



The Egyptian Society of Language Engineering

The Eighteenth Conference on Language Engineering ESOLEC'2018

PROCEEDINGS

December 5-6, 2018

Cairo, Egypt

Table of Contents

Page

I.	<u>Human Language Technologies</u>	
	1. New Trends in Developing the Human Language Technologies	1
	Prof. Dr. Mohsen Rashwan <i>Electronics & Communications Engineering Department, Faculty of Engineering, Cairo University</i>	
II.	<u>Arabic Language Technologies</u>	
	2. لغتنا العربية و التكنولوجيا	3
	أ.د/ وفاء كامل كلية الآداب - جامعة القاهرة	
III.	<u>Electronic Corpora</u>	
	3. المدونات والمفهرسات الإلكترونية وتجارب مصرية	14
	أ.د سلوى حمادة معهد بحوث الإلكترونيات	
IV.	<u>Automatic Arabic Text NLP</u>	
	4. Automatic Arabic Text Summarization: A Pilot Study	16
	Sameh Alansary Bibliotheca Alexandrina, Alexandria, Egypt Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt	
	5. LESAN: LEXical Semantic A Nnotated Resource	25
	Sameh Alansary Bibliotheca Alexandrina, Alexandria, Egypt Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt	
	6. A Syntactic Based Approach to Anaphora Resolution in Arabic	33
	Aya Nabil, Sameh Alansary Phonetics and linguistics Department, Faculty of Arts, Alexandria University	
V.	<u>Text Mining</u>	
	7. Unsupervised Emotion Detection from Text Using Word Embedding	43
	Salma Elgayar, Abdelaziz A. Abdelhamid, Zaki T. Fayed Computer Science Department, Faculty of Computer & Information Sciences, Ain Shams University, El-Khalifa El-Mamoun, Abbassia, Cairo, Egypt	
	8. Mining Publication Papers via Text Mining	54
	Ahmed S. Ibrahim, Sally Saad, Mostafa Aref Computer Science Department, Faculty of Computer & Information Sciences, Ain Shams University, Cairo 11566, Egypt	

VI. NLP Evaluation

9. **بعض برامج التشكيل الآلي والتصحيح: دراسة تقويمية** 60
مدحت يوسف السبع
كلية دار العلوم، القاهرة
10. **الإنسان والآلة في ترجمة النصوص الأدبية** 76
هل يمكن الاستغناء عن الإنسان في الترجمة؟
أسماء جعفر عبد الرسول
قسم اللغة الفرنسية، كلية الآداب، جامعة المنوفية
11. **المعجم الورقي والمعجم الآلي-دراسة لسانية مقارنة** 84
فطوم القریش
(المغرب) عضوة جمعية هندسة اللغة العربية

VII. Corpus Based NLP

12. **Ambiguity in a Corpus Based Approach for Bilingual Ontology** 97
Ahmed R. Elmahalawy *, Mostafa M. Aref **, A.A. Soliman*
*Mathematics Department, Faculty of Science, Benha University, Benha, Egypt
** Computer Science, Faculty of Computers and Information, Ain Shams University, Cairo, Egypt
13. **A Tool for Measuring Linguistic Variations in Machine Translation: A Corpus Based Study** 106
Maram Elsaadany*, Sameh Alansary**
*Institute of Applied Linguistics & Translation, Faculty of Arts, Alexandria University, Alexandria, Egypt
**Bibliotheca Alexandrina, Alexandria, Egypt
Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
14. **A Morphological Analyzed Corpus for Egyptian Child Language** 122
Heba Salama and Sameh Alansary
Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University

VIII. Education

15. **Arabic Educational System Using Augmented Reality** 129
Marwa Elgamal*, Reda Aboelezz**, Salwa Hamada***
*AASTMT, Computer Engineering Department, Cairo, Egypt
**Al Azhar University, System and Computer Department at Faculty of Engineering, Cairo, Egypt
***National Research Institute, Cairo, Egypt
16. **المدونات الإلكترونية واستكشافها والإستفادة منها في تعليم اللغات** 139
سلوى حمادة
ج.م.ع قسم بحوث المعلوماتية بمعهد بحوث الإلكترونيات

IX. Automatic Speech Recognition

17. Syllables Classification for ASR using Variable State Hidden Markov Model

162

Doaa N. Senousy *, Amr M. Gody*, Sameh F. Saad**

*Electrical Engineering Department, Fayoum University, Fayoum, EGYPT

** Modern Sciences and Arts University, 6 October City, Giza, Egypt

New Trends in Artificial Intelligence and Human Language Technologies

Mohsen Rashwan

Electronics and Communications Department, Faculty of Engineering, Cairo University

mrashwan@rdi-eg.org

My talk will start by reviewing the new trends in the area of Artificial Intelligence (AI) and their effect on the life of the humanity. A review of the main factors affecting the advances of AI field includes data, algorithms and hardware. I will discuss the value of the data relative to the other factors. New trends in algorithms that will include supervised, unsupervised and reinforcement models of training. New advances in cognitive sciences that contribute in developing new era of intelligent machines that can learn adaptively and continuously will be introduced.

Then I will focus on the most evolving trends and applications within the field of Human Language Technologies that will include: Chatbots, machine translation, data and text mining, automatic speech recognition, OCR systems and more. I will end up with the main challenges and opportunities that facing the advance of the Arabic language technologies.

اتجاهات البحث في مجال الذكاء الصناعي وتقنيات اللغة العربية

أ.د/ محسن رشوان

قسم هندسة الألكترونيات و الاتصالات، كلية الهندسة، جامعة القاهرة

سأبدأ محاضرتي بمراجعة توجهات الأبحاث في مجال الذكاء الصناعي باعتباره المظلة الأم لكثير من مجالات الأبحاث النشطة هذه الأيام وتأثير ذلك على البشرية. مع ذكر أهم اسباب التقدم في هذا المجال والذي يعزى إلى وفرة البيانات وتقدم العتاد والخوارزمات. وسوف اتوقف قليلا على أهمية وفرة البيانات مقارنة بباقي الأسباب. وسوف أراجع أهم التوجهات الحديثة لتعليم الآلة والتي تهتم بالتعليم المعزز وغير المراقب مقارنة بالتعليم المراقب. وسوف نحتاج لمراجعة النظرية المعرفية لارتباط تقنيات اللغة بها.

ثم سأراجع أهم التطورات التي تتم حاليا ومستقبلا في تقنيات اللغة وتطبيقاتها، والتي تشمل الرد الآلي للمحادثات وتقنية الترجمة الآلية والتعرف على الكلام ورقمنة الوثائق المسوحة ضوئيا.. الخ. وسوف انهي حديثي بتوصيات تخص اللغة العربية حتى لا نتخلف أكثر من ذلك عن ركب التقدم العالمي في المجال.

Biography



Prof. Mohsen A. A. Rashwan

Prof. Rashwan received the B.Sc. and M.Sc. degrees in electronics and electrical communications from the Faculty of Engineering, Cairo University, Cairo, Egypt, another M.Sc. degree in systems and computer engineering from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree in electronics and electrical communications from Queen's University, Kingston, ON, Canada.

He currently serves as a Professor in the Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, Cairo, Egypt, and as the Managing Director of RDI Corporation (www.RDI-Eg.com) that he cofounded in 1993. Over the past 25 years, he has been pursuing and leading R&D projects that focus on basic and applied research as well as building commercial products of Arabic NLP, digital speech processing, image processing, OCR, and e-learning. Among other several national and international mega projects, he has served as a Senior Scientist in the EC's FP7 projects on Arabic HLT NEMLAR and MEDAR, and as a Co-PI in the Egyptian Data Mining and Computer Modeling Center of Excellence (DMCM-CoE). Prof. Rashwan has cofounded ALTEC (an NGO to serve the Arabic language Technologies specially the Language resources: www.ALTEC-Center.org)

لغتنا العربية والتكنولوجيا

وفاء كامل

قسم اللغة العربية، كلية الآداب، جامعة القاهرة

Wafaakamel@gmail.com

❖ اللغة والهوية القومية:

اللغة قوام الفكر والثقافة. وهي أبرز مقومات الشخصية؛ إذ إنها الإطار الذي يحفظ كيان أصحابها ويحدد هويتهم، فضلا عن أنها مرآة العقل، ووعاء الأفكار والمشاعر، وأهم مظهر يتجلى فيه إبداع أبناء الأمة. كما أنها رابطة فكرية تمثل ذاكرة الأمة، فتحتزن تراثها ومفاهيمها وقيمها. واللغة من أهم مقومات وحدة الشعوب. وهي الدرع الواقي لأمتنا في مواجهة جحافل الغزو الثقافي إبان عصر المعلومات.. العصر الذي صار فيه التفوق المعلوماتي قادرا على تحييد القوة العسكرية، أو هزيمتها بتكلفة أقل. وإضعاف اللغة العربية أو ضياعها يعني تحطم الوعاء الأول للثقافة، والمخزون التاريخي للتقاليد والأعراف والفنون والإبداع. ومع مرور الوقت تذوب الهويات الحقيقية، وتطمس الملامح الذاتية. ولا يمكن أن تقوم للعلم المتميز عندنا قائمة إلا إذا تعلمناه، ورسخ في فكرنا بلغتنا القومية، وغرست بذرتة وترعرعت في أعماقنا؛ فصار منا علماء ينتجون العلم، ويضيفون إليه. وفي هذه الحالة لو انقطعت الصلة بيننا وبين العالم كله لبقى العلم مزدهرا عندنا.

❖ **دور اللغة في مجتمع المعرفة جوهري:** فاللغة محورية في منظومة الثقافة؛ لارتباطها بجملة مكونات الثقافة من فكر وإبداع، وتربية وتعليم، وتنوير وإعلام، وتراث وقيم ومعتقدات. ولا يمكن للغة العربية إلا أن تواكب التطور الهائل المتسارع في تقنيات المعلومات، وطوفان مستجدات العلم والحضارة الحديثة، لكي تلحق بركب الحضارة والعلم، ولا تتخلف عن ملاحقته. وهنا يلزم أن نتحدث عن أمرين:

- أولهما: استخدام التكنولوجيا في خدمة الثقافة العربية.
- وثانيهما: مجالات تعزيز اللغة العربية التي يمكن أن تسدّ الفجوة بينها وبين غيرها.

❖ تكنولوجيا المعلومات في خدمة الثقافة العربية:

لكل مجتمع ثقافته التي تحقق لحمته، وتشكل ذاكرته المشتركة وروحه وانتماءه. وهي أشياء لا يمكن التعبير عنها إلا من خلال اللغة الأم. واللغة لا تشكل وسيطا محايدا لنقل المعلومات أو الأفكار أو المشاعر: فهي محملة بتراث ثقافي، وبمنظومة من القيم، وبأسلوب في التفكير والمنطق، لا يمكن فصل

أي منها عن اللغة. والثقافة بشكل عام، والتعليم بشكل خاص، في قلب التنمية البشرية. سواء أكانت الثقافة تعني صقل الأدوات والمواهب، أو تعني النمط المتكامل من المعرفة والاعتقاد والسلوك، والقدرة على التعلم والتحليل.

والبحث العلمي الذي يمثل أساس دفع عجلة التنمية في أي مجتمع، يعتمد على التنمية البشرية: أهم ثروات أي مجتمع. والركيزة الأساسية للتنمية البشرية تتمثل في توفير مصادر غزيرة للمعلومات والمعرفة، في إطار اللغة والثقافة، بصورة يسهل الوصول إليها، وإتاحتها للجميع. فالمعلوماتية محور يخدم الثقافة والتعليم، ويحقق أهداف التنمية، والحفاظ على الهوية وتدعيمها.

وتشكل الشبكة المعلوماتية موسوعة ثقافية وتعليمية في جميع المجالات، بجميع اللغات العالمية: فهي وعاء لنشر الكتب والدوريات والمعارف، كما تستخدم بوصفها وسيلة إعلامية للتعريف بالشعوب والدول والثقافات. وبالرغم من أنها تشكل الطريقة الأسهل والأسرع للوصول إلى المعلومة، فإن المحتوى المعرفي العربي عليها ضعيف: حيث تشير الإحصاءات إلى أن نسبة المحتوى الرقمي للغة العربية ضئيلة، لا تتناسب مع الإرث العلمي والثقافي العربي، ولا مع عدد الناطقين بالعربية. لذا كان علينا النهوض بالمحتوى المعرفي العربي كمًّا وكيفًا؛ سواء من ناحية المحتوى اللغوي أو الأدوات المعينة له.

أهداف صناعة المحتوى الرقمي العربي:

تتمثل هذه الأهداف في أهداف مجتمعية مباشرة، وأهداف عربية بعيدة المدى، هي: تسخير المحتوى الرقمي العربي لدعم التنمية والتحول إلى مجتمع معرفي؛ وذلك ببناء قواعد البيانات التي يستفيد منها الباحثون ومطورو النظم الحاسوبية.

1. ضمان حصول جميع شرائح المجتمع على المعلومات والفرص الإلكترونية.
2. الحفاظ على الهوية العربية والإسلامية للمجتمع.
3. تعزيز المخزون الثقافي والحضاري الرقمي العربي.
4. التمكين من إنتاج محتوى إلكتروني عربي ثري؛ لخدمة المجتمعات العربية والإسلامية. من خلال إثراء المدونة العربية التي تشكل حجر الزاوية في المحتوى الرقمي العربي.

أهم ما أنجز عربيا في صناعة المحتوى العربي:

(استبعت دول المغرب العربي من رصد إنجازات المحتوى العربي؛ لأن اللغة الفرنسية هي السائدة عندهم للمحتوى الرقمي)

➤ موقع مكتبة الاسكندرية:

يحتوي على ست مكنتبات متخصصة ومايقارب العشر بلايين صفحة. ويعد هذا المحتوى أكثر من نظيره الموجود في مكتبة الكونجرس، كما وُثق في موسوعة ويكيبيديا بالنسخة العربية.

➤ موقع إلكتروني لكل أستاذ جامعي:

يساهم الأردن به في إثراء المحتوى الرقمي العربي الثقافي: إذ حرصت اللجنة الوطنية للمعلومات على إنشاء مشروع (موقع إلكتروني لكل أستاذ جامعي). تنشر فيه جميع البحوث والمؤلفات والدراسات التي أعدت في الجامعات الأردنية إلكترونيا. وهو ما يساهم في تعزيز المحتوى العربي في مختلف المجالات الثقافية والعلمية، وتوفيره أمام المهتمين والباحثين.

➤ مشروع "تعزيز صناعة المحتوى الرقمي العربي من خلال الحاضنات التكنولوجية":

أطلقت لجنة الأمم المتحدة الاقتصادية والاجتماعية لغرب آسيا "الإسكوا" ESCWA عام 2007، حيث تقوم بتنظيم عدد من المسابقات الوطنية- بالتعاون مع حاضنات تكنولوجية منتقاة- لاحتضان أفضل مشاريع المحتوى الرقمي العربي في بعض الدول العربية.

➤ برنامج (سواعد):

أطلقت "مؤسسة محمد بن راشد آل مكتوم" برنامج (سواعد)، لدعم المشاريع المبتكرة التي تهدف إلى تطوير كل ما يساهم في تطوير المحتوى العربي، وموارد التعليم والتعلم، والترويج للثقافة العربية.

➤ مبادرة الملك عبد الله للمحتوى العربي:

أطلقتها مدينة الملك عبد العزيز للعلوم والتقنية، بالمملكة العربية السعودية.

➤ المحتوى الرقمي العربي: الواقع، الدلالات، التحديات:

وهي دراسة أنجزتها «مؤسسة الفكر العربي» برعاية من (مركز الملك عبدالعزيز الثقافي العالمي)، وهي الأولى من نوعها في تناولها للمحتوى الرقمي العربي بمنهجية علمية وتقنية. وتشير هذه الدراسة إلى أن المحتوى المكتوب باللغة العربية خارج الوطن العربي يشكّل %7 من إجمالي المحتوى العربي الذي تناولته، ويُشر من خلال %2.8 من إجمالي قنوات نشر المحتوى العربي.

➤ (مجتمع المعرفة والمكتبة الرقمية العالمية):

دراسة شاملة نشرها أ. د. شريف كامل شاهين 2007، تستعرض التجارب في هذا السياق. وتلقي الضوء على مشروع المكتبة الرقمية العالمية.

❖ ثانيا: المجالات اللازمة لتعزيز اللغة العربية:

➤ أولا: التوسع في ترجمة العلوم؛ لملاحقة الاكتشافات العلمية:

لعبت الترجمة إلى العربية دورا كبيرا في العصر العباسي الأول، حين كان الخليفة هارون الرشيد يتقاضى الجزية كتبا، وكان المأمون يعطي لمن يترجمُ كتابا وزنه ذهباً.

كما قامت الحضارة الأوربية الحديثة على ما ترجم عن العلماء المسلمين، في مدارس المدن الأندلسية في إسبانيا وغيرها، واستمرت بلجيكا في تدريس كتاب القانون لابن سينا حتى القرن السابع عشر، كما ظلت الكتب المترجمة للعلماء العرب تدرّس في الجامعات الأوربية حتى القرن الثامن عشر. ومن ثم نلاحظ أن العلم في أوروبا اتكأ على الكتب العربية المترجمة؛ فدرسها واستوعبها وتمثلها، ثم انطلق بعدها إلى آفاق أوسع.

وقامت النهضة المصرية الحديثة في عهد محمد عليّ على البعثات والترجمة، ووضع كلوت بك منهجية علمية تستند إلى تعريب المؤلفات والمصطلحات الطبية وترجمتها، حتى استطاع أن يبني قاعدة طبية من الأطباء المصريين الذين حملوا على عاتقهم التدريس والتأليف بالعربية في القرن التاسع عشر.

وأصدر الوزير الأديب محمد حسين هيكل عام 1938 قرارا بتعريب التدريس في الجامعة، ولكن كلية الطب طلبت استثناء لمدة عشر سنوات؛ حتى يتم وضع المصطلحات العربية، وتأليف الكتب الجامعية. ثم جددت هذه المهلة آليا، حتى تطاولت إلى ثمانية عقود.

إن الترجمة وتعريب العلوم من القضايا المهمة التي يتأكد دورها، وتزداد أهميتها يوما بعد يوم؛ بسبب التسارع التكنولوجي الحديث، وضرورة معاصرة اللغة له. ولا بد في هذا المجال من اتباع سياسة ومنهجية عربية موحدة، تصب في قالب التضامن العربي، وتشد أواصر الوحدة العربية.

ويتعين علينا- لمواجهة طوفان الكلمات والمصطلحات الأجنبية في مختلف نواحي الحياة- أن نلجأ إلى الترجمة الشاملة التي تنقل الحركة الحياتية الكاملة في مجال العلم- بوجه خاص- إلى اللغة العربية، في مقدرة وسرعة وكفاءة؛ بحيث يصبح الذهن العربي متشعبا بحقائقها، مستوعبا ومتمثلا لها، قادرا على تطويرها؛ وهو ما يتيح للعقل العربي الإسهامَ بنصيب في التقدم العلمي العالمي، مستثمرا ملكاته الخاصة وقدراته الذاتية؛ ليضمن للكيان العربي الرسوخ والانطلاق في مجال الحضارة العالمية.

➤ ثانيا: توطين العلم باللغة العربية:

العلوم ثابتة الأصل وتنتقل بلغة ناقلها ومستخدمها: فالطب في الصين باللغة الصينية، وفي ألمانيا باللغة الألمانية وهكذا، وهذا هو التعليم. ولكن النمو والتقدم العلمي يستلزم قدرة وتمكنا من لغة أجنبية شائعة في ربوع المعرفة العلمية. وعن طريق التمكن من هذه اللغة الأجنبية تكون القدرة على استيعاب المعرفة والمعلومات، وسرعة نقلها من اللغة الأجنبية إلى اللغة الوطنية، وهذا هو التعلم. فالتعليم يكون باللغة الأم أما التعلم والتقدم العلمي والتكنولوجي فلا يكون إلا بالتمكن من اللغة الأجنبية التي كتبت بها مراجعها نطقا وكتابة؛ حيث إن بها يكون الاطلاع على المراجع الأجنبية.

إن اللغة الانجليزية تحتل مكانة متقدمة في العالم، ولكننا لا نعرف بلدا واحدا - في غير العالم العربي - أقدم على تدريس مواد العلوم والرياضيات بغير لغته القومية؛ فلا صعوبة كتابة اللغة اليابانية، ولا صغر حجم إسرائيل وبعض دويلات أوروبا - مثلا - قد حال دون أن تكون اللغة القومية هي لغة تدريس العلوم في تلك البلاد.

فباللغة الأم وحدها يتحقق للأمة التقدم العلمي والثقافي والحضاري. ومهما بلغنا مما نظنه تقدما علميا، وخاصة في مجالات العلوم والطب والهندسة، باللغات الأجنبية يظل هذا المدى العلمي محصورا تحت مظلة العلم الأجنبي، تابعا له، يستعمله ويستفيد منه، ولكنه لا يتخطاه ولا يضيف إليه؛ لأن عقول علمائنا تابعة له ومحكومة بما تتلقاه من علم أجنبي اللغة.

وقد أشار تقرير التنمية الإنسانية في الوطن العربي، الصادر في نهاية عام 2003 إلى أن " طريق التنمية لا يتحقق عبر الثقافات الوافدة، كما لا يؤدي ثماره من خلال لغات الآخرين، وإن كان يثرى من تجارب الآخرين بعد ترجمتها إلى اللغة الأم ".

ومن هنا يجب أن نحرص على **توطين العلم باللغة العربية**، وعلى ترجمة البحوث وعرضها بصورة متميزة: بأن يكون العلميون الأكفاء في كل من اللغات الأجنبية واللغة العربية، هم الذين يمكن أن يُعهد إليهم بأعمال الترجمة العلمية، أو التلخيص والعرض باللغة العربية، لنتائج بحوث إخوانهم العرب المنشورة باللغات الأجنبية، وكذلك للفيض المتدفق من الأبحاث العلمية العالمية.

➤ **ثالثا: تعزيز الدراسات الصوتية، والصرفية الصوتية العربية:**

تتفرع دراسة علم الأصوات إلى قسمين: أولهما دراسة نظرية وصفية، والآخر دراسة عملية تجريبية. ومن الأهمية بمكان تعزيز الدراسة التجريبية لعلم الأصوات النطقي باستخدام الأجهزة الحديثة، والتعامل مع البرامج الحاسوبية التي تضبط دراسته وقياساته؛ كي نخلص إلى نتائج علمية دقيقة.

❖ أهمية تعزيز الدراسات الصرفية الصوتية العربية:

تبحث الدراسات الصرفية الصوتية في الأسس والقواعد التي تحكم تكوّن الوحدات المعجمية. وتتوزع موضوعاتها بين الصوت (الفونولوجيا)، وبنية الكلمة (الصرف)، والوحدة المعجمية أي المدخل في (المعجم).

ويمكن الاستفادة من نتائج الدراسات الصرفية الصوتية العربية في اللسانيات التطبيقية واللسانيات الحاسوبية:

ففي اللسانيات التطبيقية يستفاد منها:

• في الصناعة المعجمية Lexicography :

فالربط بين أصوات الكلمة وبنيتها الصرفية يدرس تكوين الكلمة بأنواعها المختلفة؛ بما يمكن من تعميم الإجراءات والنتائج على المعجم العربي كله: فيوضح القواعد التي ينتهجها هذا المعجم في تأليف أصوات وحداته، والقواعد التي تحكم تحقق الأصوات في صيغة صرفية دون غيرها؛ ويعلل ذلك صوتيا أو دلاليا.

• في تعليم اللغة: Language teaching

في مجال تحليل الأخطاء، بضبط عين الكلمة (أو تحديد الباب الصرفي للفعل) ففي العربية أفعال يختلف معناها باختلاف صيغة الفعل المضارع منها؛ والخلط بين أبواب الفعل المضارع عند النطق بالفعل العربي من أكثر الأخطاء المرصودة في تعلم العربية.

• في علم المصطلح Terminology:

فتعميم نتائج الدراسات الصرفية الصوتية على المعجم العربي يوضح قواعد التآلف والتنافر في أصوات الكلمة العربية، فيحدد الأصوات التي يمكن أن تتجاور معا، ومدى تجاورها، والأصوات التي لا تقبل التجاور مع غيرها، وحدود هذا التنافر. كما يحدد قواعد تكوين الكلمات وبنيتها الصرفية. وفي ذلك كبير فائدة لمن يريد وضع قواعد لسك المصطلحات الجديدة، وتعريب المصطلحات الأجنبية.

وفي العمل المعجمي الحاسوبي يستفاد منها:

• في بناء قاعدة بيانات معجمية Lexical database :

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

• في تعرف الكلام Speech recognition :

وهو مبحث 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، وهو ما يسهل عملية الإدراك الآلي للأصوات.

ونورد هنا نماذج لشرائح من الدراسات الصوتية الصرفية توضح:

- أثر مخرج صوتي الفعل الثلاثي المضعف على بابهِ الصرفي.
- أثر الفاء الحلقي للفعل الثلاثي المضعف على بابهِ الصرفي.
- أثر مخرج صوتي الفعل الثلاثي المضعف وصفتهما على بابهِ الصرفي.
- أثر اتفاق صفة الإطباق أو اختلافها على الباب الصرفي للفعل الثلاثي المضعف.
- أثر حيز الفاء مع مخرج العين واللام على الباب الصرفي للفعل الثلاثي المضعف.
- استقصاء للأفعال الثلاثية المضعفة، التي لا تقع في العربية؛ بسبب تنافر أصواتها.

➤ رابعا: تحديث المعاجم العربية:

اللغة لا تقتصر على المفردات، بل تتعداها إلى التجمعات اللفظية، التي تتكون من مفردات ومركبات، تتصاحب لتكوّن مفهوما خاصا، لا يؤديه التعبير بالكلمة المفردة. ولما كانت مهمة المعجم لا تقتصر على تقديم معنى الكلمة المفردة للقارئ، بل تتعدى ذلك إلى مساعدته على فهم النص المقروء واستيعابه، والتعبير الصحيح باللغة؛ لذا كان على المعاجم أن تُدخِل التجمعات اللفظية في دائرة اهتمامها. وقد اهتمت المعاجم الغربية بهذه التجمعات، وأدرجتها في معاجم اللغات الأوربية. ولم تولها المعاجم العربية عناية كافية.

ومن أنواع التجمعات اللفظية:

أ- التصاحب، أو التلازم اللفظي Collocation :

تجمع من الكلمات المتجاورة نحويا، والمترابطة في الاستعمال، يشكل مركبا منسجما من الناحية الدلالية. والكلمات المفردة المكونة لهذه التصاحبات تحتفظ بمعانيها؛ بحيث يرتبط معنى التصاحب اللفظي بمجموع معاني الكلمات المكونة له. ومن أمثله: السوق الحرة، السيولة المرورية، الرسوب الوظيفي، البطالة المقنعة.

ب- التعبير الاصطلاحي Idiom :

مجموعة ثابتة من الكلمات، تشكل تعبيرا يحمل معنى خاصا، لا يمكن استنتاجه من مجموع معاني كلماته المفردة، مثل: السوق السوداء ، حَجَر الزاوية ، دَمُهُ أزرق.

ت- الأفعال العبارية Phrasal verbs :

الفعل العباري تعبير متعدد الكلمات، يضم الفعل وواحدا أو أكثر من الظروف وحروف الجر. ويستعمل المركب الفعلي العباري - غالبا - بمعنى مغاير لمعناه الأساسي. ومن أمثلة الأفعال العبارية: رغب فيه، رغب عنه، ركب فوق أكتاف فلان، رمى إلى كذا، رمى فلانا بكذا، رمى بثقله.

وتعاني معاجمنا العربية قصورا في معالجة التجمعات اللفظية: فالمعاجم التي خلّصت لهذا النوع، بها نقصٌ في استيعاب التجمعات اللفظية. كما أن بها خلطا في التعامل مع أنماط متباينة من التجمعات اللفظية.

مستويات التحديث في الصناعة المعجمية العربية:

➤ التجديد على مستوى التأسيس النظري:

- 1- تحديد أنماط المعاجم القطاعية التي تنقص العربية.
- 2- تحديد أنماط التجمعات اللفظية، واستخلاص معالم كل نمط.

➤ التجديد على مستوى المادة:

- 1- تستخلص هذه التجمعات من مدونة للنصوص العربية المعاصرة.
- 2- يلزم اتساع النصوص في المدونة لتشمل كل مجالات الحياة والكتابة.

➤ التجديد على مستوى تقنيات المعالجة:

1- التحليل الإحصائي Statistical Analysis

- 2- التحليل اليدوي من قبل اللسانيين المتخصصين.

➤ التجديد على مستوى المُخرَج المعجمي:

بإصدار معاجم إلكترونية، إلى جانب المعاجم الورقية.

❖ أهمية المدونات في تحديث المعاجم وإثرائها:

في اهتمامنا بإنشاء معاجم محوسبة للتجمعات اللفظية، يجب ألا نقنصر على جمعها من المعاجم التراثية العربية، التي لم تصل باللغة إلى أبعد من القرن الرابع الهجري: إذ لا بد أن تكون معاجمنا مرآة صادقة تعكس لغتنا العربية الفصيحة، التي نستخدمها، ونؤلف بها كتبنا العلمية والأدبية.

لذا يلزم اعتماد مدونة لغوية حديثة ضخمة، تُستخلص منها المصطلحات، وألفاظ الحضارة المستحدثة، والتجمعات اللفظية بأنواعها، وفي سياقاتها النصية، ثم تأتي المعالجة المعجمية لها. بهذا نكون قد حدّثنا

المعجم العربي: فأدرجنا به ما استُحدث من المصطلحات وألفاظ الحضارة، وما أقرته الاستعمالات اللغوية الفصيحة من ألفاظ وأساليب؛ وساعدنا قارئه على فهم النص المقروء واستيعابه، والتعبير الصحيح باللغة.

➤ خامسا: إطلالة على علم اللغة الحاسوبي:

علم اللغة الحاسوبي فرع من فروع علم اللغة تستخدم فيه التقنيات والأفكار الحاسوبية؛ لتوضيح المشكلات اللغوية والصوتية. وهو دراسة علمية للغة الطبيعية من منظور حاسوبي، تهدف إلى إنتاج أنظمة حاسوبية قادرة على فهم اللغة الطبيعية وإنتاجها. فعلم اللغة الحاسوبي يعد مجالا وسيطا بين علم اللغة: الذي تضم مادته جميع مظاهر اللغة: الصوتية والصرفية والنحوية، والدلالية والمعجمية، وعلم الحاسب الآلي الذي يتم من خلاله بناء نظم معالجة اللغة بمستوياتها المختلفة.

بدأ المعجم العربي دخول عصر الحاسوب في سبعينيات القرن الماضي وكانت البدايات إحصائية، ثم اكتسبت المعالجة الآلية للعربية اهتماما متزايدا من الباحثين العرب. كما اهتمت دول غير عربية بالمعالجة الآلية للغة العربية. وعلى الرغم مما اكتسبته المعالجة الآلية للغة العربية من اهتمام إلا أن الهوة لا تزال كبيرة بين المنجزات العربية ونظيراتها في أوروبا وأمريكا؛ ولذلك تبدو الحاجة ملحة إلى مزيد من الجهد والتعاون بين اللغويين والحاسوبيين؛ حتى يضاهي المنجز العربي في هذا الصدد نظيره في الغرب.

❖ المعجمية العربية الحاسوبية

أحدث دخول الحاسوب مجال العمل المعجمي ثورة عارمة في تقنيات الصناعة المعجمية، أعقبها ثورة مماثلة في المفاهيم والمعتقدات والتقاليد المعجمية، ظهرت آثارها في الجوانب الآتية:

- * في تقنيات المعجم التقليدي.
- * في إيجاد معاجم أو مصادر معجمية جديدة.
- * في إجراءات البحث المعجمي، ومنطلقاته وغاياته.
- * في إيجاد مجالات بحثية معجمية جديدة.
- * علاوة على توظيف الحاسوب باعتباره وسيطاً بين المعجم وفروع معرفية أخرى، وتطبيقات متنوعة.

ومن الأنشطة التي تدخل في نطاق المجال المعجمي الحاسوبي ما يلي:

- ❖ المعاجم في صورتها الإلكترونية.
- ❖ تشييد معاجم لدعم نظرية لسانية حاسوبية خاصة، أو نحو حاسوبي.
- ❖ دراسة النصوص المكتوبة أو المقروءة لتعزيز مداخل المعجم.

وترتبط معالجة اللغات الطبيعية Natural Language Processing بالمصادر المعجمية الحاسوبية ارتباطا وثيقا، يظهر فيما يُنتج من تطبيقات معجمية في مجال معالجة اللغة الطبيعية مثل البرامج المكتبية، وتحليل الكلام وتركيبه، والتلخيص الآلي، والفهرسة، واستخلاص المعلومات، والترجمة الآلية، إلخ، علاوة على العون الذي تقدمه المعجمية الحاسوبية في مجال تعليم اللغات، وإنشاء بنوك المصطلحات، والمولدات الآلية للمصطلحات.

إن اهتمامنا ببناء صرح حضاري علمي يلقي علينا مسئولية مزدوجة : أن نتابع التواصل مع العالم الحضاري الحديث بأسلوبه وبلغته، وأن نركز اهتمامنا لبناء صرحنا الداخلي، وتدعيمه، وتوفير البنية التحتية له ؛ كي يكون قويا متينا في مواجهة مستقبل غامض حافل بالتحديات.

السيرة الذاتية

الإسم واللقب : الدكتورة / وفاء محمد كامل أمين فايد

أستاذة متفرغة في اللغويات، بكلية الآداب جامعة القاهرة



- 1) أول سيدة تنتخب عضوة بمجمع اللغة العربية بالقاهرة، عام 2014، بعد ثمانين عاما من إنشائه.
- 2) خبيرة بمجمع اللغة العربية بالقاهرة من عام 2007 - 2014 .
- 3) عضو مراسل بمجمع اللغة العربية بدمشق من 2002.
- 4) حصلت على الجوائز التالية:
 - أ. لقب الأستاذة المثالية لجامعة القاهرة عام 2004 .
 - ب. جائزة جامعة القاهرة التشجيعية للعلوم الإنسانية والاجتماعية عام 2004.
 - ج. جائزة جامعة القاهرة التقديرية للعلوم الإنسانية والتربوية عام 2013.
 - د. جائزة التميز من جامعة القاهرة في العلوم الإنسانية عام 2016.
 - هـ. جائزة ودرع الجمعية المصرية لهندسة اللغة؛ عن بحوثها المتميزة في المجال، 2014.
 - و. درع الملحقية الثقافية السعودية بالقاهرة ومركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية؛ لفوزها في انتخابات مجمع اللغة العربية بوصفها أول عضوة فيه 2014.

ز. درع من قسم اللغة العربية وآدابها بجامعة طنطا، في اليوم العالمي للمرأة 2016.

ح. درع وزارة الثقافة ؛ لكونها أول سيدة تفوز بعضوية المجمع في تاريخه 2018.

الكتب المؤلفة والمترجمة:

- 1- (شرح عيون الإعراب للفزاري) تحقيق ودراسة.
- 2- (تراكب الأصوات في الفعل الثلاثي الصحيح).
- 3- (اتجاهات البحث اللساني) كتاب مترجم بالاشتراك.
- 4- (قصيدة الرثاء بين شعراء الاتجاه المحافظ ومدرسة الديوان: دراسة أسلوبية إحصائية).
- 5- (الباب الصرفي وصفات الأصوات: دراسة في الفعل الثلاثي المضعف).
- 6- (بحوث في العربية المعاصرة).
- 7- (المجامع العربية وقضايا اللغة).
- 8- (معجم التعابير الاصطلاحية في العربية المعاصرة).
- 9- (علاقة البنية الصرفية بمخارج الأصوات: دراسة في الفعل الثلاثي المضعف).
- 10- (مدخل إلى اللغة): فيكتوريا فرومكين - ترجمة وفاء كامل - المركز القومي للترجمة.

* * * * *



المدونات والمفهرسات الإلكترونية وتجارب مصرية

أ.د. سلوى حمادة

معهد بحوث الإلكترونيات

hesalwa@hotmail.com

الملخص

أصبح من المنفق عليه ضمناً بين العاملين في مجال دراسة اللغة وتحليلها وعلم اللغويات في الدول المتقدمة أن أية دراسة لا تتم من خلال مدونة تمثل اللغة المستخدمة فعلياً سواء المكتوبة أو المنطوقة تكون دراسة عقيمة ناقصة لا برهان لها. ومع أهمية استخدام المدونات في استقصاء العلاقات اللغوية بدأ مع بدء الدراسات اللغوية إلا أن صور وأشكال المدونات المستخدمة قد تنوعت. لقد كان على القدماء أن يستعينوا بعقولهم ومحفوظاتهم للاستشهاد والتمثيل للظواهر اللغوية التي يريدون التمثيل لها، وهو ما يمكن تسميته بالمدونات الذهنية، أو عمل قصاصات ورقية للأمثلة والشواهد، وهو ما يمكن تسميته بالمدونات الورقية. وقد تطورت المدونات من الصورة الذهنية إلى الصورة الورقية إلى الصورة الإلكترونية حتى وصلت للصور المتاحة حالياً على الشبكات سواء مجانية أو بأجر. ومن ثم نشأ علم التدوين اللغوي وهو العلم الذي يعنى بالدراسات المبنية على المدونات، وسنرى الخلاف على تسميته علماً من خلال هذا الفصل. وهذا المقال مفتاح لدراسة المدونات؛ حيث يبدأ بتعريف المدونات وينتهي بالتطبيق والاستخدام.

