the baseline. The components remained from the previous phase can be assigned in separate line easily while the removed components have not been assigned yet, components assignment takes place by calculating the distances between each component and each baseline. Each component passes through three conditions to ensure that it actually belongs to the closet baseline, the first condition checks the distances between the centroid of each component and the closest baseline once, and the separation line another, the second one benefits from the fact of there is a set of components that belongs to beneath baseline only and here diacritics classifier from the next stage could be used to check if the tested component is a part of this set or not, while the last condition comes with the failure of the past two conditions and it measures the distance between the current component and the two closest components above and below it and assign this component to the nearest component's baseline and the output is shown in Figure 6.

Algorithm 2: Line Segmentation



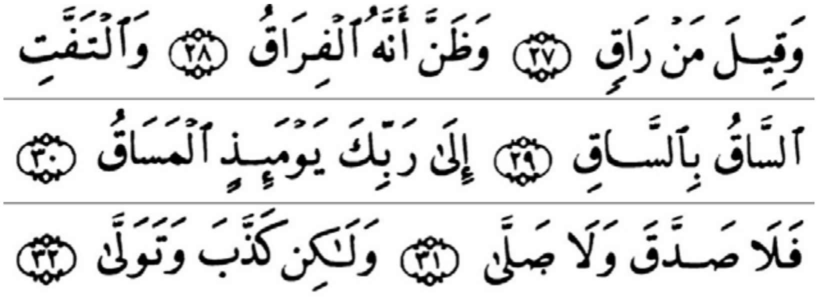

وَقِيلَ مَنْ رَاقٍ ٢٧ وَظَنَّ أَنَّهُ ٱلْفِرَاقُ ٢٨ وَٱلْتَفَّتِ
ٱلسَّاقُ بِٱلسَّاقِ ٢٩ إِلَىٰ رَبِّكَ يَوْمَئِذٍ ٱلْمَسَاقُ ٣٠
فَلَا صَدَّقَ وَلَا صَلَّىٰ ٣١ وَلَٰكِن كَذَّبَ وَتَوَلَّىٰ ٣٢

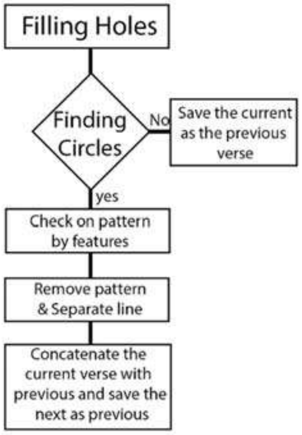
Figure 6 Output of line segmentation phase

D. Verse Detection

Verse Detection phase depends serially on the line segmentation phase to be totally completed on the whole tested qur’anic version to start. In normal situations, it should be an easy task using pattern recognition as in [10] , but for many versions, it is hard to detect all marks of different versions in a fully automated way. Some challenges faced the verse mark detection due to faults of printing and binarization of the tested version but by some morphological operations like closing and opening as discussed in [11]can solve these issues.

The Algorithm depends on the input as the whole qur’anic version in the shape of separated line, thus the verse mark detection can be detected according to algorithm 3. From observing most of the qur’anic versions, it is found that verse pattern takes a rounding shape which can be easily detected using circular Hough transform as discussed in [12] and the applied algorithm mainly depend on this feature of the verse pattern.

Algorithm 3: Verse Marking Detection


Filling Holes
Finding Circles
No
Save the current as the previous verse
yes
Check on pattern by features
Remove pattern & Separate line
Concatenate the current verse with previous and save the next as previous

Whole lines of the tested qur’anic version are received, then the progress of the algorithm takes place in a loop starting with filling the holes in the image line, then using circular Hough transform to detect the rounding circles in the line, unfortunately not all the circles detected are verse pattern as expected which leads to take some of features of the verse pattern as area of pixels, zero crossing and ratio of length to width of the chosen component, the verse marks are then removed and the line is split into the number of circles detected plus one, the previous part before the verse pattern is concatenated with the already saved verse while the next part is saved as the previous verse instead of the saved one, and this happens when circle is detected if else, the entire line is saved as it is to be concatenated with its rest on the next line. The result of the verse detection is shown in Figure 7.

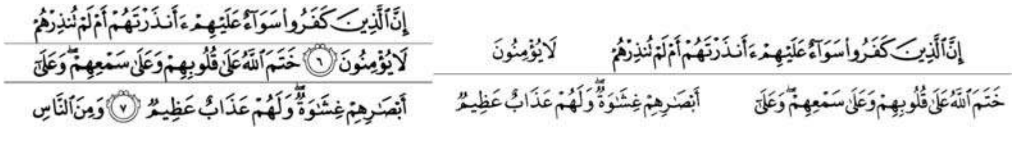

إِنَّ ٱلَّذِينَ كَفَرُوا۟ سَوَآءٌ عَلَيْهِمْ ءَأَنذَرْتَهُمْ أَمْ لَمْ تُنذِرْهُمْ
لَا يُؤْمِنُونَ ٦ خَتَمَ ٱللَّهُ عَلَىٰ قُلُوبِهِمْ وَعَلَىٰ سَمْعِهِمْ وَعَلَىٰٓ
أَبْصَٰرِهِمْ غِشَٰوَةٌ وَلَهُمْ عَذَابٌ عَظِيمٌ ٧ وَمِنَ ٱلنَّاسِ
إِنَّ ٱلَّذِينَ كَفَرُوا۟ سَوَآءٌ عَلَيْهِمْ ءَأَنذَرْتَهُمْ أَمْ لَمْ تُنذِرْهُمْ لَا يُؤْمِنُونَ
خَتَمَ ٱللَّهُ عَلَىٰ قُلُوبِهِمْ وَعَلَىٰ سَمْعِهِمْ وَعَلَىٰٓ أَبْصَٰرِهِمْ غِشَٰوَةٌ وَلَهُمْ عَذَابٌ عَظِيمٌ

Figure 7 Input and output of Verse detection

E. Separation between Main body and Diacritics

At this phase, verification may start but it is hard to withdraw the word with its diacritics from the verse, thus the simplest way is to check the word once and then check its diacritics in a separate way. By looking in Figure 8.

a, the main body can be extracted using baseline of each verse but sometimes this procedure can give wrong determination for components, in other words, a main body component can be determined as diacritic and vice versa as shown in Figure 8.b, thus baseline is not a fulfill algorithm which gives a promise separation so another check is needed to be done to determine the right category for these components, so it is needed to build Diacritics Classifier.

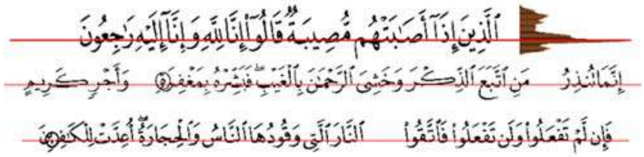


Figure 8. a Determination of the baseline **Figure 8. b Baseline Failures**

Figure 8 Baseline Algorithm

Building a classifier needs enough data to be trained with, the dataset of the diacritics is collected from different qur'anic versions manually but this collection is on offline stage only which means one-time process, so the system is still automated which means there is no entrance for a human being. Using held-out method for evaluation of each classifier which means to divide training data into two parts one for the training procedure and the other for testing the model to find the optimum values for model parameters

The applied classifier is a combination of two classifiers which are support vector machine (SVM) classifier and feed forward classifier, the both classifiers were built using held-out with ratio 7:3, 70 % percent is the training data while the rest for validation. Using SVM with adapting its kernel function to radial basis function (RBF) once and linear another one both gave the same accuracy but SVM with RBF kernel have less time consumption in testing than that with a linear kernel. While the other classifier (FFNN) is used with a back-propagation learning method with one hidden layer which contains 40 hidden neurons have been trained through hyperbolic tangent function while the output layer has been used soft-max function, the other parameters like a number of epochs, learning rate and gradient descent remained until convergence happens. The combination has been made to increase the confidence of the result, not the accuracy, so weighted majority vote technique as explained in [13]is used to establish the classifier.

The classifier has been ready to categorize between the main Body and diacritics so it will be the secondary check as illustrated in Algorithm 4.

By using the pre-trained classifier, the label of some components can be determined if they belong to main body or diacritic, the final check to avoid any wrong categorization got from the classifier, template matching is done with some pre-saved data for diacritics to make sure that labels from classifier are right.

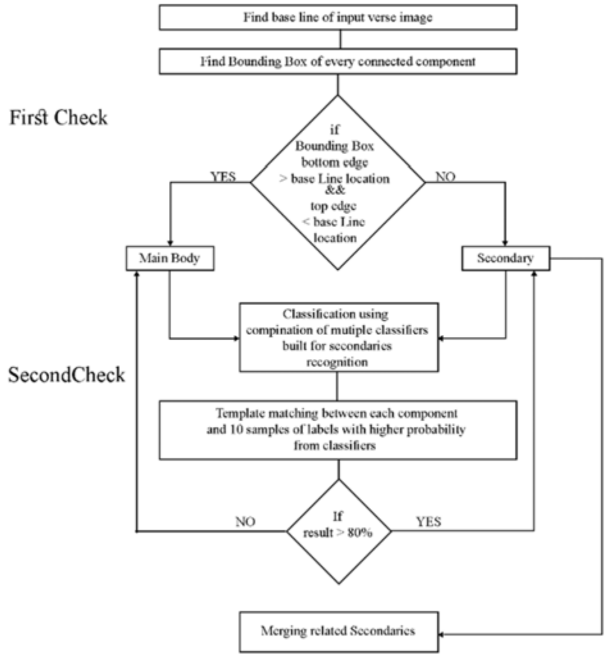
4 RESULTS

The system is giving a promising result, as it was tested on two different Quran versions, it gives an accuracy hundred percent in all phases except for the last phase of separation between main body and diacritics could be done with accuracy in 99.8%, the whole system processes a single page in 33.5 seconds only and the time varies with the changing of resolution and copies of the tested Quran. The diacritics classifier gives accuracy 99.94% from testing 140000 diacritics collected from 3 different Quran versions.

5 CONCLUSIONS AND FUTURE WORKS

The system presents a pre-processing phase using machine learning and image processing which deal with orientation correctness, text extraction, Arabic script line segmentation and verse pattern recognition. The diacritic classifier can be also used externally to recognize the diacritic symbols of the holy book or of any Arabic script. This system can open the door towards the verification process using automated ways without human entrance. The future works can be summarized in working on verification process on the scanned images of the Quran and the electronics digital versions with lower and higher resolutions, the efficiency also needs to be increased as much as possible and trying to enhance the time consumption using graphical processing units.

Algorithm 4: Separation between Main body and Diacritics


Find base line of input verse image
Find Bounding Box of every connected component
First Check
if
Bounding Box
bottom edge
> base Line location
&&
top edge
< base Line
location
YES
NO
Main Body
Secondary
Classification using
compination of mutiple classifiers
built for secondaries
recognition
SecondCheck
Template matching between each component
and 10 samples of labels with higher probability
from classifiers
If
result > 80%
NO
YES
Merging related Secondaries

REFERENCES

[1] G. R. T. H. W. Daniel S. Le, "Automated page orientation and skew angle detection for binary document images," in *Pattern Recognition*, 1994.

[2] S. S. M. Sarfraz, "An efficient scheme for tilt correction in Arabic OCR system," in *Proceedings of the Computer Graphics, Imaging and Vision: New Trends*, 2005.

[3] K. I. a. J. K. a. K. J. H. Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.,* 2003.

[4] A. Z. B. T. L. Likforman-Sulem, "Text Line Segmentation of Historical Documents: a Survey," *International Journal on Document Analysis and Recognition,* p. 2006.

[5] "Orientation Calculation," [Online]. Available: http://www.mathworks.com/matlabcentral/answers/91053-how-does-this-function-regionprops-to-find-the-orientation-of-any-object. [Accessed 01 July 2016].

[6] J. Bigun, "Optimal Orientation Detection of Linear Symmetry," in *IEEE First International Conference on Computer Vision*, 1987.

[7] "Isolation of biggest connected component," [Online]. Available: http://www.mathworks.com/help/images/ref/bwconncomp.html. [Accessed 03 July 2016].

[8] E. a. C. M. C. Bruzzone, "An Algorithm for Extracting Cursive Text Lines," in *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, 1999.

[9] M. a. L. L. a. S. C. Y. Khayyat, "Learning-based Word Spotting System for Arabic Handwritten Documents," *Pattern Recogn.,* 2014.

[10] "Pattern Recognition Concepts," [Online]. Available: https://courses.cs.washington.edu/courses/cse576/book/ch4.pdf. [Accessed 06 July 2016].

[11] "Morphological Operation," [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/HIPR2/morops.htm. [Accessed 01 July 2016].

[12] I. D. Marcin Smereka, "Circular Object Detection Using a Modified Hough Transform," *International Journal of Applied Mathematics and Computer Science,* 2008.

[13] "Combining Classifiers," [Online]. Available: https://drive.google.com/file/d/0B4TJxUk87Kr3SVc5bDVqNW85bmc/view. [Accessed 06 July 2016].

BIOGRAPHY

Ali Ramadan was born in Giza, Egypt, in 1993. He graduated from faculty of engineering Cairo university electronics and electrical communication department with degree good in 2016. He has experience in machine learning and image processing fields.

Salah Ashraf Salah was born in Cairo, Egypt, in 1993. He received the B.E. degree in Electronics and Communication Engineering from the University of Cairo, Egypt, in July2016.

In August 2016, He joined Valeo as software engineer till now. In Sep. 2016, he joined the Department of Electronics and Communication Engineering, University of Cairo, as a teaching assistant for credit hours' system.

Hazem Mamdouh Fekry was born in Cairo, Egypt, in 1993. He received the B.E degree in Electrical Electronics and Communication Engineering from Cairo University in July 2016. In September 2016, he joined trend Micro as information systems security engineer. He has experience in machine learning techniques and language processing.

Hazem Mohamed Safwat was born in 1993, he received B.Sc. in electrical electronics and communication engineering from Cairo University. He is an engineer with a great passion for open source program and currently a student in FAU Erlangen- Nirenberg in advanced optics. He will continue to work and volunteer in open source projects

إعداد و معالجة القرءان الكريم لعملية تصحيح النسخ الضوئية

محسن رشوان، على رمضان، حازم صفوت، صلاح أشرف، حازم ممدوح

كلية الهندسة- جامعة القاهرة- جمهورية مصر العربية

ملخص

يقوم هذا النظام بتأهيل النسخ الرقمية و النسخ الضوئية للقرآن الكريم لعملية التصحيح الأوتوماتيكية بدلا من الطرق اليدوية المطبقة في لجان المراجعة. تتم العملية عبر معالجة الصور للإنحرافات و الإمالات الناتجة عن طرق المسح الضوئي ثم يتم إستخراج المحتوى القرءاني عن طريق التخلص من الإطارات و الزخارف و من ثم يتم تقسيم هذا المحتوى إلى سطور ثم إلى أيات. كل العمليات السابقة تتم عن طريق عمليات مميكنة دون التدخل الآدمي. كما يوفر النظام مصنف أوتوماتيكي للتعرف على أشكال التلاوة و التشكيل بشكل عام على عدة نسخ مختلفة من القرآن الكريم.

Towards Building CECA WordNet: A Domesticated Arabic-English Lexicon of Contemporary Egyptian Colloquial Arabic Words from Twitter

Bacem A. Essam[*1], Prof Dr. Mostafa M. Aref [**2]

[*] *English Language Department, Faculty of Al-Alsun, Ain Shams University*
[1]*literaryartrans@gmail.com*

[**] *Computer Science Dept., Faculty of Computer Science and Information Sciences*
Ain Shams University, Cairo, Egypt
[2] mostafa.aref@cis.asu.edu.eg

***Abstract*— This paper aims at constructing a bilingual lexicon of the most frequent Contemporary Egyptian Colloquial Arabic (CECA) words on Twitter and their English equivalents. On the macrostructure level, it uses twenty-million-word corpora of Egyptian tweets to extract the headword entries of the lexicon. Then, it searches for the most equivalent English translation, using online informal British and American English dictionaries. On the microstructure level, the relevant senses and semantic relations of the populated headwords are defined. Therefore, the proposed lexicon, as an Arabic language resource, should provide the required raw material to build a CECA wordnet. This should promote a higher quality in machinery translation of CECA words especially the culturally-laden set of words.**

***Keywords*—** Lexicon compilation, Domestication and Foreignization, Arabic ontology, WordNet, Corpus linguistics

1 INTRODUCTION

The poor resourcing of language pairs, unavailability of competent morphological analyzers and syntactic parsers in poorly-resourced languages, specificity of the informal discourse, writing inconsistency of users and the dynamic sociolinguistic shifts may stand as the firewall against improving the machine translation quality. This rings true especially for the Arabic dialects; compared to the resources and tools applicable on the standard language variety and to the plethora of the enhanced tools for the Indo-European languages. This paper, therefore, addresses such a challenge by building a bilingual lexicon of the most frequent Contemporary Egyptian Colloquial Arabic (henceforth CECA) words on Twitter and their English Equivalents. Targeting the English equivalents of CECA words, online American and British English dictionaries are used to render the closest equivalent for each sense of the CECA headword. Enriching the output further, semantic relations - including synonyms, hypernyms, hyponymys, meronyms, and antonyms- as well as new senses are provided. Moreover, this study approaches two questions: First, what do the most frequent words in CECA reflect about society? Second, what do the Anglicized CECA words pertain to?

2 THEORETICAL BACKGROUND

Translation is a dynamic activity which takes place within a certain cultural system, translation strategies are controversially animated. Venuti has dichotomized ‘alienating/exotizing’ and ‘naturalizing’ translation approaches as ‘foreignization’ and ‘domestication’ respectively. Foreignization, on the one hand, ‘entails choosing a foreign text and developing a translation method along lines which are excluded by dominant cultural values in the target language’. Domestication, on the other hand, embraces ‘an ethnocentric reduction of the foreign text to [Anglo-American] target-language cultural values’ [1]. “Direct translation”, “overt translation”, “exotization”, “anti-illusionism”, “semantic translation”, “documentary translation”, and “observational reception” are other designations for foreignization. Domestication are also labeled as “indirect translation”, “covert translation”, “naturalization”, “illusionism”, “communicative translation”, “ instrumental translation” and “participative reception”. Although applying only one of the two effacing approaches is practically unmanageable, even in small scales, domestication enacts the lesser of two wrongs simply because translating into a language prioritizes enabling its average readers to grasp what they read [2]. Although Venuti advocates foreignizing translation, he is also aware of the necessity of domestication. This holds true in rendering a source text (ST) for a target culture when the paired languages are not indigenous [1].

Mohammad Enani [3] explains the demand for domestication in exporting the Arabic culture as the Arabic ideology and axioms are somehow alienated to the target reader. Foreignizing the translation can, therefore, make the translation ambiguous and incompetent given the negative stereotypes the Arabic culture has. Arabic Translators render the challenging culturemes at their best; however, the attempts to domesticate the Arabic texts in translation are subjective and idiosyncratic. It is imperative that domestication should be negotiated objectively. The unanswerable question pertains to the subjective mechanism by which translators render the domesticated equivalents. This crystalize the needs to unify a domesticating translation technique. Objective as it seems, we propose our ontological-based methodology of domesticating the translation of CECA words in section 3.

After the Princeton WordNet (www.wordnet.princeton.edu) gained, as an ontolexical hierarchy, widespread popularity, especially in the natural language processing community, wordnets were built in a number of different languages. Euro WordNet compiled words from eight European languages which are certainly and typologically unrelated to Indo-European languages. An important goal was to connect all wordnets to one another, so that equivalent words and meanings could easily be identified [4-5]. Pursuing this aim further, this study adopts dynamic resourcing of a bilingual lexicon given that corpora and lexicons denote, in part, dynamic language resources [6].

Ontology is a taxonomical of meanings or more precisely "specification of conceptualization". To build this Arabic ontology, the set of all Arabic terms are collected and their polysemy are identified using unique numbers and described using glosses. The Arabic Ontology is a linguistic ontology that represents the concepts of Arabic terms using formal semantic relations [7]. Interestingly, WordNet is a network of synsets (synonymous words). Since principles of sorting words ontologically, in Arabic, are based mainly on science, philosophy and linguistics respectively, the soundness measures of the Arabic ontology is a little bit different from those in the English language. English WordNet is based on the native speakers' accept and approval of the ontological relationship [8-9]. Accordingly, compiling a CECA WordNet out of Egyptian social interactions may prove profitable in terms of reflecting the Arabic mentality and ideology.

In WordNet, The main semantic relation among words is synonymy, as between the words *assume* and *suppose*, which are grouped into unordered sets (synsets). The criterion for joint synset membership is merely that the words denote the same concept. Each of WordNet's synsets is linked in turn to other synsets by means of a small number of "conceptual relations." Also, one-to-many mappings of word forms and word meanings govern polysemy and synonymy. A single word form expressing several meanings is a case of polysemy. Highly polysemous words in English are check, case, and line. Polysemy requires the reader or listener to identify the context-appropriate, intended sense of the word form. Hyponymy is a semantic relation between word meanings: e.g., tree is a hyponym of plant. Hyponymy is, therefore, a relationship found in many nouns, in quite a number of verbs, and in some adjectives. Its major significance for lexicographers is that the 'genus expression' in a definition should be the hypernym of the headword. Hyponymy holds more often between nouns while antonymy 'belongs' more to adjectives. Meronymys reflect the relationship of the part to the whole and vice versa [6].

3 RELATED WORKS

Now that computational and/or ontological compiling could prove effective in building contemporary lexicons and dictionaries has recently developed so that computational and/or ontological compiling is now the norm, this study usher the computational and ontological towards domesticating the translation of CECA words. Mustafa [10] has investigated the Arabic lexicon in the context of relational database theory. He demonstrated that lexical attributes can be classified into five categories comprising nineteen attributes, including form attributes, morphological attributes, functional attributes, meaning attributes, and referential attributes. Testing the results, he concluded that the experimental lexical database provides an efficient method for storing and retrieving Arabic lexical information. Moreover, several attempts have been succeeded in building a corpus-based hypernymy-hyponymy lexicon with hierarchical structure for Arabic from online data [11-13].

Generally, there are three approaches of processing data: manual, automatic and semi-automatic approaches. The semi-automated approach, which was devised in the last stages of database compilation, involved extracting corpus data, collocations, examples, and grammatical labels, and conducting lexicographic analysis in the dictionary-writing system rather than in the corpus tool. An evaluation that compared the manual approach with the semi-automatic approach showed that the semi-automatic approach is much quicker and presents the lexicographers with almost all the information they identified as relevant during the manual analysis, as well as additional potentially relevant information for the dictionary entry [14]. This study combines the computational and ontological approaches in building the proposed bilingual lexicon. This should enable a richer resourcing of the studied lexes.

4 METHODOLOGY

A. Data Collection

After normalization and cleaning of the extracted tweets, a corpus of 20, 252,732 tokens and 18, 050, 423 word types was compiled. All original tweets correlated to a random sample of ordinary individuals. After identifying Twitter IDs of the Egyptian organization, which publish posts in modern standard Arabic (MSA), their tweets were omitted for irrelevance. The data are normalized and cleaned to be uploaded onto Sketch Engine for further processing.

B. Data Processing Software

i. Sketch Engine

Sketch Engine is online corpus software using more than 200 corpora in 82 different languages. It offers several ready-to-use corpora, as well as tools for users to build, upload and install their own corpora. These tools include wordlist, concordancers, thesauri, and built-in dictionaries. Wordlist calculate the unique words in a text and their occurrence, concordancers identify a word's collocates, thesauri detect all the substitutive words that occur in the same or similar context of a certain word and the built-in dictionary enables searching for a specific list of words in a corpus. [15].

ii. WordNet 2.1

The Princeton WordNet (www.wordnet.princeton.edu) has gained widespread popularity, especially in the natural language processing community over the past decades. Therefore, Wordnets were built in a number of different languages. WordNet is used to provide the hierarchy of this sematic ontology using the hyponym-hypernym relations and to allocate the synsets of every lexical entry using the synonym-sorting option. This study uses WordNet 2.1 for defining the semantic relations of the English equivalents [4-6].

C. Procedures of Data Processing

This study is concerned mainly with exploring semantic relations of the CECA words. The list of the most frequent CECA words is selected before defining the lexicographic microstructure relevance. The proposed technique of rendering the most possible informal English translation is illustrated in figure 1.

Criteria of Inclusion

i. Most frequently used CECA words.

ii. Most frequently used CECA words which share MSA spelling with different senses.

iii. Most frequently used MSA words which are used in combination with CECA words and are useful in defining semantic relations. e.g.ابن كلب-ابن بنت - ابن مرة Otherwise, the phrase is excluded; e.g. ابن النادى -بنت مصر

v. Arabicized loanwords that have widely replaced the MSA equivalents are included; e.g.سيديهات ، اقراص مدمجة

vi. Under each headword, phrasal usages, even if not frequently used, are included to cover the optimal lexicographic relevance

Criteria of Exclusion

i. Words that have the same sense and spelling in CECA and MSA; e.g., حزم- أصابع –جرم

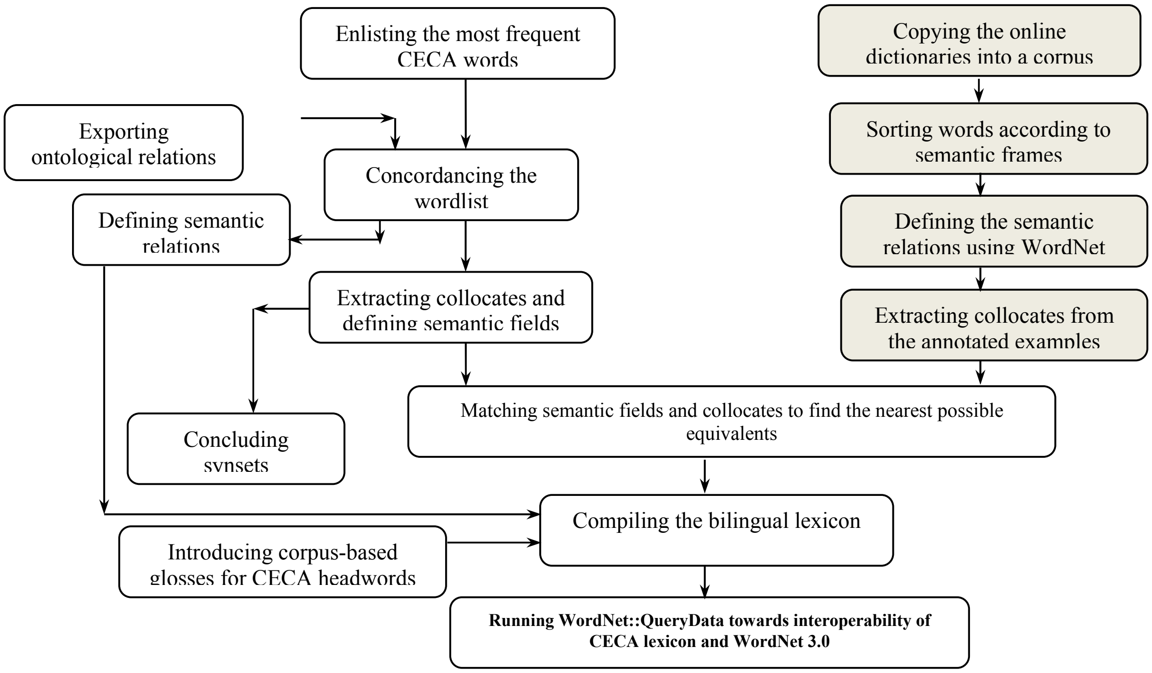
ii. Proper nouns, especially female and male given names and their derivatives; e.g., شعبان محمد - شعبولا- مرتضلا

iii. Monosyllabic non-words and vocals; e.g., هههه– ششش -تك

iv. Functional words are excluded. Only content words are enlisted

D. Defining Hypernym-hyponym Relation

The collected data are negotiated to extract the hypernym- hyponym relations using, first, some Arabic 'ontolexical markers' such as -زى - بنت - ولاد - ابن - بتاع–شوية -حتة. The second line of harvesting the hypernym-hyponym relations is screening *ath-Thalabi's* philology about the original etymology of the MSA which pertains closely to the studied CEA word because it provides an original hierarchal and ontological sense of structuring Arabic words . Words which fail both references are to be matched with the synonyms and near synonyms.



Enlisting the most frequent CECA words
Copying the online dictionaries into a corpus
Exporting ontological relations
Concordancing the wordlist
Sorting words according to semantic frames
Defining semantic relations
Defining the semantic relations using WordNet
Extracting collocates and defining semantic fields
Extracting collocates from the annotated examples
Concluding synsets
Matching semantic fields and collocates to find the nearest possible equivalents
Compiling the bilingual lexicon
Introducing corpus-based glosses for CECA headwords
Running WordNet::QueryData towards interoperability of CECA lexicon and WordNet 3.0

Figure 1: Diagramming the proposed methodology of translating CECA words into informal English

5 Results and Discussion

Appended in two pages, Table 1 exemplifies the lexicographic relevance of several headwords in our introduced CECA lexicon. Statistics indicate that the CECA words tend more to acquire new senses than the MSA peers. This should implicate about their dynamic nature. The semantic fields and ontological domains of the most frequent CECA words do center on religiosity, sexuality, nagging and satire. Although religiosity is most often populated, the concordance of the pertinent

CECA words indicates that only theological stock phrases are most often used (صلى ع النبى – ربنا معاه – ماشاء الله – ان شاء الله - الحمدلله). The same holds true when it comes to the CECA words of sexual innuendoes (متناك – شرموطة - قرنى). The least frequent CECA words tackle happiness (انا فرحان اوى – هاطير م الفرحة). The Anglicized CECA words correlate to borrowing technology-based words as well as words which are populated on social media streams ((لايك-الفيس- تويتر- تويتات- التايم لاين- فولو- بلوك- ديسك- موبايل - فلاشة - مِهنّج. Equivalent MSA Arabic words are rarely used (حظر- متابعة- أعجبنى - تغريدات - جوال- المغرد). The p-value which corresponds to chi-square measurement is lesser than 0.05. This indicates that the difference in using Arabizing loanwords into CECA discourse and equivalent MSA words are statistically significant. What intrigues about the use of loanwords on social media streams is venturing these words into the daily communication with conceptually relevant path (انا هنجت- الواحد عايز ريستارت – أجمد لايكات). This sociolinguistic and psychological folkloric orientation draws upon the rigidity of MSA in communicating day-to-day communications. Moreover, statics reveals that the use of CECA words is curvilinear. Words, such as الفشيخ- أزومل- أساحبى , have dramatically decreased. However, abandoned words can be revived or regionalized on liaising to a given sponsor. The revival of the title (ابلة) has synced with the introduction of (ابلة فاهيتا).Moreover, the noun (ازوزة) has a steady use in posts of Alexandrian populations. What also can be observed about the recently introduced words is sharing creative phonotactics (فاكس – افتكاسة- شهيص –صاورة). The adjective (فاكس) is etymologically borrowed from the transliteration of the English word fake. The noun (صاورة), which is used by quinquagenarians and hexagenarians, embraces the phonetic replacement of the interdental original sound with the connotative meaning of (صو). Thus, the word implies loudness and infidelity.

TABLE I

REPRESENTATIONS OF CECA HEADWORDS IN OUR PROPOSED LEXICON

	A	B	C	D	E	F	G	H	I
	POS	**Arabic Root**	**IPA**	**Google Translation**	**Bing Translation**	**5-year Incidence**	**Semantic field**	**Gender Preference**	**Polysemy**
ابج	Verb	-	əˀbeg	a B C	Abeg	Increasing	*Bribe*	M	N
اتاهل	Verb	اهل	əˀtʔhel	I qualify	Qualify	Steady	*Marriage*	M	N
اراجوز	N / adj	-	ʔrəʔgoʊz	Arajuz	Puppets.	Increasing	*Media*	M	N
اروبة	Adj	ارب	ʔrwbə	Arobh	Arobh	Revived	*Intelligence*	F	N
ازوزة	Noun	ازز	ʔzoʊzə	Azzouzh	Azosh	Fluctuant	*Beverage*	M=F	N
افشخانات	Noun	فشخ	əˀfʃəxənæt	Afshkhanat	Avshkhanat	Steady	*Advertising*	M	N
انتوخ	Adj	انتخ	ʔntoʊx	Antokh	Antokh	Modern	*Laziness*	M	N
انتيم	N / adj	-	ʔntim	are you	Anytime	Steady	*Intimacy*	M=F	N
حسبن	Verb	حسبن	ħæsben	Hspen	Forfeit	Fluctuant	*Revenge*	F	N
حسوك	Verb	حسك	ħəswek	Hassok	Hasok	Revived	*Sluggishness*	M	N
الش	N / adj	الش	əˀlʃ	Walsh	Walsh	Increasing	*Trivia*	M=F	N
شعنن	Verb	شعن	ʃəʕnen	Hann	Shann	Decreasing	*Anger*	M	N
شهيص	Verb	شهص	ʃəhjjʌsˤ	Shahys	Shhis	Introduced	*Flexibility*	M	N
صاورة	Noun	صور	ʌsˤ	Asawrh	Saurh	Introduced	*Political failure*	M	N
فشخ	V / adj	فشخ	fəʃx	Vchk	'Re amazing	Increasing	*Exaggeration*	M	Y
لسع	Verb	لسع	ləsʕ	sting	Sting	Increasing	*Madness*	M	N
متريش	Adj	ريش	metrɪɪʃ	Mtric	Metrish	Steady	*Wealth*	M	N
حشش	Verb	حشش	ħəʃeʃ	Hch	Mow	Steady	*Addiction*	M	N
معلش	*VP*	علش	məʕleʃ	it's okay	Forgive me,	Steady	*Apology*	F	N
وجب	Verb	وجب	wəgeb	must	I had	Revived	*Hospitality*	M	Y
هبد	Verb	هبد	həbəd	Hbd	Hebd	Fluctuant	*Violence*	M	N

Abbreviations: **Adj**: adjective; **IPA**: International phonetic alphabet **POS**: **M**: Male **F**: Female: **N**: No **VP**: verb phrase; **Y**: Yes

	J	K	L	M	N	O	P	Q
	Hypernym	**Hyponym**	**American Equivalent**	**CCM**	**British Equivalent**	**CCM**	**Glosses (brief definitions)**	**Synsets**
ابج	Bribe	Hush money	tit-for-tat, trouser	NA	*Swag, bung*	NA	***1a***. A demand to give money in order to facilitate or to do a solicited task ***1b***. A wider spectrum denoting all ill-gotten gains.	ظبط
اتاهل	Espousing	Officiate	Domesticated	NA	*Get hitched*	NA	The official completion of legal marriage procedures	اتدبس
اراجوز	broadcaster	Agenda	John Stewart	NA	*Elaine Paige*	NA	befuddling bureaucrats with ironic interview or in talk shows to show the faults in their logic.	مطبلاتى
اروبة	astuteness	Cunning	Cheeky Bitch	Ht	*snazzy*	Hm	An insanely intelligent and talented person in everything	خلبوس
ازوزة	beverage	Pepsi	cracka cola	Hm	*Coca-Cola*	Hm	Carbonated, caffeinated or decaffeinated, drink flavored with extract from Kola nuts (coca-cola) or New Coke (Pepsi).	حاجة ساقعة
افشخانات	Ostentation	Splurge	Halloween Train-wreck	NA	*Jiggery pokery*	NA	An extravagant effort ventured to show up.	حوارات
انتوخ	Disinclination	Slothfulness	sluggard	NA	*Bottler*	NA	a disinclination to work or exert oneself.	مستكنيص
انتيم	Friendship	Buddy , alter ego	Bestie	Hm	*Bestie*	NA	A very close and trusted friend who seems almost a part of yourself.	سديقى
حسبن	Agitation	Tailspin	Preach-speak	Ht	*For Pete's sake!*	Ht	An exclamation of dissatisfaction, usually by saying "*Allah (Alone) is Sufficient for me*".	-
حسوك	torpidity	hebetude	Hrach	Hm	*Croffle*	Hm	to shuffle along or walk very slowly, or take a gazillion hours for finishing a minor task.	ميتان - اتبارد
الش	Nonsensicality	Balderdash, buzzword,		Hm	*Shitty*	Hm	***1a***. trivial nonsense or foolish talk. ***1b***. stock phrases that have become nonsense through endless repetition ***1c***. a set of confused and meaningless rhetoric.	خنيق - بيض
شعنن	ire	outrage	shmobed	Hm	*shnitzen*	Hm	marked by extreme and violent energy; infuriated; nutty.	اتنرفز - اتزربن
شهيص	affect	commve	Relaxcited	NA	-	NA	puzzle over	
صاورة	group action	Egyptian revolution	Cyber-Revolution	Ht	-	NA	A term coined by the anti-arabic spring to describe the 'failing' revolution(s).	-
فشخ	1. beau ideal 2. Criticism	1. Gold standard 2. Criticism		NA	*Smashingly*	NA	1. (adj) Perfect. 2. (v) Criticizing heavily especially in political discourses.	عنب - جامد
لسع	mental illness	daftness	functionally insane	Hm	*wacky*	Ht	1. Causing a tongue or skin burn .e.g., taste a pungent, hot beverage or flame. 2. Getting crazy or insane.	هيس رايحة_منه
متريش	wealthiness	mammon	A Bill Gates	NA	*Dandyish*	NA	Extremely wealth ; jaunty	معدى
حشش	addiction	weed	dabbing	NA	*spliff*		Dronabinol is a psychoactive compound extracted from the resin of Cannabis sativa (marihuana, hashish)	برشم
معلش	apologia	sorrow	Bropology, fauxpology	NA	*soz*	Hm	Apololies, epology,	سورى
وجب	1. Interaction 2. Punishing	1. Socializing 2. Ridicule	1.Generotic 2. Ghosted	NA	*bastarding*	Ht	1. Denotes reprimanded, banquet action of being lavishly generous. 2. Ironical for punishing or giving unexpected rough time to someone.	ظبط كدر
هبد	bob	fall	splash	Hm	*By (h)eck !*	Ht	drop sharply, topple or plummet	رزع

Abbreviations: **CCM**: Cross-culture matching; **Hm**: homogenous; **Ht**: Heterogeneous; **NA**: Not Applicable.

6 Conclusions

The CECA discourse is very rich and does tackle deeply the deep layers of the societal iceberg. Harvesting and analyzing the content of written and spoken CECA discourses can produce a valuable pragmatic and sociolinguistic output. This can be ushered linguistically toward building lexicon and ontologies. On translating CECA words into English, the informal American English and slang harmonize cross-culturally more to the Egyptian culture than British English. The CECA discourse is full of controversies and express, most often, anger and frustration of the mass body.

REFERENCES

[1] Munday, J. *Introducing translation studies: Theories and applications*. Routledge. pp. 104-106. 2013
[2] Ožbot, Martina. "Foreignization and Domestication: A View from the Periphery." In *Rereading Schleiermacher: Translation, Cognition and Culture*, pp. 277-289. Springer Berlin Heidelberg, 2016.
[3] Enani, M. At-tarjjama aladabiyya. *Al-Ahram* 138 (46401). URL:<http://www.ahram.org.eg/NewsQ/249006.aspx> Accessed on Oct 5, 2016.
[4] Fellbaum, Christiane. "WordNet." In *Theory and applications of ontology: computer applications*, pp. 231-243. Springer Netherlands, 2010.
[5] Vossen, Piek. "Introduction to eurowordnet." In *EuroWordNet*: A multilingual database with lexical semantic networks, pp. 1-17. Springer Netherlands, 1998.
[6] Witt, Andreas, Ulrich Heid, Felix Sasaki, and Gilles Sérasset. "Multilingual language resources and interoperability." *Language Resources and Evaluation* 43, no. 1 (2009): 1-14.
[7] Jarrar, Mustafa, Nizar Habash, Diyam Akra, and Nasser Zalmout. "Building a corpus for palestinian arabic: a preliminary study." In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pp. 18-27. 2014.
[8] Hawwari, Abdelati, Wajdi Zaghouani, Tim O'Gorman, Ahmed Badran, and Mona Diab. "Building a lexical semantic resource for Arabic morphological Patterns." In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pp. 1-6. IEEE, 2013.
[9] Fellbaum, Christiane. *WordNet*. Blackwell Publishing Ltd, 1998.
[10] Mustafa, Suleiman Hussein. "A relational approach to the design of an Arabic lexical database." *Journal of King Saud University-Computer and Information Sciences* 14 (2002): 1-23.
[11] Abuleil, Saleem, and Martha Evens. "Extracting an Arabic lexicon from Arabic newspaper text." *Computers and the Humanities* 36, no. 2 (2002): 191-221.
[12] Elghamry, Khaled. "Using the Web in building a corpus-based hypernymy-hyponymy lexicon with hierarchical structure for Arabic." *Faculty of computers and information* (2008): 157-165.
[13] Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. "The penn arabic treebank: Building a large-scale annotated arabic corpus." In *NEMLAR conference on Arabic language resources and tools*, vol. 27, pp. 466-467. 2004.
[14] Gantar, Polona, Iztok Kosem, and Simon Krek. "Discovering automated lexicography: the case of the Slovene Lexical Database." *International Journal of Lexicography* (2016): ecw014.
[15] Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. "The Sketch Engine: ten years on." *Lexicography* 1, no. 1 (2014): 7-36.

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.

نحو بناء ووردنت للعامية المصرية المعاصرة:
بناء معجم عربى انجليزى داجن من تغريدات موقع تويتر

**[1]باسم عبدالله عصام و **[2]مصطفى محمود عارف
*قسم اللغة الأنجليزية-كلية الألسن-جامعة عين شمس
[1]literaryartrans@gmail.com
**كلية الحاسبات و المعلومات-جامعة عين شمس
mostafa.aref@cis.asu.edu.eg

ملخص
فى طيات آلاف المشاركات اليومية على مواقع التواصل الاجتماعي مثل فيسبوك وتويتر، أصبحت تلك المواقع منابر يرتادها العوام خاصة فئة المراهقين والشباب، مما حذا بتلك المواقع إدراج خاصية الترجمة الالية لترجمة المشاركات تلقائياَ من لغة إلى أخرى .يهدف البحث إلى بناء معجم ثنائى اللغة من كلمات العامية المصرية الأكثر شيوعا على موقع وتويتر وترجمتها للإنجليزية الدارجة معتمداً على تقنية الأنطولوجيا العربية و الذخائر اللغوية فى بناء معجم عامى معاصر. ويعرض المعجم تعريفًا لكل مفردة ونطقها على مواقع التواصل الاجتماعي ثم يتناول الدلالة اللفظية والمعجمية ليقترح من مجموع ذلك أقرب ترجمة ممكنة من المكافئات الإنجليزية الدارجة نحو إثراء الجزء الحاسوبي التطبيقي، الذى يستخدم فى تحسين كفاءة الترجمة الآلية وبناء أنطولوجيا عربية للعامية المصرية.

الكلمات الدلالية: بناء معجم – التدجين والتغريب – الأنطولوجيا العربية – برنامج ووردنت- علم الذخائر اللغوية

A Rule Based Method for Adding Case Ending Diacritics for Modern Standard Arabic Texts

Amany Fashwan[1], Sameh Alansary[2]

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

[1]amany.fashwan@bibalex.org

[2]sameh.alansary@bibalex.org

***Abstract*—This paper presents a rule based approach simulating the shallow parsing technique for detecting the Case Ending diacritics for Modern Standard Arabic Texts. In this system, the Case Ending diacritization problem is addressed depending on morphological analysis; the relation between each word with its Part Of Speech (POS), as well as syntactic analysis, the relation between a word and its position in the sentence. An Arabic annotated corpus of 550,000 words is used; the International Corpus of Arabic (ICA) for extracting the Arabic linguistic rules, validating the system and testing process.The output results and limitations of the system are reviewed and the Syntactic Word Error Rate (WER) has been chosen to evaluate the system.The syntactic diacritization WER achieved by the system is 9.36%.The results of the current proposed system have been evaluated in comparison with the results of the best-known systems in the literature. The best syntactic diacritization achieved is 9.97% compared to the best-published results, of [14]; 8.93%, [13] and [15]; 9.4%.**

Keywords: Automatic Diacritization · Syntactic Diacritics · Case Ending Restoration · Arabic Natural Language Processing · Arabic Shallow Parsing

1 INTRODUCTION

Modern Standard Arabic is currently the sixth most widely spoken language in the world with estimated 422 million native speakers. It is usually written without diacritics. This makes 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. However, a text without diacritics brings 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 [1]. Although undiacritized Arabic text is sufficient for Arabic speakers to use in writing and reading, this is not the case when dealing with software systems. For example, an Arabic text-to-speech system would not produce speech from undiacritized Arabic text because there is more than one way of saying the same undiacritized written Arabic word. Moreover, when searching for an Arabic word, many unrelated words would be included in the results.This suggests the need to diacritize Arabic text. Another reason for the diacritization is to permit the use of dictionaries and machine translation from and to Arabic [2].

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. The diacritization of an Arabic word consists of two components; morphology-dependent and syntax-dependent ones. While the morphological diacritization distinguishes different words with the same spelling from one another; e.g. عِلْم which means "science" and عَلَم which means "flag", the syntactic case of the word within a given sentence; i.e. its role in the parsing tree of that sentence, determines the syntax-dependent diacritic of the word. For example; دَرَسْتُ عِلْمَ الرِّيَاضِيَّاتِ implies the syntactic diacritic of the target word - which is an "object" in the parsing tree - is "Fatha", while يُفِيدُ عِلْمُ الرِّيَاضِيَّاتِ جَمِيعَ الْعُلُومِ implies the syntactic diacritic of the target word – which is a "subject" in the parsing tree - is "Damma" [3].

Due to its importance in giving the correct understanding of Arabic statement meaning, we present a rule based approach simulating the shallow parsing technique for detecting the case ending diacritics for MSA texts. In our proposed solution, the case ending problem will be addressed depending on both morphological processing level, the relation between each word with its Part Of Speech (POS), as well as syntactic processing level, the relation between a word and its position in the sentence. In what follows Section 2 reviews some related work to case ending automatic diacritization process. Section 3 details the description and processing of used corpus. Section 4 details the built Arabic diacritization system on Syntactic processing levels. Section 5 evaluates the output and compares the results with other state of the art 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], [3-17] and [33]. In addition, some software companies have developed commercial products for the automatic diacritization of Arabic; Sakhr Arabic Automatic Diacritizer [18], Xerox's Arabic Morphological Processor[19]and RDI's Automatic Arabic Phonetic Transcriptor (Diacritizer/Vowelizer) [20].

Moreover, in addition to the previous commercial products there are some trials for producing free products for the automatic diacritization of Arabic. For example, Meshkal Arabic Diacritizer [21], Harakat Arabic Diacritizer [22], Al-Du'aly [23], and Google Tashkeel which is no longer working where the tool is not available now.

It has been noticed that most of the previous systems use different statistical approaches in their diacritization systems for restoring the case ending diacritics and none of the previous systems makes use of syntax with the exception of [15]and [17] who have integrated syntactic rules.

3 Corpus Description and Processing

The used data in the diacritization system was selected from a manually annotated and disambiguated data of the International Corpus of Arabic (ICA) [24]. Each word is tagged with features, namely, Lemma, Gloss, Pr1, Pr2, Pr3, Stem, Tag, Suf1, Suf2, Gender, Number, Definiteness, Root, Stem Pattern, Case Ending, Name Entity and finally Vocalization. It contains about 500,000 manually morphologically disambiguated words. There are three kinds of data sets used here; 1) training data set which helps in extracting the linguistic rules need for restoring the case ending, 2) validation data set which helps in validating the extracted rules from the training data set, and 3) testing data set for evaluating the system. The data has been distributed as 80% for training the system, 10% for validation data set and 10% for the evaluation process.

A. Training Data Set Usage

The grammar of a language has several levels: phonological rules that govern the structure of sounds, the morphology that governs the structure of words, syntax that governs the structure of sentences and semantics that governs the meanings of words and sentences. In this paper, syntactic level is our concern.

It must be noted that extracting Arabic linguistic rules is not an easy task where these rules must be represented in a generalized format in a way that simulates the concerned component of the language. So these rules need to be constrained in a certain order to avoid overlapping among them. To extract the rules, an annotated data is needed, so the training data set is used[25].

The realization of nominal case in Arabic is complicated by its orthography, which uses optional diacritics to indicate short vowel case morphemes, and by its morphology, which does not always distinguish between all cases. Additionally, case realization in Arabic interacts heavily with the realization of definiteness, leading to different realizations depending on whether the nominal is indefinite, i.e., receiving *nunation* (تنوين), definite through the determiner Al+ (ال+) or definite through being the governor of an EDAFAH possessive construction (إضافة) [26].

In addition, case ending realization in Arabic interacts in some cases with other information:

1. **Word Patterns:** the diptote word patterns (الممنوع من الصرف) refer to a category of nouns and adjectives that have a special case ending when they are indefinite since they do not receive *nunation*. It must be noticed that when these words are definite, they receive regular case ending diacritics.
2. **Verb Transitivity:** the transitivity of the verbs helps sometimes in detecting the subject (which receives nominative case ending) and object (which receives the subjunctive case ending).
3. **Feminine Plural Word Forms:** in Arabic, the object receives the case ending for the genitive case instead of the subjunctive case; this is in the case of those words end with 'ات' 'At/NSUFF' [1]'جمع المؤنث السالم'.