والمدونة هي كم من البيانات الموثقة دائمة التطور؛ فيجب أن نتقبل أنه لا يوجد حدود يمكن وضعها للغة الطبيعية طالما أن حجم مفرداتها وأبنيتها اللغوية وتراكيبها ذات المعنى والعلاقات المختلفة بين الوحدات اللغوية يتطور باستمرار. وعليه لا توجد مدونة مهما كان حجمها ودقة تصميمها يمكن أن تمثل خواص اللغة كاللغة المستخدمة ذاتها. وتبنى المدونة وتصمم لدراسة اللغة ولأغراض لغوية تفيد التطبيقات المتعددة في (الطب والهندسة والجغرافيا). ولا بد بعد إتاحة المدونات الآلية وأدواتها من التوسع في تقييس القواعد التي قد تخالف ما اتفق عليه قبل ظهور هذه الآليات وهو الأمر الذي كان يعتمد على الخبرات والقدرات الشخصية لفرد أو جماعة. ويجب على صناعات المدونات أن يهتموا بأن تكون المدونة ممثلة للغة قدر الإمكان فكلما كانت المدونة ممثلة للغة كشفت عن مميزات في اللغة لم تكن في الحسبان عند عمل المدونة، أو على العكس قد تفشل المدونة في تمثيل الخواص اللغوية المطلوبة منها. ولا يعني ذلك أنه يجب أن يتقيد صناعات المدونات بمبدأ تجاهل الخواص اللغوية المطلوب البحث عنها ولكن عليهم التأكد من كفاية معايير التصميم في تحقيق المطلوب وقد أكدنا من قبل أنه لا توجد مدونة تمثل اللغة بمفرداتها وتراكيبها اللغوية لأن اللغة غير محددة وفي حالة تطور ونمو مستمر في المفردات والتراكيب. ومعايير الاختيار لا تكون محددة ومقيدة ولكن يدخل فيها معايير شخصية للعاملين في المجال – بخلاف النظم الأخرى- فاختيار المدونة يعتمد على الحدس والإدراك والمهارات الشخصية أيضاً. ولكن للأسف، على الرغم من مساعدة هذه المعايير بالإضافة للمعايير الأخرى إلا أنه لا يمكن تحديدها ويمكن أن تتفاوت من شخص

لآخر تبعا لثقافته وحده. بعض المدونات قد تخضع لتغيير الخصائص كالمدونات التاريخية مثلاً. وهناك المدونات المتوازية والتي تتضمن أكثر من لغة وتعرض تنوعات مختلفة لخدمة عرض المقابلة بين اللغات. والمدونة تخدم المجالات الحياتية كلها بصورة أو بأخرى كما سيتضح من العرض.

والمدونة التعليمية تخدم المتعلم من غير أبناء اللغة الذي يريد تعلم لغة غير اللغة الأم. ويجب أن يعرف أن معايير هذه اللغة سواء كانت منطوقة أو مكتوبة قد تختلف اختلافا كبيرا عن المعايير الموجودة في الكتب. هناك مثلا اللغة العامية والفصحى والخليط بينهما، فبالطبع أن إحداها لغة رئيسية تُمَثَلُ الاستخدام الأشمل. ومعرفة تلك المعايير ليس أمراً سهلاً، وبمعنى آخر؛ فإن طريقة المتكلم المحلي في التعبير عن شيء ما لن يجدها المتعلم متوافرة في كُتُب القواعد؛ حيث إنها ليست قادرة على إجابة أسئلته عن استخدام مفردات وتراكيب اللغة جميعها خصوصاً الأسئلة التي تدور حول صور استخدام الكلمة. والقواميس العربية الحديثة ليست معتمدة على البحث في مدونة اللغة العربية المعاصرة، أو مزودة بمعلومات موثوقة. ولا توجد كتب قواعد مستخلصة من البحث في المدونة حتى الآن حتى بالنسبة للغة الإنجليزية أيضاً على الرغم من أنها قطعت شوطاً في هذا المجال. وإن وجدت هذه الكتب فهي لاستعمال محترفي اللغة أكثر من المستعمل العام. ومن أمثلتها قواعدُ Longman للإنجليزية المكتوبة والمنطوقة. ولا يُمكن إنكار أن لغة المدونات أغنى بشكل ملحوظ من اللغة المؤلفة، أو الأمثلة المختارة بعناية، للإدراج في الكتب الدراسية والقواميس في عمليات البحث. هذا الغنى richness عامل مهم في تعليم اللغة.

سننتحدث عن المدونات وأهميتها، وكيفية تصميمها واستخداماتها. ونوضح أهمية الحواشي Tagging ونورد نماذج لها. ونتعرض للمفهرسات الآلية باختصار فيها يتحقق الغرض الأساسي من عمل المدونات أو بلفظ أوضح هي آلية الاستفادة من المدونات. وسنعرض في نهاية المقال تجربتنا الخاصة في عمل المدونات والمفهرسات.

Automatic Arabic Text Summarization: A Pilot Study

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt

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

sameh.alansary@bibalex.org

Abstract— one of the most relevant today’s problems is caused by the rapid growth of the web, which is called information overloading. It has increased the necessity of more sophisticated and powerful summarizers. This paper shortly introduces a historical overview along with the advancement of automatic text summarization and the most relevant approaches currently used in this area. In this paper, the proposed automatic text summarization system involves different stages: sentences extraction stage, syntactic analysis stage, and generation stage. Finally, the evaluation summarization process will be discussed.

Keywords: Text summarization (TS), Multi-document Summarization (MDS), Arabic Query-Based Text Summarization System (AQBTS), Arabic Concept-Based Text Summarization System (ACBTSS).

1 INTRODUCTION

Electronic documents are produced and made available on the internet each day, so everyone should benefit from this revolution of information. The crucial way to access these documents and get the gist of it is to be able to extract the content of these documents and utilize from the information existing in these documents. The summarized text simply has to be shorter than the original text, it can be from one or more documents and contains only the important information. Therefore, when we extract this summarized version from a certain source or sources by means of a computer, i.e. automatically, we call this Automatic Text Summarization. It is worth mentioning that automatic text processing is a research field that is currently extremely active. Automatic text summarization, aims to reduce the size of a text while preserving its information content. A summarizer is a system that produces a condensed representation of its input’s for user consumption [1].

Radev et al. (2002) defines a summary as “a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that” [2].

In order to generate a summary of a document or a sentence, we have to identify the most significant pieces of information that exist in a document or a sentence, omitting the redundant information and reducing details. A summary can be employed in an indicative way – as a pointer to some parts of the original document, or in an informative way – to cover all relevant information of the text. In both cases, the most important advantage of using a summary is its reduced reading time [1]. Summary generation by an automatic procedure has also other advantages: (i) the size of the summary can be controlled; (ii) its content is determinist; and (iii) the link between a text element in the summary and its position in the original text can be easily established [1].

Summarizing a text or a sentence requires work in natural language understanding, semantic representation, discourse models, world knowledge, and finally natural language generation.

Text summarization can be implemented in many applications that will help in facilitating life, for instance we do not have the time to read news with all its details, so we can summarize news to SMSs or to news mobile applications that will save time. Text summarization can be used to teach the computer how to read synthetically a summarized text because written texts can be boring and long. In addition, one of the most important applications of that TS can be implemented in is the search engines where a compressed description about each result is needed. In addition, businessmen need fast summaries for their reports and documents, researchers need a tool to generate fast summaries for their references, and many people with different needs need such application.

Although research on text summarization has started in the 1950s, only little work was done in the first 30 years; the progress was slow. Due to the lack of powerful computers and Natural Language Processing (NLP) techniques, early work focused on the study of text genres such as position and cue phrases. From the 70s to the early 80s, Artificial Intelligence (AI) was applied. The idea was to employ knowledge representations, such as frames or templates, to identify conceptual structures of a text and find salient concepts by inference. In the 1990s, because of information explosion in the World Wide Web, automatic text summarization became crucial to reduce information overload and this brought about its renaissance. It could be used for different purposes and different users and, thus, various types of summaries have been constructed. There were many attempts to construct an automatic text summarizer, therefore the field of automatic Text Summarization (TS) has experienced a big growth because of the new technologies and the amount of information that World Wide Web brought to us [2].

We propose a system that uses different stages in order to summarize documents. The first stage is responsible for extracting the most informative sentences of the document. The second stage is responsible for excluding the constituents that does not convey informative meaning. Each stage will be described in details.

In the following section, we will illustrate the different approaches of Automatic Text Summarization. Section 3 presents the existing summarizers systems for Arabic and other languages. Section 4 introduces the proposed system and the processing stages; it also presents the types of constituents that are considered non-essential in a certain summary. Section 5 will discuss the different methods that are used to evaluate any automated text summarizer and evaluate the proposed system. Finally, section 6 will conclude the paper.

2 AUTOMATIC TEXT SUMMARIZATION DIFFERENT APPROACHES

In the recent years, many approaches have appeared due to the huge amount of information overloaded on the Web, but there are two dominant approaches of summarization:

A. Extraction approach

This approach makes use of different properties of the text to weight the sentences by using a combination of statistical heuristics. Each sentence in the text is assigned a score using a combination of statistical heuristics [3]. The sentences are sorted in descending order according to their score values and an appropriate number of the highest scored sentences are selected from the text to form the summary according to the summarization ratio. The sentences that have the highest scores are considered very important and included in the summary.

Extraction is mainly concerned with judging the importance or the indicative power of each sentence in a given document. All sentences are first rated in terms of their importance, and then a summary is obtained by choosing a number of top scoring sentences [4].

B. Abstraction approach

The abstraction approach involves simplifying and condensing the text. When text summaries are created manually using the abstraction approach, humans read the text, reinterpret it, and rewrite it [3]. Producing abstracts is not a simple task because central topics have to be identified, topics have to be interpreted, and then the summary is generated. The abstraction approach is much more difficult to be programmed than extraction, therefore extraction is the more commonly used approach in automatic text summarization [3].

3 EXAMPLES OF EXISTING SUMMARIZATION SYSTEMS

This section illustrates different systems of automatic text summarization for different languages. Much work has been done on automatic text summarization. Natural language processing for Arabic language is more sophisticated and needs much time compared with the accomplishments in English and other European languages. These languages are discriminated from Arabic by their writing direction that flows from right –to- left, capitalization to identify proper names, acronyms, and abbreviations. Furthermore, they are rich with corpora, lexicon, and machine– readable dictionaries, which are essential to advanced research in the different areas [5]. The actually implemented systems can be divided into two categories: Systems implemented on Arabic languages and other systems implemented on other languages rather than the Arabic:

A. Arabic Text Summarization Systems

There are number of Arabic text summarization systems. The Arabic Query-Based Text Summarization System, the Arabic Concept-Based Text Summarization System [5], Sakhr summarizer, Lakhas Arabic summarizer, and the summarizer of Aramedia.

- 1) The Arabic Query-Based Text Summarization System (**AQBTSS**): It uses standard retrieval methods to map a query against a document collection and to create a summary [5]. This system gives a summary for the document in accordance to the organized query.
- 2) The Arabic Concept-Based Text Summarization System (**ACBTSS**): It creates a query-independent document summary. It uses a bag-of-words representing a certain concept as input to the system [5]. The summarization is sought consistent with the sentences that best match the query or the concept.

- 3) Sakhr Arabic Summarizer^[1]: It is a commercial online Arabic text summarization system available on the web. It should be noted that the system was only a beta release at the time we performed our experiments. It provides companies with short text summary for each item using a prioritized list of key sentences and gives the ability to select a specific level of summarization. The summarizer consists of a set of text-mining tools to identify the most relevant sentences within a document and displays them in the form of a prioritized list of key sentences [6].
- 4) Lakhas: It is a text summarization system using extraction techniques. It is the first Arabic summarization system to be formally evaluated and compared with English competitors in an evaluation competition [7].
- 5) ^[2]Aramedia: It is a summarizer used for summarizing Arabic and English documents. Based on linguistic analysis of the document, it extracts the main ideas to make it possible for the user to preview these ideas instead of reading the whole document, thus saving time and effort.
- 6) Ikhtasir: is an automatic extractive Arabic text summarization system where the user can cap the size of the summary. The system does not require learning and employs Rhetorical Structure Theory (RST) along with a sentence-scoring scheme, where individual sentences are scored. The scoring will be used to decide the importance of each sentence in the text. The system does not require any training and has the unique feature of enabling the user to define the size of the final summary itself through number of sentences, or number of words, or as a percentage of the original. The Ikhtasir algorithm is an eight steps process. These are: Sentence segmentation, Word segmentation, Stop-words removal, Root extraction, Frequency computation, Generation of RS-tree, Sentence scoring and finally, we generate the summary within the user prescribed size [8].

B. Non-Arabic Text Summarization Systems

- 1) Copernic Summarizer^[3]: It is statistically based and deals with four languages. After many years, support for Copernic Summarizer ended on April 30th 2016. The various engines behind Copernic Summarizer are no longer updated or fixed. Since October 1st 2015, all versions of Copernic Summarizer are no longer available for download or purchase. Copernic no longer supports, updates or provide technical support for Copernic Summarizer.
- 2) Pertinence summarizer^[4]: It performs linguistic processing over a document and evaluates the pertinence (the relevance) of its sentences.
- 3) Kify Text Summarizer^[5]: It uses text semantic indexing and some clever mathematics, the summarizer decides which parts of a document are important by analyzing the content.
- 4) SweSum^[6] : It is the first automatic text summarizer for Swedish. It summarizes Swedish news text in HTML/text format on the WWW. SweSum is based on both statistical and linguistic methods as well as heuristic methods. SweSum uses a 700.000 word entries dictionary. SweSum is also available for Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi (Persian) and German texts. It has been evaluated and its performance is estimated to be as good as the State of the art techniques, for English, i.e. an average of 30% summary, compression, of 2-3 pages news text gives a good summary.
- 5) MEAD^[7]: It is the most elaborate publicly available platform for multi-lingual summarization and evaluation. The platform implements multiple summarization algorithms such as position-based, centroid-based, largest common subsequence, and keywords. The methods for evaluating the quality of the summaries are both intrinsic and extrinsic. MEAD implements a battery of summarization algorithms, including baselines (lead-based and random) as well as centroid-based and query-based methods.

[1] <http://textmining.sakhr.com/>

[2] <http://aramedia.com/summarization.htm>

[3] <http://www.copernic.com/en/products/summarizer/>

[4] <https://www.office-forums.com/threads/pertinence-summarizer-multilingual-software-of-automatic-text-summarization.2217612/>

[5] <http://text.kify.com/>

[6] <http://swesum.nada.kth.se/index-eng.html>

[7] <https://people.dsv.su.se/~hercules/textsammanfattningeng.html>

4 THE ARCHITECTURE OF THE PROPOSED SYSTEM

The proposed system depends on different stages in order to generate a summarized version of documents. The first stage involves extracting of the most important and informative sentences in the document. The second stage involves excluding the less important constituents, the ones that do not contribute to the meaning of the document. Each stage will be described in details in the following two subsections:

A. Sentences extraction stage

The summarization system selects the top most important sentences to produce a summary. Many indicators should be taken into consideration while selecting the important sentences. For example, the context in which the summary is created may be helpful in deciding the importance. The type of the document (e.g. news article, email, scientific paper) is another factor which may impact selecting the sentences. Moreover, the sentence length, word frequency, and the number of topic words the sentence contain, they are all helping factors in determining the most important sentences. Then, an importance score is assigned to each sentence. The score of a sentence represents how well the sentence explains some of the most important topics of the text. The score is calculated by aggregating the evidence from different indicators. Identifying the topic could be achieved by making a list of certain words for each topic; these lists have been compiled based on how frequently the words have appeared in the training data for each topic as in Figure.1. The system depends on three linguistics resources, namely; the word-net, UNL encyclopedias, and EDGES.

EDGES: is the Entity Discovery and Graph Exploration System, a user-friendly visualization tool used for exploring semantic networks by enabling concept (words of the document) expansion, collapsing and navigation. For more information, see [9].

UNL Encyclopedia: It is also known as *UNL Example Base*; it contains semantic relations between UWs along with a degree of probability. It is built automatically by analyzing large corpora, thus, it comprises information that is related to the probability of occurrence rather than the possibility of occurrence. It is the only component in the UNLarium framework created based on the statistical approaches. For more information, see [9].

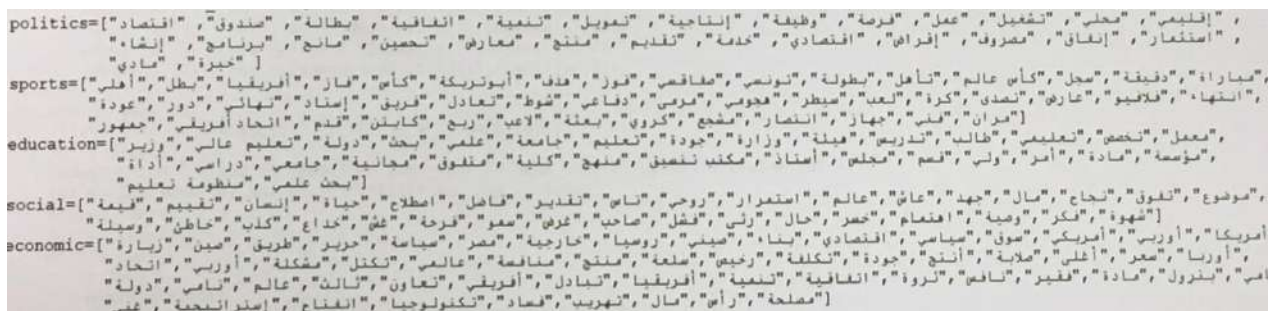


Fig.1. a list of topic words for each topic

B. Syntactic Analysis stage

This stage utilizes the output of the sentence extraction. This stage includes identifying which constituents could be omitted and which are vital, as most sentences could include two types of constituents: **incident** and **principal** constituents. The incident constituents are less important than the principal ones [5], for example a sentence like as in (1):

sentence 1) “ذهب الولد صباحاً إلى المدرسة”

The subject "الولد" and the verb "ذهب" are the principal constituents of the sentence, while 'صباحاً' is an incident constituent. Omitting the incident constituent in this sentence "صباحاً" will affect neither the grammatical soundness, nor the content.

The system represents the constituents of each sentence using X-bar theory in order to categorize them into incident and principal.

In X-bar theory, the head should be determined first, which will determine the type of phrase, and then the specifiers, complements, adjuncts, and conjunctions will be decided accordingly. X-bar postulates the similarities between different categories of lexical phrases by assigning the same structure to all phrases as shown in figure 2:

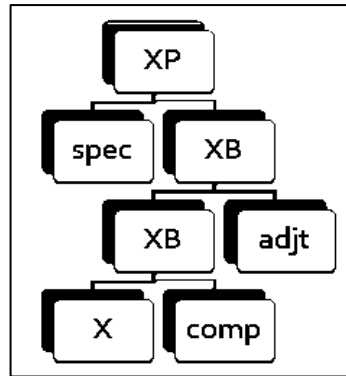


Fig.2. Basic X-bar Structure.

The claims represented in the schema in figure 2 are the following: All phrases are projected from lexical categories in the same way and a head (X) subcategorizes for all and only its sisters, moreover specifiers are optional and there may be words or phrases. There are two basic rules that cover all the lexical categories as shown in rule (1).

rule 1. Phrase Structure Rules:

- a) $XP \rightarrow \text{Specifier XB}$
- b) $XB \rightarrow X \text{ Complements}$

In the trees generated by rules (1), the top node (corresponding to the left side of the rule) is known as the mother node, with the two daughters introduced by the right side of the phrase structure rule. The daughter nodes at the same level are known as sisters, the two sister nodes in the right side of the specifier and XB (Intermediate projection) in rule 1 (a).

The head, specifier, and complement for each category should be determined in order to draw the simple and complex phrases as well as complete sentences in X bar theory.

X-bar theory focuses on drawing the tree of the deep structure of the sentence by placing specifiers, complements, and adjuncts in the subtree of their constituent. Adjuncts are removable but for complements and specifiers, the head of the sentence should be examined. For example, in nominal phrases (NPs), if there is an article, it is considered as the specifier of the NP and it cannot be removed, but in adjectival phrases (JPs), the specifier is always an adverb, and in this case it can be removed. So here, we will divide the constituents into two main divisions: complements, and modifiers which are optional [5].

For some nominal heads, complements are mandatory as we have already established, but for the other types of heads, complements may be mandatory or optional. This depends on two things, first the lexical head category, and second the head instance (lexical entry). The lexical entries are assigned with linguistic features which is a set of linguistic information developed to describe every Arabic word. Arabic words in the developed system are described on different linguistic levels: morphological information, morpho-syntactic information, syntactic information, and semantic information. Details about the dictionary and each linguistic feature is described in details in [10].

The same with prepositions which require a mandatory complement. The important thing here is that once we have all the subcategorization information for a certain head, we can decide if it has a mandatory or an optional complement. Applying the X-bar theory on Arabic has been described in details in [10] and [11].

During the syntactic analysis stage, the developed summarization grammar marks the constituent that should be deleted and could decide the constituents that should not be deleted. For example, consider sentence in (2) and its syntactic representation in figure 3:

sentence 2) “هي تستعين بالخدمة في البيت”

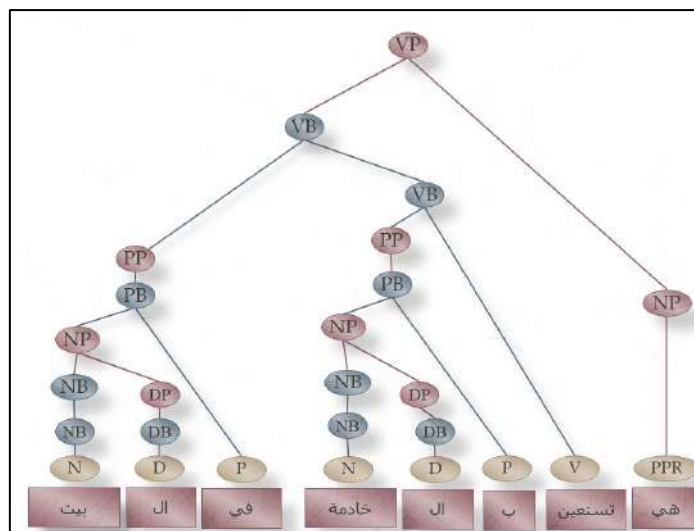


Fig.3. The tree of the sentence “هي تستعين بالخادمة في البيت” in by X-bar representation.

In figure 3, there are two prepositional phrase “بالخادمة” and “في البيت”, the summarization grammar could identify that the PP “بالخادمة” is the complement of the verb “تستعين” which should not be deleted; however, the second PP “في البيت” is the adjunct, so it should be deleted. The summarization grammar has the ability to discriminate between the two PPs in the sentence in (2) and capable of deciding which PP is incident and which one is principal.

Moreover, the summarization grammar could detect that some constituents which have specific grammatical functions could be deleted from a document during the summarization, consider the highlighted constituent in the document in figure 4 which have the following grammatical functions:

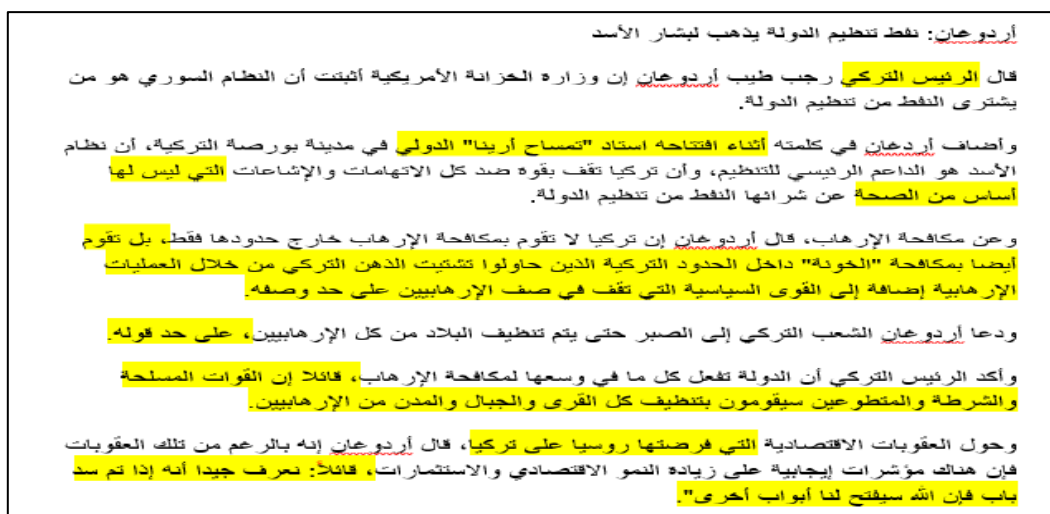


Fig. 4. Original Document

a. Appositions

The phrase “الرئيس التركي” does not give a distinct meaning when it is followed by a proper noun which is considered as the specifier of the preceding verb.

b. Adverbial phrases

The phrase “أثناء افتتاحه استاد "تمساح أرينا" الدولي” does not add additional information to the sentence’s meaning.

c. Relative clauses

The relative clause “التي فرضتها روسيا على تركيا” is mainly a redundant clause to the previous chunk or word. It is considered an incident constituent that should be deleted.

In figure.4, the highlighted sentences represent the sentences, which had been excluded because of their little to no addition or effectiveness in the meaning of the entire document. In figure.4, the original text counts 192 words. The manual reference of the document in figure.3 counts 90 words, while the automatic summarized text counts 101 words as shown in figure 5.

أردوغان: نفض تنظيم الدولة بذهب ليشار الأسد
قال رجب طيب أردوغان إن وزارة الخزانة الأمريكية أثبتت أن النظام السوري هو من يشتري النفط
من تنظيم الدولة.
وأضاف أردوغان في كلمته أن نظام الأسد هو الداعم الرئيسي للتنظيم، وأن تركيا تقف بقوة ضد كل
الانتهاكات والإشاعات.
قال أردوغان إن تركيا لا تقوم بمكافحة الإرهاب خارج حدودها فقط،
ودعا أردوغان الشعب التركي إلى الصبر حتى يتم تنظيف البلاد من كل الإرهابيين،
وأكد الرئيس التركي أن الدولة تفعل كل ما في وسعها لمكافحة الإرهاب،
وحول العقوبات الاقتصادية، قال أردوغان إنه بالرغم من تلك العقوبات فإن هناك مؤشرات إيجابية
على زيادة النمو الاقتصادي والاستثمارات.

Fig. 5. The automatic summarized document

There have been 100 documents that were chosen to be summarized manually. This data was divided into two-thirds to be used as training data and one-third to be used for testing. The training data was summarized manually by four persons. As a result, four manual references were built and were ready for testing the automatic text summarization outputs. The manual summarization texts were obtained through several steps: firstly, **reading** the whole document carefully. Secondly, **extracting** the highlights of the document. Thirdly, **assigning** to the document the category it belongs to and **choosing** a title for the topic. Finally, **selecting** only the informative sentences or phrases “principles” and ignoring the ones that are not equally informative. These manually summarized documents were used as a reference data.

There are two engines that are used in the Arabic automatic summarization during the linguistic processing stage, the first is Interactive ANalyzer (IAN) which is used in the analysis process [10], the second is dEep-to-sUrface natural language GENERator engine (EUGENE) which is used in the generation process [11]. The syntactic analysis produces the output in a certain format and the generation tool is responsible for producing an Arabic text.

5 EVALUATION

Evaluating of an automatic text summarizer is not a simple task. There are many tools and measures that can be used to do the evaluation, but most of the researches agreed that there are two main points that have to be measured in the summary: the Compression Ratio (CR) (how much shorter the summary is than the original) [12], [13] and [14] and the Retention Ratio or Omission Ratio (RP) (how much information is retained).

$$CR = \text{length of Summary} / \text{length of Full Text.}$$

$$RR = \text{information in Summary} / \text{information in Full Text.}$$

In the developed proposed system, the evaluation process consisted of two parts. The first part of the evaluation considered the extracted sentences only; the output is compared to the reference, which was manually summarized. The second part considered the omitted sentences, also the automatically omitted sentences were compared to the manually omitted sentences. The results have shown a high degree of similarity in terms of either selecting or excluding sentences.

6 CONCLUSION

The history and the definition of the automatic text summarization were discussed in details. Some examples of the existing summarization systems have been mentioned. The attention was paid to the techniques that are used and the different approaches in this field. A distinction has been made among the different types of constituents that can be removed from the original text. The evaluation of the developed summarization system has been done manually.

REFERENCES

- [1] J Larocca Neto and A. Freitas and A. A. Kaestner “Automatic Text Summarization using a Machine Learning Approach”, Pontifical Catholic University of Parana (PUCPR) Rua Imaculada Conceicao, 1155.
- [2] D Das and F.T. Martins “A Survey on Automatic Text Summarization” Language Technologies Institute Carnegie Mellon University, November 21, 2007.
- [3] A Mlynarski “AUTOMATIC TEXT SUMMARIZATION IN DIGITAL LIBRARIES” B.Sc, University of Lethbridge, 2003.
- [4] A Kiani –B and M. R. Akbarzadeh –T “Automatic Text Summarization Using: Hybrid Fuzzy GA-GP”, Department of Electrical Engineering Ferdowsi, University of Mashhad, Mashhad, Iran, July 16-21, 2006.
- [5] A Haboush and A Momani and M Al-Zoubi and M Tarazi “Arabic Text Summerization Model Using Clustering Techniques”, World of Computer Science and Information Technology Journal (WCSIT), 2012.
- [6] M El-Haj, Udo Kruschwitz, Chris Fox “Experimenting with Automatic Text Summarization for Arabic” University of Essex, School of Computer Science and Electronic Engineering.
- [7] F Soufiane Douzidia and G Lapalme “Lakhas, an Arabic summarization system” RALI-DIRO Université de Montréal.
- [8] A AZMI And S Al-THANYAN “Ikhtasir- A User Selected Compression Ratio Arabic Text Summarization System”, Riyadh, Saudi Arabia.
- [9] S. Alansary, M. Nagi & N. Adly (2010, December). UNL+ 3: The gateway to a fully operational UNL system. In 10th International Conference on Language Engineering, Ain Shams University, Cairo, Egypt.
- [10] S. Alansary, (2016, December). Improving Alserag Arabic Diacritization Grammar through Syntactic Analysis. In 16th international conference on language engineering, Cairo, Egypt.2016.
- [11] S. Alansary, M. Nagi & N. Adly (2013, February). A suite of tools for Arabic natural language processing: A UNL approach. In Communications, Signal Processing, and their Applications (ICCSIPA), 2013 1st International Conference on (pp. 1-6). IEEE.
- [12] M HASSEL “Evaluation of Automatic Text Summarization”, Licentiate Thesis Stockholm, Sweden 2004.
- [13] J Steinberger, K Ježek “EVALUATION MEASURES FOR TEXT SUMMARIZATION” Department of Computer Science and Engineering University of West Bohemia in Pilsen, 2-Mar-2009.
- [14] I MANI, G KLEIN, D HOUSE, LYNETTE HIRSCHMANS, TH FIRMIN, B SUNDHEIM, “SUMMAC: a text summarization evaluation” USA.

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

التلخيص الآلي للنصوص العربية : دراسة تجريبية

سامح الأنصاري

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

ملخص — لقد أدى انتشار استخدام الانترنت إلى نمو سريع ومتزايد للمعلومات مما أدى إلى وجود حاجة ملحة لوجود أنظمة للتلخيص الآلي للمعلومات مما يسهل من عملية الوصول إليها وفهمها بشكل أسرع. تعرض هذه الورقة البحثية مقدمة مختصرة عن تاريخ وتطور الملخصات الآلية وأكثر الاتجاهات استخداما في عملية بناء أنظمة تلخيص آلي. وتقوم هذه الورقة بعرض نظام مقترح للتلخيص الآلي يمر بمراحل متعددة منها مرحلة استخراج الجمل التي تحوي المعلومات الأساسية وغير المكررة للنص، مرحلة التحليل النحوي من أجل التمييز بين الوحدات الأساسية وغير الأساسية في الجمل المنتقاة، ومرحلة توليد النص الملخص من جديد وفي النهاية تقوم الورقة بعرض تقييم يدوي للنظام المتبع ومناقشة النتائج.

LESAN: LEXical Semantic ANnotated Resource

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt

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

sameh.alansary@bibalex.org

Abstract— Word Sense Disambiguation (WSD) is the process of selecting a sense of an ambiguous word in a given context from a set of predefined senses that usually come from a dictionary or thesaurus. In Arabic, the main cause of word ambiguity is the lack of diacritics of the majority of digital documents, so the same word can occur with different senses. In this paper, we present a Lexical Semantic Annotated Resource (LESAN) for Modern Standard Arabic which depends on the Morphological Annotation stage of ICA to help in building Arabic NLP systems. The used semantic lexicon is a lemma based lexicon and each word is assigned the suitable lexical semantic meaning according to its context and its selected lemma/tag in the morphological analysis stage.

Keywords: Word Sense Disambiguation; Arabic Text; local context; global context; Arabic WordNet; Semantic Similarity.

1 INTRODUCTION

Human language is ambiguous; many words can have more than one sense that is dependent on the context of use [10]. Resolving ambiguity is one of the most difficult Natural Language Processing (NLP) tasks which may appear due to that words can have different meanings depending on the context in which they occur [9].

Arabic is a language of rich morphology, both derivational and inflectional. Due to the fact that the Arabic script does not usually encode short vowels and omits some other important phonological distinctions, the degree of morphological ambiguity is very high. In addition to this complexity, Arabic orthography prescribes to concatenate certain word forms with the preceding or the following ones, possibly changing their spelling and not just leaving out the whitespace in between them. This convention makes the boundaries of lexical or syntactic units, which need to be retrieved as tokens for any deeper linguistic processing, obscure, for they may combine into one compact string of letters and be no more the distinct ‘words’ [14]. The average number of ambiguities for a token in Modern Standard Arabic (MSA) is 19.2, whereas it is 2.3 in most languages. Arabic is a highly structured and derivational language where morphology plays a very important role [9].

To detect the different meanings of the same word form, we need first to detect its different morphological analyses according to the context in which it occur as table 1 shows:

TABLE I
MORPHOLOGICAL ANALYSES OF THE WORD FORM ‘عين’ WITH SOME OF ITS MEANING

Morphological Analysis	Tag	Sense	Context
عَيْن	Past Verb	حَدَّدَ	عَيْنَ الْمُدِيرِ أَسَاسَ الْمَشْكَلَةِ
		وَوَظَّفَ - إِخْتَارَ - قَلَّدَ	عَيْنَ رَئِيسِ الْعَمَلِ الشَّخْصَ الْمُنَاسِبَ لِلْوِظِيْفَةِ
		خَصَّصَ	عَيْنَ الْمَلِكِ حِصَّةً مِنْ الْمَالِ لِلْفُقَرَاءِ
عَيْن	NOUN	عَضُوُ الْإِبْصَارِ لِلْإِنْسَانِ وَغَيْرِهِ مِنَ الْكَائِنَاتِ	عَيْنُ الدَّبَابَةِ مَرْكَبَةٌ
		جَاسُوسٌ	وَجَعَلَهُ عَيْنَ مِصْرَ عَلَى إِسْرَائِيلَ
		نَبْعٌ	عَيْنٌ جَارِيَةٌ

Word sense disambiguation (WSD) has long been a central question in computational linguistics, and in recent years the literature has been a large number of advances as a result of three main factors: an increased attention to machine learning techniques, widespread dissemination of sense inventories, and availability of large corpora and the means to do broad-coverage identification of relevant linguistic features in them [5].

Word sense disambiguation (WSD) is an intermediate task which is not an end in itself [7] and necessary for many natural language processing (NLP) applications such as machine translation, information retrieval, information extraction, part of speech tagging, and text categorization. It refers to the task that automatically assigns the appropriate sense, selected from a set of pre-defined senses for a polysemous word, according to a particular context where the identification of one word sense is related to the identification of neighboring word senses [9].