The extracted rules in this level simulates one of the language processing approaches that compute a basic analysis of sentence structure rather than attempting full syntactic analysis; shallow syntactic parsing. It is an analysis of a sentence which identifies the constituents (noun groups, verb groups, prepositional groups, adjectival groups, adverbial groups), but does not specify their internal structure, nor their role in the main sentence.

[1] The transliteration scheme follows that of BAMA: http://www.qamus.org/transliteration.htm [Accessed 20-5-2015]

Phrases in Arabic language: Arabic words are classified as noun “اسم”, verb “فعل”, or particles “حرف”, intended for items which are neither noun nor verb. The clear difference between the three parts is the declension “الإعراب” or syntactic parsing. The major three phrases for the Arabic language are:

- **Noun phrase:** A noun phrase is one which starts with a noun, a proper noun or a pronoun. It represents the entity of person, place, animal, etc. about which the phrase is talking.The nominal sentence is composed of “المبتدأ” “*mubtada*” “starting” which is followed by “الخبر” “*khabar*” “information/predicate”. Information is the part of the phrase to complete the information about starting.
- **Verb phrase:** A verbal phrase is one which starts with a verb in any of the three forms (imperfect verb, past verb and command verb). The verbal phrase is considered stronger than a noun phrase composition-wise. The verbs in Arabic are divided into two types [27]:
  1. Intransitive verb ‘الفعل اللازم’ which means that the verb does not need more than the subject to fulfill the meaning. For example, the verb ‘جَلَسَ’ ‘jalasa’ ‘sat’ in ‘جَلَسَ مُحَمَّدٌ عَلَى السَّرِيرِ’ ‘jalasamuHam~adNEalaYAls~ariyri’ ‘Mohammed sat on the bed’ is an intransitive verb where there is no direct object and the meaning of the sentence is completed without it.
  2. Transitive verb ‘الفعل المتعدي’ which requires a direct object to complete the meaning of the sentence as the meaning of the sentence cannot be completed without it. The action of the transitive verb is traced back to its subject and directed towards its direct object, like when saying: فَتَحَ طَارِقٌ الأَنْدَلُسَ“fataHaTAriq Al>anodalus” “Tarek conquered Al-Andalus”. So, the verb needs the subject to do the action and the direct object to receive it.

     There is a transitive verb that has two direct objects. One of them is clear/unambiguous and the other is unclear/ambiguous, like when saying: أَدُّوا الأَمَانَاتِ إِلَى أَهْلِهَا “>ad~uwAAlo>amAnAti<ilaY>aholihA” “They took/delivered what they were entrusted with to its owner”. ‘الأَمَانَاتِ’ is a direct clear/unambiguous object, while ‘أَهْلِ’ is an indirect unclear/ambiguous object. It is clearly expressed in the genitive case due to the preposition.
- **Prepositional phrase**: A prepositional phrase in Arabic is used just like in English. It is the sequence of a preposition followed by a word or phrase. These prepositions may be one-letter, two-letter and three-letter word groups.

The main target here is to detect the boundaries of each type of these phrases and then extract the suitable rule for detecting the case ending. In Arabic nouns, adjectives, adverbs and imperfect verbs are the only words that accept the case ending or mood. In addition, all nouns and adjectives have one of three cases: nominative (NOM), accusative (ACC), or genitive (GEN). What sets case in MSA apart from case in other languages is most saliently the fact that it is usually not marked in the orthography, as it is written using diacritics which are normally omitted[26].

In this stage, some Arabic linguistic rules have been extracted from the training data set and implemented to detect the case ending depending on a window of -/+ 3 words around the focused word taking into consideration the context, the selected morphological analysis (in morphological processing level), definiteness feature, stem pattern and verbs transitivity. It must be noted that the correctness of detecting the case ending depends to a big extent on the correctness of the morphological analysis output.

For restoring the case ending diacritics, the outputof the morphological processing level from a system we have developed is used. It is a multi-layered system in which each layer tries to restore the missing diacritics of some words, ignoring the syntactic or case ending diacritics using linguistic features and statistical features to help the system predict the missing internal diacritics with a high accuracy. It also uses information provided by Buckwalter Arabic Morphological Analyzer [28] to cover all words' solutions and to decrease the number of unknown words. After using these layers, the system is ready for the syntactic processing level [25].

It must be noted that not all information that BAMA can give are used while working in the morphological processing level. Only, prefix(s), stem, suffix(s) and lemma are used. The reason for neglecting the other information is that some of these information lead to having repeated morphological forms (neglecting the case ending) such as vocalization and glossaries. BAMA does not always predict correctly the definiteness that is related to syntactic processing level (so some definiteness rules are extracted). Other information related to syntactic processing, pattern and transitivity, BAMA does not give any information about them. Consequently, in order to detect the case ending correctly prior steps are done where the stem pattern, transitivity and definiteness must be detected.

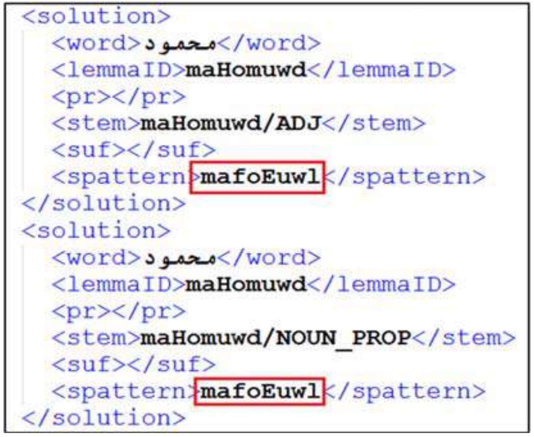
1. Stem Pattern and Transitivity Features:

BAMA does not give any information about the stem pattern so the stem pattern of each stem has been detected depending on the root, stem and lemma of each word. Root does not appear in BAMA's output, but it is found in it stem dictionary (dicStems).

For getting the stem patterns of all stems in the BAMA's dicStems, it was necessary to check and fix the roots problems that are founded in this dictionary since BAMA does not detect most of the roots correctly and it provides a root for loan or foreign stems. To do so, we were helped by the team of the Arabic Computational Center at Bibliotheca Alexandrina (BA) to review and fix all wrong roots of BAMA. After that we have made a script for detecting the stem pattern of stems that have root(s) in BAMA and review the stem pattern output. In addition, we have made another script that is able to provide the stem pattern of each stem in dicStems of BAMA to be as a part of its output as figure 1 shows. Moreover, some modifications have been done for the Perl open source file of BAMA to show the stem pattern in BAMA's output.

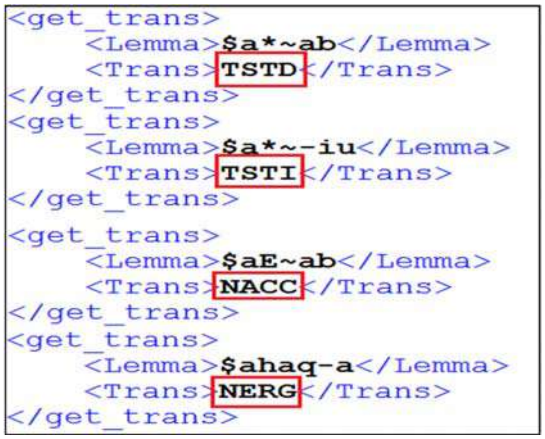
The transitivity is a feature that is assigned to the verbs which indicates the number of objects a verb requires or takes in a given instance. The transitivity of a verb can be basically classified into intransitive (NTST) and transitive (TSTD). Intransitive verbs are further classified into two types: non accusative (NACC) verb whose subject is not the agent, as the verb “تَدَفَّقَ” ‘tadaf~aqa’ ‘flow’ in "تَدَفَّقَ الْمَاءُ" ‘tadaf~aqa’ ‘Water flowed’, and non-ergative (NERG) verb whose subject is the agent, as the verb “مَشَى” ‘ma$aY’ ‘walk’ in "مَشَى الْوَلَدُ" ‘ma$aYAlowaladu’ ‘the boy walked’.

To detect the transitivity of each verb in the BAMA's dicStems, the unique lemmas of these verbs in dicStems have been used in detecting their transitivity. To get the transitivity for these lemmas we were helped by the Arabic dictionary of the Universal Network Language (UNL) [29]. Each lemma is saved with its transitivity to be used while detecting the case ending as figure 2 shows.



```
<solution>
  <word>محمود</word>
  <lemmaID>maHomuwd</lemmaID>
  <pr></pr>
  <stem>maHomuwd/ADJ</stem>
  <suf></suf>
  <spattern>mafoEuwl</spattern>
</solution>
<solution>
  <word>محمود</word>
  <lemmaID>maHomuwd</lemmaID>
  <pr></pr>
  <stem>maHomuwd/NOUN_PROP</stem>
  <suf></suf>
  <spattern>mafoEuwl</spattern>
</solution>
```

Figure 1: BAMA's Output After Adding the Stem Pattern



```
<get_trans>
     <Lemma>$a*~ab</Lemma>
     <Trans>TSTD</Trans>
</get_trans>
<get_trans>
     <Lemma>$a*~-iu</Lemma>
     <Trans>TSTI</Trans>
</get_trans>
<get_trans>
     <Lemma>$aE~ab</Lemma>
     <Trans>NACC</Trans>
</get_trans>
<get_trans>
     <Lemma>$ahaq-a</Lemma>
     <Trans>NERG</Trans>
</get_trans>
```

Figure 2: Lemmas of Verbs with Their Transitivity

Transitive verbs are classified into four types: direct transitive (TSTD); a verb which requires a subject and a single direct object as the verb “نَظَّمَ” ‘naZ~ama’ ‘organize’, indirect transitive (TSTI); a verb which requires a subject and a single indirect object as in the verb “قَضَى” ‘qaDaY’ ‘eliminate’, ditransitive (TST2); a verb which requires a subject and two objects as in the verb “جَعَلَ” ‘jaEala’ ‘make’, or tritransitive (TST3); a verb which requires a subject and three objects as in the verb ‘أَعْلَمَ’ ‘>aEolama’ ‘inform about’ [30].

In Arabic, a verb may have more than one type of transitivity, depending on its meaning for example, the verb ‘اعتمد’ ‘{iEotamada’ which is TSTI in ‘اِعْتَمَدَ وَزِيرُ الْمَالِيَّةِ الْمِيزَانِيَّةَ الْعَامَّةَ’ ‘{iEotamadawaziyruAlomAliy~apiAlomiyzAniy~apaAloEAm~apa’ ‘The Minister of Finance approved the general budget’, and TSTD in ‘اِعْتَمَدَ الْأَمْرُ عَلَى تَصْوِيتِ الْأَغْلَبِيَّةِ’‘{iEotamadaAlo>amour EalaYtaSowiytiAlo>agolabiy~api’ ‘The matter depended on the voting of the majority'. In the current system, if the verb has more than one transitivity, the best one is selected according to its context.

2. Extracted and Formulated Definiteness Rules:

As mentioned before, the definiteness information that BAMA can provide the system with are not used since BAMA is used in morphological processing level only and the definiteness is related, to a large extent, to syntactic processing level which depends on the context. So in order to set the case ending diacritics, some Arabic linguistic rules have been extracted and implemented in the system to detect the definiteness of each word depending on its context or its selected morphological analysis.

There are some rules that have been extracted and formulated for detecting the definiteness of nouns, adjective, adverbs, and pronouns. Each word of theses tags is assigned with 'DEF' which means the word is definite, 'INDEF' which means that the word is indefinite or 'EDAFAH' which means that the words are the governor of an EDAFAH possessive construction. The easiest feature to detect among these three features is when the word is definite 'DEF' and when it contains a possessive pronoun in its suffix it is assigned as 'EDAFAH'. The other cases are not easy to detect since there are a lot of cases that the indefiniteness and EDAFAH rules may overlap. So, the extracted rules here are restricted and constrained as much as possible to avoid such overlap.

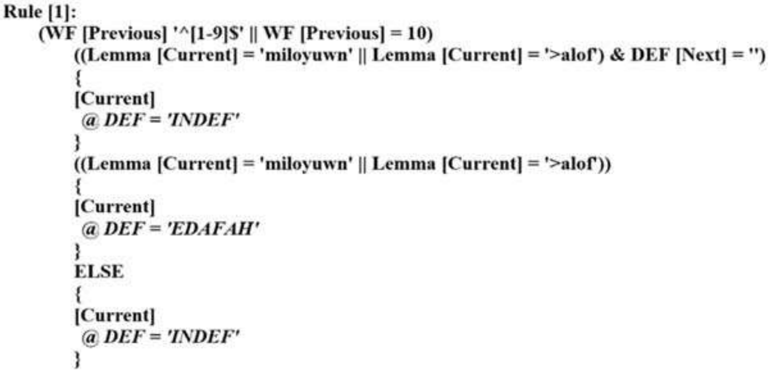
2.1 Rules for Setting the Word as Definite (DEF)

In this part, the extracted rules are easy to detect where there are limited cases in which the words are definite. If the tag of the word is proper noun (NOUN_PROP), the prefix of the word contains 'الـ' 'Al/DET' except in the word 'أمس' '>amos' that is exceptionally 'INDEF', 'mA/PRON' that act as 'الذي' 'Al~a*iy/PRON', pronouns and nouns of place or time such as 'وقتذاك' 'waqot*Ak', 'أمس' '>amos' that has no 'الـ' 'Al/DET' that is exceptionally 'DEF'.

2.2 Rules for Setting the Word as Indefinite (INDEF)

In this part, some rules for describing the word as indefinite have been extracted and formulated. As mentioned before, these rules are constrained as much as possible to avoid the overlap between the words that are assigned as EDAFAH. The simplest rule related to describing the analyzed nominal, adjectival or adverbial words as INDEF when the suffix of the word contains 'ان' 'Ani/N_SUF', 'ين''ayoni/N_SUF' 'iyna/N_SUF', 'تين' 'atayoni/N_SUF' 'تان' 'atAni/N_SUF', 'ون' 'uwna/N_SUF' or when the word ends with 'ا''A' and this 'ا' 'A' is not analyzed as 'A/N_SUF'. Examples of such words is 'رجلا' 'rajulAF' 'man', 'رجلان' 'rajulAn', 'حلقتان' 'HalaqatAni' 'rings/circles' and 'مسلمون' 'musolimuwna' 'Muslims'.

One of the extracted rules is related to the indefinite nouns that occur after the numbers from [1-10]. In this rule, there are two conditions: the first one is when the lemma of the words that occur after these numbers is '>alof' or 'miloyuwn' then this word is EDAFAH if it is followed by another word and INDEF if it is not followed by another word. The second condition is the default where the definiteness of this word is INDEF as rule 1 shows:



```
Rule [1]:
    (WF [Previous] '^[1-9]$' || WF [Previous] = 10)
        ((Lemma [Current] = 'miloyuwn' || Lemma [Current] = '>alof') & DEF [Next] = '')
        {
        [Current]
         @ DEF = 'INDEF'
        }
        ((Lemma [Current] = 'miloyuwn' || Lemma [Current] = '>alof'))
        {
        [Current]
         @ DEF = 'EDAFAH'
        }
        ELSE
        {
        [Current]
         @ DEF = 'INDEF'
        }
```

In example 1, the first condition of rule 1 is applied and the word 'ملايين' 'malAyiyn' 'millions' is INDEF. In example 2, the second condition of rule 1 is applied and the word 'ملايين' 'malAyiyn' 'millions' is EDAFAH. In example 3, the default of rule 1 is applied and the word 'جنيهات' 'junayohAt' 'pounds' is INDEF.

Ex. [1] وكانت تكلفته 9 ملايين

wakAnatotakolifathu 9 malAyiyn

Ex. [2] وكانت تكلفته 9 ملايين جنيه

wakAnatotakolifathu 9 malAyiynjunayoh

Ex. [3] وكانت تكلفته 9 جنيهات

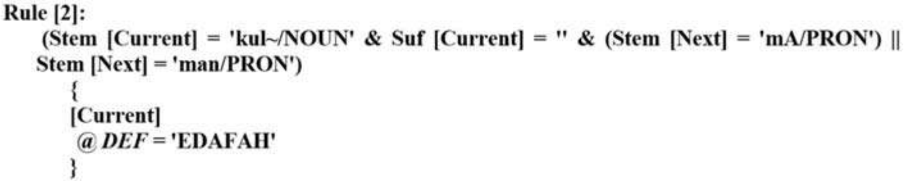
wakAnatotakolifathu 9 junayohAt

Although this part is dedicated for setting the indefiniteness feature, but in some cases to avoid the overlapping, some words that are EDAFAH are assigned here.

2.3 <u>Rules for Setting the Word as (EDAFAH)</u>

In this part, some rules for setting the word as EDAFAH have been extracted and formulated. The simplest rule related to describing the analyzed nominal, adjectival or adverbial words as EDAFAH when the suffix of the word contains ‘ا’ ‘A/N_SUF’, ‘ي’‘ayo/N_SUF’ or ‘iy/N_SUF’, ‘تي’ ‘atayo/N_SUF’ ‘تا’ ‘atA/N_SUF’, ‘و’ ‘uw/N_SUF’ or if it contains a possessive pronoun suffix ‘POSS_SUF’. Examples of these words is ‘رجلا’ ‘rajulA’ ‘man’, ‘حلقتا’ ‘HalaqatA’ ‘rings/circles’, ‘مسلمو’ ‘musolimuw’ ‘Muslims’ and ‘قلمه’ ‘qalamhu’ ‘his pen/pencil’.

One of the extracted rules related to the EDAFAH states that if the word to be assigned EDAFAH is ‘كل’ ‘kul~’ ‘all’ and the next word is ‘mA/PRON’ or ‘man/PRON’ then this word is assigned as EDAFAH as rule 2 shows:



```
Rule [2]:
    (Stem [Current] = 'kul~/NOUN' & Suf [Current] = '' & (Stem [Next] = 'mA/PRON') ||
   Stem [Next] = 'man/PRON')
        {
        [Current]
         @ DEF = 'EDAFAH'
        }
```

In example 4, the rule 2 is applied and the word 'كل' 'kul~' 'all/every' is EDAFAH.

Ex. [4] الحج فرض على كل من يستطيع

AloHaj~ufaroDNEalaYkul~imanoyasotaTiyEu

After applying all the EDAFAH rules, a default value is assigned for other words that are not governed by any rules. This default value is selected depending on what is more frequent in the data set. The selected default value is ‘EDAFAH’. For more details about the extracted rules to identify the definiteness feature, see [25].

3. Extracted and Formulated Case Ending Rules:

After detecting the stem pattern of each word that have root in Arabic, detecting the transitivity of each verb in dicStems of BAMA and extracting rules for detecting the definiteness feature, the rules for detecting the case ending are extracted and formulated in a generalized format.

The extracted rules here may be divided into five main categories; rules for detecting the imperfect verb case ending, rules for detecting the case ending of nouns, rules for detecting the case ending of some proper nouns, rules for detecting the case ending of adverbs and finally rules related to detecting the case ending of adjectives. These rules take into consideration the phrases in which each of the previous categories occur.

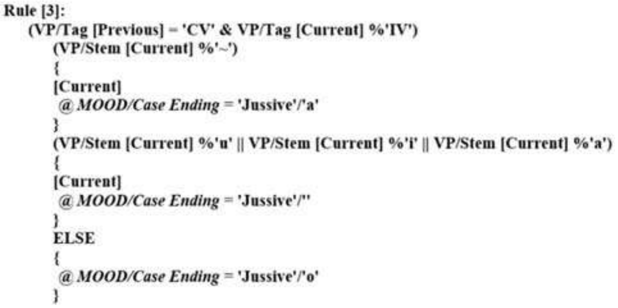
3.1 <u>Rules for Detecting Case Ending (MOOD) of the Imperfect Verb:</u>

The imperfect verbs in Arabic are influenced by subjunctive and jussive particles and command verbs. In the current proposed system, if the imperfect verb is not preceded by any of these conditions, the default mood in this case will be the indicative mood 'مرفوع'. Only imperfect verbs that contain a suffix of third masculine/feminine singular subject, its mood needs to be detected as well as its case ending diacritics.

Although it seems that the rules for imperfect verbs are easy to extract, this is not the case where there are some limitations that violate the rules since the rules are limited to use a window of -3 words before the focused word. For example, the word ‘ينجح’ ‘yanojaH’ ‘succeed’ in a sentence like ‘مَن يُذَاكِر دُرُوسهُ بِجِد دَائِما يَنْجَح’ ‘manoyu*Akirdurwshubijid~ dA}imAyanojaH’ ‘who studied his lesson hardly always will succeed’ is in jussive mood where it is influenced by the jussive particle ‘مَن’ ‘man’ ‘who’, but unfortunately it will be assigned as indicative. In addition to the previous limitation, there is another limitation where there are some jussive particles that are also used as indicative particles as ‘مَن’ ‘man’ ‘who’ and ‘لَا’ ‘lA’ ‘not’. For example, the word ‘من’ in sentence like ‘مَن يُذَاكِرْيَنْجَحْ’ ‘manoyu*AkiroyanojaH’ ‘who study will succeed’is a jussive particle while in a sentence like ‘جَاءَ مِن يُذَاكِرُ بَجِدٍ’ ‘jA'a man yu*Akirubijid~K’ ‘who study hardly came’ it is an indicative particle. So, in some cases the rules will be able to detect the mood correctly, but in other cases it will fail.

There are a number of rules that have been extracted and formulated to be implemented in the current proposed system related to the imperfect verbs. One of these rules states that if there is a verb phrase begins with an imperfect verb that its mood and case ending diacritic is not detected yet and it is preceded by another verb phrase that contains only a command verb, then the IV must be in jussive mood, but its case ending depends on the type of the verb:

1. **The geminated verb ‘مضعّف الآخر’:** the case ending of the jussive mood is fatha ‘◌َ’ ‘a’.
2. **The defective verbs ‘الفعل المنقوص’:** in the case of jussive mood any verb that ends with any of the weak letters ‘حروف العلة’; ‘و’ ‘w’ preceded by damma ‘◌ُ’ as in ‘يَغْزُو’ ‘yagozuw’ ‘invade/conquer’, ‘ي’ ‘y’ preceded by kasra ‘◌ِ’ as in ‘يَرْمِي’ ‘yaromiy’ ‘throw/fling’ or ‘ى’ ‘Y’ preceded by fatha ‘◌َ’ as in ‘يَخْشَى’ ‘fear/be afraid’ the case ending should be empty, since these weak letters are omitted, but its previous diacritics is left [31].
3. Otherwise, the default case ending diacritic of the jussive mood is sukuun ‘◌ْ’ ‘o’as rule 3 shows:



```
Rule [3]:
   (VP/Tag [Previous] = 'CV' & VP/Tag [Current] %'IV')
      (VP/Stem [Current] %'~')
      {
      [Current]
       @ MOOD/Case Ending = 'Jussive'/'a'
      }
      (VP/Stem [Current] %'u' || VP/Stem [Current] %'i' || VP/Stem [Current] %'a')
      {
      [Current]
       @ MOOD/Case Ending = 'Jussive'/''
      }
      ELSE
      {
       @ MOOD/Case Ending = 'Jussive'/'o'
      }
```

In example 5, the first condition of rule 3 is applied and the case ending diacritic of the word ‘يُحِبَّكَ’ ‘yuHib~aka’ ‘love you’ is fatha ‘◌َ’ ‘a’. In example 6, the second condition of rule 3 is applied and the case ending diacritic of the word ‘تَعْلُ’ ‘taEolu’ ‘rise/be elevated’ is empty due to the omission of the last letter ‘و’ ‘w’where it was ‘تَعْلُو’ ‘taEoluw’. In example 7, the default of rule 3 is applied and the case ending diacritic of the word ‘تَنْجَحْ’ ‘tanojaHo’ ‘succeed’ is sukuun ‘◌ْ’ ‘o’.

Ex. [5] تَصَدَّقْ يُحِبَّكَ اللهُ
taSad~aqoyuHib~aka All~`hu

Ex. [6] اِجْتَهِدْ تَعْلُ فِي الْعِلْمِ
{ijotahidotaEolufiyAloEilomi

Ex. [7] ذَاكِرْ تَنْجَحْ
*AkirotanojaHo

3.2 Rules for detecting the case ending of noun phrases:

As mentioned before, all nouns, proper nouns and adjectives have one of three cases: nominative (NOM), accusative (ACC), or genitive (GEN). It must be noted that there are some cases in which they must not receive case ending diacritic:

- Nouns and adjectives ending by plural or dual suffixes such as 'ون'‘uwna’, 'ين' 'iyna', 'ين' 'ayoni', 'و' 'uw', 'ي' 'iy', 'تان' 'atAni' … etc.
- The defective noun, proper nounor adjective, if its definiteness is 'DEF' or 'EDAFAH' and its case is genitive or nominative, for example, the word 'الْقَاضِي' 'AloqADiy' 'the lawyer'. The extracted and implemented rules here are related to detecting the case ending or syntactic diacritic for words excluding these words’ types.

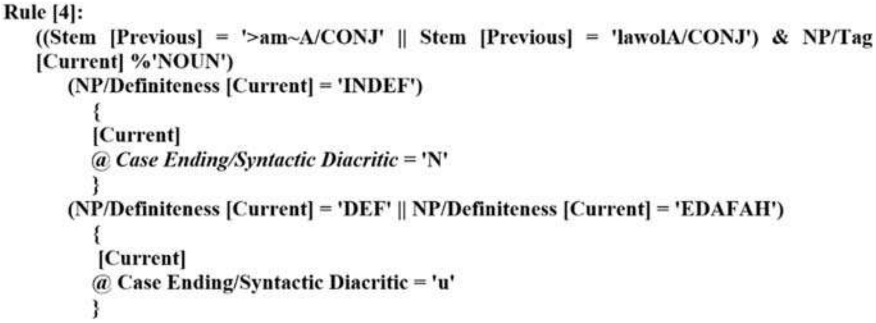
The noun phrases are those phrases starting with noun or proper noun. In this part, the extracted and implemented rules have been classified according to their function or syntactic case in the sentence.

3.2.1 Rules for Detecting the Case Ending of Nouns

In this stage, rules for assigning the case ending or syntactic case for nouns are divided into three categories; i) rules for setting the case ending as nominative, ii) rules for setting the case ending as accusative and iii) rules for setting the case ending as genitive.

One of these rules states that if there is a noun phrase preceded by 'أَمَّا' '>am~A' 'as for/concerning' or 'لَوْلَا' 'lawolA' 'if not' and the case ending diacritic is not detected yet, then the noun must be in nominative case taking into consideration the definiteness feature:

If the noun is 'INDEF' then the noun must receive *nunation* (TanweenDamma 'ٌ''), but if the noun is 'DEF' or 'EDAFAH' then the noun must receive Damma "ُ" as rule 4 shows:



```
Rule [4]:
    ((Stem [Previous] = '>am~A/CONJ' || Stem [Previous] = 'lawolA/CONJ') & NP/Tag
    [Current] %'NOUN')
        (NP/Definiteness [Current] = 'INDEF')
            {
            [Current]
            @ Case Ending/Syntactic Diacritic = 'N'
            }
        (NP/Definiteness [Current] = 'DEF' || NP/Definiteness [Current] = 'EDAFAH')
            {
             [Current]
            @ Case Ending/Syntactic Diacritic = 'u'
            }
```

In example 8, the first condition of rule 4 is applied and the case ending diacritic of the word 'صَدِيقٌ' ‘SadiyqN’ ‘friend’ is Tanween Damma 'ٌ' 'N'. In example 9, the second condition of rule 4 is applied and the case ending diacritic of the word 'صَدِيقُ' ‘Sadiyqu’ ‘friend’ is Damma 'ُ' 'u'.

Ex. [8] لَوْلَا صَدِيقٌ لِي أَخْبَرَنِي بِالْأَمْرِ

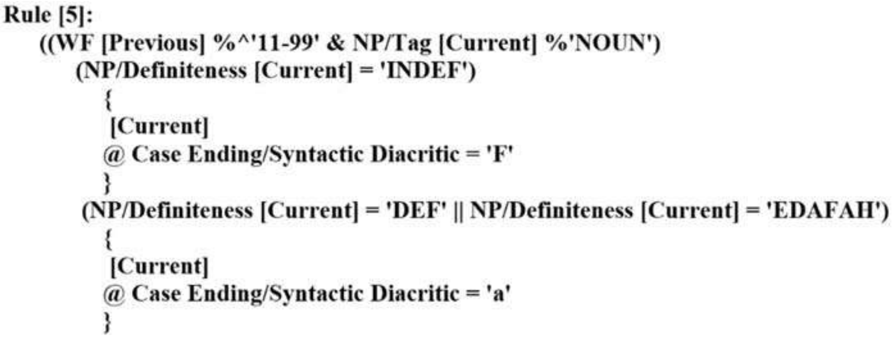
lawolASadiyqNliy>axobaraniybiAlo>amori

Ex. [9] أَمّا صَدِيقُ طُفُولَتِي فَلَهُ مِنِّي كُلُّ الْحُبِّ

>am~ASadiyquTufuwlatiyfalahumin~iykul~uAloHub~i

Another rule states that if there is a noun phrase preceded by a number ending by 11-99 and the case ending diacritic is not detected yet, then the noun must be in accusative case taking into consideration the definiteness feature:

If the noun is 'INDEF' then the noun must receive *nunation* (TanweenFatha 'ً'), but if the noun is 'DEF' or 'EDAFAH' then the noun must receive Fatha 'َ' as rule 5 shows:



```
Rule [5]:
    ((WF [Previous] %^'11-99' & NP/Tag [Current] %'NOUN')
        (NP/Definiteness [Current] = 'INDEF')
            {
             [Current]
            @ Case Ending/Syntactic Diacritic = 'F'
            }
         (NP/Definiteness [Current] = 'DEF' || NP/Definiteness [Current] = 'EDAFAH')
            {
             [Current]
            @ Case Ending/Syntactic Diacritic = 'a'
            }
```

In example 10, the first condition of rule 5 is applied and the case ending diacritic of the word 'نَبْتَةً''nabotapF' 'sprout/seedling' is TanweenFatha 'ً' 'F'. In example 11, the second condition of rule 5 is applied and the case ending of the word 'طَالِبَ' 'TAliba' ‘student’ is Fatha 'َ' 'a'.

Ex. [10] اِشْتَرَيْتُ 30 نَبْتَةً لِزِرَاعَتِهَا

{i$otarayotu 30 nabotapFlizirAEatihA

Ex. [11] جَاءَ 1556 طَالِبَ عَلْمٍ وَافِدًا إِلَى مِصْرَ

jA'a 1556 TAlibaEilomKwAfidAF<ilaYmiSora

3.2.2 Rules for Detecting the Case Ending of Proper Nouns

The Proper Noun is the definite noun that refers to a specific name of someone, something or some place. It is the given name and it does not need any other word or syllable to be added to specify it. It is definite by itself.

In Arabic, the proper nouns can be a single name (one word), such as, ‘مصر’ ‘miSor’ ‘Egypt’, ‘محمّد’ ‘muHam~ad’ ‘Mohammed’, or ‘عطارد’ ‘EuTArid’ ‘Mercury’, and it can also be a compound name (two words), such as, ‘عبد الله’ ‘EabodAll~`h’ ‘Abdu Allah’, ‘أبو بكر’ ‘>abuwbakor’ ‘Abo Bakr’.

As concerning for detecting the case ending for the proper nouns, the Arabic Language Assembly states two opinions for setting the case ending for the consecutive proper nouns [31]:

1. The first opinion states that: if the first proper noun of a sequence of consecutive proper nouns is the first name of a person and the others are the father name, grand father name, nick name, … etc., then the first one is assigned its syntactic case according to its context and the other proper names are EDAFAH, thus assigned a genitive case ending. The reason for considering these proper nouns as an EDAFAH is that in the ancient Arabic there was a word between each of these nouns; 'بن' 'bin' 'son of', but in modern Arabic it is deleted which is accepted linguistically.
2. The second opinion states that: since it is difficult in most cases to detect whether the consecutive proper nouns are belonging to the same person or more than one person, then it is preferred to deal with these proper nouns by assigning a sukuun 'o' 'ْ' at the end of each one of these proper nouns. This approach depends on an idea that suggests that these proper nouns should be dealt with as if they are followed by a pause.

"يجيز المجمع ما يجري على الألسنة من حذف (ابن) من الأعلام المتتابعة في مثل: سافر محمد علي حسن ، وتضبط هذه الأعلام على أحد الوجهين الآتيين:

الأول: يعرب العلم الأول حسب موقعه، ويجر ما يليه بالإضافة.

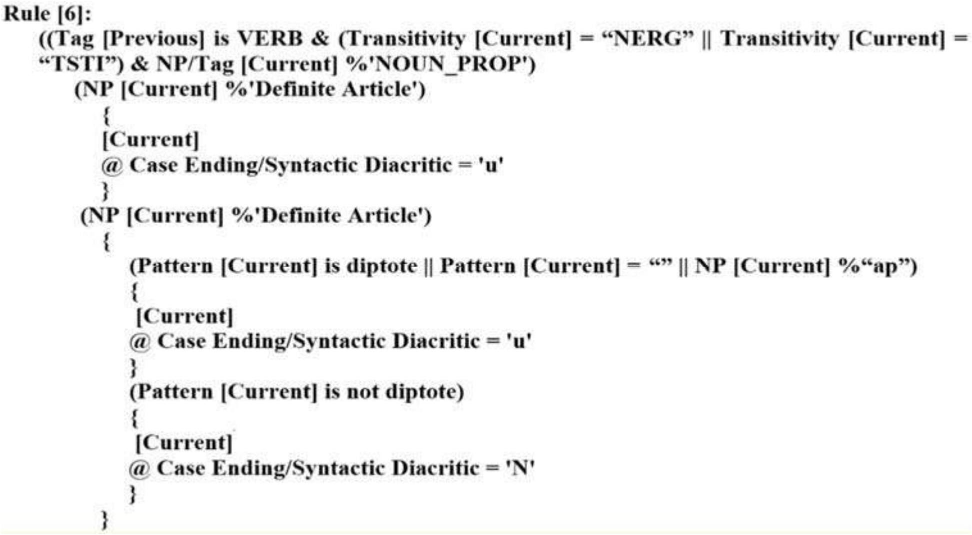
الثاني: تسكن الأعلام كلها إجراءً للوصل مجرى الوقف)."

In the process of selecting one of these approaches to be followed in assigning the case ending of the proper nouns, we found that the used corpora for both the training and testing follow the second approach for assigning the case ending of the proper nouns without assigning any case ending. Consequently, for achieving the best results the same approach has been followed.

It must be noted that although the case ending of the proper nouns will be *sukuun*, we need to extract some rules for determining the case ending or the syntactic case, since they will be used in detecting the case ending or syntactic case of the succeeding words. For example, in the appositive relations 'التّوابع' 'Alt~awAbiE'; adjectives, coordinated elements and nouns in apposition are assigned the same case ending of the preceding element.

One of these rules states that if there is an intransitive or indirect transitive verb is followed by a proper noun, which its case ending is not detected yet, then the proper noun must in the nominative case, taking into consideration the pattern feature and if there is 'ال' 'Al' definite article or not. All proper nouns are definite, however some proper nouns could be prefixed by 'ال' 'Al' definite article, such as 'القاهرة' 'AloqAhirap' 'Cairo':

If the proper noun contains ‘ال’ ‘Al’ the definite article, then the proper noun must receive Damma 'u' 'ُ'. But, if the proper noun does not contain ‘ال’ ‘Al’ the definite article and its pattern is not a diptote pattern, then it must be assigned with nominative *nunation* (Tanween Damma 'ٌ'). Moreover, if the proper noun does not contain ‘ال’ ‘Al’ the definite article, its pattern is a diptote pattern, and does not have pattern or contains taa marbouta ‘ة’, then it should not be assigned with a *nunation* and in this case its case is(Damma 'ُ') as rule 6 shows:



```
Rule [6]:
  ((Tag [Previous] is VERB & (Transitivity [Current] = “NERG” || Transitivity [Current] =
  “TSTI”) & NP/Tag [Current] %'NOUN_PROP')
     (NP [Current] %'Definite Article')
        {
        [Current]
        @ Case Ending/Syntactic Diacritic = 'u'
        }
      (NP [Current] %'Definite Article')
        {
           (Pattern [Current] is diptote || Pattern [Current] = “” || NP [Current] %“ap”)
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'u'
           }
           (Pattern [Current] is not diptote)
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'N'
           }
        }
```

In example 12, the first condition of rule 6 is applied and the case ending diacritic of the word ‘الْقَاهِرَةُ’ ‘AloqAhirapu’ ‘Cairo’ is Damma 'ُ' 'u'. In example 13, 14 and 15 the second condition of rule 6 is applied and the case ending of the words ‘أَحْمَدُ’ ‘>aHomad’ ‘Ahmed’, ‘إِبْرَاهِيمُ’ ‘<iborAhiym’ ‘Ebraheem’ and ‘أُسَامَةُ’ ‘>usAmap’ ‘Ossama’ is Damma 'ُ' 'u'. In example 16, the third condition of rule 6 is applied and the case ending of the word ‘مُحَمَّدٌ’ ‘muHam~ad’ ‘Mohammed’ is Tanween Damma ‘ٌ’ ‘N’.

Ex. [12] اِعْتَذَرَتْ الْقَاهِرَةُ عَنْ عَدَمِ اِسْتِضَافَةِ حِوَارِ الْمُصَالَحَةِ
{iEota*aratoAloqAhirapuEanoEadami {isotiDAfapiHiwAriAlomuSAlaHapi

Ex. [13] ذَهَبَ أَحْمَدُ إِلَى الْمَدْرَسَةِ
*ahaba>aHomadu<ilaYAlomdorasapi

Ex. [14] نَامَ إِبْرَاهِيمُ مُبَكِّرًا
nAma<iborAhiymumubak~irAF

Ex. [15] سَكَتَ أُسَامَةُ
sakata>usAmapu

Ex. [16] اِنْفَعَلَ مُحَمَّدٌ كَثِيرًا
{inofaEalamuHam~adNkaviyrAF

Such rules help in detecting the case ending of adjectives or appositions if they follow the proper nouns as example 17 shows:

Ex. [17] اِعْتَذَرَتْ مَادْلِينْ أُولْبِرَايِتْ وَزِيرَةُ الْخَارِجِيَّةِ الْأَمْرِيكِيَّةِ عَنْ مَا حَدَثَ
{iEota*aratomAdoliyn>uwlobiyrAyitwaziyrapuAloxArijiy~apiAlo>amoriykiy~apiEano mA Hadava

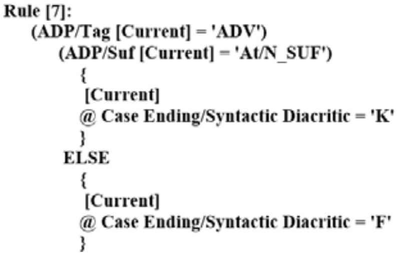
In the previous example, the noun phrase ‘وَزِيرَةُ الْخَارِجِيَّةِ الْأَمْرِيكِيَّةِ’ ‘waziyrapuAloxArijiy~apiAlo>amoriykiy~api’ ‘US Secretary of State’ is in apposition relation with the noun phrase ‘مَادْلِينْ أُولْبِرَايِتْ’ ‘mAdoliyn>uwlobiyrAyit’ ‘Madeleine Albright’. In this example, the word ‘وَزِيرَةُ’should have the same syntactic case of ‘مَادْلِينْ’ which is in this case the nominative case Damma ‘ُ’ ‘u’.

3.3 Rules for Detecting the Case Ending of Adverbs:

The circumstantial accusative in traditional Arabic grammar is known as HAl (حال). A word in this syntactic role describes the circumstances under which an action takes place. The dependent word in the HAl relation is in the accusative case [32].

In this part, the extracted rule is easy to detect where there is only one rule with two conditions for detecting the case ending for adverbs. This rule states that if there is an adverb phrase (ADP) and its case ending diacritic is not detected yet, then this adverb must be in accusative case taking into consideration that the definiteness of the adverb is INDEF and plural feminine suffix 'ات' 'At':

If the adverb contains feminine plural suffix'ات', then the noun must receive Tanween Kasra '◌ٍ' 'K' not Tanween Fatha '◌ً' 'F', else the default case is *nunation* (Tanween Fatha '◌ً' 'F') as rule 7 shows:



```
Rule [7]:
    (ADP/Tag [Current] = 'ADV')
        (ADP/Suf [Current] = 'At/N_SUF')
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'K'
           }
         ELSE
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'F'
           }
```

In example 18, the first condition of rule 7 is applied and the case ending diacritic of the word 'مُسْرِعَاتٍ''musoriEAtK' 'hurrying/in a hurry' is TanweenKasra '◌ٍ''K'. In example 19, the default condition of rule 7 is applied and the case ending of the word 'مُسْرِعَةً''musoriEapF' 'hurrying/in a hurry' is Tanween Fatha 'F' '◌ً'.

Ex. [18] جَاءَتْ الْفَتَيَاتُ مُسْرِعَاتٍ
jA'atoAlofatayAtumsoriEAtK

Ex. [19] جَاءَتْ الْبِنْتُ مُسْرِعَةً
jA'atoAlobinotumusoriEapF

3.4 Rules for Detecting the Case Ending of Adjectives

The extracted and implemented rules for adjectival phrases depend primarily on the syntactic case of the preceding nominal phrases (a proper noun or noun) since any adjective agrees with the noun it depends on in the syntactic case; nominative, genitive or accusative. It also agrees in terms of gender, number and definiteness. More than one adjective can depend on the same noun.

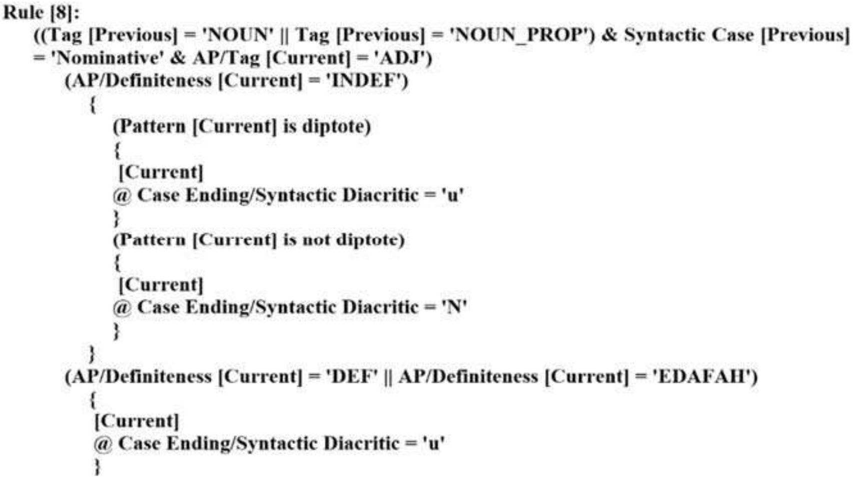
When extracting and implementing the rules for adjectival phrases it must be noted that an adjective depends primarily on the syntactic diacritic of the preceding noun phrase not the case ending diacritic.

In example 20, the word 'أَمَاكِنَ' '>amAkina' 'places' is an indefinite diptote noun that is preceded by a preposition, so its case ending must be Fatha '◌َ' 'a' notTanweenKasra 'K' '◌ٍ'. However, when trying to detect the case ending of the word 'عَدِيدَةٍ''EadiydapK' 'numerous/many' which is an adjective in this case, it must be taken into consideration that it is in genitive case as the preceding noun, but it does not receive the same case ending Fatha 'a' '◌َ'. In this example the word 'عَدِيدَةٍ' receives Tanween Kasra 'K' '◌ٍ' since it is indefinite not diptote word. Consequently, it agrees with the preceding noun in syntactic case; genitive but not in case ending diacritics.

Ex. [20] سَافَرَ فِي تِلْكَ الْفَتْرَةِ إِلَى أَمَاكِنَ عَدِيدَةٍ
sAfarafiytilokaAlofatorapi<ilaY>amAkinaEadiydap

There are a number of rules that have been extracted and implemented for detecting the case ending of the adjectival phrases. One of these rules states that if there is an adjectival phrase preceded by a nominative noun and its case ending diacritic is not detected yet, then this adjective must be in nominative case taking into consideration the definiteness and the pattern features:

If the adjective is 'INDEF' and the pattern is from the diptote patterns then the noun must receive Damma 'u' '◌ُ', but if the noun is 'INDEF' and the pattern is not from the diptote patterns then the adjective must receive *nunation* (Tanween Damma '◌ٌ'). Moreover, if the noun is 'DEF' or 'EDAFAH' then the adjective must receive Damma 'u' '◌ُ' as rule 8 shows:



```
Rule [8]:
  ((Tag [Previous] = 'NOUN' || Tag [Previous] = 'NOUN_PROP') & Syntactic Case [Previous]
  = 'Nominative' & AP/Tag [Current] = 'ADJ')
     (AP/Definiteness [Current] = 'INDEF')
        {
           (Pattern [Current] is diptote)
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'u'
           }
           (Pattern [Current] is not diptote)
           {
            [Current]
           @ Case Ending/Syntactic Diacritic = 'N'
           }
        }
     (AP/Definiteness [Current] = 'DEF' || AP/Definiteness [Current] = 'EDAFAH')
        {
         [Current]
         @ Case Ending/Syntactic Diacritic = 'u'
        }
```

In example 21, the first condition of rule 8 is applied and the case ending diacritic of the word 'خَضْرَاءُ' 'xaDorA'u' 'green' is Damma 'ُ◌' 'u'. In example 22, the second condition of rule 8 is applied and the case ending of the word 'مُسِنٌّ' 'musin~N' 'old aged/senior/superadult' is Tanween Damma 'N' 'ٌّ◌'. In example 23, the third condition of rule 8 is applied and the case ending of the word 'الطَّيِّبَةُ' 'AlT~ay~ibapu' 'good nature/goodness' is Damma 'ُ◌' 'u'.

Ex. [21] تَنْتَشِرُ بِهَا مِسَاحَاتٌ خَضْرَاءُ
Tanota$irubihAmisAHAtNxaDorA'u

Ex. [22] قَامَ رَجُلٌ مُسِنٌّ بِجَمْعِ تَبَرُّعَاتٍ كَثِيرَةٍ
qAmarajulNmusin~NbijamoEitabar~uEAtKkaviyrapK

Ex. [23] الْكَلِمَةُ الطَّيِّبَةُ كَالشَّجَرَةِ الطَّيِّبَةِ
AlokalimapuAlT~ay~ibapukaAl$~ajarapiAlT~ay~ibapi

For more details about the extracted rules for detecting the case ending, see [25]. Table 1 shows the number of the Arabic extracted linguistic rules for being used for syntactic processing levels:

TABLE 1:
EXTRACTED RULES FOR SYNTACTIC PROCESSING LEVEL.

Rules Type	Rules No.
Syntactic Rules	473
Definiteness Rules	46
Total No of Rules	**420**

4 THE CURRENT PROPOSED SYSTEM

The task of the syntactic processing level is to predict the syntactic case of a sequence of morphologically diacritized words given their POS tags, definiteness, stem pattern and/or transitivity and hence assigning the suitable case ending diacritics. Some limitations violate the rules for setting the case ending of syntactic diacritic, since the rules are limited to use a window of -/+3 words before the focused word.

As mentioned before, the output of the morphological processing level from a system we have developed is used. The morphological algorithm in this system makes use of the relations between the words and their contexts, whether the preceding or the succeeding words. In addition, the adopted syntactic algorithm is a rule-based approach that simulates one of the language processing approaches that computes a basic analysis of sentence structure rather than attempting full syntactic analysis; shallow syntactic parsing. It is an analysis of a sentence which identifies the constituents (noun groups, verb groups, prepositional groups adjectival groups, adverbial groups), but does not specify their internal structure, nor their role in the main sentence. Before diacritizing the word syntactically, its POS tag is checked first. Using the POS tag of the word, it is decided how the syntactic diacritization of this word should be handled. Figure 3 shows the general design of the adopted algorithm for the current proposed system.