The task of identifying the correct sense for the ambiguous word is not simple as it appears. What should be done to disambiguate a word? We must find a way to define the possible meanings of the word, since that we have to assign each occurrence of the ambiguous word to the appropriate sense [10].

All these issues give rise to the need to have a lexical semantic 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 language in the world with estimated 422 million native speakers. In this paper, we present a LEXical Semantic ANnotated Resource (LESAN) which is built during the process of developing the International Corpus of Arabic (ICA) and benefited from its morphological analysis stage using a built semantic lexicon. It will be available soon for all researchers to be accessed and used.

In what follows, section 2 reviews the previous related work. Section 3 reviews the description of the used semantic lexicon. Section 4 reviews the current state of the data. Finally, section 5 concludes the paper.

2 RELATED WORK

Word Sense Disambiguation (WSD) systems began with Latin Languages such as English and French since several decades. Nevertheless, the Arabic language did not get the attention until the last decade. The first system that handles WSD for Arabic words using unsupervised method using English WordNet and an English-Arabic parallel corpus for annotation sense of Arabic words. The results demonstrate that word-level translation correspondences are a valuable source of information for sense disambiguation, because words having the same translation often share some dimension of meaning [5].

The rooting algorithm with Naïve Bayes Classifier is used to disambiguate Arabic words without diacritics. The experimental study proves that using of rooting algorithm with Naïve Bayes (NB) Classifier enhances the accuracy by 16% and decreases the dimensionality of the training documents [7].

Another system presented a supervised learning method based on an Inner Product of vectors used to estimate the sense of the ambiguous word, extracts two sets of features: the set of words that have occurred frequently in the text and the set of words surrounding the ambiguous word. This approach achieves a precision of 77.1% [13].

Another contribution combined the information retrieval measures with the Lesk Algorithm to choose the most appropriate sense of the ambiguous word by returning a semantic coherence score corresponding to the context that is semantically closest to the original sentence containing the ambiguous word. This selection is based on a comparison between the glosses of the word to be disambiguated, and its different contexts of use extracted from a corpus the study proves that using of Lesk algorithm with Harman, Croft, and Okapi measures allows us to obtain an accuracy rate of 73% [16].

A comparative study has been achieved among some supervised machine learning algorithms, namely Naïve Bayes (NB), the Decision List (DL) and the K-Nearest Neighbor (KNN) for Arabic word sense disambiguation. The KNN technique outperformed the two others methods by achieving the best precision rate of 52.02% (using smoothing and stemming) compared to the NB (48.23%) and the DL (43.86%). Finally, authors confirmed that supervised machine learning algorithms required a large tagged data in order to achieve satisfactory results in Arabic word sense disambiguation ([10] & [11]).

In [12], the most appropriate sense of an ambiguous word has been estimated based on the semantic trees and a measure of collocation.

[15] presented a hybrid approach for WSD of Arabic language (called WSD-AL), that combines unsupervised and knowledge-based methods. The results found by the proposed system achieved a precision of 79%.

Later in [9], a system based on genetic and memetic algorithms and apply them to Modern Standard Arabic and comparing them against a naïve Bayes classifier. Experimental results show that genetic algorithms can achieve more precise prediction than memetic algorithms and naïve Bayes classifier, attaining 79%.

In [1], Constructed a dictionary that maps the Princeton WordNet definitions to the Arabic WordNet and an Arabic evaluation corpus to run and evaluate an adapted Ant Colony algorithm on Arabic text that used the Lesk similarity measure for disambiguating sense of Arabic text. The algorithm shows a performance of approximately 80% compared to the random baseline.

[8] made two contributions in the field of WSD; the first one, using two external resources Arabic WordNet (AWN) and WN based on term-to-term Machine Translation System (MTS). The second one consists of choosing the nearest concept for the ambiguous terms, based on more relationships with different concepts in the same local context

Unlike other systems, [4] proposed a system that considers two types of context during disambiguation process. The first one, is the local context defined by the words in the neighborhood of the ambiguous word, and the second is the global context defined by the full text. Experiments show that the system achieved an accuracy of 74%.

3 SEMANTIC LEXICON

There are two competing models of lexical processing in the literature. The first proposes that we rely on mental lexicons. The second claims there are no mental lexicons; we identify certain items as words based on semantic knowledge. Thus, the former approach – the multiple-systems view – posits that lexical and semantic processing are sub-served by separate systems, whereas the latter approach – the single-system view – holds that the two are interdependent [6].

The current goal of the ICA is to annotate the data on the lexical semantic level, thus developing LESAN began at Bibliotheca Alexandrina (BA) in 2016. We have developed a lexicon in which the possible senses for a word are defined. Various lexicographic sources have been used as references to build the inventory of senses for each word, mainly the electronic lexicon "قاموس المعاني"¹. It is a Multilingual Dictionary; it provides an educational and translation services. It includes: Arabic, English, Spanish, Portuguese, French, Turkish, Persian, Indonesian, and German. The web site provides many research services, where the searches are prepared so that the results are filtered by the use and field of interest. One of these services is, the Arabic-Arabic Dictionary. It translates the meanings of Arabic words in some famous dictionaries such as the lexicon of weeds, the lexicon of Al-Ghany, the lexicon of the modern Arabic language, the lexicon of Mokhtar al-Sahah, the glossary of terms and the lexicon of names. Less frequent meanings are discarded, together with archaic and restricted uses. This inventory of senses for each word is only preliminary, and can be modified whenever the examples found in the corpus prove the existence of a distinct sense which has not been considered.

The developed lexicon is lemma based; it contains 44358 lemmas. The lexicon entries cover all content words in ICA² [3]. Figure (1) represents an example for entries in the lexicon. The ambiguous lemma like "أزهر" has 3 tags in the lexicon; Noun, Proper noun, and Verb. Moreover, the Noun tag itself is ambiguous as its meaning is represented by 7 distinct senses. Also, the Verb tag is ambiguous too as its meaning is represented by 13 distinct senses. The average of senses for each lemma is 7 senses. About 67.1% of lexicon entries are ambiguous and %32.9 is unambiguous.

```

- <entry1>
  <arabic_lemma>أغات</arabic_lemma>
  <BW> >agAv </BW>
  - <tag1>
    VERB
    <sense id="5478">أغات / ساعد / كشف النداء</sense>
    <sense id="4888">أغات السحاب السماء : عطاها</sense>
    <sense id="5879">أغات على قات قوسين أو أفنى من العرق لو لم تخطه فزفة الإنقاذ</sense>
    <sense id="36985">أغات الداعي : أجاهه</sense>
    <sense id="9874">أغاتهم الله بالمطر : أرسله عليهم</sense>
  </tag1>
</entry1>
- <entry2>
  <arabic_lemma>أزهر</arabic_lemma>
  <BW> >azohar </BW>
  - <tag1>
    NOUN
    <sense id="9856">أزهر: الذي لونه أبيض صافٍ مشرق مضيء</sense>
    <sense id="3004">الأزهران : الشمس والقمر</sense>
    <sense id="7895">الليالي الأزهر : الليالي الثلاث من أول الشهر</sense>
    <sense id="3659">الأزهر : القمر</sense>
    <sense id="5298">جامع في القاهرة بناه جوهر الصقلي ، كان منذ قيامه أحد مراكز الفكر الإسلامي وصار حديثاً جامعة تدرس العلوم العصرية إلى جانب اللغة والتربية</sense>
    <sense id="1222">الأزهاروان : سورتنا البقرة وآل عمران</sense>
    <sense id="789">يوم الجمعة ، أكتروا الصلاة على في الليلة القراء واليوم الأزهر [ حديث ] : ليلة الجمعة ويومها</sense>
  </tag1>
  - <tag2>
    PPN
    <sense id="4003">أزهر اسم علم</sense>
  </tag2>
  - <tag3>
    VERB
    <sense id="3256">أزهرت بكه ناري : قويت بكه</sense>
    <sense id="1569">أزهرت بكه زهادي : قضيت بكه حاجتي</sense>
    <sense id="3659">أزهر القمر : أبيض وحسن وصفا لونه</sense>
    <sense id="7774">أزهرت النبات أو الشجر : أزهر طلع زهره</sense>
    <sense id="6598">أزهر الوجه : تلالاً</sense>
    <sense id="32569">أزهرت النار : اشتعلت شعلتها ، تأججت</sense>
    <sense id="7452">أزهرت فخر زهره</sense>
    <sense id="92365">أزهرت الشمس النون : عجزته</sense>
    <sense id="7453">أزهر النجم : ناعلاً وأشرق وأضاء</sense>
    <sense id="7895">أزهر نيات الخقول : أبيض ، أزهر ، نؤز</sense>
    <sense id="98753">أزهرت نازة : أضاءت ، قذفت بيرانه</sense>
    <sense id="7842">أزهرت زندي : رفعت شاتي أو قضيت حاجتي</sense>
    <sense id="69854">أزهر ، يزهر ، مصدر زهور</sense>
  </tag3>
</entry2>

```

Figure 1: Semantic Lexicon Example

As observed in figure (1), senses of the noun 'أزهر' are reviewed manually by linguists. The electronic referred lexicon "Al-Ghany" contains some morphological information like the plural 'جمع', singular 'مفرد' forms, participles 'اسم الفاعل' 'اسم المفعول'... etc. Such morphological information are already stored in BASMA's lexicon, so they were deleted from the semantic lexicon. Moreover, syntactically different senses' definitions with the same meaning are handled as in "الأزهرُ : كل ما لونه : " and "الأزهرُ : كل لون أبيض صافٍ مشرق مضيء" and "كل حيوان أو نبات برأق اللون مشرق والجمع : زهُرُ كل ما لونه : " which expresses both.

¹ <https://www.almaany.com/> [Accessed 14-10-2018]

² <https://www.bibalex.org/ica/ar/default.aspx> [Accessed 14-10-2018]

المعاني الجامعي لكل رسم معني

أزهر

2. أزهرٌ: (اسم)

○ الجمع: زُهرٌ، المؤنث: زَهْرَاءُ، و الجمع للمؤنث: زَهْرَاوَاتُ وَ زُهْرٌ

○ صفة مشبهة تدلّ على الثبوت من زهر: أبيض صافٍ مضيء مشرق

○ الأزهرُ: كلُّ لونٍ أبيضٍ صافٍ مشرقٍ مضيءٍ

○ الأزهرُ: القمر

○ وجهُ أزهرٍ: مشرقُ الوجهِ

○ يوم الجمعة، أكثرُوا الصلَاةَ عليّ في الليلة الغراء واليوم الأزهر [حديث]: ليلة الجمعة ويومها

○ الأزهرُ: كل حيوانٍ أو نباتٍ براق اللون مشرقٍ والجمع: زُهْرٌ

○ جامع في القاهرة بناه جواهر الصقلي، كان منذ قيامه أحد مراكز الفكر الإسلامي وصار حديثاً جامعة تدرس العلوم العصرية إلى جانب اللغة والشريعة

○ الأزهران: الشمس والقمر

○ الزهراوان: سورتا البقرة وآل عمران

○ الزهراء: لقب السيدة فاطمة بنت رسول الله صلى الله عليه وسلم

○ الليالي الزهراء: الليالي الثلاث من أول الشهر

3. أزهرٌ: (اسم)

○ أزهرٌ: جمع زهر

4. أزهر: (اسم)

○ أزهرٌ: فاعل من زهر

Figure 2: The lemma 'أزهر' as it occurred in the electronic lexicon

Regarding Modern Standard Arabic, some senses need to be added in the semantic lexicon and have been dealt with in the manual annotation process (section 4) due to two main reasons:

1. The word is newly used in Arabic and is not included in classical lexicons such as "أخوونة" ">axowanap" "brotherhoodness" as in example [1].
2. The senses of certain words are found in the classical Arabic lexicons, but the modern usage of these words require new senses to be added. For example, the word "فاجأ" "fAja>" has a new sense "أدهش" ">adoha\$" "surprise" as in example [2].

كَيْفَ سَقَطَتْ "أخُوْنَةُ" الدَّوْلَةَ فِي عَهْدِ مُرْسِي [1]
kayofa saqaTato ">axowanapu" Ald-awolapi fiy Eahodi murosiy
How the " brotherhoodness " of the state fell in the era of Morsi

تَمَسَّكَ التُّونِسِيِّينَ بِالإِسْلَامِ فَاجَأَ العَرَبَ [2]
tamas~uku Alt~unisiy~iyina biAl<isolAmi fAja>a Algaroba
The Tunisians' adherence to Islam surprised the West

4 MANUAL ANNOTATION AND QUALITY CONTROL

For annotating the ICA on lexical semantic level automatically, we need first to have a well-trained data to be used in extracting rules and building models for WSD. The used data for manual annotation was selected from MASAR [2]. The lexical semantic annotation procedure involves using the automatic developed interface to provide an access to the data as figure 3 shows:



Figure 3: Semantic Annotation Interface

Our semantic lexicon is used to generate a candidate list of “lemma/sense/tag” for each word. The lexical semantic annotation task is to select the suitable sense from the list of provided alternatives. Once the annotation process is done, the annotated files are saved in a database in a way where the suitable sense of each word depending on the context in which it occurs, is saved as figure 4 shows:

word	lemmaid	tags	Lexical_Semantic_Annotation
/P		BOF_Prg	
(Punc	
ساحة	sAHap	NOUN	ساحة: مكان واسع
كبيرة	kabiyr	ADJ	كبير : هائل عظيم / ذو مرتبة عالية
في	fiy	PREP	
قرية	qaroyap	NOUN	قَرْيَةٌ: عَدَدٌ قَلِيلٌ مِنَ الدُّورِ فِي بُقْعَةٍ مِنَ الأَرْضِ فِي السَّهْلِ أَوِ الجَبَلِ .
,		Punc	
نوافذ	nAfi*ap	NOUN	نُافِذَةٌ : شَبَّاكٌ
وأبواب	bAb	NOUN	باب: مدخل
تظل	>aTal~	IV	أطل: أشرف
على	EalaY	PREP	
الساحة	sAHap	NOUN	ساحة: مكان واسع
,		Punc	
بعض	baEoD	NOUN	بَعْضٌ: جزء ، نوع من ، طائفة
الأشجار	Šajarap	NOUN	شَجَرَةٌ : نَبَاتٌ يَقُومُ عَلَى ساقِ صُلْبَةٍ وَقَدْ يُطَلَّقُ عَلَى كُلِّ نَبَاتٍ غَيْرِ قَائِمِ الأشجار
الذابلة	*Abil	ADJ	ذَابِلٌ : يابِسٌ ، أَصْفَرٌ ، مَيِّتٌ
,		Punc	
وبعض	baEoD	NOUN	بَعْضٌ: جزء ، نوع من ، طائفة
الأشجار	Šajarap	NOUN	شَجَرَةٌ : نَبَاتٌ يَقُومُ عَلَى ساقِ صُلْبَةٍ وَقَدْ يُطَلَّقُ عَلَى كُلِّ نَبَاتٍ غَيْرِ قَائِمِ الأشجار
المقطوعة	maqoTuwE	ADJ	مَقْطُوعٌ : مَفْصُولٌ بَعْضُهُ عَنِ بَعْضٍ
أغصانها	guSon	NOUN	عصن: فرع ، ساق
,		Punc	
أو	>aw	CONJ	
جذوعها	ji*oE	NOUN	جذع : ساق النخلة ونحوها
,		Punc	
تفاجئ	fAja>	IV	فَاجَأَ : جَاءَ فِي وَقْتٍ غَيْرِ مَتَوَقَّعٍ ، بَأَعَثَ
عين	Eayon	NOUN	عَيْنٌ : عَضْوُ الإِبْصَارِ لِلإنْسَانِ وَغَيْرِهِ مِنَ الحَيَوَانِ
الناظر	nAZir	NOUN	ناظر : باصر بعينه
فتصدمها	Sadam-i	IV	صَدَمَ: أَفْرَعٌ
,		Punc	

Figure 4: Semantically Annotated Sample

The data of LESAN have been semantically annotated by 20 well-trained linguistic annotators. In order to make sure that the annotators follow the same guidelines and of almost the same level of professionalism, nineteen files with total of about 19,225 words (and varying numbers of senses choices per word) were annotated independently by each annotator and they were compared together. Out of 19, 225 words, only 2884 words show some disagreement. All twenty annotators agreed on 85% of the words; the pairwise agreement is at least 92.3%.

5 CURRENT STATE OF THE DATA

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

- There are about 26,447 unique lemmas that were annotated.
- 848,678 words have been annotated depending on the contexts were they occurred.
- The average of selected senses per each lemma/tag after the annotation process is 4.2.
- The minimum number of senses that have been assigned to lemma/tag is one sense and the maximum is 14 senses as figure 5 shows:

Sense	Context
الرأس: جزء أعلى من البدن ، يحوي العينين والشم والأنف والأذنين وداخله المخ للإنسان والحيوان	أبي بهوء : - أسألوا جارتنا أم محمود ! .. ومن وراء السور ، ظهر رأس أم محمود ، بدت على وجهها آثار التعب ، كانت دامعة العينين ، وفي صوتها
الرأس النووي: الجزء من السلاح النووي الذي يختزن القدرات التفجيرية	إن إلقاء روسيا على صواريخها الباليستية القارية التي تحمل رؤوساً نووية في أوروبا فعالة لعشرة أعوام ، تشكل خطراً على المنطقة
رأس: سيد	وتاريخياً فإن رأس الدولة في مصر يحتل في حياة المصريين موقفاً أكبر بكثير مما تخوله الدساتير والنظم
رأس الخيمة: إمارة من ضمن الإمارات العربية المتحدة	تمتلك إمارة رأس الخيمة إرثاً تاريخياً وثقافياً رائعا وغنيا
رأس المال : جملة المال التي تستثمر في عمل ما	قرار لوزير الاستثمار بتعديل اللائحة التنفيذية لقانون سوق رأس المال وحظر التلاعب بأسعار الأسهم والالتزام بمبادئ المنافسة
رأس بيروت: منطقة مهمة في العاصمة اللبنانية بيروت	على امتداد نصف قرن من الزمن في ذلك ' المصهر ' البشري الحضاري أي منطقة رأس بيروت إحدى أكثر مناطق المدينة الملونة التصاقاً بالذاكرة اللبنانية والعربية فكراً وأدبياً وفنياً .
رأس حربى: مهاجم - مقاتل	دواعي الأسف الشديد أن يشكل حزب الله بقيادته السياسية وأجهزته الإعلامية وقدراته اللوجستية وإمكاناته المالية رأس حربى في الهجوم على باريس 3 . . وتشارك نحو ثلاثين دولة من
رأس غالب: مدينة مصرية تتبع محافظة البحر الأحمر	النيل بالغردقة يوجد نموذج عملي نشرت عليه هيئة الطاقة الجديدة والمتجددة تمثلة مزارع الرياح بالغردقة ورأس غارب والزعفرانة ، وعلى الرغم من وجود بعض القرى السياحية التي تستخدم أجهزة تسخين المياه
رأس كامبوني: اسم معسكر	القوات الإثيوبية والصومالية أمس على مواقع لجأ إليها مقاتلو المحاكم الإسلامية قرب الحدود الكينية في رأس كامبوني أقصى جنوب الصومال .
رأس محمد: محمية طبيعية مصرية في جنوب سيناء	وهي على شكل هضبة مثلثة الشكل قاعدته على البحر المتوسط شمالاً ورأسه جنوباً في منطقة رأس محمد وخليج العقبة من الشرق وخليج السويس وقناة السويس من الغرب وتتقسم سيناء من حيث
رأس: جسم مادي	مضت شهر .. انشق التراب ، وظهر رأس أخضر ، أخذ يكبر ويكبر ، حتى صار شجرة صغيرة ، طرية الأغصان ، ناعمة
رأس: مقدمة - أول - بداية	بعد زيارة قصيرة استغرقت عدة ساعات حيث كان الرئيس مبارك على رأس مودعيه بمطار القاهرة .
مارون الرأس: قرية لبنانية	لكنه لم يستطع احتلالها ، وتراجعت قواته (عنها وعن مارون الرأس) تحمل قتلاها وإصاباتهما
مَسْقَطُ الرَّأْسِ: مكان الولادة	رحل الكاتب السوري تركي علي الربيعو ، حيث شيع أمس في مسقط رأسه بمدينة القامشلي بحضور عدد من المثقفين السوريين وأبناء عشيرته شمر ، إضافة إلى أبنائه الأربعة

“Figure 5: An Example of the Maximum Assigned Senses”

6 CONCLUSION

This paper presented an attempt to build a lexical semantic annotated resource for Modern Standard Arabic. About 848,678 words are annotated. The first release of this annotated resource will be available soon for researchers to be used in developing and testing their application. It is expected to analyze this data semantically in the future.

REFERENCES

- [1] Abdelaali, B., Tlili-Guiassa, Y., Schwab, D., & Tchechmedjiev, A. (2015). Ant colony algorithm for Arabic word sense disambiguation through English lexical information. *International Journal of Metadata, Semantics and Ontologies*.
- [2] Alansary, S. (2016b). MASAR: A Morphologically Annotated Gold Standard Arabic Resource. *In 16th International Conference on Language Engineering*. Egypt: The Egyptian Society of Language Engineering (ESOLE).
- [3] Alansary, S., Nagi, M., & Adly, N. (2007). Building an International Corpus of Arabic (ICA): progress of compilation stage. *In proceedings of the 7th International Conference on Language Engineering*. Cairo, Egypt.
- [4] Bouhriz, N., Benabbou, F., & Ben Lahmar, E. . (2016). Word sense disambiguation approach for Arabic text. *Int. J. Adv. Compt. Sci. Appl*, 1(7), 381-385.
- [5] Diab, M., & Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 255-262). Association for Computational Linguistics.
- [6] Dilkina, K. M. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain research*, 1365, 66-81.
- [7] Elmougy, S., Taher, H., & Noaman, H. . (2008). Naïve Bayes classifier for Arabic word sense disambiguation. *In proceeding of the 6th International Conference on Informatics and Systems*, (pp. 16-21).
- [8] Hadni, M., Ouatik, S. E., & Lachkar, A. (2016). Word sense disambiguation for arabic text categorization. *Int. Arab J. Inf. Technol*, 13(1A), 215-222.
- [9] Menai, M. E. (2014). Word sense disambiguation using evolutionary algorithms–Application to Arabic language. *In Computers in Human Behavior* (Vol. 41, pp. 92-103).
- [10] Merhben, L., Zouaghi, A., & Zrigui, M. (2012). Lexical Disambiguation of Arabic Language: An Experimental Study. *In Polibits* (Vol. 46, pp. 49-54).
- [11] Merhbene, L., Zouaghi, A., & Zrigui, M. (2013). An experimental study for some supervised lexical disambiguation methods of Arabic language. *In Information and Communication Technology and Accessibility (ICTA), 2013 Fourth International Conference IEEE*, (pp. 1-6).
- [12] Merhbene, L., Zouaghi, A., & Zrigui, M. (2014). Approche basée sur les arbres sémantiques pour la désambiguïisation lexicale de la langue arabe en utilisant une procédure de vote. *of the 21st conference on natural language processing (TALN 2014)*, (pp. 281-290). Marseille, France.
- [13] Nameh, M., Fakhrahmad, S. M., & Jahromi, M. Z. (2011). A new approach to word sense disambiguation based on context similarity. *the World Congress on Engineering*, 1.
- [14] Smrž, O. (2007). Functional Arabic morphology: Formal system and implementation.
- [15] Zouaghi, A., Merhbene, L., & Zrigui, M. (2012a). A hybrid approach for arabic word sense disambiguation. *International Journal of Computer Processing Of Languages*, 24(02), 133-151.
- [16] Zouaghi, A., Merhbene, L., & Zrigui, M. (2012b). Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. *In Artificial Intelligence Review* (Vol. 38(4), pp. 257-269).

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.

لسان: مصدر دلالي معجمي للغة العربية

سامح الأنصاري

مركز اللغويات الحاسوبية العربية – مكتبة الإسكندرية

قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية

sameh.alansary@bibalex.org

ملخص—فك اليبس الدلالي المعجمي هو عملية اختيار المعاني المعجمية المختلفة للكلمات تبعاً للسياقات المختلفة الواردة بها. ويمثل غياب علامات التشكيل في اللغة العربية المعاصرة المصدر الأساسي لليبس الصرفي والدلالي للكلمة حيث نجد أن الكلمة الواحدة لها العديد من المعاني الدلالية تبعاً للسياقات المختلفة الواردة بها. تعرض هذه الورقة مصدرًا محللاً على المستوى الدلالي المعجمي يعتمد على عينة لغوية محللة صرفياً من المدونة اللغوية العربية العالمية (مسار) للاستفادة بها في العديد من أنظمة المعالجة الآلية للغة العربية. وقد اعتمد على معجم دلالي يعتمد على جميع الجذوع الموجودة داخل المدونة اللغوية العربية العالمية، وقد تم الاعتماد على هذه المعجم في تحديد المعاني المختلفة للكلمات تبعاً لسياقاتها المختلفة بالاعتماد على الجذع والوسم الصرفي لكل كلمة داخل المدونة.

A Syntactic based Approach to Anaphora Resolution in Arabic

Aya Nabil Mostafa El-Said Nada, Sameh Saad Abou El-Magd Al-Ansary

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

aya.nabil253@gmail.com

Bibliotheca Alexandrina, Alexandria, Egypt

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

sameh.alansary@bibalex.org

Abstract— Anaphora is a linguistic phenomenon occurring in all natural languages and it means pointing back or referring back to an entity in a text. Although anaphora is a challenging process in modern computational linguistic studies, it is very important for a lot of Natural Language Processing (NLP) applications to reach their goals. The objective of the study is to present a new syntactic model or algorithm for anaphora resolution in Arabic, which can be applied in computerized applications. The proposed syntactic rules are generated specially for Arabic and compatible with its specific structures and features. The proposed system achieves a success rate of 86%.

Keywords: Anaphora, Arabic anaphora resolution (AAR), Natural language processing (NLP), Rule based approach.

1 INTRODUCTION

Literally, the term anaphora is an ancient Greek word which means “the act of carrying back upstream. In natural language processing (NLP), anaphora is a relation between two entities in the text being processed. Linguistically, it means the use of a linguist unit, such as a pronoun to refer to another unit in a linguistic context. The “pointing back” (reference) is called an anaphor the entity to which it refers is called an antecedent. Anaphora resolution is very important as it helps to understand the meaning of the whole context because it leads to meaning determination of an anaphor by identifying its correct antecedent. Without knowing the relations between words, we cannot know the intended meaning. Therefore, a wide range of natural language processing(NLP) applications require a successful tool for resolution of anaphora to achieve their objectives, such as machine translation, question-answer systems, text summarization, information extraction, language generation and dialog systems. The present study focuses exclusively on the resolution of four Arabic anaphor types: Nominative disjoint personal pronouns, Accusative disjoint personal pronouns, Genitive joint personal pronouns which are attached to nouns and particles not verbs and Relative pronouns. The purpose of the resolution process is to identify the correct antecedent for each anaphor. This paper presents a rule based approach especially syntactic approach to solve Arabic anaphora resolution depending on number of proposed syntactic rules which are compatible with Arabic language structures and features.

2 RELATED WORK

Several studies applied different approaches to deal with anaphora resolution in several languages; Anaphora resolution approaches are mainly divided into: Knowledge-rich approaches, Knowledge-poor approaches and Hybrid approaches. First: knowledge rich approaches which are also called traditional approaches or rule based approaches which depend heavily on linguistic knowledge such as Morphological, syntactic, discourse, semantic and pragmatic information. Traditional approaches are mainly depended on three basic steps: Determining search space, Applying constraints and applying preferences such as the studies of (Hobbs [22], 1977; Carbonell & R. Brown’s[7], 1988; Elaine & Susann[17], 1988; Mitkov, R.[37], 1997; Liang & Wu[33], 2004; Ali, Khan, & Rabbi[3], 2007).In 2017 Omar, Abdullatif, & Nazlia[49] have proposed a novel model for the Arabic pronominal anaphora resolution depended on rule based approach. Their model contains several steps. In the first step, they have identified the pronouns and removed the non-anaphoric pronouns. In the second step, they have identified a list of the candidates from the context around the anaphora. Lastly, they selected the most probable candidates for every identified anaphoric pronoun. In their study, they have determined the proper rules which can be used for this task. The different linguistic rules depended on the morphological, lexical, heuristic, syntactic, and the positional constraints. They have assessed the performance of their proposed model using the Quran corpus, which was annotated with the pronominal anaphora. Their experimental results indicated that their proposed algorithm could choose the appropriate antecedents with 84.43% accuracy.

Second Machine learning approaches: The demands of practical and inexpensive Natural language processing (NLP) systems encouraged many researchers to move away from huge linguistic knowledge to knowledge poor systems or strategies which reported good results in AR. The goal of knowledge poor or machine learning approaches is to identify anaphor-antecedent pairs through simple co-occurrence rules by using training decision trees such as the studies of (Soon, Ng, & Lim [55], 2001; Arregi, Ceberio, & Illaraza[4], 2010; Khaled, Rania, & Najwa[30], 2007; Mitkov, Orasan, & Evans [52], 2002; Baldwin[5], 1997; Nasukawa[47], 1994). In 1998 Mitkov, Belguith, & Stys [42] have presented a robust, knowledge-poor approach to resolving pronouns in technical manuals. This approach operates on texts pre-processed by a part-of-speech tagger. Input is checked against number – gender agreement and a number of antecedent indicators. Candidates are assigned scores by each indicator and the candidate with the highest score is returned as the antecedent. They proposed this approach as a platform for multilingual pronoun resolution. The robust approach was initially developed and tested for English, but they have also adapted and tested it for Polish and Arabic. For both languages, they found that adaptation required minimum modification and that further, even if used un-modified, the approach delivers acceptable success rates. Preliminary evaluation reports high success rates in the range of and over 90%.

Third Hybrid approaches: Some research techniques utilize both approaches (rule based and machine learning based) which called hybrid approaches. Hybrid approaches have achieved considerable success in anaphora resolution such as the studies of (Hinrichs, Filippova, & Wunsch [23], 2005; Dakwale, Mujadia, & Sharma [14], 2013; Kamune & Agrawal [27], 2015). In 2015 Abdullatif & Nazlia[1] have proposed a hybrid approach that combines different architectures for resolving pronominal anaphora in Arabic language. The collection of anaphora and respective possible antecedents was identified in a rule-based manner with morphological information taken into account. In addition, the selection of the most probable candidate as the antecedent of the anaphor was done by machine learning based on a k-Nearest Neighbor (k-NN) approach. In this study, the appropriate features to be used in this task were determined and their effect on the performance of anaphora resolution was investigated. Experiments of the proposed method were performed using the corpus of the Quran annotated with pronominal anaphora. The experimental results indicate that the proposed hybrid approach is completely reasonable and feasible for Arabic pronominal anaphora resolution with a precision of 71.7.

3 ARABIC ANAPHORA RESOLUTION PROGRAM

A. Input Data

The used corpus is taken from LDC Arabic Treebank Part 3 (full corpus) v 2.0 by using Microsoft Visual Studio 2012 (C# Language – Windows application program) and Microsoft SQL Server Management Studio (2012) to select specific sentences according to a criteria which is selecting verbal sentences which have only one anaphor. These anaphors refer to one noun not a clause or a sentence and the antecedents of these anaphors are orthographically visible regardless of sentence length or complexity. The selected sentences are divided into two parts. First part (2/3 from the whole sentences) is used as tuning data to extract the syntactic rules and the other part (1/3 from the whole sentences) is used for testing the proposed rules. The used database contains one table with five columns as follows: File name, Sentence ID, Sentence analysis, Sentence orthographic writing and true antecedent.

B. Arabic Anaphora Resolution Program Modules

The proposed system begins with anaphor selection module which requires determining the anaphor, its gender and number, Second candidates' determination module requires determining the possible candidates for each anaphor according to stored patterns, Third candidates' gender and number module requires determining gender and number information for each anaphor. Forth constraint rules module which filters candidates according to constraint rules (Gender and number agreement rules), Fifth antecedent selection module which selects the appropriate antecedent. The following figure illustrates these modules.

1) Anaphor selection module

The proposed system selects the anaphors that contain "PRON_3" in its syntactic analysis for each sentence in the selected corpus. For gender determination, If anaphor's analysis contains "M", its gender is masculine else if it contains "F", anaphor's gender is feminine. For Number determination, If anaphor's analysis contains "S", it is singular, yet if anaphor's analysis contains "P", anaphor's number is plural else if it contains "D", the number is dual.

2) Candidates 'determination module

The proposed system will select all possible candidates for each sentence from the list that contains all items of the sentence "listWords" and creates a list that contains possible candidates only. Candidate selection module is based on stored patterns as in the study of Omar, Abdullatif, & Nazlia [49], 2017 which depends on patterns for extracting candidates in Arabic. There are several syntactic patterns used in this research.

3) Candidates' gender and number module

Based on the previous candidate list, the gender and number information for each candidate can be determined as mentioned in the anaphor selection strategy. For gender specification, if the candidate analysis contains MASC, its gender is masculine. If the analysis contains FEM or _F, it's a feminine one. If the candidate is a proper names, store it a "Prop" in the gender list. For number specification: if the candidate analysis contains SG, its number is singular. If its analysis contains PL, it's a plural one. If it is DU, it's a dual one.

4) Constraint rules module

According to Arabic grammar, the anaphor must agree in gender and number with its antecedent. The system applies this rule which leads to decrease the number of the candidates because the new list contains only candidates that agree in gender and number with the selected anaphor.

5) Antecedent Selection Module

This module is divided into three groups:

First: When there is only one candidate after applying constraints rules (gender and number agreement), the system selects it directly as the correct antecedent.

Second: the researcher has proposed rules for direct selection without ranking. These rules are:

- Emphasis by meaning " التوكيد المعنوي "
- Prepositions with Meaning emphasis words
- Adverb " وحده " /Wahdahu/
- Nominative disjoint personal pronouns with predicate adjective phrases

Third: Preferences play an important role in selecting the correct antecedent in the case of more than one candidate. The preference rules are assigned a score (-1, 0, 1, or 2). The candidate with the highest score is the selected true antecedent. As mentioned before, the proposed linguistic rules are extracted from researcher's observation of tuning data. The most frequent rules in the tuning data take a score of 2, important but not most frequent rules take a score of 1, zero when the rule cannot be applied and -1 when the candidate should be excluded form selection. These ranking rules are categorized into four groups:

1) For nominative disjoint personal pronouns هن / هم / هما / هي / هو , a candidate takes a score of 2 if it is a/an

TABLE I
RANKING RULES FOR NOMINATIVE DISJOINT PERSONAL PRONOUNS

1.	subject
2.	object
3.	A noun in a closely related prepositional phrase (PP-CLR)
4.	A last candidate followed by an adjective in a prepositional phrase
5.	Kana noun and its sisters
6.	Anna noun and its sisters
7.	The candidate is duplicated after the disjoint pronoun

2) For accusative disjoint personal pronouns إِيَاهُ / إِيَاهَا / إِيَاهُمَا / إِيَاهُمْ / إِيَاهُنَّ , An object candidate takes a score of (2 if not 0).

3) For joint personal pronouns هُنَّ / هُمْ / هِمْ / هِمْ / هِمْ / هِمْ

A candidate takes a score of 1 or 2 according to which rule is applied.

TABLE 2
RANKING RULES FOR JOINT PERSONAL PRONOUNS

	Rule	Score
1	Conjunction	2 if not 0
2	Linguistic similarity (التشابه اللفظي)	2 if not 0
3	Specific nouns	2 if not 0
4	Plural structures (منها-بما فيها)	2 if not 0
5	Anaphor with adjective nouns (النعته المفرد)	2 if not 0
6	Anaphor in adjective clause (Nominal or verbal) "النعته الجملة: اسمية- فعلية"	2 if not 0
7	Anaphor in relative clauses	2 if not 0
8	Exception style (اسلوب الاستثناء)	2 if not 0
9	Kana nouns	1 if not 0
10	Candidates in closely related prepositional phrase (CLR-PP)	2 if not 0
11	Anaphor in Adverb phrase "الحال جملة"	2 if not 0
12	WH-Adverb phrase "الظروف"	2 if not 0
13	Subject preference	1 if not 0
14	Object preference	1 if not 0
15	Purpose	2 if not 0
16	Candidate after demonstrative pronouns	2 if not 0
17	Anaphor in a PRD PP nearest	2 if not 0
18	Anaphor in a CLR PP nearest	2 if not 0
19	Anaphor in temporal prepositional phrase (TMP-PP)	2 if not 0
20	Predicate candidates	2 if not 0
21	Anaphor in locative prepositional phrase (LOC-PP)	2 if not 0
22	EIY" and "Fy" rule	1 if not 0

After applying the previous syntactic rules on possible candidates, the system selects the candidate with the highest score to be the correct antecedent for the anaphor. If all candidates of an anaphor take a score of zero, the system chooses the closest candidate, and if two candidates have the same score, the system also selects the closest one.