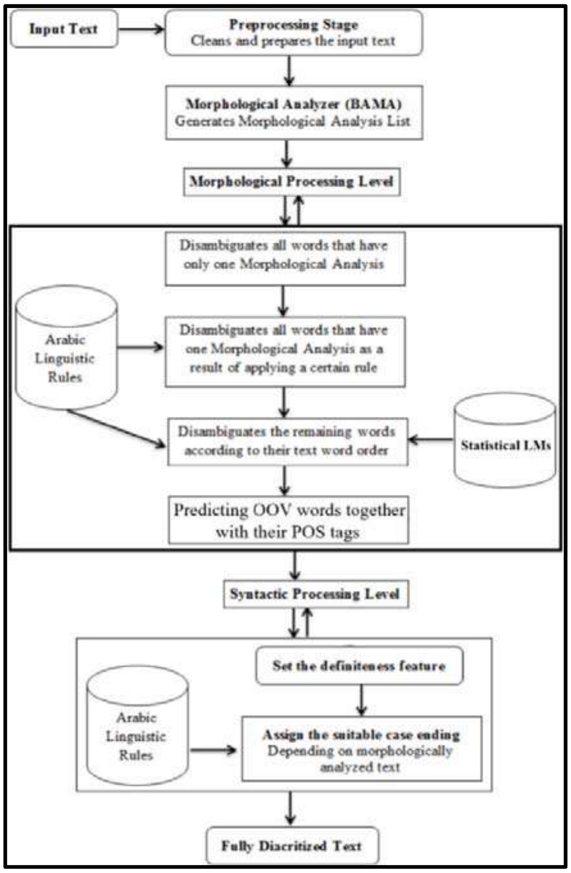
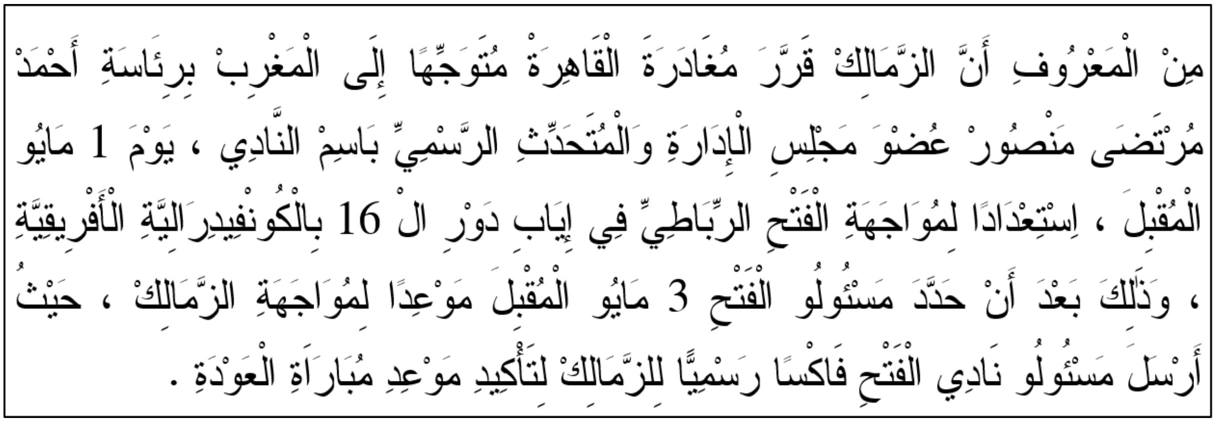

Input Text
Preprocessing Stage
Cleans and prepares the input text
Morphological Analyzer (BAMA)
Generates Morphological Analysis List
Morphological Processing Level
Disambiguates all words that have only one Morphological Analysis
Arabic Linguistic Rules
Disambiguates all words that have one Morphological Analysis as a result of applying a certain rule
Disambiguates the remaining words according to their text word order
Statistical LMs
Predicting OOV words together with their POS tags
Syntactic Processing Level
Set the definiteness feature
Arabic Linguistic Rules
Assign the suitable case ending
Depending on morphologically analyzed text
Fully Diacritized Text

Figure 3: General Design of the Adopted Algorithm for the Proposed System

The extracted rules for detecting the imperfect verb case ending, the case ending of noun, the case ending of adjectives, the case ending of adverbs and the case ending of some proper nouns have been implemented in the current proposed system, taking into consideration the phrases in which each of the previous categories occur. In this level, the validation data set is used in validating the extracted rules. Figure 4 shows the output results of an example after implementing the case ending rules:



مِنْ الْمَعْرُوفِ أَنَّ الزَّمَالِكْ قَرَّرَ مُغَادَرَةَ الْقَاهِرَةْ مُتَوَجِّهًا إِلَى الْمَغْرِبْ بِرِئَاسَةِ أَحْمَدْ
مُرْتَضَى مَنْصُورْ عُضْوَ مَجْلِسِ الْإِدَارَةِ وَالْمُتَحَدِّثِ الرَّسْمِيِّ بَاسِمْ النَّادِي ، يَوْمَ 1 مَايُو
الْمُقْبِلَ ، إِسْتِعْدَادًا لِمُوَاجَهَةِ الْفَتْحِ الرِّبَاطِيِّ فِي إِيَابِ دَوْرِ الْ 16 بِالْكُونْفِيدِرَالِيَّةِ الْأَفْرِيقِيَّةِ
، وَذَلِكَ بَعْدَ أَنْ حَدَّدَ مَسْئُولُو الْفَتْحِ 3 مَايُو الْمُقْبِلَ مَوْعِدًا لِمُوَاجَهَةِ الزَّمَالِكْ ، حَيْثُ
أَرْسَلَ مَسْئُولُو نَادِي الْفَتْحِ فَاكْسًا رَسْمِيًّا لِلزَّمَالِكْ لِتَأْكِيدِ مَوْعِدِ مُبَارَاةِ الْعَوْدَةِ .

Figure 4: The Output after Applying Syntactic Rules for Setting the Case Ending Diacritics

5 RESULTS AND EVALUATION

In this stage, the syntactic Word Error Rate are used to evaluate the system results. What is meant by the syntactic WER is the percentage of words that have at least one syntactic error but doesn'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.The testing data set has been used to evaluate the system results andthe best syntactic diacritization achieved by the system is 9.36%.

When checking some syntactic errors, it has been noticed that detecting POS, transitivity and definiteness wrongly leads to have wrong syntactic case. In addition to these features, the extracted and implemented rules from the training data set to detect the case ending depending on a window of -/+ 3 also limits the system in some cases from detecting the syntactic case correctly. In example [24], the word "هدف" "hadaf" "goal/target" has assigned definiteness as "EDAFAH" as a result of considering "وحيد" "waHiyd" "alone" as "NOUN_PROP". This problem leads to set the case ending as "◌ِ" "i" not "◌ٍ" "K".

Ex. [24] بَعْدَ عَدِمِ النَّجَاحِ فِي إِحْرَازِ هَدَفِ وَحِيدْ
baEodaEadamiAln~ajaHifiy<iHorAziHadafiwaHiydo

A. Comparing the System with Other State of the Art Systems

A comparison between the performance of syntactic results of the current proposed system and some of the most relevant state of the art systems is reported. 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 used testing data is a part of Arabic Tree Bank part 3 (ATB3) form "An-Nahar" Lebanese News Agency. It consists of 91 articles (about 52.000 word) covering the period from October 15, 2002 to December 15, 2002 [4].

Table 2 shows the comparison between the current proposed system and some of the state-of-the-art systems in terms of the syntactic diacritization.

TABLE 2:
COMPARISON OF THE SYNTACTIC DIACRITIZATION WER

Systems	Syntactic WER
Habash	9.4%
Zitouni	10.1%
Rashwan (2011)	12.5%
Rashwan (2015)	9.9%
Abandah, et al., (2015)	8.93%
Metwally, Rashwan, & Atiya (2016)	9.4%
Chennoufi & Mazroui (2016(b))	NA
Current System	9.97%

The comparison indicates that [14], [13] and [15]outperform the current proposed system's results in the syntactic diacritization. However, the current proposed system's syntactic results are still close to these results.

6 CONCLUSION

Most related works to syntactic diacritization depend in their systems on many of statistical approaches. The evaluation of the proposed system has been done using syntactic WER and it is made in comparison with other best state of the art systems. In order to have an objective evaluation of the system, the same testing data (LDC's Arabic Treebank) that was used in other systems was used to compare the results. The best syntactic diacritization achieved is 9.97% compared to the best-published results, of [14]; 8.93%, [13] and [15]; 9.4%.Since we work on the syntactic diacritization by following a rule-based approach as a trial to enhance the syntactic diacritization output, these results are expected to be enhanced by extracting more Arabic linguistic rules, adding more semantic features, using different machine learning techniques for syntactic processing level and implementing the improvements by working on larger amounts of data.

ACKNOWLEDGMENT

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

REFERENCES

[1] K. Shaalan, H. M. Abo Bakr & I. Ziedan, 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. Athens, Greece, (2009, March).

[2] M. Alghamdi, Z. Muzaffar & H. Alhakami, AUTOMATIC RESTORATION OF ARABIC DIACRITICS: A SIMPLE, PURELY STATISTICAL APPROACH. *The Arabian Journal for Science and Engineering*, Volume 35, Number 2C. (pp. 125-135), (2010, Decemcer).

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

[4] I. Zitouni, J. S. Sorensen, R. &Sarikaya, R. 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. (2006, July).

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

[6] N. Habash & O. Rambow, 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). Association for Computational Linguistics. Ann Arbor, (2005, June).

[7] N. Habash & 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 (pp. 53-56). Association for Computational Linguistics. Rochester, NY, (2007, April).

[8] R. Roth, O. Rambow, N. Habash, M. Diab & C. Rudin. 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). Association for Computational Linguistics. Columbus, Ohio, USA, (2008, June).

[9] N Habash, O. Rambow & R. Roth, 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, (2009, April).

[10] K. Shaalan, H. M. Abo Bakr & I. Ziedan, 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. Athens, Greece, (2009, March).

[11] M. A. Rashwan, A.A. Al Sallab, H. M. Raafat, & A. Rafea. 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. (2014, October).

[12] M. A. Rashwan, A.A. Al Sallab, H. M. Raafat, & A. Rafea. 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), (2015).

[13] A. S. Metwally, M. A. Rashwan, & 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).

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

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

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

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

[18] http://aramedia.com/nlp2.htm [Accessed 12-2-2015].

[19] http://aramedia.com/diacritizer.htm [Accessed 12-2-2015].

[20] http://www.rdi-eg.com/technologies/arabic_nlp.htm [Accessed 12-2-2015].

[21] http://tahadz.com/mishkal [Accessed 4-4-2015].

[22] http://harakat.ae/ [Accessed 4-4-2015].

[23] http://faraheedy.mukhtar.me/du2alee/tashkeel [Accessed 20-8-2016]

[24] S. Alansary, M. Nagi & N. Adly, Building an International Corpus of Arabic (ICA): progress of compilation stage. *In proceedings of the 7th International Conference on Language Engineering*, Cairo, Egypt, 5–6 December 2007.

[25] A. Fashwan. Automatic Diacritization ofModern Standard Arabic Texts: A Corpus Based Approach. *A Master's Thesis*. Alexandria, Egypt: Faculty of Arts, Alexanderia University, (2016).

[26] N. Habash, R. Gabbard, O. Rambow, S. Kulick & M. P. Marcus, Determining Case in Arabic: Learning Complex Linguistic Behavior Requires Complex Linguistic Features. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. (pp. 1084–1092). Prague, (2007, June).

[27] http://www.madinaharabic.com/lesson-33-2.html [Accessed 24-4-2016]

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

[29] http://www.unlweb.net/unlarium/index.php?action=export [Accessed 4-4-2016]

[30] http://www.unlweb.net/wiki/Transitivity [Accessed 4-4-2016]

[31] القرارات النحوية والتصريفية لمجمع اللغة العربية بالقاهرة: جمعًا ودراسة وتقويمًا، إلى نهاية الدورة الحادية والستينالمؤلف: خالد بن سعود بن فارس العصيمي دار التَّدْمُرِيَّة - الرياض، دار ابن حزم - لبنان، الطبعة الأولى (1424هـ - 2003 م).

[32] http://corpus.quran.com/documentation/cognateaccusative.jsp [Accessed 3-1-2016]

[33] Alansary, S. (2016). 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.

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.

منهج قاعدي لتحديد العلامات الإعرابية للنصوص في اللغة العربية المعاصرة

أماني فشوان[1]– سامح الأنصاري[2]

قسم الصوتيات واللسانيات – كلية الآداب – جامعة الإسكندرية – الإسكندرية - مصر

[1]amany.fashwan@bibalex.org

[2]sameh.alansary@bibalex.org

ملخص—تقدم هذه الورقة نظاما لتحديد العلامة الإعرابية للكلمات في اللغة العربية المعاصرة بالاعتماد على نظام قاعدي محاكٍ للتحليل السطحي النحوي للكلمات. ومن أجل معالجة مشكلة تحديد الحالة الإعرابية للكلمات تم الاعتماد على التحليل الصرفي للكلمات الذي يتم الاستفادة منه من خلال العلاقة بين الكلمة ووسمها الصرفي. كما تم الاعتماد على التحليل النحوي السطحي الذي يهتم بالعلاقة بين الكلمة وموقعها داخل الجملة من دون التعمق في تفاصيل هذه التحليل. ومن أجل بناء هذا النظام تم الاعتماد على عينة لغوية محللة تحليلا صرفيا وتحتوي على العلامة الإعرابية للكمات تتكون من حوالي 500 ألف كلمة تم اختيارها من المدونة اللغوية العربية العالمية. وقد تم استخدام هذه المدونة في عملية استخراج القواعد اللغوية التي تساهم في تحديد العلامة الإعرابية، كما تم استخدامها في عملية تطوير واختبار النظام المتبع. وقد تم تقييم النتائج الخاصة بالنظام بالاعتماد على قياس أساسي للتقييم هو معدل الخطأ على مستوى تشكيل العلامة الإعرابية والذي سجل 9.36%، كما 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 خطأ9.97%مقارنة بـ[14] الذي سجل أقل معدل خطأ 8.93% و[13]و[15]اللذين سجلا معدل خطأ 9.4%.

Lexical Growth in Egyptian Arabic Speaking Children: A corpus Based Study

Heba Salama[1], SamehAlansary[2]

Phonetics and linguistics Department, Faculty of Arts Alexandria University

[1]Heba.salama.slp@gmail.com

[2]Sameh.Alansary@bibalex.org

***Abstract*—This paper calculates developmental index of language growth in Egyptian Arabic based on a corpus consists of spontaneous speech samples from10 children 5 boys and 5 girls from 1.7 to 4 years. Depending on 30 minutes transcripts of spontaneous speech production the following properties of collected data were analysed: size of vocabulary and frequency of word use in relation to age (development of types and tokens) and individual differences in vocabulary size. The contribution of the current study lies in the use of vocabulary profile results as a measure of potential indicators of developmental language delay. The results provide a new measurement tool for lexical growth at different developmental stages.**

1 INTRODUCTION

This paper is a corpus-based study rooted in the so-called CHILDES (Child Language Data Exchange System) [1]. The objective of the present study is to describe the size of vocabulary in relation to age (development of types and tokens), identify individual differences in vocabulary size. The acquisition of the lexicon is a central and complex component of child language development. It has received growing interest within psycholinguistic research and is becoming a field of study in its own right. At the same time, lexical development also interacts with acquisition processes in other linguistic domains. Consequently, early lexical abilities might be indicative of following language skills. It is a question of crucial importance, whether certain abilities belong to the normal range of individual variation, or whether these abilities should rather be considered as an indication of a delayed or disturbed language acquisition process. To answer this question, empirically based knowledge about the normal course of vocabulary development within a particular language has first to be established. In this paper, a cross sectional study on early lexical development in Egyptian Arabic children is presented to analyze vocabulary growth. The results of this study provide an important new measure to help in assessment of language delay of Arabic language for children.

2 LEXICAL DEVELOPMENT

A. Vocabulary Growth

Normal development for young children's vocabularies has a number of applications in research design, assessment, and intervention that is very difficult to obtain before. While English and most Indo-European languages have a long tradition of examining aspects of child language production by computing different developmental indices from spontaneous language samples while, these aspects are lacking in this valuable area of research in Egyptian Arabic language. This may be partially because of the lack of longitudinal corpora of Arabic child language, large corpora or the appropriate set-up for experimental studies. Bridging this gap, this paper provides the first systematic cross-sectional study of the validity of TTR (type token ratio) developmental index in Egyptian Arabic language to measure vocabulary growth.

Lexical development is usually measured on the number of new words entering a child's vocabulary as they acquire a language. Statistical information is usually computed from spontaneous language samples of children in conversation or narrating a story. One of the first measures used in this context is the type-token ratio (TTR) or the ratio of new words (types) over the total number of words (tokens) in a speech sample. The categorical terms "types" and "tokens" are two important concepts in lexical analysis. If a text is 100 words long, we say it has 100 tokens. However within that text, there may be many words that are repeated, and as a result, there could be only 30 different words in the text in which case, we say there are 30 word types. The type vs. token ratio (TTR) is an important criterion that linguists use to evaluate lexical diversity in lexical analysis[2], introduced the index to child corpora and found a consistent ratio of around one different word for every two words uttered, independently of variables such as age range and gender. According to [3], "Templin's data are applicable only with children between the ages of 3 and 8 years". However, later work has shown that TTR depends on the size of the input transcript. That is, language samples which contain larger numbers of tokens give lower values for TTR and vice versa. As the children start producing longer utterances and language samples, a greater part of their acquired lexicon emerges and, as a result, the number of available new word types that could potentially be introduced decreases. Type and token frequency data, a major variable in psycholinguistic research, can be derived from corpora only. Language researchers and applied linguists working on a wide range of topics frequently need indices that quantify the range or number of different words in a text or conversation. Such measures are variously conceptualized as reflecting for example, lexical range and balance [4], total vocabulary size [5]. from a more negative perspective, particularly in the investigation of language disorders, measures are often seen as an index of repetitiveness manifested as perseveration in dementia and schizophrenia; speech automatisms in aphasia; echolalia in autism and mental deficiency. [3] cited [6] and admit that the values compared when determining TTR, the total number of words in a specified language sample and the total number of different words in the same language sample, are most valuable for evaluating the appropriateness of the child's vocabulary development. According to [3], "reductions in the total number of different words and the total number of words have been implicated as potential indicators of developmental language delays or disorders.

In this study the quantitative measures are used rather than qualitative to give general insight into the number of words known, but do not distinguish them from one another based on their category or frequency in language use. They have developed to make up for the widely applied measure type-token-ratio (TTR).This paper will proceed as follows: we will review a related work in Section 3. Method and procedures will be explained in Section 4. Next, we will present the results in Section 5, discuss them, and conclude with a summary of the importance of our findings in Section 6 and 7.

3 RELATED WORK

The following section presents a brief outline of some relevant findings in vocabulary development of children literatures.[7], as cited in [3]reported that at 18 months, normally-developing children had the ability to produce 22 meaningful and different words[8]and[9], as cited in[3], argued that the normally-developing 18 month-old child could produce nearly 50 different words[10], as cited by [3] asserted that the mean age at which children typically were capable of producing 50 different words was 19.75 months[11], [10] as cited by [7], as cited by [3]reported that the typically-developing 21 month-old toddler produced roughly 118 different words. [12] found that a control group of children, who were of mean age of 23 months, produced an average of 189.5 words. According to [13], toddlers between the ages of 23-25 months received total vocabulary scores of 196.24 words[14], as cited by [3]reported that the expressive vocabulary size of typically developing 2 year-old children was, at least 150 words.[15]found that at 30 months, children produced an average of 264.50 words. [16], as cited by [3] conducted a study obtaining TTR values for children under 3 years of age; however she only reported the TTR values, and not the total number of words or total number of different words necessary to compute TTR. There are a number of references showing typical TTR values for children across early development. [2], as cited in[3] norms for children between the ages of 3; 0-3; 5 shows total average of words 204.9 and total average of different words 92.5 TTR 0.45 while the mean score of the data collected from children between the ages of 2; 4 and 2; 9 is 140 and total average of different words 62.2 with TTR 0.447. In Emirati Arabic[17] found that there is no correlation between age and TTR was found in six children.

4 METHOD

The pre-compiled corpora include cross-sectional studies investigating the speech of 10 children in certain activity contexts (e.g. toy-playing-asking question, describe pictures). Five boys and five girls were selected randomly with no language delay from a nursery in Alexandria ranged from 1.6 to 4 years with a mean age 2.77 transcribed in chat format [1]. The children were split into five age groups each group contain (one boy-one girl) as shown in Table1. All children were normal and their first language is Arabic. The corpora contain 25,645 transcribed words from all 10 transcribed chat files. The summary of statistics of the corpus data is shown in Table2.The command in CLAN program that was used in the current research is FREQ (frequency) command. This command stands for Frequency Analysis. It is powerful and quite flexible, permitting frequency analysis. FREQ counts the frequencies of words used in selected files. FREQ produces a list of all the words used in the file, along with their frequency counts, and calculates a type–token ratio. The type–token ratio found by calculating the total number of unique words used by a selected speaker (or speakers) and dividing that number by the total number of words used by the same speaker(s). It is generally used as a rough measure of lexical development. The following command **freq +t*CHI farah.cha** looks specifically at a child's tier. The output printed in the CLAN window comes in alphabetical order. Using this command, researchers can count the number of words appearing in selected files; in addition, the ratio of different words (Types) to the total number of words (Tokens) Type-Token ratio (TTR) of words can be reported.

TABLE I
AGE RAND GROUPS

No	Age Range	Number of Children
1	1.6 -2	One boy and one girl
2	2 – 2. 6	One boy and one girl
3	2.6 – 3	One boy and one girl
4	3 – 3.6	One boy and one girl
5	3.6 – 4	One boy and one girl
Total	10	
mean	2.77	

TABLE 2
THE SUMMARY OF STATISTICS OF THE CORPUS DATA

Age	Number of investigator items	Number of child items	Total
19	1070 Mot, Inv311	376	1765
21	1122	385	1507
26	1448 Mot, Inv 760	699	2914
28	1882	1109	2987
34	1351	627	1978
36	1949	1269	3226
41	657	2407	3059
42	1380	1323	2701
43	1686	1162	2844
44	1252	1414	2664
Total	14,868	10,777	25,645

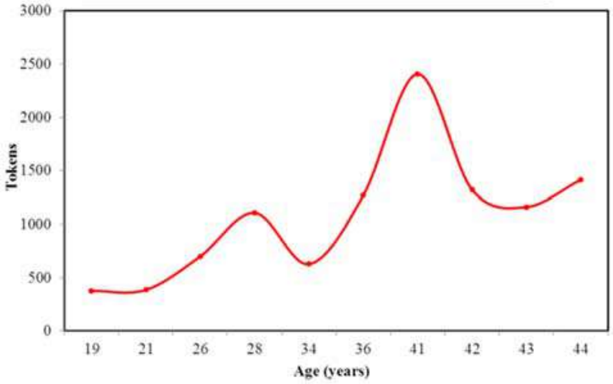


Figure 1: Growth patterns of age and word(tokens)

TABLE 3
FONT SIZES FOR PAPERS

Age	Types	Tokens	TTR
19	165	376	0.43
21	104	385	0.27
26	251	699	0.35
28	378	1109	0.34
34	333	627	0.53
36	373	1269	0.29
41	847	2407	0.35
42	547	1323	0.41
43	455	1162	0.39
44	546	1414	0.38

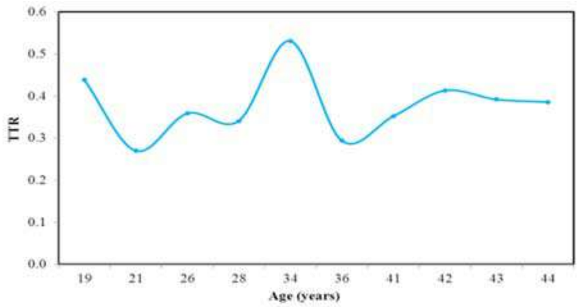


Figure 2: Growth patterns of age and TTR

5 DISCUSSION OF RESULTS

The results presented in Table 3 on spontaneous use of words in 10 Egyptian children contribute to show a picture of the process of lexical development in Egyptian children from the 19 months to four years. The results of vocabulary size in the groups can be illustrated by the language profiles of single cases. The rate at which children continue to accumulate new words in the second and third year of life differs individually. Looking at single cases as well as larger samples, varying growth patterns have been demonstrated. The vocabulary size increased from19 to 26 months gradually. A sudden acceleration in the rate of vocabulary growth has been reported around the age of 28 months (1109). This widespread phenomenon of a rapid and sudden growth is referred to as vocabulary spurt. The observed exponential increase in spontaneous word production in the 28 months, followed by a further deceleration, described as vocabulary growth within a 'region of acceleration' as [18] explained. Although the findings clearly support a general trend of a vocabulary spurt phase in the second year. The vocabularies continue increase from 36 to 41months and decrease from 42to43 months and continue to increase by the age of 44 months. The individual patterns are showing linear phases of vocabularies growth or a spurt in the third and half year. Supporting the findings of [19] has found that children vary a great deal in the rate at which they acquire vocabularies. Finally, the development of vocabularies in this research is independent of variables such as gender and parental influence.

By comparing our results with the previous work we found that 21 months produce 384 words while [12] found 23 months, produced an average of 189.5 words. The children between the ages of 21-26 months show mean vocabulary score 738 words where According to [13], toddlers between the ages of 23-25 months received total vocabulary scores of 196.24 words. The 26 months produce 706 words where [14], as cited by [3] reported that the expressive vocabulary size of typically developing 2 year-old children was, at least 150 words. The overall findings show the vocabulary increase gradually from 1,7 to 3,8 years. It appears that increase in early lexical abilities considered an indicator for later grammatical complexity. Accordingly, the study shows that lexical limitations successfully serve as a reliable, early predictor of potential language acquisition problems and sometimes, of severe and continuing disorders. This evaluation supports the results of other studies in which lexical development is taken to be a valid predictor of further language acquisition. A satisfactory level of lexical development is a prerequisite for grammatical development. The lexical limitation is an indicator of a problem in the other linguistic areas such as syntax and morphology.

The results of TTR counts of all children shows that TTR size of different files in the corpus is not constant as shown in Fig. 2. The correlation between tokens and TTR is non-linear. The vocabulary spurt affect TTR result and thus for each child, 19 months child with 318 words show a higher TTR count 0.43score where, 44 month child with 1323 words show TTR count 0.41score. There is a correlation between tokens and TTR, the large sample size give small TTR. For example, the 41 months with 2407 words give 0.35 score TTR and 44 months with 1414 words give 0.38 score.TTR is declined with increasing sample size. Therefore, any single value of TTR lacks reliability as it will depend on the length in words of the language sample used. A graph of TTR against tokens for a transcript will lie in a curve beginning at the point (1,1) and falling with a negative slope that becomes progressively less steep. (TTR) developmental index is not alone a valid measure for vocabulary growth in Egyptian Arabic language. Further measure such as VOCD (measurement of vocabulary diversity) should incorporate with TTR.

6 CONCLUSIONS

The obtained results in this research can form the base on which further research on figuring out the developmental stages of Egyptian Arabic can be based. This is an important research project because of the fundamental lack of work in this area of Arabic linguistics. It is obvious that further work needs to be done on Large-scale corpus-based to allow us shed light on the mechanisms of early lexical development. This is an important step towards establishing robust developmental stages in Egyptian Arabic language.

REFERENCES

[1] B. MacWhinney The CHILDS project. Tool for analyzing talk Electronic Edition. part 2: the CLAN programs. Carnegie Mellon university available on line at , http://childs.psy.cmu.edu/manuals/clan(2012).

[2] M. C.Templin Certain language skills in children. Minneapolis: University of Minnesota Press (1957).

[3] K. Retherford, Guide to analysis of language transcripts (3rd ed.). Eau Claire, WI: Thinking Publication (2000).

[4] D. Crystal, Profiling linguistic disability. London: Edward Arnol (1982).

[5] G. H., & Thompson, J. R. Thomson Outlines of a method for the quantitative analysis of writing vocabularies. British Journal of Psychology, 1904-1920, 8(1), 52-69. (1915).

[6] J. F. Miller, Assessing language production in children: experimental procedures. London: Edward Arnold (1981).

[7] P. S. Dale, Language Development: Structure and Function. New York: Holt Rinehart and Winston (1976).

[8] R.E. Owens, Language disorders: A functional approach to assessment and intervention. New York: Merrill/Macmillan (1991).

[9] H. Benedict, Early lexical development: Comprehension and production. Journal of child language, 6(02), 183-200. (1979).

[10] K. Nelson, Structure and strategy in learning to talk. Monographs of the society for research in child development, 1-135. (1973).

[11] M. E. Smith, An investigation of the development of the sentence and the extent of vocabulary in young children. University of Iowa Studies: Child Welfare. (1926).

[12] D., &Tobias, S.Thal, Relationships between language and gesture in normally developing and late-talking toddlers. Journal of Speech and Hearing Research, 37, 157-170 f (1994).

[13] L. Rescorla, N. B. Ratner, P. Jusczyk, & A. M. Jusczyk Concurrent Validity of the Language Development Survey: Associations with the MacArthur—Bates Communicative Development Inventories Words and Sentences. American Journal of Speech-Language Pathology, 14(2), 156-163, (2005).

[14] A. Mehrabian, The development and validation of measures of affiliative tendency and sensitivity to rejection. Educational and psychological measurement (1970).

[15] J. Heilmann, S. E. Weismer, J. Evans & C. Hollar, Utility of the MacArthur—Bates Communicative Development Inventory in Identifying Language Abilities of Late-Talking and Typically Developing Toddlers. American Journal of Speech-Language Pathology, 14(1), 40-51. (2005).

[16] J. R Phillips Syntax and vocabulary of mothers' speech to young children: age and sex comparisons. Child Development 44. 182. (1973).

[17] D. Ntelitheos & A. Idrissi. Language Growth in Child Emirati Arabic. In 29th Annual Symposium on Arabic Linguistics (pp. 9-11) (2015).

[18] E. Bates, P.S. Dale & D. Thal. Individual Differences and their Implications for Theories of Language Development. In P. Fletcher, & B. MacWhinney, (eds.), The Handbook of Child Language (pp. 96-152). Cambridge: Basil Blackwell. (1995).

[19] L. Fenson, E. Bates, P.S. Dale, S.J. Pethick, J.S. Reznick & D. Thal. Variability in Early Communicative Development. (Monographs of the society for research in child development, 59/4). Chicago: The University of Chicago Press. (1994).

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.

Heba Salama has a master's degree in corpus linguistics from the faculty of Arts phonetics and linguistics department Alexandria University 2015. A PhD student is building a morphologically analysed corpus for Egyptian children. She is interested in child language research. Her main interest is to collect corpus data to study child language development. She is searching for standard criteria to collect and transcribe data.

نمو المفردات في عربية الاطفال المصريين: دراسة مؤسسة علي مدونة لغوية

هبه سلامة - سامح الانصاري
كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية

ملخص
تهدف الدراسة الي وصف تطور المفردات وعلاقتها بالعمر لدي الأطفال المصريين ومعرفة الأختلافات الفردية لدي الأطفال في نمو المفردات وحساب نسبة كلمات الطفل علي الكلمات الكليةTTR. يعد اكتساب الأطفال للمفردات محور رئيسي ومعقد في تطور اللغة عند الأطفال وقد لاقي اهتماما كبيرا في الأبحاث اللغوية النفسية واصبح مجال للدراسة في حد ذاته.وتقوم هذه الدراسة علي مدونة لغوية للأطفال مسجلة من الكلام التلقائي لمدة 30 دقيقة لعشرة أطفال 5اولاد-5 بنات من عمر1,7 الي 4 سنواتوتحليل معدل تطور الكلمات مع تطور العمر.يساهم هذا البحث في معرفة المعدل الطبيعي لنمو المفردات لتميزومعرفة مدي تاخر اللغة عند الأطفال ،ويعد من الأسهامات الأولي في اللغة المصرية للأطفال.

نحو إنجاز قاعدة بيانات للاسم في اللغة العربية

دكتورة فطوم القريبش
أستاذة التعليم الثانوي التأهيلي
عضوة مع فريق ابتكارات بالمدرسة المحمدية للمهندسين بالرباط
Fettoum.krieche@gmail.com

ملخص:

تهدف هذه الدراسة إلى تقديم نظرة أولية حول إنجاز قاعدة بيانات للاسم؛ للتمكن من إنجاز معجم آلي للاسم في اللغة العربية، انطلاقا من اعتماد قاعدة بيانات الفعل بكل أنواعه(الثلاثي الصحيح السالم، الثلاثي الصحيح المهموز، الثلاثي الصحيح المضعف، الثلاثي المعتل، الرباعي المضعف والغير مضعف) وذلك بالاعتماد على مجموعة من الخطوات اللسانية والحاسوبية لإنتاج قاعدة بيانات منتظمة تضم معلومات صرفية وصوتية وصرف تركيبية ودلالية، وذلك بتوخي مفهوم الشمولية التي برمجنا بها معجما آليا للفعل، وقد تمكنا من خلال هذه الدراسة من التوصل إلى جملة من الاستنتاجات اللسانية.

الكلمات المفاتيح: البرمجة الآلية- المعجم الآلي- المعالجة الآلية- اللغة العربية- قاعدة البيانات.

المقدمة

أصبحت الحاجة ملحة إلى ضرورة إنجاز برامج لمعالجة اللغات، وخاصة اللغة العربية، مع الثورة التكنولوجية التي صار يعيشها العالم في الوقت الراهن، فكل تبادل لمعلومات أو أفكار صار يستدعي بالضرورة بنكا من المعلومات التي تقوم على برامج آلية للغة المستعملة في التواصل الآلي، وهذا يستوجب تواصلا فعالا و بنّاء بين اللساني والحاسوبي؛ اللساني الذي يدرك عمق اللغة بكل مستوياتها اللسانية، والحاسوبي الذي يعلم خبايا الآلة وميكانيزماتها.

لقد قمنا في مراحل إنجاز بحث لنيل شهادة الدكتوراه بإنجاز معجم آلي للفعل في اللغة العربية، باعتماد عدة مراحل تراوحت بين خطوات لسانية وأخرى معلوماتية، وذلك ما سنبرز في هذا المقال في القسم الأول، وتأسيسا على هذه القاعدة سوف نقدم نظرة لسانية لإنجاز قاعدة أخرى للاسم في اللغة العربية باعتبار الاتجاه النحوي القائل بأن الأصل في الاشتقاق في اللغة العربية هو الفعل (مدرسة الكوفة)، وذلك ما سنبين في القسم الثاني من هذا المقال، وذلك حتى يتسنى لنا في المراحل اللاحقة إنجاز قاعدة بيانات للاسم في اللغة العربية، تأسيسا على المعجم الآلي المنجز للفعل.

أولا: توصيف قاعدة بيانات للفعل في اللغة العربية

1_بناء قاعدة بيانات الفعل في اللغة العربية

(1)قاعدة بيانات الفعل الثلاثي[1]

قمنا بإنجاز قاعدة بيانات تخص الفعل الثلاثي بكل أنواعه باعتماد مجموعة من المراحل تتأرجح بين معطيات لسانية وآليات معلوماتية، وهي كالآتي:

(أ) مرحلة توليد التوليفات

لقد توخينا في إنشاء المتن اللغوي للفعل في اللغة العربية مفهوم الشمولية عن طريق توليد كل الجذور الممكنة من25 حرفا، وذلك بخلق نظام معلوماتي يسمح بتوليد كل هذه التوليفات بطريقة أوتوماتيكية (لوغاريتمC) فحصلنا على قاعدة بيانات تضم 15625 جذرا، وقمنا بفرز الجذور المستعملة من المهملة في اللغة العربية ثم إقصاء هذه الأخيرة من قاعدة البيانات، باعتماد مرحلتين أساسيتين هما:

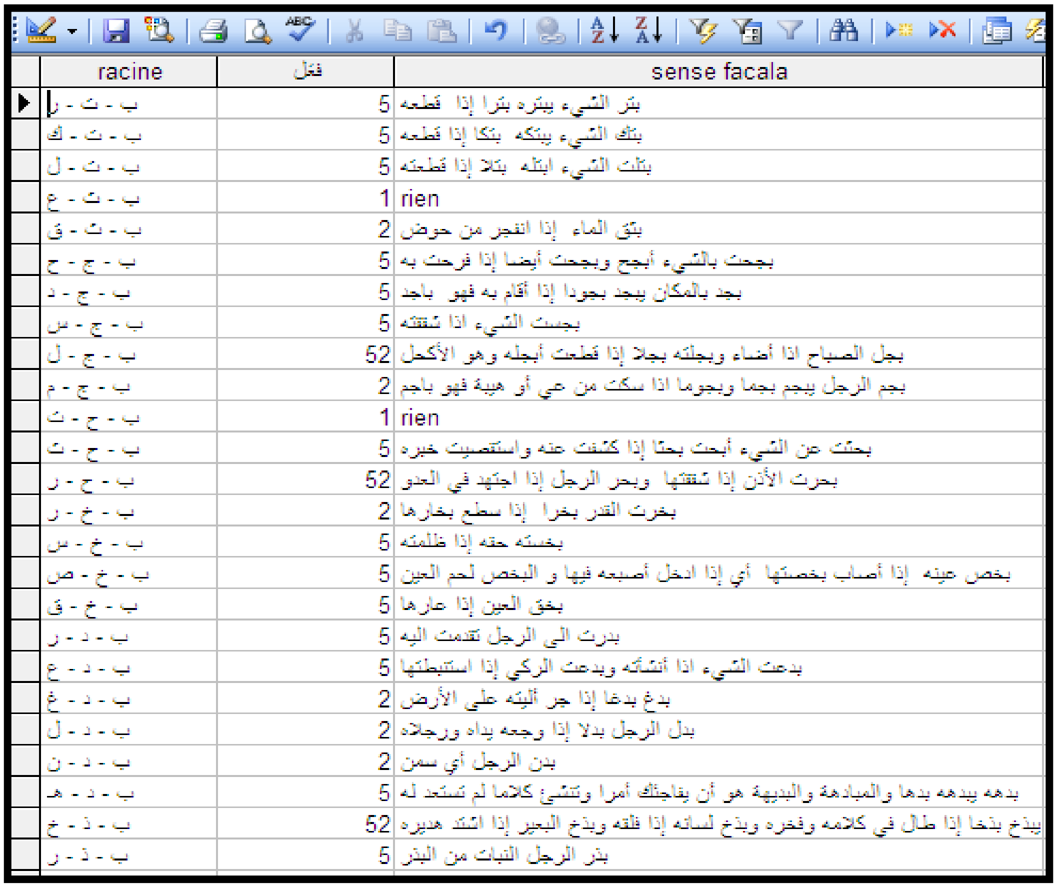
(ب) مرحلة التصفية الصواتية

وتأسست على استقصاء كل القواعد الصوتية التي تقيد تركيب البنية الصرفية للكلمة في اللغة العربية عامة والفعل على وجه الخصوص، ثم استثمارها لحذف كل الجذور التي تخالف هذه القواعد، وذلك مثل: عدم تآلف حروف الصفير في كلمة واحدة، يقول ابن جني: "حروف الصفير وهي الصاد والسين والزاي لا تتركب بعضها مع بعض ليس في الكلام مثل: سص وصس وسز"**[2]** ، وذلك باعتماد requêtes SQL وهي كالآتي: (التي تخص حروف الصفير)، وبالتالي بقي في قاعدة البيانات بعد تطبيق هذه المرحلة **11942** جذرا.**(أحد عشر ألف وتسعمائة واثنان وأربعون جذرا)**

(ج) مرحلة التصفية الصرف_تركيبية والدلالية

وتقوم هذه المرحلة على فرز، وانتقاء، الجذور التي تنتمي للنسق اللغوي العربي باعتبار دلالتها من جهة، وخصائصها الصرف-تركيبية من جهة ثانية؛ لنتوصل إلى قاعدة بيانات تضم سوى الأفعال المتداولة في اللغة العربية بتحقق شرط المناسبة الصوتية والدلالية والصرف-تركيبية مع كل جذر على حدة"compatibilities racines-shéme"، وكذا الحصول على قاعدة بيانات أخرى تضم توليفات تستجيب للقواعد الصوتية لكنها لا تحمل معنى، وبالتالي لا تتوفر على خصائص صرف-تركيبية. ونقدم هنا جزءا من القاعدة التي أنجزنا للفعل الثلاثي الصحيح السالم المتداول في اللغة العربية كالآتي:

الجدول (1): واجهة تمثل جزءا من قاعدة بيانات الأفعال المتداولة



racine	فَعَل	sense facala
ب - ت - ر	5	بتر الشيء يبتره بترا إذا قطعه
ب - ت - ك	5	بتك الشيء يبتكه بتكا إذا قطعه
ب - ت - ل	5	بتلت الشيء ابتله بتلا إذا قطعته
ب - ت - ع	1	rien
ب - ت - ق	2	بتق الماء إذا انفجر من حوض
ب - ج - ح	5	بجحت بالشيء أبجح وبجحت أيضا إذا فرحت به
ب - ج - د	5	بجد بالمكان يبجد بجودا إذا أقام به فهو باجد
ب - ج - س	5	بجست الشيء اذا شققته
ب - ج - ل	52	بجل الصباح اذا أضاء وبجلته بجلا إذا قطعت أبجله وهو الأكحل
ب - ج - م	2	بجم الرجل يبجم بجما وبجوما اذا سكت من عي أو هيبة فهو باجم
ب - ح - ث	1	rien
ب - ح - ث	5	بحثت عن الشيء أبحث بحثا إذا كشفت عنه واستقصيت خبره
ب - ح - ر	52	بحرت الأذن إذا شققتها وبحر الرجل إذا اجتهد في العدو
ب - خ - ر	2	بخرت القدر بخرا إذا سطع بخارها
ب - خ - س	5	بخسته حقه إذا ظلمته
ب - خ - ص	5	بخص عينه إذا أصاب بخصتها أي إذا ادخل أصبعه فيها و البخص لحم العين
ب - خ - ق	5	بخق العين إذا عارها
ب - د - ر	5	بدرت الى الرجل تقدمت اليه
ب - د - ع	5	بدعت الشيء اذا أنشأته وبدعت الركي إذا استنبطتها
ب - د - غ	2	بدغ بدغا إذا جر أليته على الأرض
ب - د - ل	2	بدل الرجل بدلا إذا وجعه يداه ورجلاه
ب - د - ن	2	بدن الرجل أي سمن
ب - د - ه	5	بدهه يبدهه بدها والمبادهة والبديهة هو أن يفاجئك أمرا وتنشئ كلاما لم تستعد له
ب - ذ - خ	52	يبذخ بذخا إذا طال في كلامه وفخره وبذخ لسانه إذا فلقه وبذخ البعير إذا اشتد هديره
ب - ذ - ر	5	بذر الرجل النبات من البذر

إن اللائحة أعلاه عبارة عن مصفوفة، تضم معلومات لسانية تتمثل في إيراد معنى لكل فعل تناسب جذره مع وزن معين، كتناسب الجذر "ب، ت، ر" مع الوزن "فَعَلَ" يؤدي معنى للفعل "بَتَر" وهو: **بتر الشيء يبتره إذا قطعه**، كما ينتج عن هذا التناسب أيضا إسناد خصائص صرف-تركيبية (مجرد، متعد، تام التصرف) لهذا الفعل وقد اختزلنا هذه الخصائص في الرقم"5"، مما استوجب علينا تشفير كل مجموعة من الخصائص الصرف-التركيبية برقم معين، حسب البنية الصرفية والتركيبية والدلالية للفعل المعالج، فبقي في قاعدة البيانات التي اشتغلنا بها بعد هذه المرحلة ما يناهز **8794 (ثمانية ألاف وسبعمائة وأربعة وتسعون)** فعلا متداولا في اللغة العربية (أي يستجيب للقواعد الصوتية، وله خصائص صرف_تركيبية، وله معنى).

وقد أنجزنا قاعدة بيانات للفعل الثلاثي المضعف والمهموز والمعتل بإتباع المراحل نفسها.

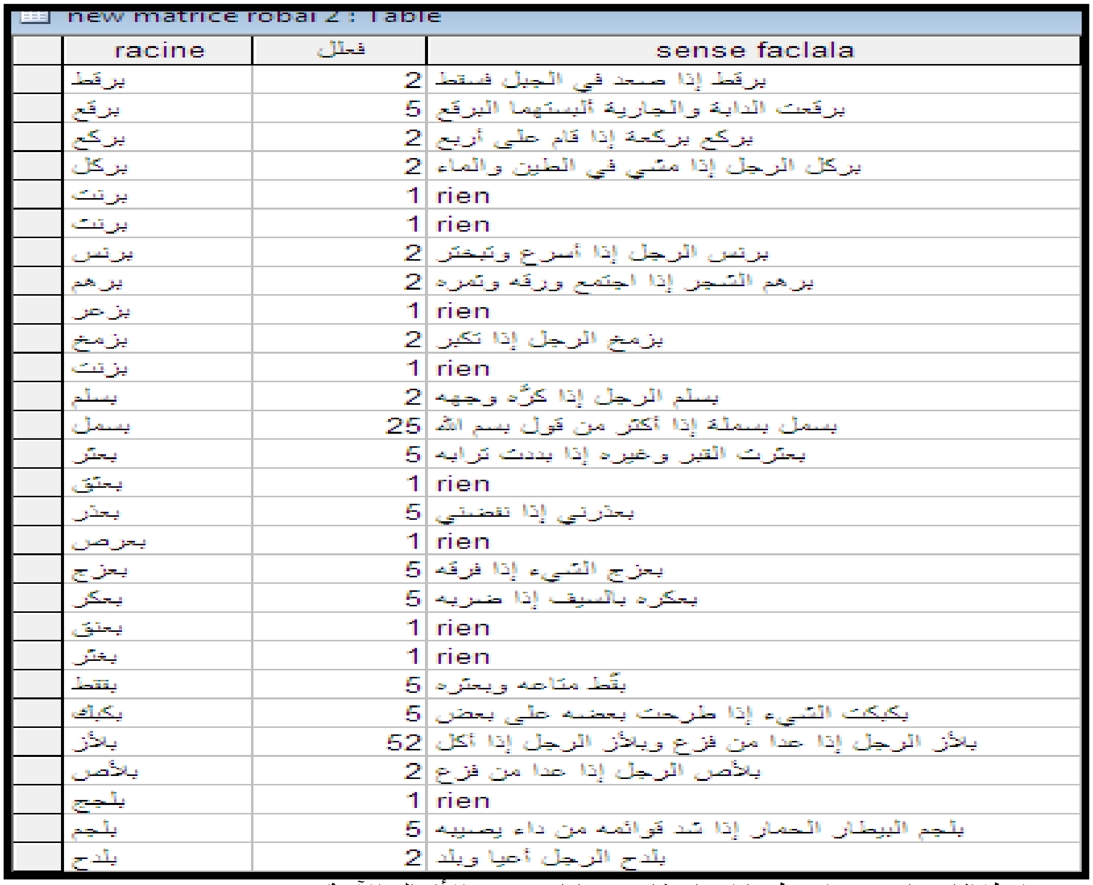
(2) قاعدة بيانات الفعل الرباعي

تبنينا في إنشاء قاعدة البيانات التي تخص الفعل الرباعي المبدأ نفسه الذي أنجزنا به قاعدة البيانات التي تخص الفعل الثلاثي والذي يتمثل في توخي الشمولية، وذلك حتى نستطيع الإحاطة بكل الجذور الرباعية بوصفها خطوة أولية، ثم فرز المستعملة من المهملة في مرحلة لاحقة، إلا أن ذلك كان موافقا للبنية الصرفية الرباعية المضعفة[3]؛ لأن عدد التوليفات الممكنة من 26 حرفا هو: 26x26=676 جذرا، وهو عدد قابل للمعالجة في وقت وجيز، ولما أردنا تبني المبدأ نفسه في خلق قاعدة بيانات تخص البنية

الصرفية للفعل الرباعي غير المضعف[3] حصلنا على عدد كبير من التوليفات:26x26x26x26=456976 جذرا، مما سيتطلب وقتا كبيرا في المعالجة. أضف إلى ذلك أن هذا النوع من الأفعال قليل جدا في النسق اللغوي العربي، لذلك ارتأينا أن نجمع هذه الأفعال الرباعية غير المضعفة من المعاجم العربية (مثل معجم الجمهرة ومعجم كتاب الأفعال) ونبرمجها في قاعدة بيانات مستقلة، دون اعتماد طريقة الشمولية في ذلك.