4) For relative anaphora

The system selects the head noun of the immediately preceding noun phrase to be the correct antecedent.

C. Example of the whole Arabic Anaphora Resolution (AAR) Algorithm

Sentence:

عرض رئيس الجمهورية اميل لحود الاوضاع العامة في البلاد مع النائب ميشال المر، وتم خلال اللقاء تقويم التطورات الراهنة وموقف لبنان منها، اضافة الى طرح عدد من المواضيع السياسية الداخلية.

1) *Anaphor selection module*: The following table illustrates this module according to the previous example.

TABLE 3
ANAPHOR SELECTION MODULE OUTPUT

Anaphor	Gender	Number
(NP (PRON_3FS hA)))	Feminine	Singular

2) *Candidates' determination module*

3) *candidates' gender and number module*

following table illustrates these two modules according to the previous sentence

TABLE 4
CANDIDATES' DETERMINATION MODULE AND CANDIDATES' GENDER AND NUMBER MODULE OUTPUT

Candidate	Syntactic analysis with its transliteration and location id	Gender	Number
رئيس	(NP-SBJ (NP (NOUN+CASE_DEF_NOM r}ys):1	Masculine	Singular
الجمهورية	(NP (DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN Aljmhwrp)):2	Feminine	Singular
الايضاع	(NP-OBJ (NP (DET+NOUN+CASE_DEF_ACC AlAwDAE):5	Feminine	Plural
البلاد	(NP (DET+NOUN+CASE_DEF_GEN AlblAd)):8	Feminine	Plural
النائب	(NP (NP (DET+NOUN+CASE_DEF_GEN AlnA}b)):10	Masculine	Singular
اللقاء	(NP (DET+NOUN+CASE_DEF_GEN AllqA')):17	Masculine	Singular
التطورات	(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN AltTwrAt):19	Feminine	Plural

4) Constraint rules module

The following table illustrates this module

TABLE 5
CONSTRAINT RULES MODULE OUTPUT

Candidate	Syntactic analysis with its transliteration and location id	Gender	Number
رئيس	(NP-SBJ (NP (NOUN+CASE_DEF_NOM r}ys):1	Masculine	Singular
الجمهورية	(NP (DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN Aljmhwrp)):2	Feminine	Singular
الايضاع	(NP-OBJ (NP (DET+NOUN+CASE_DEF_ACC AlAwDAE):5	Feminine	Plural
البلاد	(NP (DET+NOUN+CASE_DEF_GEN AlblAd)):8	Feminine	Plural
النائب	(NP (NP (DET+NOUN+CASE_DEF_GEN AlnA}b)):10	Masculine	Singular
اللقاء	(NP (DET+NOUN+CASE_DEF_GEN AllqA')):17	Masculine	Singular
التطورات	(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN AltTwrAt):19	Feminine	Plural

5) Antecedent selection module

The following table illustrates this module

TABLE 6
ANTECEDENT SELECTION MODULE OUTPUT

Candidate	Syntactic analysis with its transliteration and location id	Score	Ranking rule	Selected Antecedent
الجمهورية	(NP (DET+NOUN+NSUFF_FEM_SG+CASE_DEF_GEN Aljmhwrp)):2	0		
الاضاع	(NP-OBJ (NP (DET+NOUN+CASE_DEF_ACC AlAwDAE)):5	1	Object preference	
البلاد	(NP (DET+NOUN+CASE_DEF_GEN AlblAd)):8	0		
التطورات	(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN AltWrAt):19	2	Conjunction /w/	2(Highest score)

The following figure illustrates these five modules

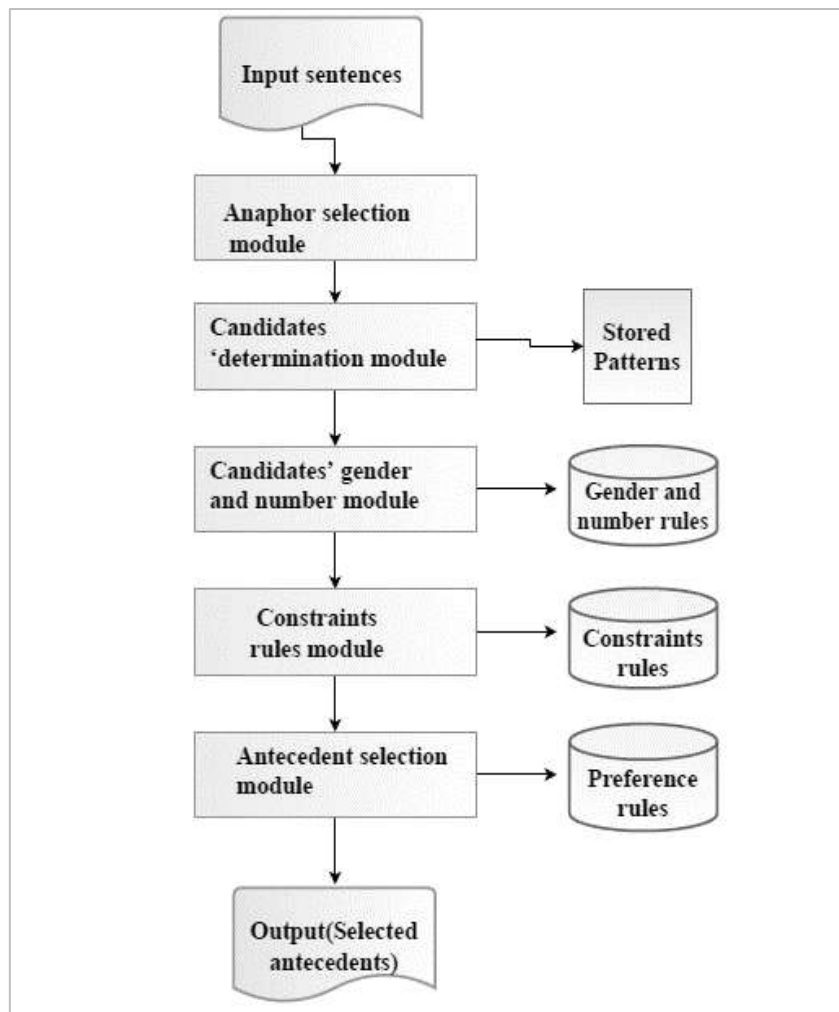


Figure 1: Proposed Arabic anaphora resolution (AAR) system architecture

4. Results

Evaluating the performance of the system by calculating the success rate according to this equation (Number of resolved anaphora \ number of all anaphors). These results can be concluded as follows:

Group one: For nominative disjoint personal pronouns, the proposed system achieved a success rate of 80 %.the following table illustrates the results of the first group.

TABLE 7
RESULTS FOR NOMINATIVE DISJOINT PERSONAL PRONOUNS

Number of nominative anaphors in testing data	Correctly resolved	Incorrectly resolved
24	19	5

Group two: For accusative disjoint personal pronouns, the proposed system solved it correctly because this type is a very rare anaphora type and appeared only one time in testing data.

Group three: For joint personal pronouns, the proposed system achieved a success rate of 87 % for this type. The following table illustrates the results of the third group.

TABLE 8
RESULTS FOR JOINT PERSONAL PRONOUNS

Number of Joint anaphors in testing data	Correctly resolved	Incorrectly resolved
325	282	43

For the whole system (Previous three types), the proposed system achieved a success rate of 86 %(302 out of 350 sentences).the following table illustrates the results for the whole system.

TABLE 9
Results for whole proposed system

Number of all testing sentences	Correctly resolved	Incorrectly resolved
350	302	48

Group four: Relative pronouns

The proposed system achieved 88% success rate. The following table illustrates the results for forth group

TABLE 10
RESULTS FOR JOINT RELATIVE PRONOUNS

Number of relative anaphors in testing sentences	Correctly resolved	Incorrectly resolved
162	143	19

From the previous distribution of the types of Arabic anaphora, it can be concluded that: Dative and accusative joint personal pronouns are the most frequent anaphors in the selected corpus and also in Modern Standard Arabic (MSA). Accusative disjoint personal pronouns are the least frequent in the selected corpus and in MSA as they appear once in testing data and three times in training data. The other types (Nominative disjoint personal pronouns and relative pronouns) are used frequently in the selected corpus and in MSA.

According to the success rate results, that the following can be concluded: The proposed system achieved a success rate up to 85% in most studied anaphora types (Four types), and so it can be inferred that a rule based approach (syntactic based approach) can resolve Arabic anaphora resolution with a great success. According to the extracted rules : These rules do not exist in any other previous works done in this field. The proposed rules are generated specially for Arabic language and compatible with its specific features and structures.

5. Conclusions

Results of the study, based on syntactic rules, encourage the move forward other studies and works in Arabic anaphora resolution according to syntactic based approach. The study is based on Arabic verbal sentences. In future work, the researcher seems to apply this syntactic model on nominal sentences, sentences that have one or more anaphors from different types of anaphora as well as on texts so as to show the efficiency of the proposed model or algorithm.

ACKNOWLEDGMENT

First, praise is to Allah for giving me the blessing, the strength, the chance and endurance to complete this study. I would like to express my gratitude to all those who gave me the possibility to complete this study. I wish to express my sincere thanks and appreciation to Professor Sameh Al-Ansary for his expertise, comments, stimulating feedback and support during the research and for giving me the opportunity to participate in this conference. I also thank the organizers of the conference and the audience

REFERENCES

- [1] Abdullatif, A., & Nazlia, O. (2015). A Hybrid Approach to Pronominal Anaphora Resolution in Arabic. *Journal of Computer Sciences*, 764:771.
- [2] ALJMER, K., & ALTENBERG, B. (2014). Introduction. In G. Kennedy, *An Introduction to Corpus Linguistics* (pp. 1-35). New York: Routledge.
- [3] Ali, R., Khan, M. A., & Rabbi, I. (2007). Strong Personal Anaphora Resolution in Pashto Discourse. *Proceedings of the International Conference on Emerging Technologies* (pp. 148-153). Islamabad: IEEE Xplore Press.
- [4] Arregi, O., Ceberio, K., & Illaraza, A. D. (2010). A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque. *Proceedings of the 12th Ibero-American conference on Advances in Artificial Intelligence* (pp. 234-243). Berlin: Springer Berlin Heidelberg.
- [5] Baldwin, B. (1997). CogNIAC: high precision coreference with limited knowledge and linguistic resources. *Proceedings of the ACL'97/EACL'97 workshop on operational factors in practical robust anaphora resolution*, (pp. 38-45). Madrid.
- [6] Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics*. London; New York: Routledge.
- [7] Carbonell, J. G., & Brown, R. D. (1988). Anaphora Resolution: A Multi-Strategy Approach. In *Proceedings of the 12th conference on Computational linguistics* - (pp. 96-101). USA: Association for Computational Linguistics.
- [8] Carter DM (1986) A shallow processing approach to anaphor resolution. PhD thesis, University of Cambridge
- [9] Chaves, A. R., & Rino, L. H. (2008). The Mitkov Algorithm for Anaphora Resolution in Portuguese. *Computational Processing of the Portuguese Language* (pp. 51-60). Portugal: Springer, Berlin, Heidelberg.
- [10] Chowdhury, G. (2003) Natural language processing. *Annual Review of Information Science and Technology*, 37. pp. 51-89. ISSN 0066-4200
- [11] Converse, S. P. (2005). Resolving Pronominal References in Chinese with the Hobbs Algorithm. . *Proceedings of SIGHAN work-shop on Chinese language processing*, 116-122.
- [12] D'Souza, J., & Vincent, N. (2012). Anaphora Resolution in Biomedical Literature: A Hybrid Approach. *Computational Biology and Biomedicine*, 113-122.
- [13] Dagan, I., & Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In: *Proceedings of the 13th international conference on computational linguistics (COLING'90)*, (pp. 1-3). Helsinki.
- [14] Dakwale, P., Mujadia, V., & Sharma, D. M. (October 2013). A Hybrid Approach for Anaphora Resolution in Hindi. *International Joint Conference on Natural Language Processing* (pp. 977-981). Japan: Nagoya.
- [15] Deoskar, T. (2004). *Techniques for Anaphora Resolution A Survey* Unpublished, Cornell University, USA, 2004.
- [16] Dutta K, P. N. (2008). Resolving Pronominal Anaphora in Hindi using Hobbs' Algorithm. *Web Journal of Formal Computation and Cognitive Linguistics*.
- [17] Elaine, R., & Susann, L. (1988). An architecture for anaphora resolution. *Proceedings of the second conference on applied natural language processing (ANLP-2)*, (pp. 18-24). Austin.
- [18] Fallahi, F., & Shamsfard, M. (2011). Recognizing Anaphora Reference in Persian Sentences. *International Journal of Computer Science Issues (IJCSI)*, 324-29.
- [19] Halliday, M., Hasan, R. (1976). *Cohesion in English*, Longman
- [20] Hammami, S., Belguith, L., & Hamadou, A. B. (2009). Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links. *The International Arab Journal of Information Technology*, 481-489.
- [21] Hammami, S., Belguith, L., & Hamadou, A. B. (2008). Anaphora Resolution in Arabic Language: Developing a Corpora Annotating Tool For Anaphoric Links, LARIS-MIRACL Laboratory, University of Sfax, Tunisia
- [22] Hobbs, Jerry, 1977, Resolving pronoun references, in *Readings in Natural Language Processing*, Grasz, Jones and Webber, eds., Morgan Kaufman Publishers, Inc. Los Altos, California, USA.
- [23] Hinrichs, E. W., Filippova, K., & Wunsch, H. (2005). What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German. Germany: University of Tübingen.
- [24] Ho, H., Min, K., & Yeap, W.-K. (2004). Pronominal Anaphora Resolution Using a Shallow Meaning Representation of Sentences. *Pacific Rim International Conference on Artificial Intelligence (PRICAI)* (pp. 862-871). Springer, Berlin, Heidelberg.
- [25] J.Leffa, V. (2003). Anaphora Resolution Without World Knowledge. *D.E.L.T.A*, v. 19 (1), 181 - 200.
- [26] Jurafsky, D., & Martin, J. (2000). *Speech and Language Processing; An Introduction to Natural Language Processing*. In *Computational Linguistics and Speech*. New Jersey: Prentice Hall.
- [27] Kamune, K. P., & Agrawal, A. (2015). Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text. *Modern education and computer science*, 56-64.
- [28] Kennedy, G. (2014). *An Introduction to Corpus Linguistics*. New York: Routledge.
- [29] Khadiga Mahmoud Seddik & Ali Farghaly. (2014). Anaphora Resolution. In *Natural Language processing of semitic languages* (pp. 247-275). Heidelberg: Springer-Verlag Berlin.
- [30] Khaled, E., Rania, A.-S., & Najwa, E.-Z. (2007). Arabic Anaphora Resolution Using the Web as Corpus. Cairo.
- [31] Ku'cov', L., & Zabokrtsk', Z. ˇ. (2005). Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution. *Proceedings of 8th International Conference on Text, Speech and Dialogue* (pp. 93-98). Springer-Verlag Berlin Heidelberg.
- [32] Kupsc, A., Mitamura, T., Durme, B. V., & Nyberg, E. (2004). Pronominal Anaphora Resolution for Unrestricted Text. *Language Resources and Evaluation Conference (LREC)*, (pp. 1495-1498).
- [33] Liang, T., & Wu, D.-S. (2004). Automatic Pronominal Anaphora Resolution in English Texts. *Computational Linguistics and Chinese Language Processing*, 21-40.
- [34] Lust, B. (1986). *Barbara Lust, Introduction to Studies in the Acquisition of Anaphora*. USA: Reidel Publishing Company.
- [35] Maamouri, M., Bies, A., Buckwalter, T., Jin, H., & Mekki, W. (2005, May 24). *Linguistic Data Consortium*. Retrieved from <https://catalog.ldc.upenn.edu/docs/LDC2005T20/>
- [36] N'em'cik, V. (2006) Anaphora resolution. Master's thesis, Masaryk University, Faculty of Informatics, Brno.

- [37] Mitkov, R. (1997). Factors in anaphora resolution: They are not the only things that matter: A case study based on two different approaches. Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, (RUT '97), Association for Computational Linguistics, USA, pp: 14-21.
- [38] Mitkov, R. (1999). Anaphora Resolution: The state of the art Technical Report based on COLING'98 and ACL'98 Tutorial on Anaphora Resolution. University of Wolverhampton: School of Languages and European Studies.
- [39] Mitkov, R. (2001). Outstanding Issues in Anaphora Resolution. Computational Linguistics and Intelligent Text Processing. CICLing 2001. Lecture Notes in Computer Science. vol 2004, pp. 110-125. Mexico: Springer, Berlin, Heidelberg.
- [40] Mitkov, R. (2003). Anaphora Resolution. In The oxford Handbook of computational linguistics (pp. 266-283). New York: Oxford university Press Inc.
- [41] Mitkov, R. (2013). Anaphora resolution Third Avenue. New York: Routledge.
- [42] Mitkov, R., Belguith, L., & Stys, M. (1998). Multilingual robust anaphora resolution. In Proceedings of the 3rd Conference on Empirical Methods in Natural, (pp. 7-16). Spain.
- [43] Mitkov, R., Evans, R., Orasan, C., & Pekar, V. (2007). Anaphora Resolution: To What Extent Does It Help NLP Applications? 6th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC) (pp. 179-190). Portugal: Springer-Verlag Berlin Heidelberg.
- [44] Muñoz, R., Saiz-Noeda, M., & Montoyo, A. (2002). Semantic Information in Anaphora Resolution. Advances in Natural Language Processing (pp. 63-70.). Germany: Springer-Verlag Berlin Heidelberg.
- [45] Mohammed, R. M. (2008). Pronominal Anaphora Resolution in Arabic/English Machine Translation Systems Using Al-Ahram Newspaper 1998-2006 as an Input. Ain Shams University.
- [46] Naing, M. T., & Thida, A. (2014). Pronominal Anaphora Resolution Algorithm in Myanmar Text. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2795-2800.
- [47] Nasukawa, T. (1994). Robust method of pronoun resolution using full-text information. In: Proceedings of the 15th international conference on computational linguistics (COLING'94), (pp. 1157-1163). Kyoto.
- [48] Nemeč, V. (2006). Anaphora Resolution. Brno: MASARYKOVA University.
- [49] Omar, Abdullatif, A., & Nazlia. (2017). A Computational Model for Resolving Arabic Anaphora using Linguistic Criteria. Indian Journal of Science and Technology..
- [50] Pınar Tüfekçi & Yılmaz Kılıçaslan. (2007). A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm. International Journal of Computer and Information Engineering, 1416-20.
- [51] Poesio, M. (2016). Linguistic and Cognitive Evidence About Anaphora. In M. Poesio, Roland, & Y. Versley, Anaphora Resolution: Algorithms, Resources and Application (pp. 23-54). Springer-Verlag Berlin Heidelberg.
- [52] Mitkov, Orasan, & Evans. (2002). A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method. Computational Linguistics and Intelligent Text Processing: Third International Conference. vol 2276, pp. 69-83. Mexico: Springer, Berlin, Heidelberg.
- [53] Seddik, K. M., Farghaly, A., & Fahmy, A. A. (2011). Arabic anaphora resolution using Holy Qur'an text as corpus. ALTIC.
- [54] Shalom, L., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. Computer Linguistics, 535-561.
- [55] Soon, W. M., Ng, H. T., & Lim, D. C. (2001). A machine learning approach to coreference resolution of noun phrases. Computer linguistics, 27: 521-544.
- [56] Steinberger, J., Poesio, M., Kabadjov, M. A., & Jezek, K. (2007). Two uses of anaphora resolution in summarization. Information Processing and Management, 43, 1663-1680.
- [57] Svartvik, J. ed.. (1992). Directions in corpus linguistics. Proceedings of the Nobel Symposium 82. The Hague: Mouton de Gruyter.
- [58] Vicedo, J. L., & Ferrandez, A. (2000). Importance of Pronominal Anaphora Resolution in Question- Answering System. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, (pp. 555-562).
- [59] نائل محمد إسماعيل. (2011). الإحالة بالضمائر ودورها في تحقيق الترابط في النص القرآني. مجلة جامعة الأزهر بيزة، سلسلة العلوم الإنسانية. 1100- 1061.
- [60] نوال بنت سليمان الثنيان. (2010). الإحالة الضميرية في اللغة العربية: مقارنة تطبيقية في ضوء نحو النص: مقالات اتحاد المالك في الحوار والاختلاف أنموذجاً. علوم اللغة - مصر، 165-256.
- [61] سهال مز غني الحمادي & لمياء هديش بلغيث. (2011). التحليل الآلي للضمائر العائدة ودورها في المعالجة الآلية للغة العربية. مجلة أبحاث الحاسوب. 1-9.
- [62] محمد خطايي. (1991). لسانيات النص، المركز الثقافي العربي، بيروت، الدار البيضاء.

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.

Aya Nabil: Demonstrator at Phonetics and Linguistics department, Faculty of Arts, Alexandria University



She got a Bachelor of Arts in 2012 (Phonetics and Linguistics department – Faculty of Arts - Alexandria university – Alexandria - Egypt). She has graduated from Information technology institute (ITI) in 2013 intake 33. She works as a demonstrator at Phonetics and Linguistics department , Faculty of Arts ,Alexandria university form 2013 till now .She has discussed her Master thesis in 13-10-2018.

TRANSLATED ABSTRACT

الإحالة الضميرية القبلية في اللغة العربية اعتماداً على المنهج التركيبي

آية نبيل مصطفى السيد ندا¹ ، سامح سعد أبو المجد الأنصاري¹
قسم الصوتيات واللسانيات، كلية الآداب، جامعة الإسكندرية، الشاطبي، الإسكندرية، مصر

¹Aya.nabil253@gmail.com

²sameh.alansary@bibalex.org

ملخص— الإحالة هي ظاهرة لغوية تحدث في جميع اللغات الطبيعية وهي تعنى العودة إلى الورا أو الرجوع إلى كيان في النص. على الرغم من كون هذه الظاهرة من التحديات في الدراسات اللغوية الحديثة ، فهي في غاية الأهمية في تطبيقات معالجة اللغات الطبيعية وذلك لأن هذه التطبيقات لا تستطيع التوصل إلى أهدافها بدون حل مشكلة الإحالة. وهذه الدراسة تقدم نموذجاً تركيبياً جديداً لحل مشكلة الإحالة الضميرية القبلية في اللغة العربية ، ويمكن استخدام هذا النموذج التركيبى في التطبيقات اللغوية الحاسوبية المختلفة ، ومجموعة القواعد التركيبية المقترحة والتي يتكون منها النموذج التركيبى قد صممت خصيصاً للغة العربية ومتوافقة مع بنيتها ومميزاتها الخاصة، وهذه الدراسة هي محاولة لحل الإحالة الضميرية في اللغة العربية وقد حقق النموذج التركيبى المقترح نسبة نجاح تصل إلى 86٪.

Unsupervised Emotion Detection from Text Using Word Embedding

Salma Elgayar¹, Abdelaziz A. Abdelhamid², Zaki T. Fayed³

¹ *Computer Science Department, Faculty of Computer & Information Science, Ain Shams University
El-Khalifa El-Mamoun, Abbassia, Cairo, Egypt*

¹salma.elgayar@cis.asu.edu.eg

²abdelaziz@cis.asu.edu.eg

³ztfayed@hotmail.com

Abstract: Text usually not only includes information but also emotional content. This paper proposes a completely unsupervised model for textual emotion detection using hybrid technique of lexicon and word-embedding concept. The proposed model represents text sentences and their meanings in terms of word vectors. To enhance the overall accuracy, emotion ratios were assigned to short sentences and word lexicon. The proposed approach has been validated using the International Survey on Emotion Detection Antecedents and Reactions (ISEAR) and twitter datasets. The evaluation results show that the proposed approach successfully classifies ISEAR sentences based on hybrid technique of lexicon and word embedding with an overall accuracy of 81%.

Key words: Natural Language Processing, Artificial Intelligence, Word Embedding, Lexicon based, Text Mining, Human Computer Interaction.

1 INTRODUCTION

THE relationship between emotions and text is considered from the enthralling topics along centuries. This is because of the textual data importance in the form of communication. Recently, a lot of online social interaction happens with the use of computers, mobiles and the social media applications. In fact, the availability of such great number of platforms have multiplied the availability of unstructured data that could be used in NLP to extract some useful information.

Emotions have an important role in effective social media interaction, so the increase of social communication and interaction means an increase in affective social communication. Indeed, emotions are complex, ambiguous and easily misunderstood entities. Moreover, emotions in text are even sophisticated. For instance, Facial expressions and vocal tone can be easier, to some extent, to detect the underlying emotional tone of the speaker when he/she states a sentence. Imagine all these aforementioned features removed away and all we had were the words only [1].

In the survey presented in [2], authors discussed the recent advances in emotion models and techniques. That survey answered some of the research questions; such as how people feel when they read written text. How writers could transfer their emotional feelings to the readers through the text? What is the best way for writing emotional text to send clear message? And this led to the motivation of more human-computer interaction [3].

When detecting emotions, opinion or sentiment from text can be overlapping. However, some efforts in this direction were employed and resulted in many information retrieval applications which allow businesses, researchers, governments, politicians and organizations to know about people's sentiments [4]. These gains play an important role in decision-making processes such as:

- Businesses realized the potential of emotional advertisements as emotions make users to buy brands and get more services.
- Corporations want to evaluate their services, products and customers feedback.
- Emotional mining in text works for automatic answering and in chat dialogue system based on your current mood.
- Sentiments analysis concerning voters and public opinion extracted from politician's tweets review [5].
- Online classes teachers will be able to connect in a better way with their students by automatically identifying their current affective.

Authors in [6] focused on studying people's reactions, opinions and reactions towards specific events, topics and situations to get the opinion mining and computational analysis concepts.

A lot of research on automatic text analysis typically try to extract the overall sentiment revealed in a document, such as positive or negative. In addition, authors in paper [7] assessed the impact of emotion detection on investor opinion of management announcements, press releases and third-party news. In paper [8, 9, 10] authors used text-mining techniques to obtain reputation of products.

There are two different techniques to perform emotional analysis: supervised and unsupervised algorithms [11]. Supervised methods require training the machine-learning task for every input with corresponding target or output. This will provide answer for any new input after sufficient training. While unsupervised methods do not require a ground truth, which will be able to find the structure or relationships between different inputs [12].

In social media, using supervised sentiment analysis to obtain sentiment labels requires a lot of time, training, memory and effort consuming. This makes unsupervised algorithms for sentimental analysis more reliable and essential. The authors in [13] employed unsupervised learning methods to automatically detect emotions from text. However, different attempts were held on several different standard unsupervised techniques such as Latent Semantic Analysis (LSA) [14] and Non-negative Matrix Factorization (NMF) [15] to tag each sentence with four emotions (Anger, Fear, Joy, Sadness) that are common to all the used datasets.

Recently, a new unsupervised learning technique for natural language processing, called word-embedding technique, was emerged. This technique is a type of word representation that allows words with similar meaning to have a similar representation. By mapping words and phrases to vectors of real numbers that captures similarities between them. There are several models perform word embedding such as Word2Vec model [16] and Glove [17]. These models employed a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term document information to differentiate positive vs. negative sentiment [18].

In this paper, we propose a hybrid approach of word embedding and lexicon. This approach exploits the advance of word embedding that convert main sentence features to numerical vector that describe the semantics of this words. In addition to the advantage of lexicon that is a predefined list of emotions and its corresponding words that can describe each emotion category [19]. To increase the overall emotion detection ratios.

The remainder of this paper is organized as follows. Section 2 formally presents the problem and motivation. Section 3 mentions some of the previous related work and current methods, Sections 4 and 5 describe briefly word vectors and word embedding concept respectively. In Section 6, sentence-preprocessing steps, then section 7 describe our developed detection methodology. Moving to section 8, the used datasets, experimental results and discussion were highlighted. Lastly, Section 9 devoted to concluding remarks and future developments.

2 MOTIVATION

Most of current researches in emotion detection are based on the supervised approach of emotion detection from text [2]. This approach usually requires large annotated training data regardless it requires a lot of time and effort to train large dataset to acquire high accuracy. In addition, the model that is trained on data set of specific domain does not work well on different domains. On the other hand, unsupervised learning can give us a solution to these difficulties. In this paper, we propose an unsupervised approach based on word embedding and lexicon to detect seven ISEAR's emotion categories.

The proposed approach is realized in an application which is capable of getting the percentage of each emotion contained in a sentence ("Soft information") or assigning the sentence to the most dominant emotion percentage ("hard information"). It can also obtain the sentimental analysis of the negative emotions such as anger, disgust, sadness, shame and fear categories or positive emotion such as happiness category.

3 RELATED WORK AND METHODS

Sentimental classification or emotion detection methods are mainly divided into lexicon, deep learning, and hybrid based methods. These methods are briefly discussed in the following sections.

A. Lexicon-based Methods

The lexicon-based sentiment analysis approach depends mainly on a dictionary of opinion words to find their synonyms and antonyms as to conclude the semantic direction of adjective [19-21]. It is widely used because of its simplicity, scalability and to great extend efficient. However, this method suffers from some drawbacks such as low coverage of domain variety [22] and dependence on human effort in preparing hand-labeled documents [23].

Authors found using the machine learning methods would give more accurate results of textual classification than the lexicon methods only [24]. Mudinas et al. [25] have combined machine learning techniques i.e., support vector machine

(SVM) with lexicon-based method, which increased the accuracy of sentiment analysis. Basari et al. [26] used SVM method and Particle Swarm Optimization (PSO) technique that used for solving dual optimization problem, in sentiment analysis of movie reviews.

B. Deep Learning method

Deep learning methods are using the most recent techniques and technology to challenge machine-learning problems; such as natural language processing, vector representations of words [27].

The Word2Vec and GloVe algorithms play an important role in deep learning of textual data; such as text classification, information retrieval and text clustering techniques, which can convert words into meaningful vectors. Learning based methods apply several theories to get the closed nearest emotion category of input text.

C. Hybrid Method

Since lexicon based emotion detection only could not get satisfied accuracy results, some systems then are using hybrid methodology for combining both keyword and deep learning together. Lin and Chuang [28] used a rule based approach and lexicon ontology. While authors in paper [29] used hybrid model which depends on ontology with keywords semantic similarity. Haji Binali and Chen wu [30] implemented hybrid emotion-detection methodology and validated his architecture using support vector algorithm.

4 WORD VECTORS

A traditional way of representing words called one-hot vector. It is essentially a vector with only one target element being "1" and the others being "0" as mentioned in [31]. Though this representation of words is simple and easy to implement, there are several issues. First, you cannot infer any relationship between two words given their one-hot representation. For instance, the word endure and tolerate, although have similar meaning, their targets 1 are far from each other.

In addition, sparsity is another issue as there are numerous redundant "0" in the vectors. This means that we are wasting a lot of space. Therefore, the proposed model need a better representation of words to solve these issues here comes Word embedding technique.

5 WORD EMBEDDING

Word embedding is one of the learning feature techniques in Natural Language Processing (NLP). It maps words or phrases to vectors of real numbers to be converted to Word2Vec models. It introduced first in [32], in which a parameterized function is used to map words (W) to vectors values. A word is converted to vector using a lookup table parametrized by a matrix. The vectors are learned through the contribution of similar representations of words that appear in same context. The context is usually defined as terms, words, sentence or passage that follow or precede a particular word that affect its meaning or make a change.

The authors in [16] and [33] proposed an efficient algorithm called Word2Vec to calculate vector words representations. There are two different models for learning neural networks of Word2Vec architecture. The first one is the Continuous Bag of Words (CBOW) model, as shown in Figure 1. It is a model that tries to drive a word given the context of a few words before and after the target word.

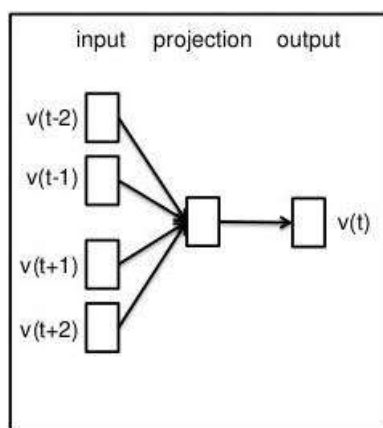


Figure 1: The CBOW architecture predicts the current word (w_j) based on the context C (w_j).

Where Skip-gram model is the second architecture, below in Figure (2). This model is the opposite of the First model where its goal is to get the context words C (w_j) around the given word (w_j). The input to this model is the target word.

Same hidden layer in two models while the output neural network layer propagate many times to conform number of context words. According to the author note¹, the advantage of CBOW model that it is faster than skip-gram model but the second one get better results for infrequent word. According to paper [33], authors recommended 10 words dimensions for CBOW model while only five words for skip-gram model. There is a prodigious feature of word embedding which is the analogies between words can be encoded between words in different vectors. Such as, there seems to be a constant woman-man difference vector:

$$W(\text{female}) - W(\text{male}) \approx W(\text{sister}) - W(\text{brother})$$

$$W(\text{female}) - W(\text{male}) \approx W(\text{wife}) - W(\text{husband})$$

The advantage of word embedding is ability to understand the semantic meaning between words. In which the words of the same context of the sentence will be in the same vector space. This indicates the emotion degree in particular text.

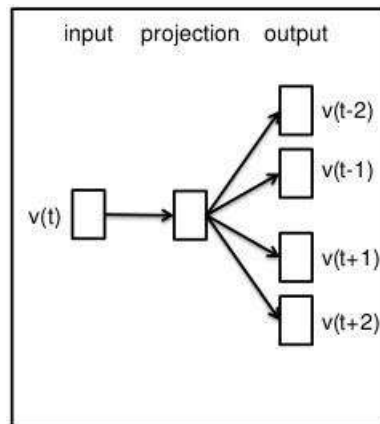


Figure 2: The Skip-gram model architecture learns word vector representations that are good at Predicting the nearby words

6 SENTENCE PREPROCESSING STEPS

The preliminary steps mainly for preprocessing and cleaning the raw input text data, which is the most important phase in machine learning. The output of this stage is a cleaned sentence t from a raw input text φt . The following steps summarize the entire preprocessing process:

A. Minimization Rules

The proposed model first applies some rules on input sentence to focus on the emotional parts and removing non-emotional part of the sentence as mentioned in [34].

- Rule 1: if the input sentence contains word "but" or words that have the same meaning then ignore the sentence before "but". For example, "It was a bit hard, but we enjoyed the day". Then consider only the remaining sentence "we enjoyed the day".
- Rule 2: if the input sentence contains word "as" or words that have the same meaning followed by pronoun then ignore the sentence after "as". We can remove it, because it is counted as meaning complement of the sentence. For example, "She is amazing as she can be.", then the remaining sentence is "She is amazing".

B. Replace Apostrophes

Some sentences in datasets use apostrophes on tokening words. Like "isn't" are taken as one word we need to separate this token to two "is not". These can be replaced using regular expression.

C. Tokenization

The next step is to apply tokenizing which means splitting sentences and words from the body of text. This step is applied at blank space between words in the given sentence.

We did not apply lowercase step, as capital letters can detect emotions such as anger.

D. Stop words removal

¹<https://code.google.com/archive/p/word2vec/>.

Not all the words in a given sentence are the subjects or intent objects. There are a lot of connecting parts in text. Common stop words removed from the text by comparing text to common English words. Those present in list of Python library NLTK [18].

Words like (the) or (and) can be removed. While negative words such as ("not") or ("no") not removed since they are indicative of sentiment. Question (?) and exclamation mark (!) not removed for the same reason.

E. Correcting words

Correcting words is another step in the preprocessing phase and is divided into two sub-steps:

Firstly: Spelling correction step which applies the process of correcting words spelling for example (lizr) instead of (list) using Textblob correct library feature.

Secondly: Word lengthening step which is also a type of spelling mistake in which characters within a word are repeated wrongly for example (amazzzzing) instead of (amazing).

F. Non-English words removal

Using Python library PyEnchant to search for English words and remove non-English words from the sentence. As they not found in the word-embedding dictionary, also single letters and numbers are removed.

G. Emoticon substitution

Emoticons speak thousands than written words can do and is usually used to transfer feelings through social media perfectly so translating emoticons used in messages (if exists) will help to enhance the overall accuracy of emotion detection from text. In paper [35], authors build a skip gram word embedding model by mapping both words and emotion in the same vector space.

We mined this emoticon in UTF-8 encoding. The proposed model can assign each emoticon to one of the seven categories we had then, substitute Unicode with the word that belongs to the representative emotion.

In the following table, we show some emoticon and their Unicode as used in our experiments.

TABLE I
EMOTICON UNICODE.

Emoticon	Unicode
Smile face	U0001F601
Rational face	U0001F914
Surprised face	U0001F632
Furious face	U0001F621

H. Word Stemming

Stemming is only done in condition that the English word was not present in the embedding dictionary phase. Then check if the stemmed word available in the embedding or discard it from the sentence.