ونقدم هنا جزءا من قاعدة البيانات التي قمنا ببرمجة محتواها باعتماد المراحل نفسها المتبعة في إنجاز قاعدة بيانات الفعل الثلاثي وهي كالآتي:

الجدول (2): واجهة تمثل جزءا من قاعدة البيانات للفعل الرباعي غير المضعف



	racine	فعلل	sense faclala
	برقط	2	برقط إذا صعد في الجبل فسقط
	برقع	5	برقعت الدابة والجارية ألبستهما البرقع
	بركع	2	بركع بركعة إذا قام على أربع
	بركل	2	بركل الرجل إذا مشي في الطين والماء
	برتك	1	rien
	برتك	1	rien
	برتس	2	برتس الرجل إذا أسرع وتبختر
	برهم	2	برهم الشجر إذا اجتمع ورقه وتمره
	بزعر	1	rien
	بزمخ	2	بزمخ الرجل إذا تكبر
	بزتك	1	rien
	بسلم	2	بسلم الرجل إذا كرّه وجهه
	بسمل	25	بسمل بسملة إذا أكثر من قول بسم الله
	بعثر	5	بعثرت القبر وغيره إذا بددت ترابه
	بعثق	1	rien
	بعذر	5	بعذرني إذا نقضني
	بعرص	1	rien
	بعزج	5	بعزج الشيء إذا فرقه
	بعكر	5	بعكره بالسيف إذا ضربه
	بعنق	1	rien
	بعثر	1	rien
	بقلط	5	بقّط متاعه وبعثره
	بكبك	5	بكبكت الشيء إذا طرحت بعضه على بعض
	بلأز	52	بلأز الرجل إذا عدا من فزع وبلأز الرجل إذا أكل
	بلأص	2	بلأص الرجل إذا عدا من فزع
	بلجج	1	rien
	بلجم	5	بلجم البيطار الحمار إذا شد قوائمه من داء يصيبه
	بلدح	2	بلدح الرجل أعيا وبلد

وانطلاقا مما سبق، استطعنا إنجاز قاعدة بيانات تضم الأفعال الآتية:

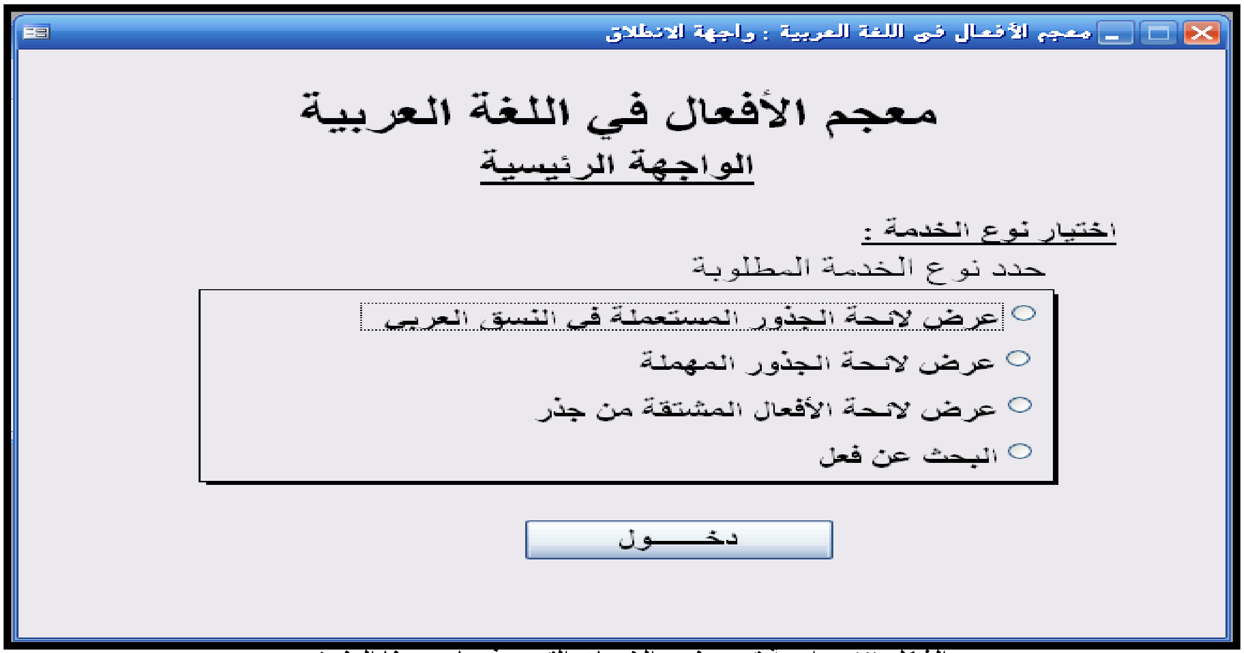
- الفعل الثلاثي السالم (3262 فعلا متداولا).
- الفعل الثلاثي المضعف (470 فعلا متداولا).
- الفعل الثلاثي المهموز (606 فعلا متداولا).
- الفعل الثلاثي المعتل (1892 فعلا متداولا).
- الفعل الرباعي المضعف (433 فعلا متداولا).

- الفعل الرباعي غير المضعف (873 فعلا متداولا).

2- توصيف المعجم الآلي المنجز للفعل

(1) واجهة الانطلاق

لقد قمنا بتفريغ محتويات قاعدة البيانات التي أنجزنا للفعل في معجم آلي باعتماد آليات معلوماتية حتى نبين قيمة العمل المنجز(قاعدة بيانات شاملة لكل المعلومات اللسانية التي تخص مقولة الفعل في اللغة العربية) وهذا المعجم يقوم على واجهات عدة أولها واجهة الانطلاق، وفي هذه الواجهة يحدد المستخدم لمعجمنا نوع الخدمة المراد تحقيقها، وشكل هذه الواجهة كالآتي:



معجم الأفعال في اللغة العربية

الواجهة الرئيسية

اختيار نوع الخدمة :

حدد نوع الخدمة المطلوبة

○ عرض لائحة الجذور المستعملة في النسق العربي
○ عرض لائحة الجذور المهملة
○ عرض لائحة الأفعال المشتقة من جذر
○ البحث عن فعل

الشكل (1): واجهة تحديد نوع الخدمات التي يوفرها معجمنا المنجز

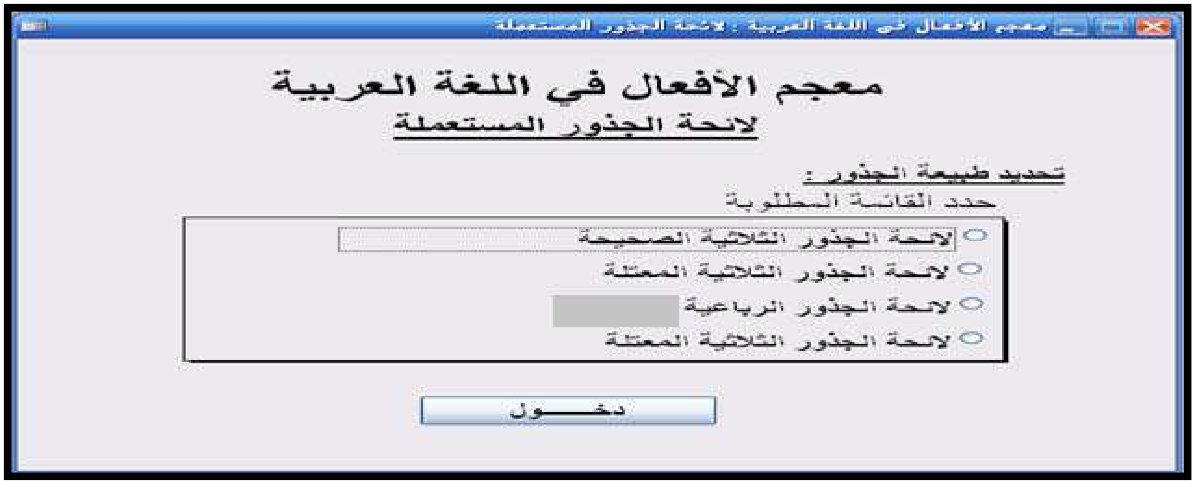
كما يتبين لنا من خلال الواجهة أعلاه؛ فإن معجمنا يمكن أن يوفر لنا أربع خدمات، هي:

- يمكننا التعرف على جذر حكمنا على وجوده في المتن اللغوي العربي، انطلاقا من الخطوات التي ذكرناها في المبحث الأول من هذا الفصل، مع تقديم طبيعة الفعل (ثلاثي، رباعي) ونوعه (صحيح، معتل، سالم، أجوف....) وخصائصه الصرف-التركيبية والدلالية.
- يزودنا المعجم أيضا بالجذور المهملة أوالمفترضة؛ نظرا لأنها تستجيب للخصائص الصوتية في اللغة العربية ولكنها غير متداولة. (إنها جذور ليس لها معنى وليس لها خصائص صرف-تركيبية في اللغة العربية).

- من الممكن أيضا أن نتعرف كل الأفعال التي تتوافق مع كل وزن على حدة، ونتبين معانيها وخصائصها الصرف-التركيبية، انطلاقا من إدخال الجذر في الواجهة.
- وأخيرا يمكن للمعجم أن يمدنا بجذر الفعل ومعناه وخصائصه الصرف-التركيبية، انطلاقا من إدخال الفعل.

(2) واجهة الأفعال المستعملة

عندما يتم تحديد نوع الخدمة (اختيار الخدمة الأولى) ننتقل إلى الواجهة الثانية و التي سنحدد من خلالها طبيعة الجذر المراد البحث عنه، ونقدم الواجهة كالآتي :



معجم الأفعال في اللغة العربية

لائحة الجذور المستعملة

تحديد طبيعة الجذور :

حدد القائمة المطلوبة

○ لائحة الجذور الثلاثية الصحيحة

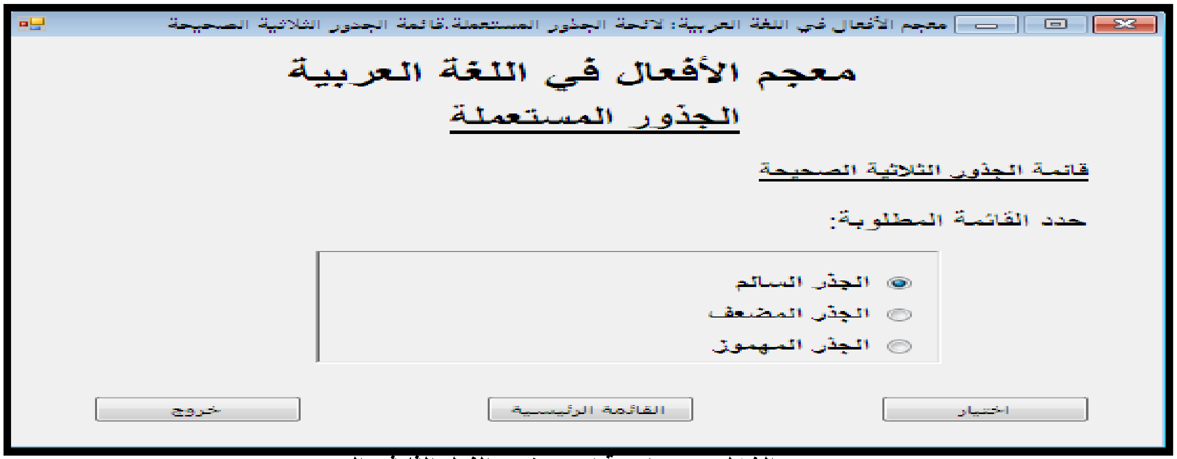
○ لائحة الجذور الثلاثية المعتلة

○ لائحة الجذور الرباعية

○ لائحة الجذور الثلاثية المعتلة

الشكل(2): واجهة تحديد طبيعة الفعل المراد التعرف على خصائصه الصرف-التركيبية والدلالية

وهذه اللائحة كما نلاحظ تضم كل أنواع الفعل في اللغة العربية وأقسامها، ويمكننا بفضلها تحديد طبيعة الجذر المراد التعرف عليه ، فإذا ما أردنا الفعل الثلاثي الصحيح سوف يقدم المعجم الواجهة الآتية:



معجم الأفعال في اللغة العربية: لائحة الجذور المستعملة.قائمة الجذور الثلاثية الصحيحة

معجم الأفعال في اللغة العربية

الجذور المستعملة

قائمة الجذور الثلاثية الصحيحة

حدد القائمة المطلوبة:

● الجذر السالم

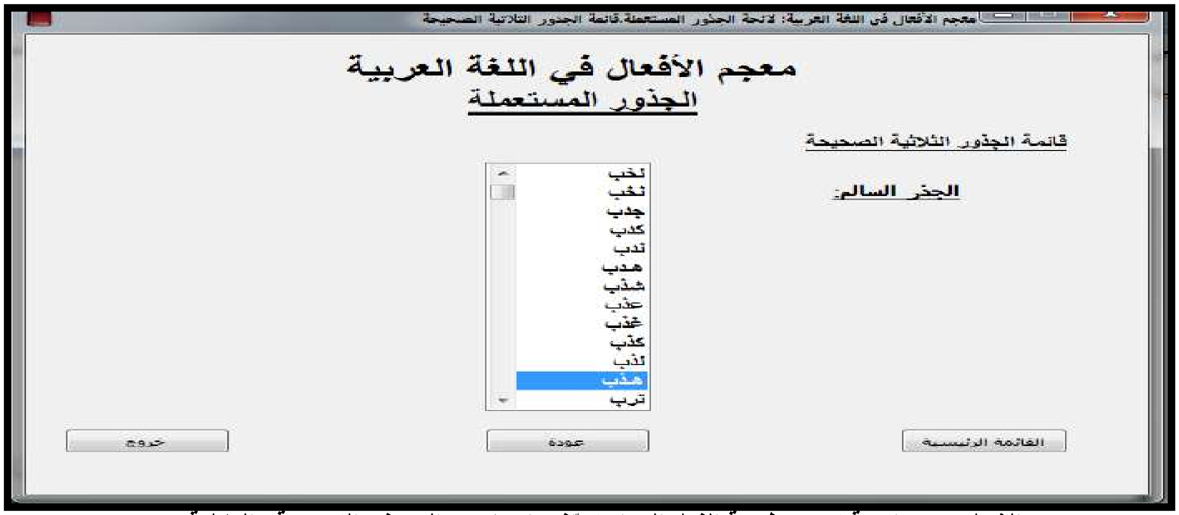
○ الجذر المضعف

○ الجذر المهموز

اختيار
القائمة الرئيسية
خروج

الشكل(3): واجهة تحديد نوع الفعل الثلاثي الصحيح

عند تحديد طبيعة هذا الفعل يزودنا المعجم بلائحة لكل الأفعال المبرمجة في قاعدة بيانات الفعل الثلاثي السالم، وهذا بطبيعة الحال إذا ما تم تحديد الفعل الثلاثي السالم، مع تقديم كل الخصائص الصرف-التركيبية والدلالية للجذر المحدد. ونقدم هذه الواجهة كالآتي:



معجم الأفعال في اللغة العربية: لائحة الجذور المستعملة.قائمة الجذور الثلاثية الصحيحة

معجم الأفعال في اللغة العربية

الجذور المستعملة

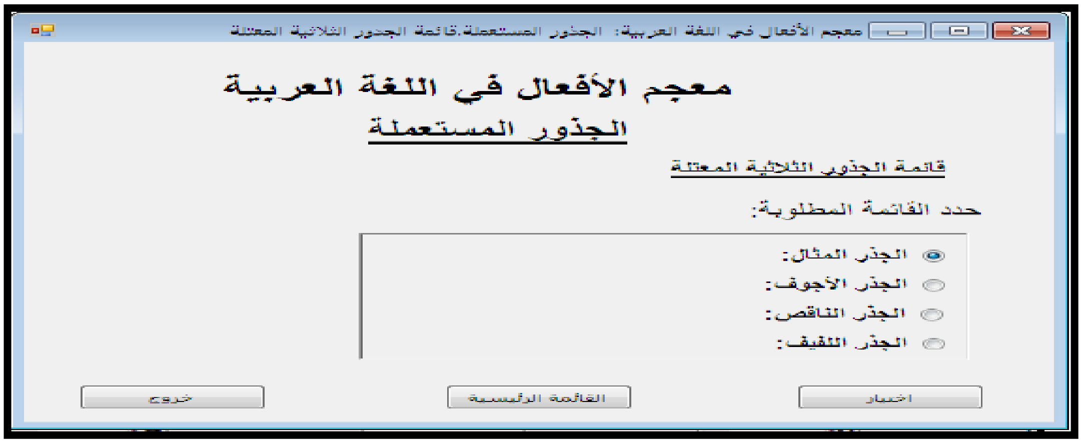
قائمة الجذور الثلاثية الصحيحة

الجذر السالم:

لخب
نخب
جدب
كدب
تدب
هدب
شذب
عذب
غذب
كذب
لذب
هذب
ترب

الشكل(4): واجهة تحديد طبيعة الفعل المراد تعرّف خصائصه الصرف-التركيبية والدلالية

أما إذا ما اخترنا الفعل الثلاثي المعتل، فسيقدم المعجم لائحة بكل الأفعال الثلاثية المعتلة كالآتي:



معجم الأفعال في اللغة العربية: الجذور المستعملة.قائمة الجذور الثلاثية المعتلة

معجم الأفعال في اللغة العربية

الجذور المستعملة

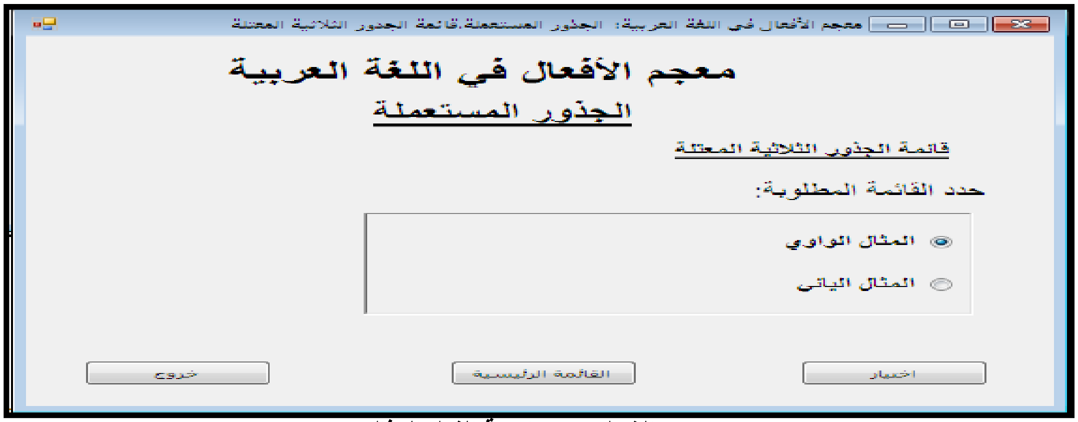
قائمة الجذور الثلاثية المعتلة

حدد القائمة المطلوبة:

● الجذر المثال:
○ الجذر الأجوف:
○ الجذر الناقص:
○ الجذر اللفيف:

الشكل(5): واجهة الأفعال الثلاثية المعتلة

وعند اختيارنا للفعل المثال سيقدم البرنامج الواجهة الآتية:



معجم الأفعال في اللغة العربية: الجذور المستعملة.قائمة الجذور الثلاثية المعتلة

معجم الأفعال في اللغة العربية

الجذور المستعملة

قائمة الجذور الثلاثية المعتلة

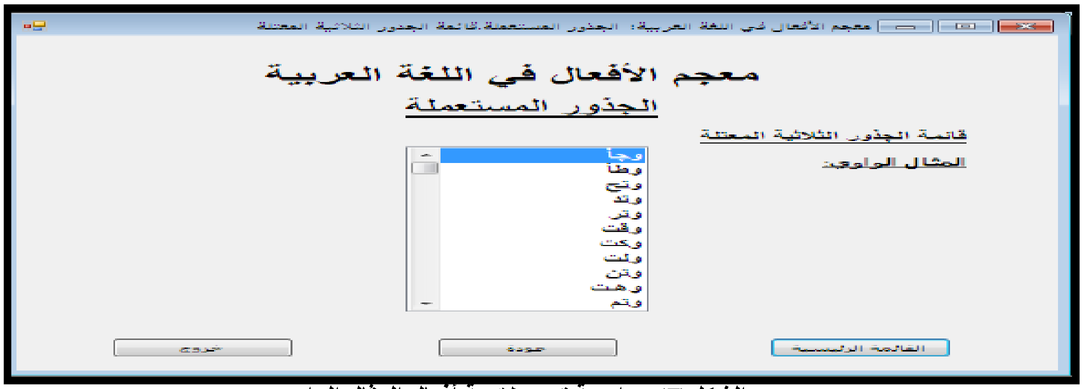
حدد القائمة المطلوبة:

● المثال الواوي

○ المثال الياتي

الشكل(6): واجهة الفعل المثال

ومن أجل التعرف على لائحة أفعال المثال الواوي مثلا يقدم الحاسوب الواجهة الآتية:



معجم الأفعال في اللغة العربية: الجذور المستعملة.قائمة الجذور الثلاثية المعتلة

معجم الأفعال في اللغة العربية

الجذور المستعملة

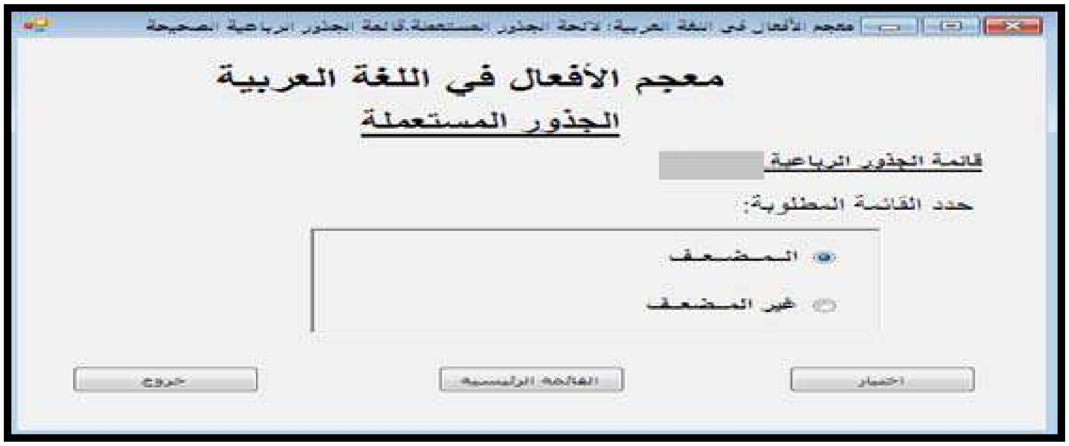
قائمة الجذور الثلاثية المعتلة

المثال الواوي

وجأ
وطأ
وتج
وتد
وتر
وقت
وكت
ولت
وتن
وهت
وتم

الشكل(7): واجهة تحدد لائحة أفعال المثال الواوي

ونفس الشيء بالنسبة للفعل الرباعي أيضا، ونقدم هنا مثالا يخص واجهة الأفعال الرباعية كالآتي:



معجم الأفعال في اللغة العربية: لائحة الجذور المستعملة.قائمة الجذور الرباعية الصحيحة

معجم الأفعال في اللغة العربية

الجذور المستعملة

قائمة الجذور الرباعية

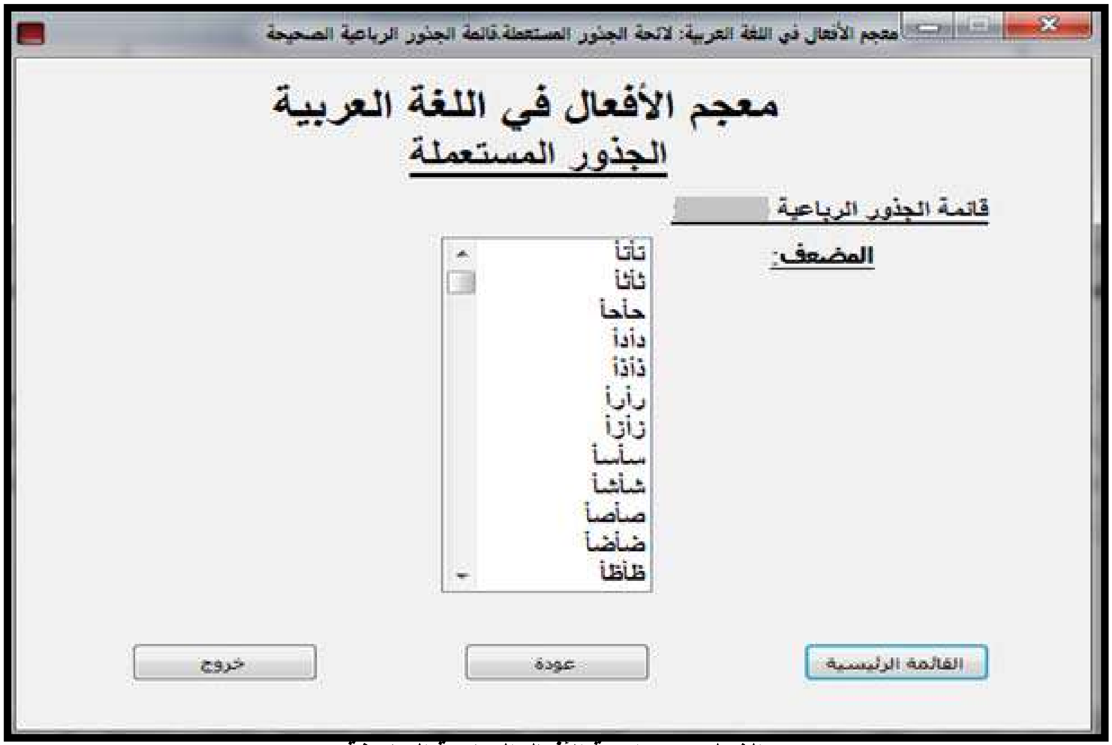
حدد القائمة المطلوبة:

● المـضـعـف

○ غير المـضـعـف

الشكل(8): واجهة الأفعال الرباعية

وإذا ما اخترنا مثلا الفعل الرباعي المضعف، نجد الواجهة الآتية:



معجم الأفعال في اللغة العربية: لائحة الجذور المستعملة.قائمة الجذور الرباعية الصحيحة

معجم الأفعال في اللغة العربية

الجذور المستعملة

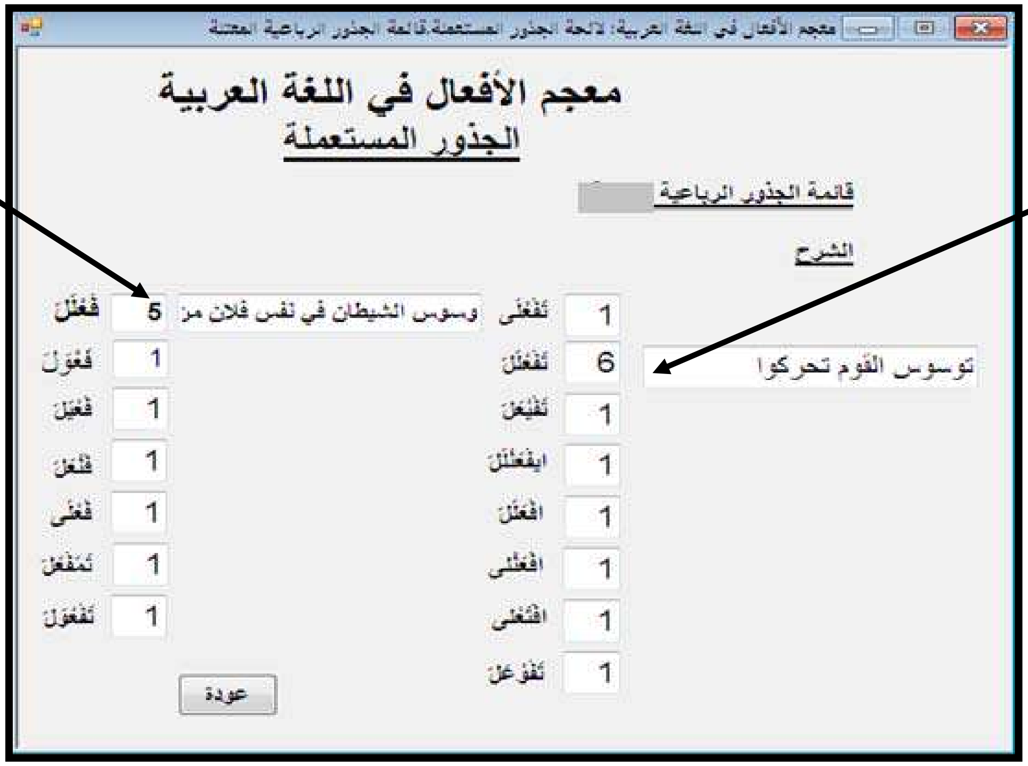
قائمة الجذور الرباعية

المضعف:

تأتأ
ثأثأ
حأحأ
دأدأ
ذأذأ
رأرأ
زأزأ
سأسأ
شأشأ
صأصأ
ضأضأ
ظأظأ

الشكل(9): واجهة الأفعال الرباعية المضعفة

كما نلاحظ فإن الواجهة أعلاه زودتنا بكل الجذور الرباعية المضعفة، بحيث يمكن تحديد أحدها ليقدم المعجم كل التناسبات المتوافقة مع الجذر المحدد ويقدم أيضا كل الخصائص الصرف-التركيبية والدلالية لهذا الفعل الناتج عن هذا التناسب. ولتوضيح ذلك نقدم الواجهة أسفله:



معجم الأفعال في اللغة العربية: لائحة الجذور المستعملة:قائمة الجذور الرباعية المعتلة

معجم الأفعال في اللغة العربية

الجذور المستعملة

قائمة الجذور الرباعية

الشرح

تَفَعْلَى	1	فَعْلَلَ	5
تَفَعْلَلَ	6	فَعْوَلَ	1
تَفَيْعَلَ	1	فَعْيَلَ	1
ايفَعْلَلَ	1	فَنْعَلَ	1
افْعَلَلَ	1	فَعْلَى	1
افْعَلَى	1	تَمَفْعَلَ	1
افْتَعْلَى	1	تَفَعْوَلَ	1
تَفَوْعَلَ	1		

وسوس الشيطان في نفس فلان مز

توسوس القوم تحركوا

يحيل هذا الرقم على الخصائص الآتية:

*لازم

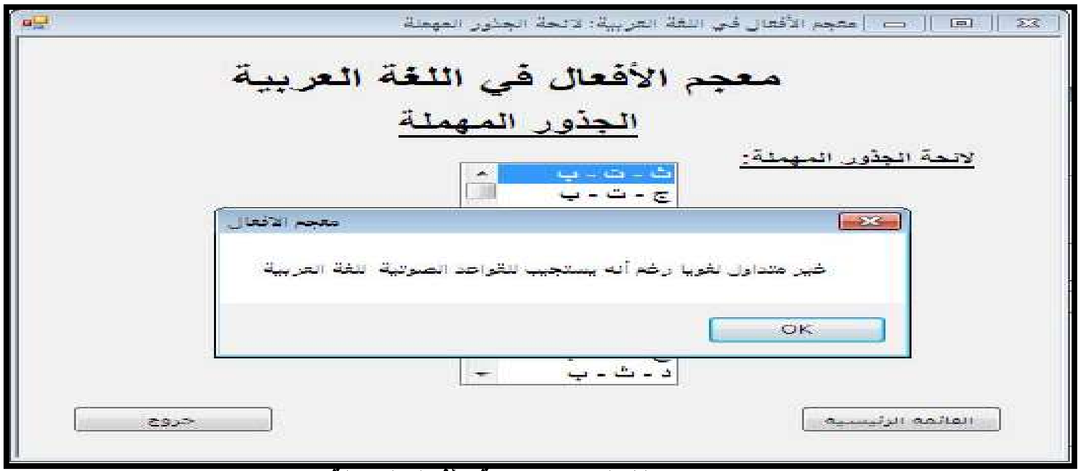
*مزيد

*تام

يحمل هذا الفعل الخصائص الآتية:
*متعدي
*مجرد
*تام التصرف.

الشكل(11): واجهة الفعل "وسوس" الناتج عن توافق الجذر مع الصيغة[4] وخصائصه الصرف التركيبية والدلالية

وهكذا بخصوص الواجهات السابقة، فبتعيين الجذر يقدم لك المعجم كل المعلومات اللسانية المرتبطة به، أما إذا أردنا تعرف الأفعال المهملة - التي تحقق فيها شرط التقيد بالقواعد الصوتية التي تحكم تجاور الحروف في اللغة العربية، إلا أنها لا تحمل معنى في هذه اللغة- فعلينا أن نحدد هذه الخدمة في واجهة الانطلاق ثم نضغط على "دخول"، فيزودنا المعجم برسالة مفادها أن الفعل المحدد غير متداول في اللغة العربية رغم أنه يستجيب للقواعد الصوتية التي تحكم تجاور الحروف (ويتعلق الأمر بالمصفوفة التي سنوردها في خاتمة البحث). ونقدم جزءا من هذه الواجهة كالآتي:



معجم الأفعال في اللغة العربية: لائحة الجذور المهملة

معجم الأفعال في اللغة العربية

الجذور المهملة

لائحة الجذور المهملة:

ث - ت - ب
ج - ت - ب
د - ث - ب

معجم الأفعال

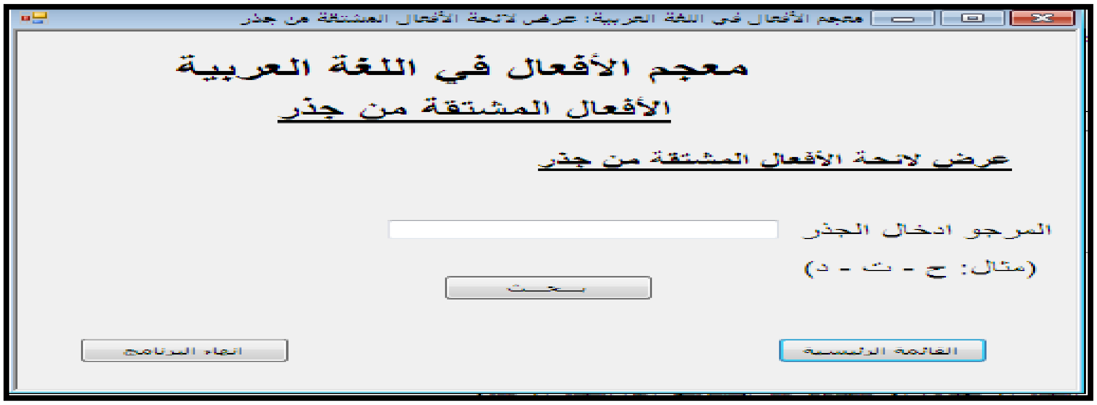
غير متداول لغويا رغم أنه يستجيب للقواعد الصوتية للغة العربية

الشكل(12) واجهة الأفعال المهملة

وهذه اللائحة تقدم مجموع التناسبات التي تنتج أفعالا يفترض إمكانية وجودها في اللغة العربية، ما دامت تحقق شرط التقيد للقواعد الصوتية (إن هذه الأفعال المفترض وجودها في النسق اللغوي العربي، ستسعفنا على إضافة كل الأفعال التي تنتمي للغة العربية المعاصرة، في مرحلة لاحقة، في معجمنا الذي أنجزنا).

(3) واجهة التعرف على الفعل انطلاقا من الجذر

أما إذا ما انتقلنا إلى الخدمة الثالثة، فيمكننا بفضلها أن ندخل جذرا ثلاثيا أو رباعيا على هذا المنوال: ك – ت – ب ⟵ فبين كل حرف وآخر نترك فراغا وحتى بين الحرف والعارضة وهذه الأخيرة تكتب بالرقم"6" (مثال لجذر ثلاثي:ك، ت، ب). لنتعرف كل الصيغ الصرفية التي تناسبت مع هذا الجذر والذي نتج عنه تكوين فعل ثلاثي صحيح أو ثلاثي معتل أو رباعي مع تقديم كل الخصائص الصرف-التركيبية والدلالية المرتبطة بكل تناسب على حدة، ونقدم هذه الواجهة كالآتي:



معجم الأفعال في اللغة العربية: عرض لائحة الأفعال المشتقة من جذر

معجم الأفعال في اللغة العربية

الأفعال المشتقة من جذر

عرض لائحة الأفعال المشتقة من جذر

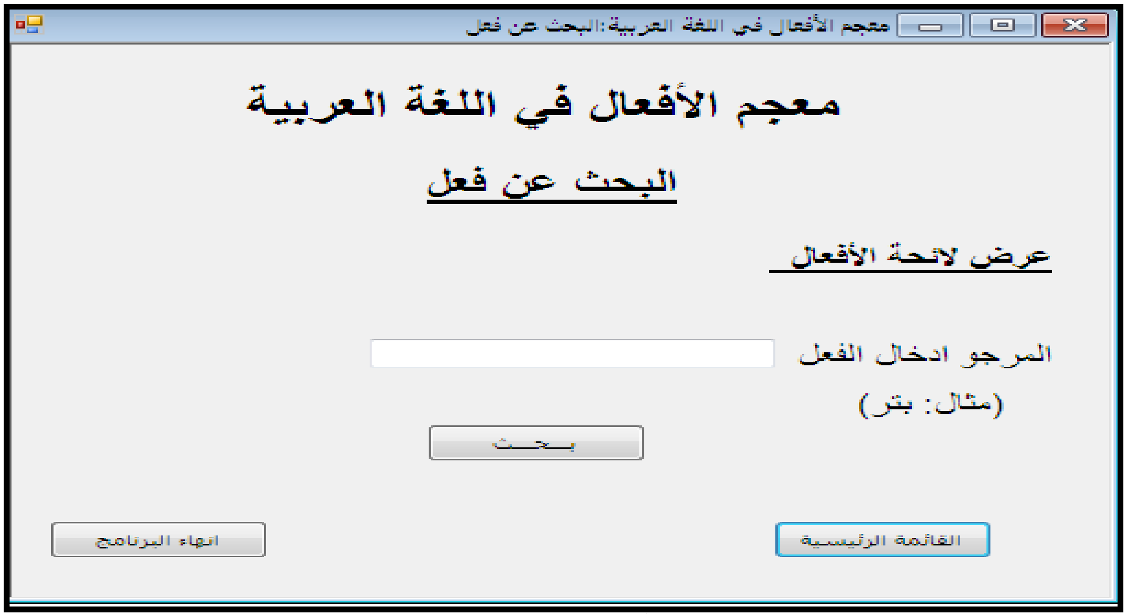
المرجو ادخال الجذر

(مثال: ج - ث - د)

الشكل(13) واجهة الأفعال المشتقة من الجذر

(4) واجهة التعرف على الجذر انطلاقا من الفعل

أما إذا أردنا التعرف على الجذر انطلاقا من إدخال فعل، فإننا نختار الخدمة الرابعة في واجهة الانطلاق، فيقدم لنا المعجم الواجهة الآتية:



الشكل(14)واجهة البحث عن الجذر انطلاقا من الفعل

ومن خلال ما سبق نقول بأن هذا البرنامج سيساعد المعلوماتيين في تطوير أنظمة أتوماتيكية موازية له (معجم التصريف الذي نحن بصدد الاشتغال عليه، معجم الأسماء، مصحح إملائي....)، وسيساعد اللغويين في البحث عن الأفعال في اللغة العربية بطريقة سهلة وموجزة.

ثانيا: نحو إنجاز قاعدة بيانات تخص الاسم في اللغة العربية

1- الخلاف النحوي بين مدرسة الكوفة والبصرة

لقد تعددت الأبحاث والمؤلفات (من هذه الكتب نذكر على سبيل المثال لا الحصر: - الإنصاف في مسائل الخلاف بين البصريين والكوفيين، لأبي البركات بن الأنباري- التبيين عن مذاهب النحويين البصريين والكوفيين، لأبي البقاء العُكبري- ائتلاف النصرة في اختلاف نحاة الكوفة والبصرة لعبد اللطيف الزَبيْدي- ما فات الإنصاف في مسائل الخلاف، لفتحي حمودة- همع الهوامع في شرح جمع الجوامع لجلال الدين السيوطي، الخ....) في مسألة الخلاف النحوي الذي كان بين المدرستين النحويتين الكوفية والبصرية، وهذا الخلاف كان في الأصل سياسيا ولا يسعنا المقام لذكره، كما اختلفوا في عدة مسائل نحوية منها الاختلاف في أصل الاشتقاق، حيث ذهب البصريون إلى أن المصدر هو الأصل، في حين يقول الكوفيون إن الفعل هو الأصل، وقد استدل كل من الفريقين بأدلة قوية، وحاول كل منهما تفنيد أدلة الآخر. وما يفيد بحثنا هذا

هو قول الكوفيين بأن أصل الاشتقاق هو الفعل حتى نتمكن من إنجاز قاعدة بيانات للاسم على غرار قاعدة بيانات الفعل المنجزة سابقا بسهولة ويسر، هذا إذا أحطناكم علما بأن اشتغالنا على قاعدة بيانات الفعل دام لأكثر من ثلاث سنوات.

2- نظرة أولية لسانية لإنجاز معجم آلي للاسم

تأخذ الأسماء حيزا وافرا في اللغة العربية فهناك أسماء الأفعال والمفعول والآلة والزمان والمكان والصفة المشبهة واسم التفضيل....سوف نشتغل في المرحلة الأولى على اسم الفاعل واسم المفعول باعتماد الجذر الثلاثي وغير الثلاثي. يصاغ اسم الفاعل من الثلاثي على وزن فاعل(مثال: كتب ← كاتب)، ويصاغ من غير الثلاثي بإبدال حرف المضارعة ميما وكسر ما قبل الآخر(مثال: علَّم ← معلِّم)، ويصاغ اسم المفعول من الثلاثي على وزن مفعول(مثال: كتب ← مكتوب) ومن غير الثلاثي بإبدال حرف المضارعة ميما مضمومة وفتح ما قبل الآخر(مثال: علَّم ← معلَّم). وذلك باعتماد قاعدة بيانات الفعل المنجزة سابقا وذلك من خلال ترجمة القاعدة اللغوية إلى لوغاريتم.

خاتمة

قمنا في هذا المقال بوصف المعجم الآلي الذي أنجزنا للفعل في اللغة العربية. لكن قبل ذلك قدمنا الخطوات الإجرائية (اللسانية والحاسوبية) التي اعتمدناها في إنجاز قاعدة بيانات منضبطة تحتوى على كل المعلومات اللسانية (الصوتية- الصرفية-التركيبية-الدلالية) التي ترتبط بمقولة الفعل في اللغة العربية، فتوصلنا -انطلاقا من ذلك- إلى مجموعة من الاستنتاجات التي تنم عن مدى أهمية قاعدة البيانات التي أنشأنا، بالإضافة إلى فعاليتها الناجعة في أننا استثمرناها في المعجم الذي أنشأنا؛ ومرد ذلك إلى أننا اعتمدنا في إنجازها على مفهوم الشمولية. ومن أهم ما توصلنا إليه ما يلي:

- تقوم البرمجة الآلية لقاعدة البيانات الخاصة بالفعل وغيره على معاجم لغوية ورقية عن طريق التخزين الأتوماتيكي للمعلومات اللسانية.

- إن المعالجة الآلية للغات الطبيعية أو ما يعرف باللسانيات الحاسوبية - وكما هو معلوم- تقوم بالأساس على منطقة البنية الصرفية للكلمة أي الصورة التمثيلية والنهائية بعد صوغ كل قواعد التأليف (قواعد صرف-صوتية) للكلمة، وذلك كما هو الشأن بالنسبة للأفعال المعتلة مثلا التي تتقيد بقواعد الإعلال، وبالتالي فلا يتطلب ذلك، البحث في أصل الكلمة والإجراءات التي تعرضت لها من حذف وقلب وإبدال، بل الاعتماد فقط على تخزين الصورة النهائية للكلمة.

- ولإنجاز معجم آلي للفعل لابد من اعتماد قاعدة بيانات منتظمة تضم كل المميزات الصرفية للفعل من جذور وصيغ صرفية وجذوع وخصائص صرف- تركيبية، وكذا الخصائص الدلالية للأفعال المتداولة في اللغة العربية الناتجة عن تناسب جذر مع صيغة صرفية معينة، وقد تمكنا باتباع مراحل كثيرة من إنشاء معجم آلي للفعل، الذي وصفنا عناصره في القسم الأول من هذا المقال، وقد أفضى بنا ذلك إلى التوصل إلى الاستنتاجات الآتية:

- هناك مجموعة من الأفعال التي تنتمي إلى اللغة العربية الكلاسيكية(القديمة) اندثرت من اللغة المتداولة في العصور الحالية منها:

☒ "حتد" و"عهن" بالمكان حتودا وعهونا، إذا أقام به.

☒ "رنق" الماء، إذا كدر.

☒ "وبطت حظ الرجل أبِطه"، إذا أخسَسْته أو وضعت من قدره و"وبط الرجل" ضعف.

☒ "بخدعه" إذا ضربه بالسيف، و"بخذعه" و"ذخعه" إذا قطعه بالسيف.

☒ "خيعر الرجل"، والخيعرة خفة وطيش.

☒ "عيدنت النخلة"، إذا صارت عيدانة أي طويلة ملساء.

- هناك أفعال تستجيب للقواعد الصوتية التي تحكم تجاور الحروف في اللغة العربية لكنها لا تتداول في اللغة العربية لا الكلاسيكية ولا المعيارية، وقد حذفناها من قاعدة البيانات التي اشتغلنا بها لإنجاز معجم آلي للفعل، ووضعناها في قاعدة أخرى ليتمكن الباحثون من الاستفادة منها في هذا المجال، ويصل عددها زهاء 3418 جذرا.
- لقد واجهنا مشكلا بخصوص الصناعة المعجمية في العالم العربي وذلك لأنه مازال هناك تأخر كبير في هذا الجانب، فلم يتم بعد إنتاج معجم عربي محض يضم كل الكلمات المتداولة في اللغة العربية في العصور الحالية، ويمكن أن يكون مرد ذلك إلى كون تأليف معجم ورقي للغة العربية عامة يقوم على شرطين أساسين هما: سماع اللفظة، وسماع معناها من أفواه متكلميها، وفي هذا يقول الأسترابادي"... يحتاج إلى سماع استعمال اللفظ المعين، وكذا استعماله في المعنى المعين " **[5].** إلا أن الشرط الأول مقبول، لكن الشرط الثاني تظل مقبوليته نسبية غير قطعية، لأنه يصعب رصد الملكة اللغوية معجميا للغة العربية المعاصرة نوعا ما، وذلك بسبب حلول ثقافات أخرى(كلمات الشات، لغة الرسائل الهاتفية...) في لغتنا الحالية.

وقد مكنتنا هذه الدراسة من خلق أفكار متعددة بخصوص برمجة كل المقولات اللغوية التي تنتمي إلى النسق اللغوي العربي، من قبيل مثلا إنجاز قاعدة بيانات تخص الاسم على غرار قاعدة البيانات التي أنجزناها للفعل وذلك من خلال اعتماد الفكرة النحوية التي تقول بأن أصل اشتقاق الكلمة هو الفعل وليس المصدر وذلك ما تعرضنا له في القسم الثاني من هذا المقال.

المراجع

[1] فطوم القرييش- يوسف طاهر- زهور حوتي، الندوة الدولية الرابعة حول المعالجة الآلية للغة العربية، ص: 79-90، 2و3 ماي (2012).

[2] أبو الفتح عثمان بن جني، الخصائص، تحقيق محمد علي النجار، الجزءII، الناشر دار الكتاب العربي بيروت لبنان.

[3] دة.خديجة الحديثي، أبنية الصرف في كتاب سيبويه معجم ودراسة، مكتبة لبنان ناشرون، بيروت لبنان، الطبعة الأولى، (2003).

[4] المقصود بتوافق الجذر مع الصيغة مثل: تناسب الجذر "ب، ت، ر" مع الوزن "فَعَلَ" يؤدي معنى للفعل "بَتَر" وهو: بتر الشيء يبتره إذا قطعه، كما ينتج عن هذا التناسب أيضا إسناد خصائص صرف-تركيبية(مجرد، متعد، تام التصرف).

[5] رضى الدين الأسترابادي، شرح الرضي على الكافية، تح وتع يوسف حسن عمر، سنة الطبع: (1395 – 1975 م).

المعاجم: (استخدمنا هذه المعاجم في ملأ قاعدة البيانات)

- معجم العين لأبي عبد الرحمن الخليل بن أحمد الفراهيدي (145-100هـ) تحقيق الدكتور مهدي المخزومي، د. إبراهيم السامرائي الأجزاء 1 و2 و3 و4 و5 و6 و7 و8.
- معجم جمهرة اللغة لابن دريد أبي بكر محمد بن الحسن الأزدي البصري (ت:سنة 321هـ) الجزء I وII.

السيرة الذاتية

فطوم القرييش، أستاذة السلك الثانوي التأهيلي بالجهة الشرقية بالمغرب، حاصلة على دكتوراه وطنية بميزة مشرف جدا في محور اللغة العربية تخصص اللسانيات الحاسوبية، بجامعة سيدي محمد بن عبد الله، كلية الآداب والعلوم الإنسانية، فاس، شاركت في عدة ندوات منها على سبيل المثال لا للحصر: ندوة وطنية حول موضوع: الكتاب الأمازيغي إشكاليات التأليف والقراءة،ومقال نشر في مجلة وقائع الندوة الدولية الرابعة للمعالجة الآلية للغة العربية CITALA 12 بالرباط، و مقال شاركنا به في رومانيا ونشر بجريدة Journal of Modern Education Review (usa) بعنوان:" Arabic wordnet :new content and new applications"، عضوة مع فريق ابتكارات بالمدرسة المحمدية للمهندسين بالرباط، عضوة مع منظمة الألكسو بتونس، عضوة في جمعية هندسة اللغة العربية بالرباط.

Towards the Achievement of A Database of Names in Arabic language

Fettoum Krieche
Secondary school teacher qualifying
Member of Innovations Team in Mohamadia School of Engineering in Rabat
Fettoum.krieche@gmail.com

***Abstract:** The study aims to provide a preliminary overview about the implementing of an automatic database of names, so that to be able to implement an automatic lexicon of names in Arabic language depending on the automatic verbs database of all kinds of verbs(True Triple-Salem,True Triple- Almahmoz,True Triple Modaaf,Triple Moatal, Quadruple Modaaf and Non- Modaaf),and that by relying on a set of computational and linguistics steps; to produce a database that contains a range of syntactic, phonological, morphologic, morphosyntactic and semantical information by adopting a holistic concept about the lexicon automatic of verbs which I had programmed. I have managed out of this study to achieve a series of linguistic conclusions.*

Keywords

***Machine Programming, Automatic Lexicon, Arabic Language, Database**.*