7 EMOTION DETECTON METHODOLOGY

In this paper, we used hybrid detection of two main phases, which are lexicon phase and word embedding phase.

In Fig.3 the flow chart of main phases of our used methodology is depicted.

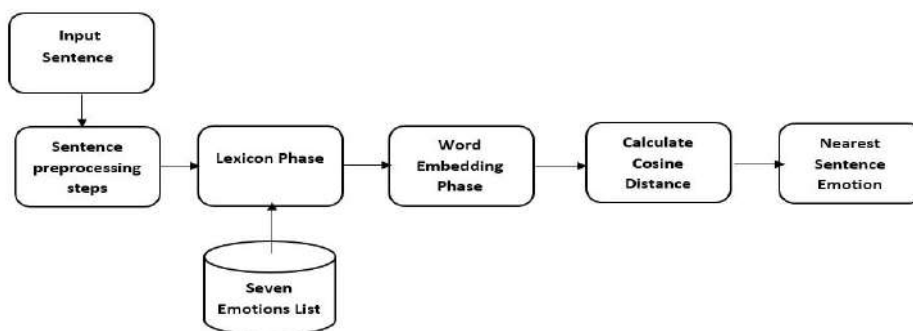


Figure 3: Flowchart of unsupervised emotion detection from text using hybrid technique

A. Lexicon based Phase

The primary phase of emotion detection is trying to discover keyword or phrase. Which indicates certain emotion category. For each emotion category there is a list of related keywords that express its emotion.

Each emotion category expressed through different words. After preprocessing phase, the emotional keywords, from *Seven categories* database, are looked up.

We found in paper [36] this phase is straightforward and growing emotion lexicon will enhance the accuracy but this phase is not successful enough in detecting sentence emotion. Therefore, it is only a helper phase for accuracy enhancement. There may be words call different emotion in different context so the overall emotion of the sentence not simply be the addition of the emotions evoked. The proposed model will move to the next phase, which is the word embedding calculation.

B. Word Embedding Phase

The Second phase in our unsupervised emotion detection is to calculate word embedding input sentence and emotion categories and match the sentence to the nearest cosine distance.

A pretrained word model vectors used: implemented by Google, trained on a GoogleNews corpus with the CBOW architecture we used python 2.7 programming language. Where the input to the system is dataset of raw sentences for emotion detection while the output array of the seven ISEARs emotions weights.

1) Sentiment computation:

Our proposed method after preprocessing steps the cleaned sentence (t) formed of all the remaining words. Then we need to calculate the representative vector of each sentence in our dataset. Therefore, the proposed model used word-embedding concept to represent each string's word (W_j) of the cleaned sentence (t) in a d -dimensional vector as below equation (1).

$$V_{i,t} \in R^{d*1} \quad (1)$$

Then sum all word vectors of sentence to get the entire sentence's vector as mention in equation (2).

$$X_t = \sum_{j=1}^{n_t} V_{j,t}, \text{ where } X_t \in R^{d*1} \quad (2)$$

(n_t) number of words in the cleaned sentence (t).

It is possible to sum several word or phrase vectors to form representation of short sentences as mentioned in [16]. The obtained sentence vector result have to normalize. In proposed model, we used the normal normalization method summation of sentence features vectors over total number n of these features as in equation (3).

$$NX_t = \frac{X_t}{N} \quad (3)$$

2) Representative of each emotion category word set:

A simple solution to get the emotion vector after extracting sentence features is to calculate cosine distance score between sentence's words (W_i) and the word representing an emotion concept.

However, for each emotion concept there are various words such as glad, joy... etc. can also express the same emotion "happiness".

Therefore following authors in [37] proposing to sum word embedding to a few words rather than just one generic word representing the entire emotion category.

$$E_t = \sum_{j=1}^{n_t} V_{j,t}, \text{ where } E_t \in R^{d*1} \quad (4)$$

Where (n_t) the number of representative words of each category (t). Table II shows examples of some representative words for each emotion concept that are most commonly used. The idea behind choosing some representative words of each emotion category instead of represent each emotion category with just one keyword, represent each emotion with at least five words related to each other and describe the same emotion category.

TABLE II
REPRESENTATIVE WORDS OF EACH EMOTION.

Emotion	Representative words
Joy	Happy – glad – joy – good – love.
Sad	Sad– sorrow – hurt– cry– bad.
Angry	Angry – irritate– stupid– annoy– frustrate.
Fear	Fear – afraid– frighten– scare– terrify.
Astonish	Surprise – amazing – astonish – incredible wonder.
Disgust	Disgust– dislike – hate – sick– ill.

3) Cosine distance calculation:

To get the nearest emotion to the input sentence (t). The proposed model used the cosine distance for calculating each emotion of the seven ISEAR's emotions. Using emotion category embedding vector $E_t \in R^{d*1}$ and each sentence vector (NXt) to calculate the corresponding percentage of each emotion in every sentence in the dataset.

8 RESULTS AND DISCUSSION

A The International Survey on Emotion Detection Antecedents and Reactions (ISEAR) Dataset

The proposed model used (ISEAR) dataset have a huge number of emotional sentences reported from questioned persons taken from previous reactions and experiences to seven emotional situations (*sadness– anger – joy – fear – disgust – guilt*) with total Sentences of (3250).

ISEAR database mined, where Table III shows the distribution of these sentences of each category. A summary of multiclass classification is below in Table IV. The report contains the precision, recall and F1 output of the proposed model.

TABLE III
NUMBER OF SENTENCES OF EACH CATEGORY.

Emotion	Number of sentences
Joy	799
Fear	461
Anger	734
Sadness	304
Disgust	252
Shame	261
Guilt	439

Another result that it is possible to exploit is the binary classification case. In this setting, only two classes were presented positive and negative sentiment. In the positive class are grouped the sentences labeled with the category *Happiness*. While the negative class includes *Anger, Disgust, Fear, Shame and Sadness* categories. Table V reports the dichotomy classification results.

TABLE IV
RESULTS OF MULTICLASS USING ISEAR DATABASE.

MULTICLASS RESULTS			
Emotion	Prec.	Rec.	F1
Joy	87.3	57.5	69.4
Fear	71.8	40.5	51.8
Anger	91.8	54.8	68.6
Sadness	61.1	27.2	37.6
Disgust	43.2	18.7	26.1
Shame	95.0	29.1	44.6
Guilt	88.6	40.7	55.8

²<http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/>

TABLE V
RESULTS OF BINARY CLASS USING ISEAR DATABASE.

BINARY RESULTS			
Emotion	Prec.	Rec.	F1
Positive	87.3	57.5	69.4
Negative	79.0	95.0	86.2

B Twitter Dataset

Comparing our methodology with previous work of MirkoMazzoleni and Gabriele Maroni of paper [38]. After getting their twitter dataset, which downloaded by specifying a keyword "Christmas" of total of 64 tweets were manually labeled of the six Ekman's basic ones emotions 1-Anger, 2-Disgust 3-Fear, 4-Happiness, 5-Sadness and 6-Surprise.

A multiclass classification comparison for each emotion category shown in Table VI. Which obvious by combining word embedding and lexicon based classification models. The proposed model got better performance instead of converting sentence to embedding matrix directly. The algorithm check first if any keyword in the input sentence available in one of each category lists.

This will save processing time of the computational calculations of embedding matrix. It's very clear the directly proportional relationship between getting better performance and adding more keywords in each list of each category.

TABLE VI
RESULTS OF MULTICLASS USING TWITTER DATABASE.

TWITTER MULTICLASS COMPARISON				
Emotion	Prec.	Rec.	F1	Number
Anger(Embedding)	0	0	0	5
Anger(Lexicon + Embedding)	0	0	0	5
Disgust(Embedding)	100	20.0	33.0	5
Disgust(Lexicon + Embedding)	100	26.3	41.6	5
Fear(Embedding)	0	0	0	0
Fear(Lexicon + Embedding)	0	0	0	0
Happiness(Embedding)	76.0	65.0	70.0	43
Happiness(Lexicon + Embedding)	88.3	80.5	84.4	43
Sadness(Embedding)	33.0	40.0	36.0	5
Sadness(Lexicon+ Embedding)	60.0	20.0	30.0	5
Surprise(Embedding)	22.0	33.0	27.0	6
Surprise(Lexicon+ Embedding)	66.6	25.0	36.3	6

CONCLUSIONS

Emotion detection from text plays an important role in effecting computing, natural language processing and artificial intelligence. It is also a motivating field and have a wide number of researches on this field as it influences wide massive range of real-life applications. There are previous approaches that target emotion detection from text supervised approach that require large dataset for training phase to get a good accuracy and unsupervised approach that do not focus on the all sentence meaning but only each word individually.

The proposed model is a completely unsupervised algorithm that does not require pre-training data. The evaluation results show that the proposed model successfully classifies ISEAR sentences based on hybrid technique of lexicon and word embedding with and overall accuracy of 81%. Also with comparison to the previous work of classifying, twitter dataset as shown in table VI, the proposed model got better results by adding lexicon phase before converting sentence to word embedding vectors. For example, the happiness emotion in the previous work reached 76.0 precision, 65 recall and 70 F-score, whereas the proposed model got 88.3, 80.5 recall and 84.4 F-score.

The proposed model can work also with high computational efficiency with different linguistic styles such as misspelling or translating other languages using Textblob python libraries. Adding more functionality to the system, the proposed model can detect the Seven ISEAR's emotions from any other language messages also, by first detected message language then translating it to English one using Textblob python libraries, then processed the same steps mentioned before.

We are seeking to detect more than one specific emotion in compound sentences. The proposed model can detect up to the seven emotions with high accuracy however also we are seeking to apply more unsupervised methodologies for higher accuracy and we aim to apply this software in a real life applications. In the future work we are seeking also to consider the context meaning of the next and previous sentences of a paragraph.

REFERENCES

- [1] Eliot Masezi Sibuyi. “*The analysis of the impact of nonverbal communication Xitsonga discourse*”. PhD thesis. University of Limpopo (Turfloop Campus), 2011.
- [2] Salma Elgayar, Abdel ElAziz A. Abdelhamid, and Zaki T. Fayed. “*Emotion Detection from Text : Survey*”. In: 2017.
- [3] Roddy Cowie et al. “*Emotion recognition in human-computer interaction*”. In: *IEEE Signal processing magazine* 18.1 (2001), pp. 32–80.
- [4] Irene Lopatovska and Ioannis Arapakis. “*Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction*”. In: *Information Processing & Management* 47.4 (2011), pp. 575–592.
- [5] Brendan O’Connor et al. “*From tweets to polls: Linking text sentiment to public opinion time series.*” In: *Icwsm* 11.122-129 (2010), pp. 1–2.
- [6] Bing Liu and Lei Zhang. “*A survey of opinion mining and sentiment analysis*”. In: *Mining text data*. Springer, 2012, pp. 415–463.
- [7] Sanjiv Das and Mike Chen. “*Yahoo! for Amazon: Extracting market sentiment from stock message boards*”. In: *Proceedings of the Asia Pacific finance association annual conference (APFA)*. Vol. 35. Bangkok, Thailand. 2001, p. 43.
- [8] Pero Subasic and Alison Huettner. “*Affect analysis of text using fuzzy semantic typing*”. In: *IEEE Transactions on Fuzzy systems* 9.4 (2001), pp. 483–496.
- [9] Satoshi Morinaga et al. “*Mining product reputations on the web*”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 341–349.
- [10] Paul Ekman. “*Facial expression and emotion.*” In: *American psychologist* 48.4 (1993), p. 384.
- [11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.
- [12] James Dougherty, Ron Kohavi, and Mehran Sahami. “*Supervised and unsupervised discretization of continuous features*”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 194–202.
- [13] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. “*Evaluation of unsupervised emotion models to textual affect recognition*”. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics. 2010, pp. 62–70.
- [14] Thomas K Landauer. *Latent semantic analysis*. Wiley Online Library, 2006.
- [15] Daniel D Lee and H Sebastian Seung. “*Algorithms for non-negative matrix factorization*”. In: *Advances in neural information processing systems*. 2001, pp. 556–562.
- [16] Tomas Mikolov et al. “*Distributed representations of words and phrases and their compositionality*”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.
- [17] Jeffrey Pennington, Richard Socher, and Christopher Manning. “*Glove: Global vectors for word representation*”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532– 1543.
- [18] Andrew L Maas et al. “*Learning word vectors for sentiment analysis*”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics. 2011, pp. 142–150.
- [19] MaiteTaboada et al. “*Lexicon-based methods for sentiment analysis*”. In: *Computational linguistics* 37.2 (2011), pp. 267– 307.
- [20] Xiaowen Ding, Bing Liu, and Philip S Yu. “*A holistic lexicon-based approach to opinion mining*”. In: *Proceedings of the 2008 international conference on web search and data mining*. ACM. 2008, pp. 231–240.
- [21] Minqing Hu and Bing Liu. “*Mining and summarizing customer reviews*”. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pp. 168–177.
- [22] Maria Giatsoglou et al. “*Sentiment analysis leveraging emotions and word embeddings*”. In: *Expert Systems with Applications* 69 (2017), pp. 214–224.
- [23] Zhang Hailong, Gan Wenyan, and Jiang Bo. “*Machine learning and lexicon based methods for sentiment classification: A survey*”. In: *Web Information System and Application Conference (WISA)*, 2014 11th. IEEE. 2014, pp. 262–265.

- [24] Kumar Ravi and Vadlamani Ravi. “A survey on opinion mining and sentiment analysis: tasks, approaches and applications”. In: *Knowledge-Based Systems* 89 (2015), pp. 14–46.
- [25] Andrius Mudinas, Dell Zhang, and Mark Levene. “Combining lexicon and learning based approaches for concept-level sentiment analysis”. In: *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*. ACM. 2012, p. 5.
- [26] AbdSamad Hasan Basari et al. “Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization”. In: *Procedia Engineering* 53 (2013), pp. 453–462.
- [27] Oscar Araque et al. “Enhancing deep learning sentiment analysis with ensemble techniques in social applications”. In: *Expert Systems with Applications* 77 (2017), pp. 236–246.
- [28] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. “Emotion recognition from text using semantic labels and separable mixture models”. In: *ACM transactions on Asian language information processing (TALIP)* 5.2 (2006), pp. 165–183.
- [29] Samar Fathy, Nahla El-Haggag, and Mohamed H Haggag. “A hybrid model for emotion detection from text”. In: *International Journal of Information Retrieval Research (IJIRR)* 7.1 (2017), pp. 32–48.
- [30] Haji Binali, Chen Wu, and Vidyasagar Potdar. “Computational approaches for emotion detection in text”. In: *Proceedings of the IEEE international conference on digital ecosystems and technologies (DEST 2010)*. IEEE. 2010, pp. 172–177.
- [31] Georgios Balikas and Massih-Reza Amini. “An empirical study on large scale text classification with skip-gram embeddings”. In: *arXiv preprint arXiv:1606.06623* (2016).
- [32] Yoshua Bengio et al. “A neural probabilistic language model”. In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155.
- [33] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
- [34] Shadi Shaheen et al. “Emotion recognition from text based on automatically generated rules”. In: *2014 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2014, pp. 383–392.
- [35] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. “What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis.” In: *LREC*. 2016.
- [36] Abdul Hannan. “Emotion Detection from Text”. In: *International Journal of Engineering Research and Development* 11.7 (2015), pp. 23–34.
- [37] Ameeta Agrawal and Aijun An. “Unsupervised emotion detection from text using semantic and syntactic relations”. In: *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society. 2012, pp. 346–353.
- [38] MirkoMazzoleni, Gabriele Maroni, and Fabio Previdi. “Unsupervised learning of fundamental emotional states via word embeddings”. In: *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*. IEEE. 2017, pp. 1–6.

BIOGRAPHY



Salma Mohamed Osama Elgayar, Teaching Assistant at Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University.



Abdelaziz Holds a PhD Degree in Electrical and Computer Engineering, from the University of Auckland, New Zealand and MSc Degree in Computer Science from Ain Shams University, Cairo, Egypt. He is currently a computer science assistant professor, Ain Shams University, Cairo, Egypt. Abdelaziz was a member of Healthbots research project sponsored by the University of Auckland and ETRI (Korea). He published more than 20 publications in international conferences and journals. He received the best paper award from the international conference on social robotics, China. His research interest includes Deep Learning, End-to-End Speech Recognition, End-to-End Speech Synthesis, and Natural Language Processing.



Prof. Dr. Zaki Taha Fayed, Professor in Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University.

كشف الإنفعال من النص دون رقيب باستخدام المعنى الضمني للكلمة

اسلمي محمد اسامه الجيار،² عبد العزيز عبد المنعم عبد الحميد،³ ذكي طه فايد

قسم علوم الحاسب، كلية حاسبات و معلومات، جامعه عين شمس

الخليفة المأمون، القاهرة، مصر.

salma.elgayar@cis.asu.edu.eg¹

abdelaziz@cis.asu.edu.eg²

ztfayed@hotmail.com³

الملخص

الذكاء الاصطناعي ليس فقط قدره الحاسب الآلي إن يحل و يتفاعل بذكاء مع المستخدم النهائي و لكن أيضاً إن يتفاعل بإنسانيه و عقلانيه . لذلك كشف الإنفعال من النص يلعب دوراً فعالاً في مجال تفاعل الانسان مع الحاسب الآلي. مؤخراً كشف الإنفعال من النص حظي علي إنتباه الكثير من الباحثين. بفضل الثورة الكبيره لوجود الكثير من البيانات المتاحة التي تحمل الإنفعالات علي المواقع التواصل الاجتماعي و أكثر علي تطبيقات التليفون المحمول و مواقع الويب. برغم من بعض المناهج المتبعه سابقاً في هذا المجال و لكن يتبقي القليل من الجهد ليبدل للكشف عن الانفعال دون رقيب كلياً او تدريب نهائياً.

في هذا البحث نحن نقدم إقتراح للكشف عن الإنفعال من النص دون رقيب . عن طريق دمج طريقتين و هما البحث في القاموس و المعنى الضمني للكلمة. حيث يمكن تحويل الخصائص المميزة في الجملة الي مصفوفة رياضية من الأرقام التي تعبر عن معني كل كلمة في الجملة. يركز هذا العمل بشكل أساسي علي تخصيص نسبه لكل إنفعال من الإنفعالات المتاحة للجملة, و من ثم اختيار أعلي نسبه إنفعال لتمثل الجملة. بينما طريقه البحث عن خصائص الجملة في قاموس كل إنفعال لتحسين نسبه دقه المنهجيه المقدمه.

تم اختبار المنهجيه المتبعه في هذا البحث علي قاعدة البيانات:- دراسة استقصائية دولية عن عوامل الكشف عن التفاعلات العاطفية وردود الفعل (ISEAR) و أيضاً قاعدة البيانات الخاصه بالتواصل الاجتماعي تويتتر لمقارنه هذا العمل بعمل سابق. يمكن اختصار التقنيه المقترحه في هذه الخطوات تجهيز الجملة ، البحث في القاموس ، حساب متوسط مرصوصه الجملة و اخيرا اختيار اقرب مسافه بين الجملة و جميع الانفعالات المتاحة. نتائج التقييم توضح نجاح المنهج المقترح في تصنيف كل جملة من قاعده البيانات (ISEAR) بنسبه واحد و ثمانين بالمئه.

الكلمات الداله

معالجة اللغة الطبيعية ، الذكاء الاصطناعي ، تضمين الكلمة ، معجم الكلمة ، نص التعدين ، تفاعل الإنسان والحاسوب.

Mining Publication Papers via Text Mining

Ahmed S. Ibrahim¹, Sally Saad², Mostafa Aref³

Computer Science Department, Faculty of Computer & Information Sciences,
Ain Shams University, Cairo 11566, Egypt

¹a.saeedibr@cis.asu.edu.eg

²sallysaad@gmail.com

³aref_99@yahoo.com

Abstract-Text mining has become one of trendy fields due to most of data in a format of text and has been incorporated in several research fields. Text Mining (known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT)) is the process of extracting or gathering structured information automatically from unstructured text, via text mining tools. In this paper, some of text mining techniques will be discussed then the proposed methods for mining the publication papers using text mining approaches will be discussed. Three methods have been presented in this paper, first method is searching for related papers using keywords, second method is recognizing the named entities in the papers using named entity recognition and third method is categorizing the paper using machine learning.

Keywords: Text Mining, Named Entity Recognition (NER), Mining Publication Papers, Machine Learning (ML)

1. INTRODUCTION

There are a lot of data produced every-day and increasing of this data makes it important to process and analyze this data to come up with new information that nobody or no one talks about it. Text is considered as one type of data so mining the text one of important task to extract information especially for unstructured text. As all of publication papers documents contain a text so mining the text would help a lot to find the undiscovered relationships among the papers that is related to specific paper and that result to save the effort and time for the researcher. We used a machine learning algorithm and text mining techniques to build the system. This paper is organized in the following manner: Section II for Related Work, Section III Proposed Technique, Section IV Discussion, finally followed by Conclusion and References.

2. RELATED WORK

S.-H. Liao [1] described the text mining process steps as following: collecting, extracting, pre-processing, text diversion, feature extraction, pattern election, and evaluation. In addition, various widely used text mining approaches, i.e., clustering, categorization, natural language processing, information extraction, topic tracking, text summarization, and their application in diverse fields are surveyed.

A. Henriksson and H. Moen [6] discussed integrating a framework Medline biomedical database that helps to eliminate unnecessary details and extract valuable information. Integrating this framework for named entity recognition (NER), classification of text, hypothesis generation and testing, extract abbreviations, relationship and synonym extraction.

K. Sumathy and M. Chidambaram [7] discussed the type of text documents that might be structured; semi structured or unstructured and extracting useful information technique. Giving an overview of applications, tools and issues appear to mine the text. A generic framework has been presented in this paper for concept-based mining which can be visualized as knowledge distillation phases and text refinement. There is a dependency between intermediate form of entity representation mining and a specific domain.

R. Rajendra and V. Saransh [9] presented a method to combine the similar text using k-mean clustering technique for bottom up approach. However, there are two different approach for web-based text mining process, a top down and bottom up approach. The document TF-IDF (Term Frequency- Inverse Document Frequency) algorithm to identify the similarity within the document and discover information regarding specific subjects.

Chau et. al [8] aimed to extract meaningful entities from police narrative reports, such as person, location and organization using machine learning (ML) with feedforward /backpropagation neural network the system Performed well for some entities, and not so well for others, Person names (74.1% precision), narcotic drugs (85.4% precision), Personal property, addresses not as good as expected.

3. PROPOSED TECHNIQUE

This paper proposes a technique for mining the publication papers. This technique depends on using text mining approaches and machine learning. The method goal is to search for scientific papers using similar keywords that used in a specific paper resulting from this process is a set of papers which might relate to each other and that would help in discovering the hidden relationships among papers. After that recognizing of the named entities for each paper which help to find the relation among keywords, finally classifying all documents to specific class (See Figure 1).

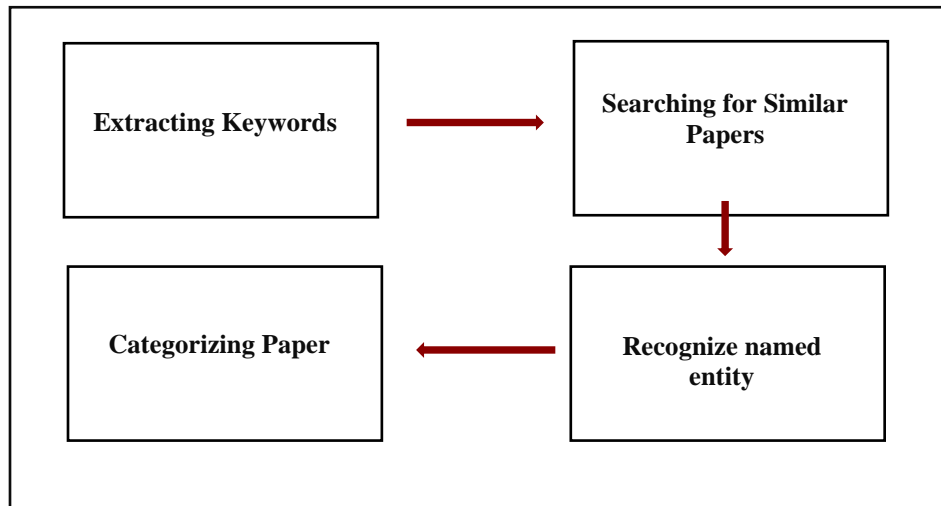


Figure 1: Workflow of Mining the Scientific Papers

A. Extracting Keywords and Searching for Similar Papers

1) *Extracting Relationships among keywords*

Extracting keywords will be based on the similarity and relatedness score among paper terms. So, each term is assigned to a score that points to the likeness of their meaning content. There are two main methods for extracting keywords that can be classified as following; quantitative and qualitative methods. Quantitative method is based on a set of statistical relations



Figure 2: Process of extracting the keywords and relations from paper

where it considered as the simplest keyword extraction models. Calculating the term frequency with TF*IDF is one of the models to identify the keywords in the documents, however the method is not enough to address a proper keyword (See Figure 2).

Qualitative method is based on the semantic relation and analysis that uses the semantic similarity to measure the similarity between two terms according to their meaning. There is a tiny difference between semantic similarity and semantic relatedness both can measure same thing terms. the difference in how to calculate the similarity between two terms. Semantic relatedness is measured using vector space model (VSM) with some of similarity matric like cosine similarity. on the other hand, the semantic similarity measures the similarity among terms using ontologies where it is considered as a directed graph that consists of the terms and the definition of relationships between terms.

To extract the keywords in the documents and prepare the document for mining we used text mining process that considered as a set of steps where each document goes through each step (see Figure 3):

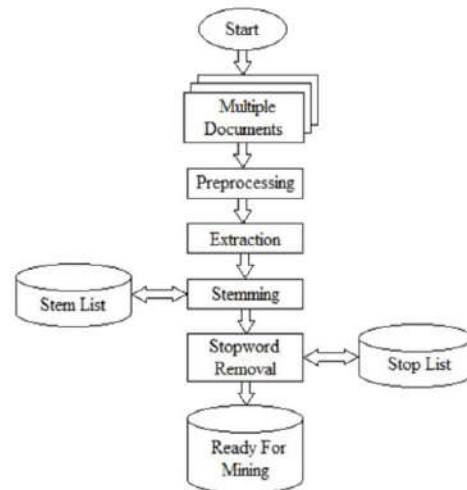


Figure 3: Work flow of Pre-processing stage of Documents

1) Document Gathering:

In this step we will collect a set of text documents with a different category which each document might be in different format i.e. word, pdf, html etc.

2) Pre-processing

In this step we will prepare the document throughout some processes each document is processed for removing inconsistencies, redundancies, stemming, separating words to be ready for next step.

- **Tokenization:**

The process of identifying the document as a set of words by splitting or breaking each sentence into a sequence of pieces (token) such as words, keywords, symbol, phrase. Sentences are splitted by whitespace, punctuation marks or line breaks.

- **Removal of Stop word:**

The process of removing useless words that will not affect the meaning of the sentence those words called stop words such as a, an, but, and, of, the etc. also the search engine had been programmed to ignore or filtering those words.

- **Stemming:**

The process of converting the words into their root it also called rooting and that is done by chopping of the beginning or ending of the words where commonly are called prefix and suffix respectively.

3) Text transformation

In this step the document would be represented as a collection of words and their occurrences and these words would be the features of the document. there is method that can be used to represent the documents:

- **Vector Space Model:**

It's the vector that holds multiset of keywords that represents the documents in addition to it multiplicity [3]. That is done using three phases:

- 1- Document indexing in this phase the keywords are extracted from the text documents.
- 2- Indexed keywords weighting using the TF-IDF method.
- 3- Ranks the documents according similarity measure, measuring similarity can be determined using cosine similarity, Dice' Coefficient and Jaccard's coefficient.

4) Feature Selection (Attribute Selection):

This step is responsible for disposing irrelevant features (keywords) from the input publication paper and this would reduce the amount of database space required. There are two methods in feature selection

- **Filtering:**
Base on the certain statistical criteria all features are ranked the features with highest ranking are value are selected and low ranked feature are omitted.
- **Wrapping methods:**
It is based on selecting a subset of features and try the model with this subset and based on the inferences we decide to add or remove feature from this subset.

5) Searching for Similar Papers

After the documents is preprocessed and represented in a vector of keywords, we will begin start to use this vector to search for scientific papers containing similar keywords (See Figure 4).

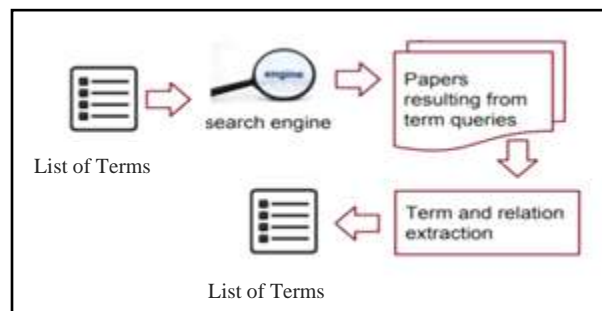


Figure 4: Process of using extracted keywords to search for related papers

A. Named Entity Recognition:

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is the method of identify all main concepts and entities in the publication paper and then attaching attributes to a specific entity or predefined category, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Named-entity recognition (NER) is a subtask of information extraction which considered as one of the techniques for text mining. This step would help in organizing the data and finding the relationships among keywords. We will use the machine learning Recurrent Neural Network (RNN) in this task.

- **Training phase:**

We will train the Recurrent neural network model using back -propagation through time (BPTT) algorithm where it is similar to tradition back propagation (BP) algorithm with a little difference, the gradient at each output depends not only on the calculations of the current time step, but also the previous time steps.

After training our model, we will evaluate it on different groups of datasets. These groups will be collected from different online research papers databases such as: Google Scholar, Web of Science, etc... After that, we will compare the results of our model using the evaluation metrics mentioned below.

B. Publication Papers Classification:

The process of classifying the documents using machine learning (ML), the goal of this process is to assign a specific publication paper to specific class (See Figure 5). There are three types of learning: supervised learning, unsupervised learning and semi-supervised learning. All these types could be used in any classification problem; however, it differs in whether it uses labeled data or not. Document Categorization or clustering is considered as one of the techniques of text mining [5].

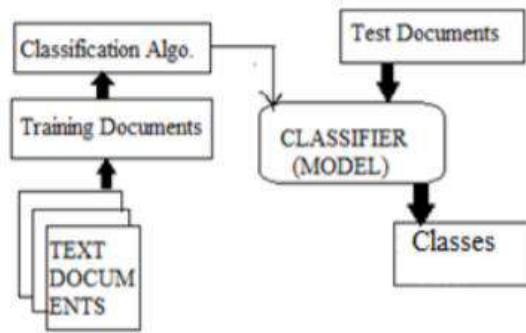


Figure 5: Document Classification Model.

4. DISCUSSION

The evaluation metric that will be used are precision and recall both could be interpreted as possibilities not ratios their equations are as the following: -

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Where the **tp** is stands for true positive, **fp** false positive and **fn** false negative.

This the paper introduced an architecture that concerns with mining the publication papers using text mining techniques and machine learning. The proposed system consists of the following features. First extracting keywords using natural language processing methodologies. Second categorizing documents using machine learning. Last identify the main theme for the documents using also machine learning. Many of the researches that using text mining involved in mining data for business wise to predict the changes in different sectors based on the given data. However, number of researches that concerns mining the scientific papers is very limited. Therefore, focusing on this point is very important where it has a great impact especially for researchers throughout saving efforts and time for researchers and collecting most of the papers that are related in a specific domain.

5. CONCLUSION

In this paper, methods have been introduced for mining the publication papers based on text mining techniques and machine learning algorithms. The system is divided into three main tasks, first extracting the keywords of the publication papers after it preprocessed by the mentioned methods above, Second task is to recognize named entity for the publication papers using machine learning algorithm and it would be the Recurrent Neural Network (RRN) and last task it to categorize the publication papers throughout supervised learning(Classification) or unsupervised learning (Clustering).

REFERENCES

- [1] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications—a decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11 303–11 311, 2012.
- [2] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 1, pp. 30–44, 2012.
- [3] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," *International Journal of Computer Applications*, pp. 159–171, vol.3, 2013.
- [4] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word feature with text classification," in *Proceedings of the 51st Annual Meeting of the ISSS-2007*, Tokyo, Japan, vol. 51, 2007, p. 45.
- [5] K. Thilagavathi, V. Shanmuga "A Survey on Text Mining Techniques", *International Journal of Advanced Research in Computer Science and Robotics* , ISSN: 2320 7345 Volume 2, Issue 10, Oct. 2014 pp 41-50, in IJRCAAR.

- [6] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," *Journal of biomedical semantics*, vol. 5, no. 1, p. 1, 2014.
- [7] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications, tools and issues-an overview," *International Journal of Computer Applications*, vol. 80, no. 4, 2013.
- [8] Chau, M., Xu, J. J., & Chen, H. (2002, May). Extracting meaningful entities from police narrative reports. In Proceedings of the 2002 annual national conference on Digital government research (pp. 1-5). Digital Government Society of North America.
- [9] Roul, Rajendra Kumar, Saransh Varshneya, Ashu Kalra, and Sanjay Kumar Sahay. "A novel modified apriori approach for web document clustering." In *Computational Intelligence in Data Mining-Volume 3*, pp. 159-171. Springer, New Delhi, 2015.

BIOGRAPHIES



Ahmed Saeed, Teaching Assistant, Faculty of Computer and Information Sciences, Ain Shams University. Graduate from Faculty of Computer and Information Sciences, Ain Shams University, Computer Science Department 2014.



Sally Saad Ismail, Lecturer in Computer Science Department (Coordinator of Software Engineering Program), Faculty of Computer and Information Sciences, Ain Shams University.



Mostafa Aref is a professor at the Faculty of Computer and Information Sciences of Ain Shams University in Cairo, Egypt. Ph.D. in Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. M.Sc. in Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask. Canada. B.Sc. in Electrical Engineering-Computer and Automatic Control section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, Egypt.

Translated Abstract

تنقيب أوراق النشر عن طريق التنقيب النصي

احمد سعيد¹, سالي سعد², مصطفى عارف³
كلية الحاسبات و المعلومات جامعة عين شمس
a.saeedibr@cis.asu.edu.eg
sallysaad@gmail.com
aref_99@yahoo.com

الملخص

أصبح التنقيب النصي واحداً من الحقول العصرية نظراً لمعظم البيانات التي في شكل نص تم دمجها في العديد من المجالات البحثية. تنقيب النص (المعروف باسم تحليل النص الذكي ، تنقيب بيانات النص أو اكتشاف المعرفة في النص (KDT)) هو عملية استخراج أو تجميع المعلومات المهيكلة تلقائياً من نص غير منظم ، عبر أدوات تنقيب النص. في هذه الورقة ، بعض تقنيات التنقيب النصي ستناقش ثم سيتم مناقشة الطرق المقترحة للتنقيب أوراق النشر باستخدام أساليب التنقيب عن النص. تم تقديم ثلاث طرق في هذه الورقة ، الطريقة الأولى هي البحث عن الأوراق ذات الصلة باستخدام الكلمات الرئيسية ، الطريقة الثانية هي التعرف على الكيانات المسماة في الأوراق باستخدام التعرف على الكيان المحدد والطريقة الثالثة هي تصنيف الورقة باستخدام التعلم الآلي.

بعض برامج التشكيل الآلي والتصحيح: دراسة تقويمية

مدحت يوسف السبع

كلية التربية بالوادمي، جامعة شقراء، المملكة العربية السعودية

myalsabaa@su.edu.sa

المستخلص: موضوع هذا البحث هو تقويم أداء بعض برامج التشكيل الآلي والتصحيح من خلال مناقشة بعض الأسس العلمية – الحاسوبية واللغوية - التي تقوم عليها، من منظور لغوي حاسوبي. وقد أجرى البحث تطبيقاته على البرنامجين: (الحركات) و(مشكال)، واختار البحث هذين البرنامجين للتطبيق عليهما؛ لأن كلا منهما يمثل اتجاهاً معيناً في التشكيل الآلي والتصحيح، ونظراً لذبوع صيتهما في هذا المجال. والبحث مكون من مقدمة، وبها التعريف بمشكلة البحث، والبرامج مجال التطبيق، والأسس الحاسوبية واللغوية المهمة التي تقوم عليها برامج التشكيل الآلي والتصحيح، وبلي ذلك تقويم أداء بعض برامج التشكيل الآلي والتصحيح ومناقشة بعض الأسس العلمية التي تقوم عليها، ثم تأتي نتائج البحث، وخلصته.

الكلمات المفتاحية: نظام حاسوبي للمعالجة الآلية، برامج التشكيل الآلي والتصحيح، معالجة الأساليب اللغوية، صفّ قوائم لغوية.

1 المقدمة:

1-1 برامج التشكيل الآلي Automated configuration programs:

أصبح حقيقةً وجود بعض البرامج الحاسوبية التي تقوم بالتشكيل الآلي والتصحيح اللغوي، وقد زاد الإقبال عليها مؤخراً؛ لما أصبح للجانب اللغوي من شأن في الأونة الأخيرة، التي شهدت نوعاً من الوعي الحضاري، الذي دفع إلى العودة والبحث عن الأصول والجذور، ومحاولة التشبث بها؛ خوفاً من الانجراف بتأثير عواصف الفكر الحديث. وقد دعا هذا إلى محاولة تعرف مدى الفائدة المرجوة من استعمال هذه البرامج، وقدرتها على تلبية حاجات النهم اللغوي الملاحظ انتشاره.

غير أن التعامل مع مثل هذه البرامج له بعض السمات، أهمها:

- 1- أنه يعتمد على الحدس (Heuristic) ومبدأ معاودة المحاولة للاكتشاف؛ إذ ليس مثل هذه البرامج مما ينشر طريقة عمله وقواعد بنائه، وإنما على الباحث أن يحاول اكتشاف ذلك بنفسه.
- 2- تداخل مراحل العمل وتشابكه؛ فقد أكون في طور اختبار أحد البرنامجين في قاعدة ما؛ فإذا بالجملة المدخلة يحدث فيها خطأ أو أكثر يلفت النظر إلى أهمية اختبار قاعدة أخرى غير التي أدخلت الجملة لاختبارها.
- 3- كثرة محاولات إدخال الجمل المختبرة؛ إذ اتساع اللغة العربية يفرض ذلك، وللتأكد من صحة النتائج.

1-2 مجال تطبيق البحث:

أجرى البحث تطبيقاته على البرنامجين: (الحركات) و(مشكال) في الفترة من: 1/ 12/ 2017م الموافق 13/ 3/ 1439هـ إلى: 30/ 12/ 2017م الموافق 12/ 4/ 1439هـ. [1]

واختار البحث هذين البرنامجين للتطبيق عليهما؛ نظراً لانتشارهما، وذبوع صيتهما في مجال التشكيل الآلي والتصحيح، ولأن كلا منهما يمثل اتجاهاً معيناً في التشكيل الآلي والتصحيح كما سيبيّن البحث.

1-3 أسس بناء برامج التشكيل الآلي والتصحيح:

بمتابعة العمل في مجال حوسبة اللغة، والمعالجة الآلية لها؛ اتضح أنه يجب توافر عدد من الأسس تُبنى عليها برامج التشكيل الآلي والتصحيح، وهي أسس حاسوبية ولغوية، ومن هذه الأسس:

إعداد نظام حاسوبي للمعالجة الآلية، وضع قاعدة بيانات لغوية، صفّ قوائم لغوية ترتب فيها الألفاظ المتشابهة حسب الأكثر استعمالاً، تنميط الجمل، معرفة وسائل طول الجملة، حصر السوابق واللاحق، رصد الفرائز اللغوية والمقامية، التعامل مع اللغة على تنوع مستويات استعمالها، معالجة علامات الترقيم.

وسأبين هذه الأسس، وسأقوم بمحاولة الكشف عن تحققها في البرنامجين الحاسوبيين محل تطبيق البحث؛ وذلك من منظور لغوي حاسوبي.

2- تقويم أداء بعض برامج التشكيل الآلي والتصحيح ومناقشة بعض الأسس العلمية التي تقوم عليها:

1-2 إعداد نظام حاسوبي للمعالجة الآلية:

توجد عدة أنظمة للتشكيل الآلي والتصحيح والإعراب تبني عليها برامج المعالجة الحاسوبية، منها:

1- تحليل الألفاظ صرفياً ووضع تشكيل البنية، واختيار معانيها، ومن ثم وضع تشكيل الإعراب، ويتم ذلك عن طريق إدخال النص اللغوي إلى النظام المعالج آلياً، فيجري عليه عدداً من عمليات التحليل والتركيب، ومن ثم ينتج عدد من الصور اللغوية المقبولة أو المرفوضة للنص اللغوي، بعدها يحاول النظام المعالج آلياً أن يصل إلى الصورة الأصح من بين الاحتمالات الناتجة من عمليات التحليل والتركيب؛ مراعي الجانب الدلالي "تم تطوير عدة معالجات آلية للصرف العربي قادرة على القيام بجميع عمليات التحليل والتركيب لكلمات اللغة العربية؛ يقوم الشق التحليلي بتفكيك الكلمة إلى عناصرها الأولية الاشتقاقية والتصريفية والإعرابية والواصق السابقة واللاحقة (مثال تحليل كلمة: (وإيجادهم) إلى حرف العطف (و)، وساق الكلمة (إيجاد)، والضمير المتصل (هم)، ثم تحليل ساق الكلمة (إيجاد) إلى الجذر (وجد) على صيغة (إفعال) بعد عكس عمليات إبدال (و) إلى (ي). أما التركيب الصرفي فيمثل العملية العكسية لتكوين الكلمات من عناصرها الأولية؛ كأن يغذى المعالج الصرفي الآلي بالجذر (ق و م)، ويطلب منه تركيب فعله المضارع على صيغة (استفعل) لجمع المؤنث الغائب، فيقوم المعالج بإخراج الكلمة النهائية (يستقمن)، يتم ذلك من خلال قيام معالج الصرف الآلي بصهر الجذر في قالب الاشتقاق المطلوب، والقيام أتوماتيكياً بجميع عمليات الإبدال والإعلال والحذف. ويعد المعالج الآلي مقوماً أساسياً في تحليل النصوص العربية واسترجاعها، وكذلك في عمليات الإعراب الآلي للجمل العربية." [2]

فالمعالج الآلي يقوم بعمليتي التحليل والتركيب، وبعد ذلك تبدأ مرحلة وضع تشكيل الإعراب؛ حيث "يقوم نظام المعالج النحوي الآلي (parsing) بتفكيك الجمل إلى عناصرها الأولية من أفعال وأسماء وأشبه جمل وظروف وما شابه، وتحديد الوظائف النحوية لكل عنصر (فاعل، مفعول، خبر، صفة، حال...)، وربط الضمانات بمرجعها، والتعويض عن المحذوف، وذلك تمهيداً لتمثيل بنية الجملة بصورة تفصيلية." [3]

2- ترجيح احتمال واحد من عدد من صور التحليل المدخلة سلفاً؛ بناء على عدد من الأسس، ثم وضع تشكيل الإعراب، دون معرفة المعنى؛ اكتفاءً بتوظيف بعض القرائن اللفظية، ويتم ذلك عن طريق إدخال النص اللغوي إلى النظام المعالج آلياً، فيجري عدداً من عمليات التعرف الشكلي على بعض الألفاظ، ومن خلال ذلك يُعمل الحدس (Heuristic)، فيصل إلى تشكيل ما تبقى من الألفاظ النص اللغوي، وذلك عن طريق تعرف الألفاظ شكلياً، ثم اختيار التحليل الأول في قائمة بها ترتيب للألفاظ المتشابهة الشكل حسب كثرة الاستعمال، بحيث يوضع الأكثر استعمالاً أولاً ثم الأقل وهكذا؛ فكلمة مثل : (كتب) لها أكثر من تحليل (كُتِبَ/ كَتَبَ/ كَتَّبَ/ كُتِبَ/ كُتِبَ/ كُتِبَ/ كُتِبَ/ كُتِبَ) فترتب حسب الأكثر استعمالاً "إعادة ترتيب المداخل المعجمية للأفعال مسألة حيوية للغاية؛ لكونها ركناً أساسياً في التحليل النحوي والدلالي، يزيل كثيراً من الالتباس والاضطراب، ويعيد الدقة والوضوح، وبخاصة حين تختلط الأمثلة وتتشابه معانيها إلى حد يجب معه ضرورة اللجوء إلى تحليل آخر إلى جانب التحليل التركيبي." [4]، وليس هذا خاصاً بالأفعال فحسب، بل هو لكل أقسام الكلام. وغني عن التنبيه أن هذا المسلك في التعامل مع النص اللغوي العربي قد يجانبه الصواب؛ لما يتمتع به النص اللغوي العربي من ثراء دلالي قد لا يفصح عنه التعامل السطحي مع النص. ويبين ذلك تشكيل البرنامجين لبعض الأمثلة في الجدول رقم (1):

الجدول رقم (1)

طرق المعالجة الآلية للنص العربي

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- كتب أحمد كتابه	كُتِبَ أَحْمَدُ كِتَابَهُ	كُتِبَ أَحْمَدُ كِتَابَهُ	برنامج الحركات حلل (كتب) فعلاً ماضياً ثلاثياً، وكوّن جملة صحيحة التشكيل.
2- كتب أحمد كتابه جيداً	كُتِبَ أَحْمَدُ كِتَابَهُ جَيِّدًا	كُتِبَ أَحْمَدُ كِتَابَهُ جَيِّدًا	أما برنامج مشكال فشكل بنية (كتب) على أنها جمع تكسير، وهذا خطأ، وأخطأ في تشكيل إعراب كلمة (أحمد). والبرنامجان كلاهما أصاب في تشكيل كلمة (كتابه) وكلمة (جيداً).
3- كتب أحمد كتابه جيداً	كُتِبَ أَحْمَدُ كِتَابَهُ جَيِّدَةً	كُتِبَ أَحْمَدُ كِتَابَهُ جَيِّدِهِ	برنامج الحركات صحح الخطأ في كلمة (أحمد) فجعل همزة الوصل همزة قطع، وكذلك في كلمة (كتابه) وكلمة (جيداً)؛ حيث جعل الهاء تاءً مربوطة، وأنتج جملة صحيحة التشكيل إعراباً وبنيةً. أما برنامج مشكال فشكل كلمة (كتب) فعلاً ماضياً، ولم يصحح الخطأ في (أحمد)، بل اعتبره فعل أمر، ولم يصحح الخطأ في (كتابه)، بل اعتبرها اسماً مضافاً للضمير، وكذلك لم يصحح الخطأ في (جيداً)، بل اعتبرها صفة مضافة للضمير.

الاستنتاج:

1- برنامج الحركات يقوم بعملية تحليل الألفاظ، ومن ثم تعرف المفردات المدخلة، وأجاد في تشكيل البنية، وتكوين جمل، ووضع تشكيل الإعراب.

أما برنامج مشكال فلا يقوم بعملية تحليل الألفاظ، ولكن يعمل عن طريق تعرف الألفاظ شكلياً، ثم اختيار التحليل الأول في قائمة بها الألفاظ المتشابهة الشكل مرتبة حسب كثرة الاستعمال، بحيث يوضع الأكثر استعمالاً أولاً ثم الأقل وهكذا. وقد رُتبت ألفاظ القائمة بطريقة خاطئة؛ إذ قُدمت (كُتِب) جمع التكسير في الترتيب على (كُتِب) الفعل الماضي، ومن ثم أخطأ في تشكيل جملة: كتب أحمد كتابه، ولكن لما جاء بعدها كلمة تعرفها البرنامج على أنها فعل ماضٍ (أحمد)؛ اختار (كُتِب) فعلاً ماضياً. ولم يقدّم برنامج مشكال - كذلك - بتكوين جمل؛ لأنه لو كان يكون جملاً لجا (أحمد) مجروراً لأنه مضاف؛ فهو اسم جاء بعد اسم.

2- برنامج الحركات استطاع أن يصحح الخطأ في: (أحمد/ كتابه/ جیده)، وينتج جملة صحيحة التشكيل؛ لأنه يكون جملاً بناءً على تحليلات لغوية للألفاظ، ومن ثم رفض التحليلات الخاطئة التي لا تتوافق مع الفهم الآلي لوضع الجملة الصحيح. لكن برنامج مشكال ليس عنده قدرة على تصحيح الخطأ، وإنما يتعامل مع النص المدخل فحسب.

3- رَبَط برنامج الحركات ربطاً جيداً بين (كتابه وجيدا) وبين (كتابه وجيدة)؛ حتى إنه صحح الكتابة المدخلة خطأً ليحدث التوافق. ولم يستطع ذلك برنامج مشكال.

2-2 وضع قاعدة بيانات لغوية Language database :

تعرف قاعدة البيانات بأنها "مجموعة من البيانات المرتبطة والمنظمة في الصورة الإلكترونية التي يمكن الدخول عليها ومعالجتها بواسطة برمجيات كمبيوتر متخصصة" [5].

والفائدة المترتبة على تغذية البرامج المعالجة آلياً بقواعد البيانات هي تكوين ملكة اصطناعية، تستطيع الفهم الآلي للغة، ومن ثم التشكيل، أو على الأقل تكون مرجعاً معرفياً تستفيد منه البرامج المعالجة آلياً. إنه سعي إلى الوصول للكفاية اللغوية عند تشومسكي N Chomsky . ، ويُعرفها بأنها [6] " نظام ثابت من المبادئ المولدة " والتي تُمكن كل فرد من إنتاج عدد لا نهائي من الجمل ذات المعنى في لغته ، كما تمكنه من التعرف التلقائي على الجمل ، على اعتبار أنها تنتمي إلى هذه اللغة ، حتى وإن كان غير قادر على معرفة لماذا ، و غير قادر على تقديم تفسير لها.

ويُتابع تغذية قواعد البيانات دوماً؛ لما يطرأ على اللغة من تطور في الأساليب، وزيادة في المفردات، أو لتدارك النقص الحاصل. وقد لاحظت أن برنامج الحركات لم يدرج في قاعدة بياناته الفعل (مات) رغم شيوع استعماله، والجدول رقم (2) يوضح ذلك.

الجدول رقم (2)

متابعة تزويد قاعدة البيانات اللغوية

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
إن مات الرجل حوسب	إنَّ مَاتَ الرَّجُلُ حُوسِبَ	إنَّ مَاتَ الرَّجُلُ حُوسِبَ	برنامج الحركات لم يتعرف الفعل (مات)؛ ولذلك اعتبر (إن) مؤكدة. برنامج مشكال اعتبر (إن) مؤكدة، وإن كان قد تعرف الفعل (مات).

الاستنتاج:

برنامج الحركات نقص في قواعد بيانات الأفعال؛ فلم يدخل الفعل (مات) في قواعد البيانات بدليل عدم تشكيل آخره، ومن ثم عدم اعتبار (إن) شرطية.

والبرنامجان كلاهما به نقص في قواعد الجملة الاسمية المؤكدة؛ إذ شكلا (إن) حرف توكيد رغم أن ما بعدها ليس جملة اسمية.

2-3 صفّ قوائم لغوية ترتب فيها الألفاظ المتشابهة:

من المبادئ التي تقوم عليها برامج المعالجة الآلية للغة - أية لغة - ترتيب قائمة الألفاظ المتشابهة حسب الأكثر استعمالاً والأشهر (frequency)؛ بحيث يوضع الأكثر استعمالاً والأشهر في أعلى القائمة نزولاً إلى الأقل استعمالاً وشهرةً، وهذا يفيد عندما يجد البرنامج المعالج حاسوبياً أمامه أكثر من تحليل لغوي للفظ الواحد، ولا توجد قرينة ترجح تحليلاً على آخر؛ فيلجأ البرنامج لاختيار التحليل الأول في القائمة. ويجب ترتيب هذه التحليلات بطريقة صحيحة؛ لأن الخطأ في ذلك ينتج تحليلات خاطئة وتشكيلاً خاطئاً "وهكذا فإن إعادة ترتيب المداخل المعجمية للأفعال مسألة حيوية للغاية؛ لكونها ركناً أساسياً في التحليل النحوي والدلالي، يزيل كثيراً

من الالتباس والاضطراب، ويعيد الدقة والوضوح، وبخاصة حين تختلط الأمثلة وتتشابه معانيها إلى حد يجب معه ضرورة اللجوء إلى تحليل آخر إلى جانب التحليل التركيبي." [7]

وغني عن الذكر أن مثل هذه القوائم التي ترتب فيها الألفاظ المتشابهة بحيث يوضع اللفظ الأكثر شهرة أولاً، ويليه الأقل شهرة، وهكذا دواليك-تحتاج إلى مراجعة دورية؛ لما يطرأ على ألفاظ اللغة من تغير في نسبة الاستعمال والشيوع. وسأوضح ذلك بالأمثلة:

- كلمة المدرسة:

كلمة (المدرسة) لها تحليلان لغويان، هما: (المُدْرَسَة) مؤنث المدرس أي المعلم، و(المُدْرَسَة) المكان الذي تقدم فيه الخدمة التعليمية، وفي (الوسيط): "المدرس: الكثير الدرس والتلاوة في الكتاب. والمعلم" [8]، أما المُدْرَسَة فهي: "مكان الدرس والتعليم. وجماعة من الفلاسفة أو المفكرين أو الباحثين، تعتنق مذهباً معيناً، أو تقول برأي مشترك" [9].

فكيف يفاضل البرنامج الحاسوبي المعالج بين التحليلات المطروحة المتشابهة اللفظ؟ في مناقشة تشكيل البرنامج الحاسوبيين للجملة الآتية في الجدول رقم (3) توضيح لذلك:

الجدول رقم (3)

صف قوائم لغوية ترتب فيها الألفاظ المتشابهة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- ذهب محمد إلى المدرسة	ذَهَبَ مُحَمَّدٌ إِلَى الْمَدْرَسَةِ	ذَهَبَ مُحَمَّدٌ إِلَى الْمَدْرَسَةِ	التشكيل صحيح لدى البرنامجين كليهما؛ فقد اختار البرنامجان كلاهما (ذهب) فعلاً ماضياً، و(محمد) فاعلاً، و(إلى) حرف جرّ، و(المدرسة) اسماً مجروراً، وهي مكان الدرس والتعليم، وليس المعلمة.
2- ذهب محمد إلى المدرسة التي غابت أمس	ذَهَبَ مُحَمَّدٌ إِلَى الْمَدْرَسَةِ الَّتِي غَابَتْ أَمْسٌ	ذَهَبَ مُحَمَّدٌ إِلَى الْمَدْرَسَةِ الَّتِي غَابَتْ أَمْسٌ	شكل البرنامجين كلاهما (المدرسة) على أنها مكان الدرس والتعليم، وليس المعلمة. وباقي الجملة تشكيله صحيح لدى البرنامجين.
3- ذهبت فاطمة إلى زميلتها التي غابت أمس	ذَهَبَتْ فَاطِمَةٌ إِلَى زَمِيلَتِهَا الْمَدْرَسَةِ الَّتِي غَابَتْ أَمْسٌ	ذَهَبَتْ فَاطِمَةٌ إِلَى زَمِيلَتِهَا الْمَدْرَسَةِ الَّتِي غَابَتْ أَمْسٌ	شكل برنامج الحركات الجملة تشكيلاً صحيحاً، لكن شكل كلمة (المدرسة) على أنها مكان الدرس والتعليم، وليست المعلمة، وأخطأ في إعرابها. أما برنامج مشكال فبنى الفعل للمجهول، وأخطأ في إعراب فاطمة، وشكل كلمة (المدرسة) على أنها مكان الدرس والتعليم، وليس المعلمة.

الاستنتاج:

- 1- في الجملة الأولى (ذهب محمد إلى المدرسة) التشكيل صحيح؛ لكن ماذا لو كان المراد: المدرسة أي المعلمة؟
 - 2- في الجملة الثانية (ذهب محمد إلى المدرسة التي غابت أمس) الخطأ في ترتيب قائمة الألفاظ المتشابهة؛ حيث اختار البرنامجين كلاهما تشكيل (المدرسة) بمعنى مكان الدرس والتعليم، ولم يختاراً تشكيلها بمعنى المعلمة. وكان المفترض أن الفعل (غاب) يكون دليلاً على أن تشكيل كلمة (المدرسة) بمعنى المعلمة هو الصواب، وليس مكان الدرس والتعليم؛ لأن إسناد الفعل (غاب) إلى كلمة (المدرسة) بمعنى المعلمة أكثر استعمالاً من إسناده إلى كلمة (المدرسة) بمعنى مكان الدرس والتعليم.
 - 3- في الجملة الثالثة (ذهبت فاطمة إلى زميلتها التي غابت أمس) برنامج الحركات أخطأ في تشكيل كلمة (المدرسة) وفي إعرابها، ولو اعتمد على كلمة (زميلتها)، واستفاد منها قرينة، وكانت عنده قواعد البديل قوية؛ لاختار تشكيل (المدرسة) بمعنى المعلمة، ولأعرابها إعراباً صحيحاً. أما برنامج مشكال فاعتبر تاء التانيث نائب فاعل، وهذا خطأ، وترتب عليه الخطأ في إعراب كلمة (فاطمة). وخطؤه في تحليل (المدرسة) دليل على عدم فهمه المعنى؛ لأنه أعرابها بدلاً مجروراً، لكن لم ينجح في اختيار التحليل الصواب.
- إذن توجد مشكلة في تحليل كلمة (المدرسة)؛ فهل البرنامجين ليس عندهما إلا تحليل واحد، وهو (المدرسة) بمعنى مكان الدرس والتعليم؟

حاولت التأكد عن طريق إدخال عدد من الجمل الأخرى التي تحتوي على كلمة (المدرسة) إلى البرنامجين، بحيث يكون الميل أكثر إلى اختيار تحليل (المدرسة) بمعنى المعلمة، كما سيأتي في الجدول رقم (4):

الجدول رقم (4)

التأكد من إدراج بعض الألفاظ في قوائم لغوية

الملاحظات	برنامج مشكال	برنامج الحركات	المدخل
برنامج الحركات اعتبر الفاء عاطفة، والفعل ثلاثيا مجردا مبنيا للمعلوم، وشكل (المدرسة) على أنها المعلمة. برنامج مشكال اعتبر الفعل ثلاثيا مجردا مبنيا للمجهول، والتاء نائب فاعل، والمدرسة مفعولا به ثانيا، وتشكيل بنيتها يدل على أنها مكان الدرس والتعليم، وليست المعلمة.	فَهَمَّتِ الْمَدْرَسَةُ	فَهَمَّتِ الْمَدْرَسَةُ	1- فهمت المدرسة
برنامج الحركات اعتبر الفاء عاطفة، والفعل ثلاثيا مجردا مبنيا للمعلوم، وشكل (المدرسة) على أنها مكان الدرس والتعليم، ونصب (الطلاب) و(الدرس).	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ	2- فهمت المدرسة الطلاب
برنامج مشكال أخطؤه هي أخطاء إعراب الجملة السابقة، ولكن شكل (الطلاب) و(الدرس) تشكيلا صحيحا.	فَهَمَّتِ الْمَدْرَسَةُ الدَّرْسُ	فَهَمَّتِ الْمَدْرَسَةُ الدَّرْسُ	3- فهمت المدرسة الدرس
برنامج الحركات تعامل مع هذه الجملة تعامله مع الجملتين السابقتين، ولكن شكل (الدرس) مرفوعاً. برنامج مشكال تعامل مع هذه الجملة تعامله مع الجملتين السابقتين، ونصب (الدرس).	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ الدَّرْسُ	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ الدَّرْسُ	4- فهمت المدرسة الطلاب الدرس
برنامج الحركات تعامل مع هذه الجملة تعامله مع الجمل السابقة، وأخطأ في إعراب المعطوف (المنهج). برنامج مشكال تعامل مع هذه الجملة تعامله مع الجمل السابقة، ولكن أصاب في تشكيل المعطوف (المنهج) تشكيل إعراب، أما تشكيل البنية فقد اختار غير الأشهر؛ حيث كسر الميم.	فَهَمَّتِ الْمَدْرَسَةُ الدَّرْسُ وَالْمَنْهَجُ	فَهَمَّتِ الْمَدْرَسَةُ الدَّرْسُ وَالْمَنْهَجُ	5- فهمت المدرسة الدرس والمنهج
برنامج الحركات تعامل مع هذه الجملة تعامله مع الجمل السابقة، ولكن أصاب في عطف (المنهج) على (الدرس)؛ وإن كانت علامة المعطوف عليه خطأ، وأخطأ في تشكيل ضمير الغائب المضاف إلى كلمة (كل)؛ حيث جره في حين أنه مبني على الضم. برنامج مشكال تعامل مع هذه الجملة تعامله مع الجمل السابقة، ولم يتعرف ضمير الغائب المضاف إلى كلمة (كل)، ومن ثم لم يشكله.	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ الدَّرْسُ وَالْمَنْهَجُ كُلُّهُ.	فَهَمَّتِ الْمَدْرَسَةُ الطُّلَّابُ الدَّرْسُ وَالْمَنْهَجُ كُلُّهُ	6- فهمت المدرسة الطلاب الدرس والمنهج كله

الاستنتاج:

1- يتضح من الجملة الأولى (فهمت المدرسة) أن برنامج الحركات مدرج عنده تحليل (المدرسة) بمعنى المعلمة، ولكن عنده إشكال في ترتيب الألفاظ المتشابهة؛ فقدم التحليل الذي اختاره، وهو تحليل (همّ) المسبوق بالفاء العاطفة، على كونه فعلا ماضيا دون حرف عطف (فهمت).

أما برنامج مشكال فترتيب قوائم الألفاظ المتشابهة عنده به خطأ؛ إذ إنه قدم المبني للمجهول على المبني للمعلوم. وبه نقص في قواعد البناء للمجهول.

2- يؤكد تشكيل الجملتين الثانية (فهمت المدرسة الطلاب) والثالثة (فهمت المدرسة الدرس) أن البرنامجين كليهما لم يفهم الجملتين المدخلتين.

وبرنامج الحركات به خطأ في ترتيب قائمة الألفاظ المتشابهة؛ حيث قدم تحليل (همّ) المسبوق بالفاء العاطفة على كونه فعلا ماضيا دون حرف عطف (فهمت)، وقدم تحليل (المدرسة) بمعنى مكان الدرس والتعلم عليها بمعنى معلمة، والكلمتان اللتان زيدتا على الجملتين، وهما: (الطلاب) و(الدرس) ترجحان كون (المدرسة) بمعنى المعلمة.

أما برنامج مشكال فتشكيل هاتين الجملتين يؤكد الخطأ في ترتيب قوائم الألفاظ المتشابهة والنقص في قواعد البناء للمجهول. ولم يتضح الأساس الذي شكل بناءً عليه كلا من: (الطلاب) و(الدرس) منصوبتين.

وتشكيل الجملة التالية يبين موقف كل من البرنامجين من المفعول الثاني للفعل (فهم).

3- يؤكد تشكيل الجملة الرابعة (فهمت المدرسة الطلاب الدرس) أن أسباب خطأ كل من برنامج الحركات وبرنامج مشكال في تشكيل هذه الجملة هي الأسباب نفسها في الجملتين السابقتين، ورفع برنامج الحركات كلمة (الدرس) غير مبرر، وكذلك نصب برنامج مشكال لها.

4- يؤكد تشكيل الجملة الخامسة (فهتت المدرسة الدرس والمنهج) على ما سبق من أخطاء برنامج الحركات، ويضاف إليه أن به نقصاً في قواعد العطف وتكوين الجملة؛ إذ لم يعطف (المنهج) على (الدرس). أما برنامج مشكال فلا يشكل اعتماداً على المعنى بدليل اختياره مجموعة تحليلات غير متوافقة.

5- يتضح من تشكيل الجملة السادسة (فهتت المدرسة الطلاب الدرس والمنهج) أن كلا من برنامج الحركات وبرنامج مشكال - إضافة لما سبق - بهما نقص في قواعد إضافة كلمة (كل) للضمير، فضلاً عن أن برنامج مشكال به نقص في قائمة الألفاظ المتشابهة؛ إذ إن برنامج الحركات مدرج عنده تحليل (المدرسة) بمعنى المعلمة، وليس كذلك برنامج مشكال.

2-4 تنميط الجمل:

يتم تزويد البرنامج الحاسوبي المعالج بأنماط الجمل في العربية؛ ليتمكن من بناء الجمل على ضوء ذلك بعد الانتهاء من تحليل المفردات وفك اللبس الصرفي "النماذج اللغوية المتاحة لصياغة قواعد النحو لأغراض المعالجة الآلية قد صممت أصلاً لتلائم مطالب اللغة الإنجليزية التي تتسم بالصرامة النسبية لرتبة الكلمات strict word order داخل جملها، ويعني ذلك احتياجنا إلى بحوث أساسية وتطبيقية لكتابة قواعد النحو الصوري formal grammar اللازم لمعالجة النحو العربي ألياً؛ حيث يجب أن يراعى في كتابته جميع البدائل الممكنة لأنماط الجملة العربية بفعل عمليات التقديم والتأخير، والحذف، والإبدال، والإضمار. ولتوضيح الفرق يحتاج نحو الإنجليزية إلى ما يقرب من (1000) قاعدة رياضية، في حين وصل عدد القواعد لنحو اللغة العربية غير المشكولة الذي قام بصياغته الكاتب إلى ما يزيد على (12) ألف قاعدة" [10]. وسيعرض البحث عدداً من أنماط الجملة الاسمية على البرنامجين محل تطبيق البحث ليتضح مستوى الأداء، وهذا ما يوضحه الجدول رقم (5).

الجدول رقم (5)

تنميط الجمل في قاعدة بيانات البرنامج المعالج

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- محمد مجتهد	مُحَمَّدٌ مُجْتَهِدٌ	مُحَمَّدٌ مُجْتَهِدٌ	برنامج الحركات تشكيله صحيح بنية وإعراباً. برنامج مشكال تشكيله صحيح بنية، ولكنه أخطأ في تشكيل إعراب (مجتهد)، ولم ينون كلمة (محمد).
2- في الكلية طلاب	فِي الكَلِيَّةِ طُلَّابٌ	فِي الكَلِيَّةِ طُلَّابٌ	برنامج الحركات تشكيله صحيح بنية وإعراباً. برنامج مشكال تشكيله صحيح بنية، ولكن أخطأ في تشكيل إعراب كلمة (طلاب).
3- منا رجال	مِنَّا رِجَالٌ	مِنَّا رُجَالٌ	برنامج الحركات تشكيله صحيح بنية وإعراباً. برنامج مشكال تشكيله خاطئ؛ حيث اعتبر (منا) أمراً للمثنى، و(رُجَالٌ) جمع (راجل)، ونصبها، وحقها الرفع.
4- في البيت رجال	فِي البَيْتِ رِجَالٌ	فِي البَيْتِ رُجَالٌ	برنامج الحركات تشكيله صحيح بنية وإعراباً. برنامج مشكال أصاب في تشكيل الجار والمجرور، ولكن أخطأ في تشكيل ما بعدهما (رُجَالٌ)؛ حيث اعتبرها جمع (راجل)، ونصبها، وحقها الرفع.
5- من المؤمنين رجال	مِنَ الْمُؤْمِنِينَ رِجَالٌ	مِنَ الْمُؤْمِنِينَ رُجَالٌ	

الاستنتاج:

يتضح من تشكيل هذه الجمل الخمسة أن برنامج الحركات استطاع أن يشكل تشكيلاً صحيحاً، وكوّن جملاً صحيحة؛ لأن أنماط الجمل: (محمد مجتهد)، وجملة (في الكلية طلاب)، وجملة (منا رجال)، وجملة (في البيت رجال)، وجملة (من المؤمنين رجال) مدرجة فيه.

أما برنامج مشكال فبه نقص في قواعد تركيب الجملة الاسمية؛ إذ لم ينون كلمة (محمد)، وأضاف إليها كلمة (مجتهد)، ولم يرفع (مجتهد) و(طلاب) و(رجال) أخباراً في الجمل المذكورة؛ لأن أنماط هذه الجمل الاسمية غير مدرجة فيه. وبرنامج مشكال خطأ - أيضاً- في ترتيب الأشهر من الألفاظ المتشابهة؛ ف(منا) الأشهر أن تكون جاراً ومجروراً، وليس أمراً للثنتين. و(رجال) الأشهر أن تكون جمع (رجل)، وليست جمع (راجل).

2-5 معرفة وسائل طول الجملة:

طول الجملة يضع صعوبات لغوية أمام النظام المعالج حاسوبياً، وكذلك الجمل المتداخلة، ومن ثم يجب تزويد البرنامج المعالج حاسوبياً بوسائل طول الجملة لمعالجتها وتجنب الخطأ، ومن وسائل طول الجملة العربية:

5-2-1 كون الخبر جملة:

مجيء الخبر جملة من وسائل طول الجملة في العربية، وإن لم يكن البرنامج مدرجا فيه هذا النمط، فلن يتمكن من تشكيل الجملة، ويتضح ذلك من الجدول رقم (6):

الجدول رقم (6)

طول الجملة العربية عن طريق كون الخبر جملة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
داء محمد سهل علاجه	دَاءٌ مُحَمَّدٌ سَهْلٌ عِلَاجُهُ	دَاءٌ مُحَمَّدٌ سَهْلٌ عِلَاجُهُ	برنامج الحركات أخطأ في تشكيل كلمة (سهل)؛ حيث إنه لم ينونه، وفي تشكيل (علاجه)؛ حيث إنه لم يرفعه، وفي الضمير المضاف؛ إذ لم يبينه على الضمّ. برنامج مشكال أخطأ في تشكيل كلمة (محمد) حيث إنه لم يجره، بل رفعه غير منون، وأخطأ أخطاء برنامج الحركات نفسها.

الاستنتاج:

يتضح من تشكيل الجملة أن كلا من برنامج الحركات وبرنامج مشكال به نقص في القواعد الخاصة بكون الخبر جملة، وفي قواعد إعراب الصفة المشبهة العاملة عمل فعلها، وفي قواعد الإضافة للضمير. أما برنامج مشكال فبه - كذلك - نقص في قواعد الإضافة للاسم الظاهر؛ إذ لم يصف (محمد) إلى داء.

5-2-2 عوارض تؤثر تركيباً ودلالة على الجملة الاسمية:

من العوارض التي تطيل الجملة الاسمية وتؤثر تركيباً ودلالة (كان) الفعل الناقص، وقد اضطرب البرنامجان في التعامل معها، ويبينه الجدول رقم (7).

الجدول رقم (7)

تعرف عوارض الجملة الاسمية المؤثرة تركيباً ودلالة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- ما كان محمد	مَا كَانَ مُحَمَّدًا	مَا كَانَ مُحَمَّدٌ	برنامج الحركات اعتبر (محمد) خبراً لكان؛ فنصبه، وأضاف إليه ألف تنوين النصب. برنامج مشكال شكل (محمد) بالضمّة، ولم ينونه.
2- ما كان محمد خائناً	مَا كَانَ مُحَمَّدٌ خَائِنًا	مَا كَانَ مُحَمَّدٌ خَائِنًا	برنامج الحركات شكّل الألفاظ كلها تشكيلاً صحيحاً. برنامج مشكال لم ينون (محمد)، واعتبر (خائناً) مثني مرفوعاً.
3- ما كان محمد أباً	مَا كَانَ مُحَمَّدًا أَبًا	مَا كَانَ مُحَمَّدٌ أَبًا	برنامج الحركات أضاف ألفاً إلى (محمد) وأعربها خبراً لكان، وأخطأ في إعراب (أباً)؛ حيث إنه لم ينونها منصوباً بالفتحة، وإنما نصبها بالألف. برنامج مشكال اعتبر (محمد) اسماً لكان، لكن أخطأ في تشكيل (أباً)؛ حيث إنه لم ينونها منصوباً بالفتحة، وإنما نصبها بالألف.

الاستنتاج:

- 1- الجملة الأولى تبين أن برنامج الحركات يفترض أن الجملة المدخلة إليه كاملة بدليل أنه أعرب (محمد) خبراً لكان منصوباً، ولم يفترض أنه اسمها، وأن الخبر لم يأت بعد. أما برنامج مشكال فنجد في أنه عدّ هذه الجملة ناقصة، لكن لم ينون (محمد).
- 2- الجملة الثانية فهم برنامج الحركات أن الخبر (خائنا)، ومن ثم نجح في تشكيلها تشكيلاً صحيحاً، أما برنامج مشكال فيه نقص في قواعد عمل (كان)؛ إذ لم ينون (محمد)، واعتبر (خائنا) مثني مرفوعاً.
- 3- الجملة الثالثة تثبت أن كلا من برنامج الحركات وبرنامج مشكال بهما نقص في القواعد؛ إذ اعتبرا أن كلمة (أبا) اسم من الأسماء الستة تنصب بالألف؛ ومن ثم لم ينوناها منصوبةً بالفتحة، في حين أنها لا تنصب بالألف إلا إذا كانت مضافة إلى اسم ظاهر، وهذا الشرط غير متحقق. وعدّ برنامج الحركات اسم كان مستتراً، ومن ثم نصب (محمد)، وهذا خلاف الأولى.

5-2-3 عوارض تؤثر تركيباً ودلالة على الجملة الفعلية:

من العوارض التي تدخل على الجملة الفعلية وتؤثر تركيباً ودلالة (لن) التي ينصب بعدها الفعل المضارع، و(لم) التي يجزم بعدها الفعل المضارع، ويبين الجدول رقم (8) تصرف البرنامجين معهما.

الجدول رقم (8)

تعرف عوارض الجملة الفعلية المؤثرة تركيباً ودلالة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
لن يذهب محمد	لَنْ يَذْهَبَ مُحَمَّدٌ	لَنْ يَذْهَبَ مُحَمَّدٌ	برنامج الحركات لم ينصب الفعل المضارع (يذهب) بعد (لن). برنامج مشكال نصب الفعل المضارع (يذهب) لوقوعه بعد (لن)، ولكن لم ينون (محمد).
لم يذهب محمد	لَمْ يَذْهَبْ مُحَمَّدٌ	لَمْ يَذْهَبْ مُحَمَّدٌ	برنامج الحركات لم يجزم الفعل المضارع (يذهب) بعد (لم). برنامج مشكال جزم الفعل (يذهب)، ولكن لم ينون (محمد).

الاستنتاج:

- 1- برنامج الحركات به نقص في قواعد الفعل المضارع؛ إذ إنه لم يفهم كلا من العارض (لن) الذي يُنصب الفعل المضارع بعده، والعارض (لم) الذي يُجزم الفعل المضارع بعده.
أما برنامج مشكال ففهم كلا من العارضين (لن) و(لم)، ولكن به نقص في قواعد تكوين الجملة؛ إذ إنه لم ينون (محمد) وهي فاعل.

5-2-4 الصفة:

من وسائل طول الجملة في العربية دخول الصفة في الجملة، ويوضح الجدول رقم (9) تصرف البرنامجين معهما:

الجدول رقم (9)

طول الجملة العربية عن طريق احتوائها على صفة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- عندنا رجل	عَدْنَا رَجُلٌ	عَدْنَا رَجُلٌ	برنامج الحركات أخطأ في اختيار الضمير المضاف للظرف (عند). برنامج مشكال نجح في اختيار الضمير المضاف للظرف (عند)، لكن أخطأ في إعراب كلمة (رجل)؛ حيث نصبها رغم أنها مبتدأ مؤخر.
2- عندنا رجل عادل	عَدْنَا رَجُلٌ عَادِلٌ	عَدْنَا رَجُلٌ عَادِلٌ	برنامج الحركات لم يفهم ضمير المتكلمين، ولم ينون (رجل)، ولم يعرب (عادل) صفة، بل نصبها رغم أنها صفة لمرفوع. برنامج مشكال لم ينون (رجل)، ولم ينون (عادل).

الاستنتاج:

- 1- يتضح من الجملة الأولى أن برنامج الحركات به نقص في قواعد إضافة الضمير للظرف. أما برنامج مشكال فيه نقص في قواعد تركيب الجملة الاسمية؛ إذ أخطأ في تشكيل كلمة (رجل)، ونصبها رغم أنها مبتدأ مؤخر. فماذا سيكون تشكيل هذه الجملة إذا دخلتها الصفة؟
- 2- يوضح تشكيل الجملة الثانية أن برنامج الحركات به نقص في قواعد الإضافة للضمائر، والخبر، والصفة. ولما زاد طول الجملة زاد الخطأ؛ إذ لم ينون المبتدأ المؤخر، ونصب صفته. أما برنامج مشكال فبه نقص في قواعد الخبر.

5-2-5-2-5 البديل:

البديل في العربية لغةً هو: الخلف والعوض [11]، واصطلاحاً هو: التابع المقصود بالحكم بلا واسطة [12]، ويأخذ البديل حكم المبدل منه الإعرابي رفعا ونصبا وجرا. ويعدّ وسيلة من وسائل طول الجملة في العربية، ويراعى ذلك في المعالجة الحاسوبية للغة العربية. والجدول رقم (10) يوضح موقف البرنامجين من البديل.

الجدول رقم (10)

طول الجملة العربية عن طريق البديل

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
شارك عضو هيئة التدريس محمد بن عبدالله في مؤتمر المعالجة الآلية	شَارَكَ عَضُو هَيْئَةِ التَّدْرِيسِ مُحَمَّدُ بْنُ عَبْدِ اللَّهِ فِي مُؤْتَمَرِ الْمُعَالَجَةِ الْأَلْيَةِ	شَارَكَ عَضُو هَيْئَةِ التَّدْرِيسِ مُحَمَّدُ بْنُ عَبْدِ اللَّهِ فِي مُؤْتَمَرِ الْمُعَالَجَةِ الْأَلْيَةِ	برنامج الحركات أخطأ في تشكيل: كلمة (محمد)؛ حيث نونها، وكلمة (بن)، وكلمة (عبد)، وكلمة (الله)؛ حيث لم تعرب (ابن) صفة، وبالتالي لم تعرب كلمة (عبد)، وكلمة (الله) سواء كانت متصلة بالكلمة السابقة أو منفصلة. برنامج مشكال أخطأ في: تشكيل كلمة (محمد).

الاستنتاج:

برنامج الحركات رفع كلمة (محمد)، ومن ثم يمكن أن نتوقع أنه أعربها بدلا من كلمة (عضو)، ولكن به نقص في قواعد النعت؛ حيث نون كلمة (محمد)، ولم يشكل كلمة (بن) صفة في كثير من الجمل التي أدخلتها إليه، وكذلك به نقص في قواعد الإضافة؛ إذ لم يشكل كلمة (عبد)، وكلمة (الله).

وبرنامج مشكال به نقص في قواعد البديل؛ لأنه أخطأ في تشكيل كلمة (محمد) في هذه الجملة، في حين أنه لم يخطئ فيها في جملة: (جاءَ مُحَمَّدٌ)، وإن لم ينونها. ورفع كلمة (بن) مع أنها صفة لـ(محمد) التي نصبها.

2-6-2-6 حصر السوابق واللواحق:

دخول السوابق واللواحق قد يحدث تغييرا في بنية الألفاظ بالزيادة أو الحذف، وعلى البرنامج المعالج معالجة التغيير اللفظي الناتج عن دخول السوابق واللواحق، وأن يزود بمعلومات عن ذلك، وسيوضح البحث أثر ذلك:

1-2-6-2-6 السوابق:

تباينت ترجمات اللغويين لمصطلح Prefixes فمنهم من ترجمه بالسوابق أو الصدور [13] أو اللواحق القبلية [14]، ويسمى بعض اللغويين هذه العملية بالوصل أو الضم [15]. ويوضح الجدول رقم (11) تصرف البرنامجين مع دخول لام الجر على المعرف (ب-أل):

الجدول رقم (11)

حصر السوابق

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
شارك عضو هيئة التدريس محمد في مؤتمر المعالجة الآلية للغة العربية	شَارَكَ عَضُو هَيْئَةِ التَّدْرِيسِ مُحَمَّدٌ فِي مُؤْتَمَرِ المُعَالِجَةِ الآليَّةِ لِلغَةِ العَرَبِيَّةِ	شَارَكَ عَضُو هَيْئَةِ التَّدْرِيسِ مُحَمَّدٌ فِي مُؤْتَمَرِ المُعَالِجَةِ الآليَّةِ لِلغَةِ العَرَبِيَّةِ	برنامج الحركات شكّل الجملة تشكيلاً صحيحاً. برنامج مشكال أخطأ في: تشكيل كلمة (محمد)؛ إذ نصبها دون وضع الألف. وأخطأ في إدخال السابقة اللام على كلمة (اللغة)؛ إذ لم يشدد لام كلمة (اللغة).

الاستنتاج:

تثبت الجملة المدخلة: (شارك عضو هيئة التدريس محمد في مؤتمر المعالجة الآلية للغة العربية) أن برنامج مشكال به نقص في قاعدة إدخال اللام على ما يبدأ ب(أل) بدليل أنه أخطأ كذلك في: (أعطى للغة حقها، إنتظر للليل)؛ في حين أن برنامج الحركات لم يقع في هذه الأخطاء. وبرنامج مشكال - كذلك - نقص في قواعد البذل، وسبق التنبيه إلى ذلك.

2-2-6 اللواحق:

اللواحق هي: ما يضاف من الحروف إلى آخر الكلمة لاشتقاق كلمة أخرى [16]، وقد تنوعت ترجمات مصطلح Suffixes فترجم بالأعجاز والكسع. [17]. ويوضح الجدول رقم (12) تصرف البرنامجين مع دخول الضمير (هم) على الاسم وحرف الجر:

الجدول رقم (12)

حصر اللواحق

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
أشعر بهم وبمشاكلهم	أشْعُرُ بِهِمْ وَبِمَشَاكِلِهِمْ	أشْعُرُ بِهِمْ وَبِمَشَاكِلِهِمْ	برنامج الحركات شكّل الجملة تشكيلاً صحيحاً. برنامج مشكال أخطأ في: تركيب (اللاحقة: هم) مع كل من: حرف الجر (الباء) وكلمة (مشاكل)؛ إذ لم يكسر هاء الضمير في (هم) معهما.

الاستنتاج:

برنامج الحركات لم يخطئ في إلحاق الضمير (هم) بحرف الجر ولا بجمع التكسير، أما برنامج مشكال فقد أخطأ في ذلك، ومن ثم فيه نقص في قواعد اتصال اللاحقة (هم) بالكلمات.

2-2-7 رصد القرائن اللغوية والمقامية:

يُستفاد من القرائن في فك اللبس الصرفي والنحوي والدلالي؛ وذلك لأن البرنامج المعالج يكون أمامه أكثر من تحليل صرفي وتشكيل إعرابي وطرح دلالي، وعليه أن يفاضل بينها بالاعتماد على مجموعة وسائل، منها القرائن، وهي نوعان "تصنيف أي اسم ينطوي على نوعين متكاملين من القرائن سنطلق عليهما اسم (القرينة الداخلية) و(القرينة الخارجية)". [18]

أ- القرينة الداخلية: لها صور تتجلى فيها "تستمد القرينة الداخلية من داخل تسلسل الكلمات التي يتكون منها الاسم. وقد يكون هنالك معايير قاطعة لذلك، مثل: وجود مصطلحات تأسيس الشركات التي تدل على أسماء الشركات، مثل G.M.B.H Ltd، وقد تكون هذه المعايير حدسية Heuristic criterial، مثل: الاختصارات أو الأسماء الأولى المعروفة، والتي عادة ما تشير إلى أشخاص". [19]

ب- القرينة الخارجية: ومصدرها السياق "إن القرينة الخارجية يتم الحصول عليها من السياق الذي يظهر فيه الاسم. والأساس الذي تقوم عليه هذه القرينة هو الملاحظة الواضحة القائلة بأن الأسماء ما هي إلا طرق للإرشاد إلى أفراد من نوع معين (كالأشخاص، دور العبادة، مجموعات الصخور... إلخ)، وإن هذه الأنواع لها خواص مميزة، وتشارك في أحداث مميزة أيضاً.

وجود هذه الخواص أو الأحداث في علاقة تركيبية مع أحد أسماء الأعلام يمكن أن يستخدم لتقديم دليل تأكيدي أو معياري على التصنيف الذي يندرج تحته الاسم، ويتم تحليل القرينة الخارجية External Evidence في حدود ما يتعلق الأمر بأسبقية الإحلال، ويتم تفعيلها في إطار قواعد إعادة الكتابة المعتمدة على السياق. " [20]

وحرف الجر الذي يتعدى به الفعل قرينة "وتحديد التعدي عن طريق الحرف مهم، ويجب تمثيله في Verb Preterminals ؛ لأنه سيتم ربط الجار والمجرور المطابق للمسجل في الـ(Preterminals) بالفعل (ركن الإسناد)، في حين لا يتم ربط الجار والمجرور غير المسجل به؛ ولذلك يتعلق (إلى الشجرة) في قولنا: (صعد الرجل إلى الشجرة) بالفعل (صعد)، في حين لا يرتبط الجار والمجرور (على حين غفلة) بالفعل في قولنا: (صعد الرجل إلى الشجرة على حين غفلة من الحاضرين)... ولا يمكن التنبؤ بحرف الجر بالاعتماد على أي أسس أو اعتبارات دلالية، ولكنها خاصية وحقيقة معجمية لهذا الفعل. وضرب لذلك مثلاً من الإنجليزية؛ فالفعل depends يأخذ حرف الجر (on) لا غير "[21].

وكذلك الاسم الذي يأتي بعد حرف جرٍ ما يكون به سمة تناسب بعض سمات هذا الحرف " إن سمة (+/- مكان) التي يتضمنها - على سبيل المثال لا الحصر - حرف الجر (إلى) تحدد استعمال الركن الاسمي الذي يتبع هذا الحرف كظرف مكان؛ لذلك لا بد من أن يتضمن الاسم الذي يقع بعد حرف الجر (إلى) سمة (+مكان)، ويجب أن يتوافق الفعل مع الاسم الذي يحتوي على هذه السمة "[22]. ومناقشة تشكيل الجمل الواردة في الجدول رقم (13) توضيح لدور القرينة:

الجدول رقم (13)

دور القران اللغوية

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- عندنا دين	عَدْنَا دِينَ	عَدْنَا دِينَ	برنامج الحركات أخطأ في اختيار تحليل الضمير المضاف للظرف (عند). برنامج مشكال نجح في اختيار تحليل الضمير المضاف للظرف (عند)، لكن أخطأ في تشكيل إعراب كلمة (دين)؛ حيث نصبها رغم أنها مبتدأ مؤخر.
2- عندنا دين عظيم	عَدْنَا دِينَ عَظِيمًا	عَدْنَا دِينَ عَظِيمًا	برنامج الحركات أخطأ في اختيار تحليل الضمير كما سبق في الجملة السابقة، ولم يعرب كلمة (دين) لما وصفت بـ(عظيم) رغم أنه تعرفها وأعربها في الجملة السابقة. برنامج مشكال نجح في اختيار تحليل الضمير المضاف للظرف (عند)، لكن لم ينون كلمة (دين) ولا كلمة (عظيم).
3- علينا دين	عَلَيْنَا دِينَ	عَلَيْنَا دِينَ	كل من برنامج الحركات وبرنامج مشكال أخطأ في تشكيل كلمة (دين)؛ لأن الجار والمجرور (علينا) يقتضي أن يكون التشكيل (دِينَ). وبرنامج مشكال أخطأ في تشكيل إعراب كلمة (دين)؛ إذ نصبها وحقها الرفع.
4- علينا دين كبير	عَلَيْنَا دِينَ كَبِيرًا	عَلَيْنَا دِينَ كَبِيرًا	برنامج الحركات أخطأ في تشكيل كلمة (دين)؛ لأن الجار والمجرور (علينا) يقتضي أن يكون التحليل (دِينَ)، ولكن تشكيلها الإعرابي صحيح، وكذا تشكيل ما بعدها. برنامج مشكال أخطأ في اختيار تحليل كلمة (دين)، ولم ينونها، ولم ينون ما بعدها.
5- الدين هم بالليل	الِدِينُ هُمُ بِاللَّيْلِ	الِدِينُ هُمُ بِاللَّيْلِ	التحليل واحد لدى البرنامجين، وأخطأ كل منهما في: اختيار تشكيل كلمة (الدين)، وفي اختيار تحليل (هم) ضميراً منفصلاً.
6- الدين ذل بالنهار	الِدِينُ ذُلٌّ بِالنَّهَارِ	الِدِينُ ذُلٌّ بِالنَّهَارِ	برنامج الحركات أخطأ في اختيار تشكيل كلمة (الدين)، وفي تحليل (ذل) فعلاً ماضياً. برنامج مشكال أخطأ في تشكيل كلمة (دين)، وفي تحليل (ذل) اسماً منصوباً.
7- هذا دين في رقبتي	هَذَا دِينَ فِي رَقَبَتِي	هَذَا دِينَ فِي رَقَبَتِي	البرنامجان كلاهما أخطأ في تشكيل كلمة (دين)، وكلمة (الدين)، وكلمة (حل)، ومن ثم أخطأ في اختيار تشكيل كلمة (سداد).
8- حل سداد الدين	حَلُّ سَدَادِ الدِّينِ	حَلُّ سَدَادِ الدِّينِ	فضلاً عن أن برنامج مشكال لم يحلل كلمة (رقبتي).

9- هذا الدين واجب السداد	هَذَا الدِّينُ وَاجِبُ السَّدَادِ	هَذَا الدِّينُ وَاجِبُ السَّدَادِ	
-----------------------------	--------------------------------------	--------------------------------------	--

الاستنتاج:

- 1- في الجملة الأولى (عدنا دين) شكل برنامج الحركات كلمة (دين) بكسر الدال، ولكن به نقص في قواعد الضمير؛ إذ لم يتعرف الضمير (نا) المضاف للظرف، وبه نقص في قواعد إضافة الضمير للظرف. أما برنامج مشكال فشكل كلمة (دين) بكسر الدال، وتعرف الضمير، وأضافه للظرف، ولكن به نقص في قواعد تركيب الجملة الاسمية؛ فقد نصب كلمة (دين) رغم أنه قد سبقها شبه جملة.
- 2- في الجملة الثانية (عدنا دين عظيم) وُصفت كلمة (دين) بالوصف (عظيم)، وبرنامج الحركات به – إضافة لما سبق - نقص في البيانات الخاصة بتركيب الجملة الاسمية والصفة؛ حيث لم يشكل كلمة (دين) لما وصفت بـ(عظيم)؛ مع أنه شكلها في الجملة السابقة وهي غير موصوفة. ولم يتضح لي على أي أساس نون بالرفع كلمة (عظيم). وبرنامج مشكال به نقص في قواعد تركيب الجملة الاسمية؛ إذ لم ينون كلمة (دين) ولا كلمة (عظيم) مع أنه تعرف الضمير المضاف للظرف، ورفع كلمة (دين) وكلمة (عظيم).
- 3- من تشكيل الجملة الثالثة (علينا دين) يتضح أن البرنامجين كليهما بهما نقص في القدرة على الفهم الآلي؛ حيث إن شبه الجملة قرينة في توجيه تشكيل بنية كلمة (دين)؛ فإن كان شبه الجملة (علينا) فالأولى أن يكون ما بعدها (دَيْنٌ)، وإن كان شبه الجملة (عدنا) فالأولى أن يكون ما بعدها (دَيْنٌ). وسأحاول – في الجملة التالية - إدخال صفة فارقة ترجح كون تحليل كلمة (دين) هو (دَيْنٌ)، وليس: (دَيْنٌ). أما برنامج مشكال فعنده نقص في قواعد تركيب الجملة؛ إذ نصب كلمة (دَيْنٌ) بعد شبه الجملة.
- 4- في الجملة الرابعة (علينا دين كبير) البرنامجان كلاهما أخطأ في تحليل كلمة (دين)، وكلمة (كبير) تساعد في ترجيح تحليل (دَيْنٌ)؛ إضافة إلى ما سبق مما ذكرته في الجملة السابقة. وبرنامج مشكال لم ينون كلمة (دين) ولا كلمة (كبير)، وهذا دليل على نقص قواعد تركيب الجملة عنده.
- مما سبق يتضح عدم ورود التحليل (دَيْنٌ) بفتح الدال؛ فهل هو غير مدرج في قاعدة بيانات كل من البرنامجين؟ يبين ذلك محاولة إدخال بعض الجمل الصريحة في كون التحليل (دَيْنٌ)، وهي بعض الأقوال المشهورة والمأثورات كما سيوضح في الجملتين: الخامسة والسادسة التاليتين.
- 5- في الجملتين: الخامسة (الدين هم بالليل) والسادسة (الدين ذل بالنهار) تأكيد على عدم إدراج تحليل (دَيْنٌ) في قاعدة بيانات البرنامجين؛ إذ سياق كل من الجملتين يفرض أن يكون (دَيْنٌ) هو التحليل الصحيح. وترتيب قائمة الألفاظ المتشابهة خاطئ لدى البرنامجين؛ فالضمير (هُم) وُضع في القائمة قبل (هَمَّ) المصدر. وفي برنامج الحركات الفعل الماضي (ذَل) يسبق المصدر في ترتيب قائمة الألفاظ المتشابهة، وخطأ برنامج مشكال في نصبه (ذَل) يدل على قصور في قواعد الجملة الاسمية لدى البرنامج.
- 6- في الجمل السابعة والثامنة والتاسعة أخطأ البرنامجان كلاهما في تشكيل كلمة (دين) وكلمة (الدين)، وهذا يؤكد على صحة الاستنتاج السابق. وأخطأ البرنامجان كلاهما في تشكيل كلمة (حَلٌّ)؛ نظراً لأنهما قدّما المصدر على الفعل الماضي في قائمة الألفاظ المتشابهة. وعدم تعرف برنامج مشكال كلمة (رَقِبْتِي) يدل على نقص في قواعد بيانات البرنامج.

2- القدرة على التعامل مع اللغة على تنوع مستويات استعمالها:

فالبرنامج الحاسوبي المعالج يجب أن يكون مؤهلاً للتعامل مع اللغة التراثية كما يتعامل مع اللغة المعاصرة، وسأقوم بإدخال بعض الجمل التي لغتها لغة معاصرة وأخرى لغتها لغة تراثية؛ لأرى كيف يتعامل معها كل من البرنامجين، والجدول رقم (14) يوضح ذلك:

الجدول رقم (14)

القدرة على التعامل مع اللغة على تنوع مستويات استعمالها

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- تلك إذن قسمة صحيحة	تلك إذن قسمة صحيحة	تلك إذن قسمة صحيحة	برنامج الحركات أخطأ في تشكيل كلمة (قسمة)، وإن كان تشكيل الإعراب صحيحاً.
2- تلك إذن قسمة ضيزى	تلك إذن قسمة ضيزى	تلك إذن قسمة ضيزى	برنامج الحركات لم يتعرف كلمة (ضيزى)؛ ولذلك لم يشكلها، ومن ثم لم ينون كلمة (قسمة)، وأخطأ في تشكيل بنيتها. برنامج مشكال لم ينون كلمة (قسمة)، وأخطأ في تشكيل بنيتها.

3- رضوا بأن يكونوا مع الخوالف	رَضُوا بِأَنْ يَكُونُوا مَعَ الخوالف	كل من البرنامجين لم يتعرف كلمة (الخوالف). وبرنامج مشكال نون كلمة (مع) رغم أنها بدون الألف.
-------------------------------	--------------------------------------	--

الاستنتاج:

- 1- تثبت الجملة الأولى (تلك إذن قسمة صحيحة) أن البرنامجين كليهما تعرف كلمة (صحيحة)، وإن كان برنامج مشكال به نقص في قواعد تركيب الجمل وقواعد التنوين؛ إذ لم ينون كلمة (قسمة) ولا كلمة (صحيحة). أما برنامج الحركات فأخطأ في تشكيل كلمة (قسمة)؛ إذ شكلها: (قَسَمَة).
- 2- تثبت الجملتان الثانية (تلك إذن قسمة ضيزى) والثالثة (رضوا بأن يكونوا مع الخوالف) أن كلا من برنامج الحركات وبرنامج مشكال لا يجيد التعامل مع اللغة التراثية بكفاءة عالية؛ إذ لم يتعرفا على كلمة (الخوالف)، ولم يتعرف برنامج الحركات على كلمة (ضيزى)؛ لأنهما كلمتان تراثيتان، ومن ثم حدث الخطأ في تحليل كلمة (قسمة) وفي عدم تنوينها. وبرنامج مشكال به نقص في قواعد إعراب الظرف (مع)؛ إذ نونه رغم أنه بدون الألف.

2-9 معالجة علامات الترقيم:

يجب تزويد البرنامج الحاسوبي المعالج بعلامات الترقيم؛ ليضع لها قواعد تعالجها حتى يستطيع التعرفا أثناء المعالجة، ومن ذلك:

2-9-1 علامة الاستفهام: يوضح الجدول رقم (15) معالجة البرنامجين لعلامة الاستفهام.

الجدول رقم (15)

معالجة علامة الاستفهام

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
ماذا فعلت بنا؟	مَاذَا فَعَلْتِ بِنَا؟	مَاذَا فَعَلْتِ بِنَا؟	برنامج الحركات يقلب علامة الاستفهام العربية إلى إنجليزية. برنامج مشكال بنى الفعل للمجهول، وأسند للمؤنث، لكن لم يقلب علامة الاستفهام العربية إلى إنجليزية.

الاستنتاج:

برنامج الحركات به نقص في قواعد علامات الترقيم؛ لأنه يجعل علامة الاستفهام العربية إنجليزية، ولا يفعل ذلك برنامج مشكال. وبرنامج مشكال به خطأ في ترتيب قائمة الكلمات المتشابهة اللفظ، وقد سبق التنبيه إلى ذلك.

2-9-2 الفاصلة: يبين الجدول رقم (16) معالجة البرنامجين للفاصلة.

الجدول رقم (16)

معالجة الفاصلة

المدخل	برنامج الحركات	برنامج مشكال	الملاحظات
1- ذهب محمد وعامل العدو	ذَهَبَ مُحَمَّدٌ وَعَامِلُ الْعَدُوِّ	ذَهَبَ مُحَمَّدٌ وَعَامِلُ الْعَدُوِّ	برنامج الحركات عدّ (عامل) فعلاً، ونصب كلمة (العدوّ)، وعطف جملة على جملة. برنامج مشكال رفع (محمد) ولم ينونه، وعطف (عامل) عليه، وعدّه اسم فاعل لا فعلاً، ومن ثم جرّ كلمة (العدوّ).
2- ذهب محمد، وعامل العدو	ذَهَبَ مُحَمَّدٌ، وَعَامِلُ الْعَدُوِّ	ذَهَبَ مُحَمَّدٌ، وَعَامِلُ الْعَدُوِّ	برنامج الحركات يقلب الفاصلة العربية إلى إنجليزية. لما وُضعت الفاصلة عدّ الجملة بعدها مستقلة عن التي قبلها، ورفع كلمة (العدوّ) فاعلاً. برنامج مشكال نون كلمة (محمد)، ورفع كلمة (عامل)، وأضاف إليه

كلمة (العدوّ).

الاستنتاج:

برنامج الحركات به نقص في قواعد علامات الترقيم؛ لأنه يجعلها الفاصلة إنجليزية، وبه نقص في قواعد عطف الجمل إذا فصلت بينها الفاصلة؛ إذ عد - في الجملة الأولى (ذهب محمد وعامل العدو) - (عامل) فعلاً، ونصب كلمة (العدوّ) مفعولاً، وعطف جملة على جملة. ولكنه في الجملة الثانية (ذهب محمد، وعامل العدو)، التي دخلت فيها الفاصلة، عدّ الجملة بعد الفاصلة مستقلة عن التي قبلها، ورفع كلمة (العدوّ) فاعلاً. والأولى أن يكون إعرابها كما في الجملة الأولى؛ لتكون جملة مكتلمة.

أما برنامج مشكال فيه نقص في قواعد التنوين؛ لأنه لم ينون كلمة (محمد) في الجملة الأولى، وفي قواعد عطف الجمل إذا فصلت بينها علامة الترقيم؛ إذ لم يكون جملة كاملة مما بعد الفاصلة كما فعل برنامج الحركات.

3- الخاتمة:

بعد عرض ما قدمه هذا البحث من جهد نظري وعمل تطبيقي يحسن به أن يسجل أهم النتائج التي توصل إليها:

- الأولى أن يقوم برنامج التشكيل الآلي والتصحيح على الفهم الآلي للنص، وألا يقوم على ترجيح أحد مجموعة احتمالات حسب القرائن اللغوية والمقامية؛ لأن ذلك يجنبه كثيراً من الأخطاء، بعضها بدهي.
- اللبس ambiguity في الألفاظ العربية نسبته عالية؛ ويحد منه الاعتماد على القرائن الداخلية والخارجية.
- القرائن اللغوية والمقامية ذات أهمية كبيرة في المعالجة الآلية للغة العربية؛ لذلك يحسن رصدها وتغذية البرنامج المعالج بها، لأنه يرد أمامه أكثر من تحليل صرفي وتشكيل إعرابي وطرح دلالي، وعليه أن يفاضل بينها بالاعتماد على مجموعة وسائل، منها القرائن. ومن هذه القرائن اللغوية حروف الجر التي تتعدى بها بعض الأفعال والمشتقات، والألفاظ ذات الرتبة الثابتة، والألفاظ التي تلزم حركة إعرابية ثابتة؛ كالتي تلزم الرفع أو النصب وهكذا.
- من المفيد في المعالجة الآلية للغة - أية لغة - وضع قوائم بالألفاظ المتشابهة، ترتب فيها الألفاظ حسب الأكثر استعمالاً والأشهر (frequency)؛ بحيث يوضع الأكثر استعمالاً والأشهر في أعلى القائمة نزولاً إلى الأقل استعمالاً وشهرة، وهذا يفيد عندما يجد البرنامج المعالج حاسوبياً أمامه أكثر من تحليل لغوي للفظ الواحد، ولا توجد قرينة ترجح تحليلاً على آخر؛ فليجأ البرنامج لاختيار التحليل الأول في القائمة. وغني عن الذكر أن مثل هذه القوائم اللغوية تحتاج إلى مراجعة دائمة وتعديل؛ نظراً لأن لفظاً قد يشيع استعماله في فترة ما لسبب، ثم يتراجع استعماله مفسحاً المجال للفظ آخر، وهكذا دواليك.
- برامج التشكيل الآلي والتصحيح بها نقص واضح في تحديد أنماط الجملة العربية.
- الجمل القصيرة يسهل - غالباً - معالجتها حاسوبياً، وتزداد الصعوبة بطولها؛ لما قد يؤدي إليه ذلك من عدم إدراج نمط الجملة في البرنامج المعالج، أو احتوائها على قاعدة لغوية غير مدرجة فيه...إلخ.
- يحسن ألا يقوم برنامج التشكيل والتصحيح الآلي بتصحيح أخطاء النص المدخل، كما يفعل برنامج الحركات؛ لأن ذلك قد يورطه في افتراض صور لغوية غير مقصودة، ولأن الأصل أن تكون الجمل المدخلة صحيحة.
- برامج التشكيل الآلي والتصحيح لا تولي اهتماماً كبيراً لمعالجة اللغة التراثية؛ ولذا لم تتعرف بعض ألفاظها.
- يؤكد البحث على ما ثبت من أهمية دور برامج التشكيل الآلي والتصحيح، خاصة مع الطلب المتزايد والاحتياج الملح لتشكيل كثير من الكلمات للضعف اللغوي العام المنتشر بين الناطقين بالعربية.

4- الخلاصة.

تبنى برامج التشكيل الآلي والتصحيح على أسس حاسوبية ولغوية، وقد حاول هذا البحث أن يقف على بعض هذه الأسس المهمة، ويحاول أن يرى مدى تحققها في بعض هذه البرامج من منظور لغوي حاسوبي. وفي سبيل ذلك قام البحث بإدخال عدد كبير من الجمل العربية إلى البرنامجين مجال تطبيق البحث، واتضح له كثير من وجوه القصور، وحاول أن يقترح بعض الحلول لتداركها.

قائمة المصادر والمراجع

[1] نص البحث على الفترة التي أجري فيها التطبيق على البرنامجين لأنه قد يحدث تطوير لهما؛ فتختلف النتائج.

[2] العرب وعصر المعلومات، د. نبيل علي، عالم المعرفة، العدد 184، الكويت، 1994م:ص 371

- [3] العرب وعصر المعلومات، د. نبيل علي، عالم المعرفة، العدد 184، الكويت، 1994م:ص 371
- [4] عناصر النظرية النحوية، د. سعيد حسن بحيري، مكتبة الأنجلو المصرية، الطبعة الأولى، 1410هـ/1989م:ص 34
- [5] قواعد البيانات الرقمية وأهميتها في محركات البحث، محمد محمود زين الدين، مجلة المعلوماتية، العدد 29، صفر 1431هـ.
- [6] نعوم شومسكي، جوانب من نظرية النحو، ترجمة: مرتضى جواد باقر، البصرة، 1985م : 28
- [7] عناصر النظرية النحوية، د. سعيد حسن بحيري، مكتبة الأنجلو المصرية، الطبعة الأولى، 1410هـ/1989م:ص 34
- [8] المعجم الوسيط، مجمع اللغة العربية، ط4، مكتبة الشروق الدولية، القاهرة، 1425هـ / 2004م: درس
- [9] المعجم الوسيط، مجمع اللغة العربية، ط4، مكتبة الشروق الدولية، القاهرة، 1425هـ / 2004م: درس
- [10] العرب وعصر المعلومات، د. نبيل علي، عالم المعرفة، العدد 184، الكويت، 1994م:ص 354 ، 355
- [11] المعجم الوسيط، مجمع اللغة العربية، ط4، مكتبة الشروق الدولية، القاهرة، 1425هـ / 2004م: بدل
- [12] حاشية الصبان على (شرح الأشموني): محمد بن علي الصبان، دار إحياء الكتب العربية: 3 / 123
- [13] مباحث تأسيسية في اللسانيات، عبدالسلام المسدي، مؤسسة عبدالكريم للنشر والتوزيع، تونس، 1997 م : ص 73
- [14] كيفية تعريب السوابق واللاحق في اللغة العربية، التهامي الراجي الهاشمي، مجلة اللسان العربي، الرباط، عدد 21:ص 27
- [15] دور الكلمة في اللغة، ستيفن أولمان، ترجمة: كمال بشر، مكتبة الشباب للطباعة والنشر، القاهرة، ط1، 1975م: ص 69
- [16] المعجم الوسيط، مجمع اللغة العربية، ط4، مكتبة الشروق الدولية، القاهرة، 1425هـ / 2004م: سبق
- [17] المصطلحات العلمية في اللغة العربية في القديم والحديث، مصطفى الشهابي، دار إيزيس للطبع والنشر والتوزيع، ط1: 12
- [18] Semantics, second edition , F.r Palmer ,Combridge University, 1981, p. 22/
- [19] Semantics, second edition , F.r Palmer ,Combridge University, 1981, p. 22
- [20] Semantics, second edition , F.r Palmer ,Combridge University, 1981, p. 22
- [21] Models of Natural Language Processing, www.ai.mit.edu/
- [22] الألسنية التوليدية والتحويلية وقواعد اللغة العربية (الجملة البسيطة)، د. ميشال زكريا، المؤسسة الجامعية للدراسات والنشر والتوزيع، ط2، 1406هـ / 1986م: 63

السيرة الذاتية المختصرة:

مدحت يوسف السبع



- يعمل أستاذا مشاركا في جامعة شقراء، وباحثا في المجلس الأعلى للشؤون الإسلامية.
- حاصل على الدكتوراه من كلية دار العلوم، جامعة القاهرة، 2004م، بمرتبة الشرف الأولى، بإشراف الدكتور نبيل علي، والأستاذ الدكتور علي أبو المكارم.
- شارك ببحوث علمية في مؤتمرات بعدد من الجامعات العربية والأمريكية والجمعيات اللغوية المختصة.
- له عدد (8) أبحاث لغوية محكمة ومنشورة في مصر وخارجها.
- شارك من خلال العمل في شركة (صخر) وغيرها من الشركات المتخصصة في عدد من الأعمال في مجال حوسبة اللغة العربية، وهي: المحلل النحوي، المحلل الصرفي، إعداد معجم حاسوبي، المكنز العربي، المصحح الآلي، والمشكل الآلي، وغيرها.

- نشرت له دور نشر كبرى، ومنها دار المعارف؛ حيث نشرت له كتاب: مع القرآن الكريم في دراسة مستلهمة: ج2، سلسلة "اقرأ"، 2009م. وطبع له كتاب ضمن منشورات: (مهرجان القراءة للجميع) بوزارة الثقافة بمصر.

Some Configuration and Automatic Correction Programs: Evaluation Research

Medhat Yousef Al-Sabaa

College of Education- Shaqra University - Saudi Arabia

myalsabaa@su.edu.sa

Abstract: The subject of this research is to evaluate the performance of some programs of automatic configuration and correction by discussing some of the scientific basics - computational and linguistic - on which it is based, from computer language perspective. The research conducted its applications on the two programs (Multillect) and (mishkal). The research selected these two application programs; both represent a specific trend in automated configuration and correction, given their popularity in this field. The research consists of an introduction, the definition of the problem of research, the two programs of application, and the important computer and linguistic basics underlying the automated configuration and correction programs, This is followed by evaluating the performance of some automated configuration and correction programs, and discuss some of the scientific basics on which it is based, then produce the search results, and its conclusion.

Key words: Computer system for automated processing, Configuration and automatic correction programs, Treatment of linguistic methods, Sort language lists.

الإنسان والآلة فى ترجمة النصوص الأدبية

هل يمكن الاستغناء عن الإنسان فى الترجمة ؟

أسماء جعفر عبد الرسول
قسم اللغة الفرنسية، كلية الآداب، جامعة المنوفية
gmasmaa@yahoo.com

المخلص:

عند ترجمة النصوص الأدبية يحتاج المترجم لمهارات عدة من بينها الثقافة والحصيلة المعرفية وليست اللغوية فقط. سوف نستعرض فى هذا البحث هل هذه المهارات تتوفر فى مواقع الترجمة والقواميس الإلكترونية كجوجل ترجمة وقاموس REVERSO . وسينصب اهتمامنا على النواحي الأسلوبية التى تشكل فارقاً كبيراً من حيث إبداع المترجم وفهمه للنص فهماً صحيحاً. وسوف نطرح بعض الأمثلة للمقارنة بين الترجمات المختلفة. الكلمات المفتاحية : النصوص الأدبية، القواميس الإلكترونية، الأسلوبية، الحصيلة المعرفية واللغوية، إبداع المترجم.

المقدمة

سوف نتناول فى هذا البحث نصاً أدبياً مترجماً فى النصف الثانى من القرن التاسع عشر [1] ومقارنته مع الترجمات الإلكترونية [2]، مما يعنى عدم تواجد التكنولوجيا وخدمات المواقع والقواميس الإلكترونية التى تترجم النصوص. وكان هذا مقصوداً من باب دراسة الإشكالية الآتية : هل تستطيع الآلة الإلكترونية أن تحل محل الإنسان يوماً ما بخصوص عملية الترجمة، وخاصة أن الترجمة الإلكترونية تحتل مكاناً كبيراً حالياً أم أن عملية الترجمة تتطلب مهارات تكمن فى عقلية وشخصية وثقافة المترجم. هذا ما سنحاول الإجابة عليه من خلال هذا البحث. ولقد اخترنا بعض الأمثلة الخاصة بالأسلوبية لأنه من خلال هذا الأخير نستطيع الحكم على المترجم من حيث حصيلته المعرفية واللغوية معاً.

أولاً : الصور البلاغية المرتبطة بالمعنى

(1) التشبيه (La Comparaison)

لا نستطيع إنكار الدور التفسيري الذى يلعبه التشبيه : فإنه يفسح المجال لاكتشاف عالمين مختلفين حقيقى وخيالى. ومن جانب آخر، فالتشبيه يساعد على تجنب الغموض الذى قد يحتويه النص بين هذين العالمين السابق ذكرهما، مما يتطلب وجود موهبة إبداعية من قبل المترجم. ولقد فرق بورجواز فى قاموسه [3] بين التشبيه والاستعارة. بالنسبة للأولى، فإنه ذكر أنه هناك تشبيه مباشر وآخر مجازى. وقد عرف التشبيه المباشر أنه لا يوجد

فيه مشاكل مقارنة بالتشبيه المجازي حيث أن عنصرى التشبيه (المشبه والمشبّه به) لا ينتمون إلى حقل معجمى واحد والصورة البلاغية التى تنتج حاضرة.

Ex1 : «Vous voyez que gloire et fortune tombent sur moi dru comme grêle».

SUE (Eugène), *L'Orgueil*, p. 6.

«اترين صيب النعم علىّ من كل الجوانب». ديمترى أفندى خلاط، عزّة النفس، الاهرام، عدد 1322، 1882.

«ترى أن المجد والثروة يقعان علىّ مثل البرد». (جوجل ترجمة).

«ترى المجد والثروة علىّ ترحيب لى، الكثيف».

(REVERSO)

نلاحظ فى هذا المثال أنه هناك ثلاث ترجمات مختلفة لنفس المثال. فنجد فى ترجمة جوجل أنها ترجمة تقترب من الترجمة الصحيحة لأنه يفتصها مهارة المترجم ومهاراته فنجد أنه ترجم «*tomber sur moi*» حرفياً «يقعان علىّ»، وترجم «*grêle*» بصفة من صفاتها وهى «البرد» ولم يترجمها «بالثلج». أما بالنسبة لترجمة REVERSO فإنها ترجمة غير صحيحة تماماً وسيئة للغاية إذ أنها تحتوى على كلمات موجودة فى القواميس العربية ولكن لا تعطى معنى.

أما بالنسبة لترجمة المترجم الذى ينتمى للقرن التاسع عشر فإننا نلاحظ إبداعه الحقيقى فى هذه الترجمة. لقد ترجم الاثنى معاً «*gloire et fortune*» بكلمة واحدة وهى «النعم». وهنا نجده لجأ إلى حقل معجمى شاسع عند ترجمتهما لأنه تكلم عن كل النعم بدون تحديد النعم التى ذُكرت فى المثال فقط. أما بخصوص ترجمة «*tomber sur moi*» فإنه نقلها «بالصيب». وهنا نستطيع القول بأن المترجم استطاع أن يفهم المعنى الموافق للسياق لأن هذه الأخيرة تعنى مطراً شديداً الانصباب [4]. وبالتالي فإن المترجم اختار صورة موافقة للثقافة الهدف وهى الثقافة الإسلامية والعربية لأن المطر يعنى فى هذه الثقافة ذروة الخير. ونجد أن [5] Seleskovitch et Lederer يؤيدون هذا القول الأخير بأن المعنى يتشكل باستمرار على مدار قراءتنا للنص.

Ex2 : «Mais, une fois que les mères veulent quelque chose... dans l'intérêt de leurs fils... elles deviennent des lionnes, des tigresses...». SUE (Eugène), *L'Orgueil*, p. 39.

«وانا موقن ان سعى الام بصالح ابنها اشبه بسعى اللبوة لخير اشبالها». ديمترى أفندى خلاط، عزّة النفس،

الاهرام، عدد 1186، 1881.

« ولكن بمجرد أن تريد الأمهات شيئاً ... لصالح أبنائهن ... يصبحون اللبوات والنجوم ...». جوجل ترجمة.
« ولكن عندما الامهات يريدون مصلحة ابنائهم (التصريف) تصبح *tigresses, lionesses* ...».

REVERSO

نرى في ترجمة هذا المثال تفوق جوجل ترجمة على REVERSO أيضاً : فقد قام جوجل بترجمة الجملة حرفياً ولكن صحيحة فيما عدا نقطتين أساسيتين : أولاً، الفعل «يصبحون»، فإنه حسب قواعد العربية، الجملة تتحدث عن الأمهات أى عن كائن مؤنث، إذاً فالترجمة الصحيحة «يصبحن». ثانياً، قام جوجل بترجمة «*tigresses*» «بالنجوم» وهذا معنى خاطيء تماماً وهو أن هذا الموقع لم يستدل على المقابل الرئيسى لهذه الكلمة وهى موجودة فى كافة المعاجم بمعنى أنثى النمر. أما بالنسبة لترجمة REVERSO جاء الشق الأول من الجملة فقط صحيحاً أما بقية الجملة فقد نقل نفس الكلمات الواردة فى النص الأصيل. وعندما نقارن هاتان الترجمتين بترجمة المترجم التى بين أيدينا، سنجد أن المترجم لم يترجم الجملة حرفياً ولكنه تصرف فى ترجمتها بما لا يخالف الإخلال بالمعنى الأصيل أو الخروج عن سياق النص. فقد قام بترجمة «*lionnes*» فقط «باللبوة» ولم يترجم الجزء الثانى من الجملة، فإن المترجم قد نقل الصورة الخيالية الموجودة فى الجملة المصدر لعلمه أن الأسد أشد قوة من النمر.

(2) الكناية (La Métonymie)

الكناية تلعب دوراً هاماً فى التقارب بين عنصرين ينتمون إلى حقل معجمى مشترك.

Ex1 : «*Tout Paris sera là, on s'arracha les billets de tribune... car, lorsque M. de Mornand parle... c'est un événement*». SUE (Eugène), *L'Orgueil*, p. 42.

«فالناس تزدهم على التقاط اوراق المجلس لورد منهل علمه والمنهل العذب كثير الزحام». ديمترى أفندى
خلاط، عزة النفس، الاهرام، عدد 1881، 1190.

« جميع باريس ستكون هناك ، ونحن ننتزع تذاكر المنبر ... لأنه عندما يتحدث السيد مورناند ... إنه حدث».
جوجل ترجمة.

«ستكون باريس هناك قد اصاحب (استخراجها) الفواتير (تذاكر) ومنتدى (حامل) لانه عندما Mornand

mister دى تكون حدثاً». REVERSO.

سنجد نفس الحال بالنسبة لترجمة REVERSO فإن الترجمة عبارة عن كلمات عربية ولكن لا تعطى معنى واضحاً ولكن بالنسبة لجوجل فإن الترجمة صحيحة إلا أنه لم يفهم المعنى المجازى لكلمة «Paris» بأنها كناية

عن «الناس». بالنسبة للمترجم، فإنه استطاع وفقاً لمهاراته اللغوية أن يترجمها بمعناها المجازى وأدرك أن باريس هي الحاوى الذى يحتوى على المحتوى.

(3) المجاز المرسل (La Synecdoque)

يمثل المجاز المرسل ليس فقط علاقة تقارب بين عنصرين ولكن علاقة اندماج لكليهما البعض.

Ex1 : « En face de la cheminée, on voyait le piano d’Herminie, son gagne-pain». SUE (Eugène), *L’Orgueil*, p. 59.

«وكان البيانو موضوعاً بازاء المصطفى». ديمترى أفندى خلاط، عزة النفس، الاهرام، عدد1205، 1881. « أمام الموقد ، يمكنك أن ترى بيان Herminie ، مصدر رزقها». جوجل ترجمة.

« امام الموقد, راينا من البيانو Herminie, (فى بلدها)(الخبز الوظائف)». REVERSO.

عند المقارنة بين الترجمات الثلاث، لفت انتباهنا قوة ترجمة جوجل والذى فهم المجاز المرسل فى الكلمة الفرنسية وترجمها كما هى، على الرغم من ترجمة REVERSO الذى مازال يفاجئنا بالترجمة السيئة. أما عند ملاحظة ترجمة المترجم، فإنه لم ينقل التعبير المصدر إلى اللغة الهدف وهذا يعتبر اعتداء على النص الأسمى لأن فقر أرمينيى يلعب دوراً محورياً على مدار الرواية. وبالتالي، نجد أن الحذف هنا يعكس انقطاعاً مع الموضوع الأساسى للرواية.

ثانيا : الصور البلاغية المرتبطة بالفكر

(1) التورية (La Périphrase)

الهدف من التورية هو الإشارة إلى الشيء ولكن بطريقة غير مباشرة.

Ex1 : «Je tâcherai de repêcher le godelureau ; sinon je mettrai ses Louis au tronc des pauvres... ». SUE (Eugène), *L’Orgueil*, p. 65.

«والان افتش على الشاب وارد له نقوده والا اضعها وديعة فى صندوق البر...». ديمترى أفندى خلاط، عزة النفس، الاهرام، عدد1211، 1881.

« سأحاول التقاط godelureau ؛ وإلا سأضع له لويس على جذع الفقراء». جوجل ترجمة.

« ساحاول ان الاسماك تصل السيدات رجل والا اطرح له (فى بلدها)(لويس فى صندوق الفقراء) REVERSO

من الملاحظ في هذا المثال أن ترجمة جوجل غير صحيحة تماماً فهي ترجمة حرفية صرفة ولكن نلاحظ أن الترجمة الأخرى قد ترجمت التعبير المقصود كما هو من غير أخطاء إلا أن الجزء الأول من الجملة غير صحيح تماماً. أما بالنسبة للمترجم محل الدراسة فإننا نجد لهجاً إلى الترجمة الحرة عند نقل هذا التعبير إلى اللغة الهدف لأنه أدرك معناه المجازي لأن التورية تمهد الطريق للإيحاءات والكلام المجازي الغير مباشر. ومن جانب آخر، فإن التورية تمهد الطريق للمنهج التفاعلي الذي يؤدي بدوره إلى فهم الانعطافات اللغوية في النص.

(2) التضاد (L'Oxymore)

سوف نوضح فكرة التضاد من خلال المثال التالي :

Ex1 : «Ernestine, de son côté, resta quelques moments silencieuse, pensive, pour deux motifs ; elle était rêveuse, d'abord parce qu'elle se rappelait les regards singuliers qu'Olivier avait jetés sur elle en apprenant qu'il était officier... regards... dont Ernestine croyait comprendre la touchante et généreuse signification, puis la jeune fille ressentait un_mélancolique Bonheur en songeant que sa nouvelle amie était la jeune artiste que l'on avait appelée auprès de madame de Beaumesnil pendant ses derniers moments». SUE (Eugène), *L'Orgueil*, p. 96.

«وانشغل فكر ارنسة بما لقيته في عيون اوليفيه من لواظ الانعطاف والحب لما بلغه خبر ترقيه عدا عن حيرتها بكيفية سياق الحديث مع ارمنا لتصل الى النتيجة المرغوبة " عن اجتماعها واغتائها بالاميرة المرحومة قبل وفاتها" ولهذين السببين بقيت صامته كارمنا». ديمتري أفندي خلاط، عزة النفس، الاهرام، عدد1271، 1881.

«من جانبها ، بقيت إرنستين صامته لبضع لحظات ، بهدوء ، لسببين. كانت حاملة ، أولاً وقبل كل شيء لأنها تذكرت النظرات الفردية التي ألقاها عليها أوليفيه عندما سمعت أنه ضابط ... يبدو أن إرنستين اعتقدت أنها تفهم معنى اللمس والسخاء ، ثم شعرت الفتاة بأنها حزينة. السعادة في التفكير في أن صديقها الجديد كان الفنان الشاب الذي تم استدعاؤه إلى Madame de Beaumesnil خلال لحظاتها الأخيرة». جوجل ترجمة.

« Ernestine, » من الجزء (على الجانب الايمن), لا تزال بعض لحظات صمت مكتنبا وهو يجيب, تدفعان(); كان حالم, اولا بسبب تتذكر نفسه بصيغة المفرد المظهر ما القى (لقى اوليفر) من قبل علمه بانته يتراس جلسة (وركه تبدو التي تتحرك وكريم Ernestine يفهم معنى, ثم شعرت الفتاة افكاره السوداوية احد السعادة بتذكر

ان صديقنا الجديد الفنان الشاب الذى دعى Beaumesnil سيدتى من خلال جولتها فى ((اللحظات الاخيرة)).

REVERSO

سنبدأ فى هذا المثال بالتعليق على ترجمة REVERSO والتي نرى أنها غير ملائمة تماما ولا يمكن أن نطلق عليها ترجمة من الأساس. ولكن نعود إلى ترجمة جوجل والتي مازالت تبهنا فهى ترجمة فيها من الذوق الجمالى ما لا يختلف عليه اثنين ما عدا التعبير الذى أريد التركيز عليه. فهذا التعبير «un mélancolique» «Bonheur» يعد كلمتين متضادتان متجاورتان ولا يفصل بينهما شيئاً وإذا سلطنا الضوء على ترجمة ديمترى أفندى سنجد أنه قام بحذف ترجمة هذا التعبير، مما انعكس بالسلب على معنى هذه الفقرة لأن هذا التعبير يعنى المشاعر التي شعرت بها أرنستين عندما علمت أن صديقتها هى نفسها من كانت تخفف من آلام والدتها بالموسيقى.

ومن ثم نستطيع القول أن هذا النوع من التضاد يكمل المعنى ويوضح حالة مشاعرها المضطربة عندما تذكرت أرمينى وأوليفيه.

وفى هذا الصدد، سنجد أن كارس [6] Karras تذكر بأن للحذف تأثيراً أدبياً ولغوياً.

ومن هنا فإننا نقترح ترجمة أخرى :

«من جانبها، ظلت أرنسة لبعض لحظات صامته وشاردة الفكر لسببين : أولهما أنها كانت حاملة فى بادىء الأمر لأنها كانت تتذكر النظرات الغريبة التي كان أوليفيه يوجهها ناحيتها وهى تعلم بترقيته ضابطاً وكانت تعتقد أنها تفهم مغزى هذه النظرات المؤثرة والقوية. وكانت تشعر بسعادة يغمرها الحزن عندما تتذكر أن صديقتها الجديدة كانت هى القينة [7]الشابة التي كانت قد استدعتها مدام بومسنيل لديها قبل أن تلفظ أنفاسها الأخيرة».

الخاتمة

نستخلص من هذا البحث أن ترجمة المواقع الإلكترونية لا تستطيع محو دور المترجم على الرغم من أن المترجم يقع هو الآخر فى أخطاء ولكنها ليست كما هو الحال بالنسبة لهذه المواقع لأنها تحتوى على بعض الأخطاء اللغوية كما اتضح لنا إلى جانب تلك المعرفية والثقافية. لا نستطيع الجزم بأن كافة المواقع الإلكترونية غير صالحة للترجمة وأن تلك التي تجيد الترجمة تحتاج أيضاً لمهارات المترجم لكي تكتمل. وهناك ملحوظة أخرى أكثر أهمية وتكمن فى أنه على الرغم من التفاوت الزمنى بين الترجمات محل الدراسة، إلا أننا لاحظنا مهارة المترجم وحسن استخدامه لذكائه المهني فى ظل عدم توافر الثورة التكنولوجية التي يشهدها عصرنا فى حقبته التاريخية. ونود أن نشير أيضاً بأن غياب التطور التكنولوجي الذى نشهده حالياً خلال القرن التاسع عشر كان له تأثيراً على كتابة بعض الحروف مثل الهمزة نظراً لعدم تطور آلات الطباعة آنذاك.

المراجع

[1] SUE (Eugène), (1865), *L'Orgueil*, Le Siècle, Paris.

(خلاط) ديمترى أفندى، *عزة النفس*، الأهرام، 1881، 1882.

[2] <https://translate.google.com/?hl=ar&tab=wT>

<https://dictionnaire.reverso.net/francais-arabe/>

[3] POUJEOISE (Michel), (2006), *Dictionnaire de poésie*, Belin, pp. 108-306.

[4] www.almaany.com/ar/dict/ar-ar/الصيب/ (تم الاطلاع في 1-5-2013)

«صَيْبٌ : مطر شديد الانصباب».

[5] SELESKOVITCH (Danica), LEDERER (Marianne), (1983), *Interpréter pour traduire*,

3ème édition, revue et corrigée, publications de la Sorbonne, Littératures I 10, Didier érudition, Collection “ Traductologie 1”, Paris, p. 18.

[6] KARAS (Hilla), (2007), «Le Statut de la traduction dans les éditions bilingues, de l'interprétation au commentaire» in *Palimpsestes*, N. 20, pp. 137-160.

[7] قينة- أى «مغنية فى الاصل العبرانى معناها الرائية».

العيسى (طوبيا)، (1965، 1964)، *تفسير الألفاظ الدخيلة فى اللغة العربية مع نكر أصلها بحروفه*، دار العرب للبيئتانى، ، ص 60. ولقد أوضح «لسان العرب» بأن الليث ذكر بأن «القينة» هى : «قال الليث عوام الناس يقولون القينة المغنية».

In <http://www.maajim.com/dictionary/قينة/> (تم الاطلاع في 10-5-2013).

السيرة الذاتية :



أسماء جعفر عبد الرسول

حاصلة على ليسانس آداب وتربية عام 2007 ثم على ليسانس آداب عام 2009. حصلت على الماجستير فى الترجمة وتعمل حالياً مدرس مساعد بكلية الآداب، جامعة المنوفية. لقد شاركت بالبحث المعنون «الاختلاف الثقافى» فى المؤتمر الدولى للترجمة الذى أقيم فى المجلس الأعلى للثقافة والمجلس القومى للترجمة فى نوفمبر 2016. ومن جانب آخر، شاركت بالبحث المعنون «حركة الترجمة وتأثيرها على الأدب العربى» فى ملتقى العلاقات الثقافية الفرنسية- المصرية والذى أقيم فى المجلس الأعلى للثقافة يومى 21 و22 مايو 2017. ولها بحث منشور فى مجلة كلية الآداب، جامعة المنوفية، تحت عنوان «إعادة قراءة الروايات الفرنسية المترجمة إلى العربية فى الصحافة المصرية فى الفترة من 1881 حتى 1893. دراسة فى ترجمة الثقافة». وشاركت فى المؤتمر الدولى الثانى للغات الأوروبية بكلية الآداب، جامعة المنوفية المنعقد فى الفترة من 3 إلى 5 ديسمبر 2017، بالبحث المعنون «المترجم والوساطة الثقافية». وشاركت أيضاً فى المؤتمر السابع عشر لهندسة اللغة والذى أقيم فى جامعة مصر الدولية يومى السادس والسابع من ديسمبر 2017 بالبحث المعنون «دور السياق فى صياغة المعنى فى الترجمة». وشاركت فى المؤتمر الدولى الثانى عن التراث العربى والإسلامى، الذى أقيم بمعهد المخطوطات العربية يومى 21 و22 فبراير 2018، بالبحث الموسوم «حركة الترجمة فى جريدة الأهرام فى الفترة من 1881 حتى 1893». وشاركت فى مؤتمر «اتجاهات معاصرة فى دراسات المستعربين» والذى أقيم بكلية الآداب، جامعة القاهرة فى الفترة من 3 إلى 5 أبريل 2018 بالبحث المعنون «التباعد الزمنى والترجمة». عضوة فى عدة نشاطات تابعة للجمعية المصرية لأساتذة اللغة الفرنسية، وقد حضرت

الكثير من المؤتمرات والندوات واللقاءات على هامش هذه الجمعية. وقد حصلت على شهادة تفيد بإجادتها للمستوى اللغوى Delf B2 من وزارة التربية والتعليم الفرنسية.

The Translator and Machine-Assisted Translation on Translating the Literary Texts

Can the Machine-Assisted Translation Replace the Human Role?

Asmaa Gaafar Abdel-Rassoul

Faculty of Arts, Menoufia University

gmasmaa@yahoo.com

Abstract: In the translation of literary texts, translator needs several skills including cognitive and linguistic culture toll not only. We will review in this research do these skills available in translation sites electronic dictionaries as GOOGLE Translate and dictionary REVERSO. Our attention will focus on the stylistic aspects that constitute a significant difference in terms of the creativity of the translator and his understanding of the proper understanding of the text. They will ask some examples of comparison between different translations.

Keywords: literary texts, electronic dictionaries, cognitive and linguistic stylistic outcome, the creativity of the interpreter.

المعجم الورقي والمعجم الآلي-دراسة لسانية مقارنة-

فطوم القريش

عضوة مع فريق ابتكارات بالمدرسة المحمدية للمهندسين بالرباط
عضوة مع جمعية هندسة اللغة العربية (المغرب)
Fettoum.krieche@gmail.com

ملخص:

تكشف هذه الدراسة المقارنة الاختلاف والتباين الحاصل بين المعجم الورقية والمعجم الآلية على مستوى المادة العلمية وعلى مستوى التداول، كما هو معلوم فإن كل مفردات اللغات الطبيعية تخزن في معجم عامة أو معجم خاصة، فمنذ القدم واللغويون يهتمون بدراسة اللغة من خلال جمعها من أفواه متكلميها وتصنيفها في كتب لغوية ورقية، إلا أنه مع تطور الوسائل التكنولوجية وأنظمة معالجة اللغات الطبيعية تغيرت مناهج وطرق دراسة اللغة وصارت تخزن مفرداتها في معجم آلية بعدما كانت تخزن في معجم ورقية.

الكلمات المفتاح: المعجم الورقية- المعجم الآلية- المعالجة الآلية- قاعدة بيانات- التصنيف - التشفير.

مقدمة:

يعتبر المعجم من المؤلفات العلمية المهمة في حياة الباحث باعتباره يمثل جزءا من معرفة المتكلم بلغته، وقد حظي هذا المستوى من مستويات الدرس اللساني بأهمية كبرى في جل النظريات اللغوية، منذ أن ظهرت فكرة جمع اللغة من أفواه متكلميها في شبه الجزيرة العربية، وقد تنوعت المعجم بتنوع حاجة الناس إليها؛ فبعدما كانت الغاية من تأليف المعجم حفظ اللغة العربية وتدوينها أولا في رسائل ثم في معجم، أصبح الهدف من تأليفها في العصور الحالية، ومع تطور التكنولوجيا، تبسيط تقنيات استعمال المعجم والاستفادة منه أكثر في وقت وجيز فيما عرف بالمعالجة الآلية للغات الطبيعية "Le Traitement Automatique des Langues Naturelles" أو ما يعرف باللسانيات الحاسوبية "La Linguistique Calculatoire"، والتي تهتم بإنجاز وتطوير أنظمة آلية لمعالجة اللغات الطبيعية.

وبما أن هناك فوارق بين المعجم الورقي والمعجم الآلي، فقد ارتأينا في هذا المقال أن نعقد مقارنة بينهما؛ من خلال التطرق إلى أوجه الاختلاف والانتلاف بين المعجم الورقي والمعجم الآلي على مستوى المادة العلمية وعلى مستوى التداول أيضا.

أولا: المعجم الورقي: تعريفه وخصائصه

1- تعريف المعجم الورقي

نقصد بالمعجم الورقي كل كتاب ورقي يحمل بين دفتيه مجموعة من المفردات اللغوية التي تنتمي للنسق اللغوي العربي والتي يمتلكها متكلم هذه اللغة، وقد عرفه الدكتور سالم الرامي بأنه: "مخزون المفردات الذي يمثل

جزءاً من قدرة المتكلم والمستمع"^[1]. مثل: معجم العين للخليل بن أحمد الفراهيدي، وجمهرة اللغة لابن دريد، ولسان العرب لابن منظور،... فما هي مميزات هذا المعجم على مستوى المادة العلمية وعلى مستوى التداول؟

2- خصائص المعجم الورقي

(1) على مستوى المادة العلمية

لقد تبين لنا من خلال تصفحنا لمجموعة من المعاجم اللغوية القديمة أن مؤلفي هذه المعاجم عالجوا مجموعة من المسائل اللغوية أثناء شرحهم للمفردة وهي: مسائل معجمية لغوية، وتتمثل في إيراد المفردات اللغوية التي تكون نظام اللغة العربية، ومسائل صوتية، وتتجسد في تفسير مجموعة من الظواهر الصوتية للمفردة (الإعلال، القلب...)، ومسائل صرفية، وتتجلى في الأوزان والصيغ الصرفية، ومسائل تركيبية وتظهر في إيراد جمل تضم المفردة المدروسة، ثم مسائل دلالية تتمثل في شرح المفردة بآيات قرآنية وأبيات شعرية وأمثلة عربية، وتتم الإشارة أحياناً إلى بعض المسائل التداولية مما مكن صناع المعاجم العربية من المحافظة على الإرث اللغوي باعتماد الآيات القرآنية والأحاديث النبوية والأشعار والأقوال العربية، ونورد هنا مثلاً يعالج المفردة من الناحية الصوتية في معجم العين كالاتي:

"في زكرياء أربع لغات :

زكرياء بالمد، وفي التثنية : زكرياءان، وزكرياوان، وفي الجمع : زكرياءون.

وزكريا، بطرح الهمزة، وفي التثنية : زكريّاءان، وفي الجمع : زكريّيون.

وزكريّ، وفي التثنية: زكريّان، والجمع : زكريّون، مثل: مدني ومدنيان [مدنيون].

وزكري، بطرح الألف، وتخفيف الياء، وفي التثنية زكريان، وفي الجمع زكرون بطرح الياء"^[2].

إن المتأمل في هذا المثال يلاحظ أنه يحتوي على ثلاث ظواهر صوتية هي: المد، والوقف الذي يتجلى في طرح الهمزة في "زكريا"، والإعلال ويتمظهر في إبدال الياء واوا في "زكرون".

ونورد هنا مثلاً يتعلق بمعالجة معجم العين للجانب الصرفي أيضاً كالاتي:

" دَرِي يَدْرِي وَدِرْيَةٌ وَدِرْيَانَا وَدِرْيَانَا، ويقال أتى فلان الأمر من غير درية أي من غير علم، والعرب

ربما حذفوا الياء من قولهم: لا أدري، في موضع لا أدري، يكتفون بالكسرة فيها كقول الله عز وجل:(الليل إذا

يسر) الفجر:4، والأصل يسري"^[3].

¹-سالم الرامي، المعجم العربي العصري وإشكالاته "حوسبة المعجم العربي التقليدي" ص235.

²-الخليل بن أحمد الفراهيدي، المعجم العين، مادة(ز ك ر) ج: 2 ص: 188.

³-معجم العين، مادة (د،ر،ي) ج 2 ص: 64.

يتبين لنا من خلال هذا المثال أن الخليل شرح المادة بإيراد كل الصور الصرفية التي تتشكل فيها وهي: الميزان الصرفي للفعل (دري يدري على وزن فَعِل يفعل) والمصدر (دريّة، دِرْيَا، دِرْيَانَا، وِدْرَايَة) وصورة الفعل في حالة النفي (لا أدري).

كما نقدم هنا مثالا آخرًا يتعلق بالجانب التركيبي:

"عبط: عَبَطَتِ النَّاقَةُ عَبَطًا، وَاعْتَبَطَتْهَا اعْتِبَاطًا إِذَا نَحَرْتَهَا مِنْ غَيْرِ دَاءٍ وَهِيَ سَمِينَةٌ فَتِيَةٌ، وَاعْتَبَطَ فُلَانٌ: مَاتَ فَجَاءَ مِنْ غَيْرِ عِلَّةٍ وَلَا مَرَضٍ وَقَوْلُهُمْ: الرَّجُلُ يَعْبُطُ بِسَيْفِهِ فِي الْحَرْبِ عِبْطًا، اشْتَقَّ مِنْ ذَلِكَ، وَيَعْبُطُ نَفْسَهُ فِي الْحَرْبِ إِذَا

أَلْقَاهَا فِيهَا، غَيْرُ مَكْرَهٍ (.....)

واحد العُبطُ: عبيط، والرجل يعبط الأرض عبطًا، ويعتبطها إذا حفر مرضعا لم يحفره قبل ذلك، وكل مبتدأ من حفر أو نحر أو ذبح أو جرح فهو عبيط (...). ومات فلان عبطًا، أي شابا صحيحا (.....). واعتبطه الموت. ولحم عبيط: طري وكذلك دم عبيط وزعفران عبيط شبيه بالدم، بيّن العبط، وعبطته الدواهي، أي نالته من غير استحقاق بذلك (.....).

والعبيطة: الشاة أو الناقة المُعْتَبَطَةُ ، ويجمع عبائط (.....)[4].

لقد بين المؤلف في المثال أعلاه كل التراكيب التي يمكن أن ترد فيما مادة (ع، ب، ط) وهي كما يلي: فبالنسبة للفعل الذي يصاغ من هذه المادة تارة يكون لازما في التركيب الآتي: اعتبط فلان وتارة أخرى متعد في: عبطت الناقة عبطا- اعتبطتها اعتباطا- يعبط نفسه في الحرب- الرجل يعبط الأرض عبطا- يعتبطها- اعتبطه الموت- عبطته الدواهي.

أما الاسم فتارة يكون حالا في: مات فلان عبطة وتارة نعتا في: لحم عبيط- دم عبيط- زعفران عبيط. وقد تعرض أصحاب المعاجم الورقية في مؤلفاتهم المعجمية لمعالجة الجانب التداولي ونقدم أسفله مثلا في ذلك:

"رثأت اللبن أرثؤه رثاء، إذا حلبت حلبيا على حامض.

والرثينة: اللبن الخائر

وأهل اليمن يقولون: رثأت الميت في معنى رثيته"[5].

وللدلالة على إحاطة المعاجم العربية بكل المعلومات اللغوية (الصوتية، والصرفية، والتركيبية...) نقدم مثالا

آخر من معجم جمهرة اللغة في جذر (م، ط، ر):

"والمطر: معروف؛ مَطَرَتِ السَّمَاءُ تَمَطَّرُ مَطَرًا، وَرَبِمَا قَالُوا: مَطَرًا، فَجَعَلُوهُ مَصْدَرًا.

4- معجم العين، مادة (ع، ب، ط) ج 3 ص ص: 86 87 .

5- ابن دريد، جمهرة اللغة، مادة (ر، ث، أ) ج 2 ص: 564.

وأمرت السماء لغة فصيحة لم يتكلم فيها الأصمعي لأنه جاء في القران: ﴿عَارِضٌ مُّمْطِرُنَا﴾[الأحقاف: الآية 24] و(وأمطرنا عليهم)[الأعراف: الآية 84؛ والحجر: الآية 74؛ والشعراء: الآية: 173؛ والنمل: الآية 58]. وأرض مطيرة وممطرة، ويوم ماطر وممطر.

ومر الفرس يمْطِرُ مَطْرًا، إذا عدا عدوا شديدا، وكذلك البعير. قال الراجز:

أما ترى القَرْطِيَّ يَفْرِي مَطْرًا (.....)" [6]

وهنا نلاحظ أن المؤلف قدم كل الصيغ الصرفية والتركيبية التي يمكن أن يرد فيها الجذر المدروس، مع شرح معناه باعتماد آيات قرآنية وبيتا شعريا، وتوجد في هذين المعجمين (العين وجمهرة اللغة) وغيرهما أمثلة أخرى لا تعد ولا تحصى تؤكد ما قلناه.

(2) على مستوى التداول

إن الباحث في أحد المعاجم العربية الورقية تعترضه بعض الصعوبات من قبيل: عليه أن يكون قادرا على التمييز بين معاجم المعاني ومعاجم المصطلحات ومعاجم المفردات ومعاجم الموضوعات...، وأن يكون ملما بطرق ترتيب أهم المعاجم ولا نقول كلها؛ فمثلا معجم العين مرتب ترتيبا صوتيا، بالاعتماد على مخارج الحروف؛ حيث عرض فيه الخليل حروف الهجاء على أعضاء النطق حرفا حرفا، وقاس مدارجها، ورأى أنها تصدر من أعضاء النطق متدرجة من أقصى الحلق نازلة إلى أسفل، انتهاء بالحروف الشفوية، وبذلك أدرك الخليل أن أول حرف ينتجه أقصى الحلق هو حرف العين ولذلك سمي معجمه "بالعين"، ولا يمكن أن يعرف هذا إلا دارس اللغة العربية مما يخلق مشكلا عند الباحث الذي لا يعرف بعض المفاهيم في علم الأصوات، بينما نجد معجم الجمهرة مرتب ترتيبا هجائيا باعتماد أول حرف للمفردة ثم ثاني حرف وهكذا دواليك... إلا أن ابن منظور رتب معجمه على نظام الأبواب والفصول حيث يعالج كل باب حرفا من حروف الهجاء وفقا لآخر جذر الكلمة، ثم يورد في كل باب فصلا لكل حرف وفقا لأوائل جذور الكلمة، وبالتالي فالمعجم الورقي لا يستغل إلا في البحث عن معنى لمفردة فقط، بمعنى لا يكون شاملا لكل الخصوصيات اللغوية المرتبطة بالمفردة (مثل تصريف الفعل المشروح....)

أضف إلى ذلك، فالمستعمل لهذه المعاجم يواجه صعوبة في حملها معه دوما لكثرة أجزاء المعجم الواحد، إلا أن هذا لا ينفي أهمية الاستعانة والاستفادة من المعاجم اللغوية الورقية مما تزخر به من إرث لغوي ضخم.

ثانيا- المعجم الآلي: تعريفه وخصائصه

⁶ - ابن دريد، جمهرة اللغة، مادة (م، ط، ر) ص: 77.

1- تعريف المعجم الآلي

نقصد بالمعجم الآلي كل معجم يمكن أن يتضمن كل المفردات اللغوية التي تنتمي للغة الطبيعية المدروسة لمستوى من مستويات الدرس اللغوي أو أكثر (الصواتة أو الصرف أو التركيب أو الدلالة)، إلا أن المعلومات اللسانية المرتبطة به تخزن بطريقة أتوماتيكية باعتماد لغة الترميز أو التشفير "codage"، لأنه يعتمد على بناء نظام آلي يقوم على قاعدة معطيات تبرمج فيها معلومات لسانية دقيقة وموجزة تخص جانبا من جوانب الدرس اللغوي، باعتماد بنية المفردة لا باعتماد معناها لأن ما يتم حوسبته هو بنية المفردة وشكلها باعتبار أن الحاسوب يخزن المفردة في شكل سلسلة من الأرقام. إذن فما هي خصائص هذا المعجم على مستوى المادة العلمية وعلى مستوى التداول؟

2- خصائص المعجم الآلي

(1) على مستوى المادة العلمية

لا يقوم المعجم الآلي على جمع المادة العلمية من أفواه متكلميها كما هو الشأن بالنسبة للمعجم الورقي، بل يعتمد على مادة لسانية جاهزة مسبقا، وعلى الذكاء الاصطناعي والمنطق والرياضيات؛ لتصير اللغة طيبة لبرامج المعالجة المعلوماتية باعتماد آلية البرمجة أو التخزين (تخزين معطيات لغوية في ذاكرة الحاسوب) حتى تسهل عملية إدخال واسترجاع هذه المعطيات المخزنة.

نقدم هنا كمثال لذلك جزءا من قاعدة البيانات التي قمنا بإنجازها، وهي تخص إنشاء "معجم آلي للفعل في اللغة العربية":

racine	فعل	sense faclala
برقط	2	برقط إذا صد في الجبل فسقط
برقع	5	برقعت الدابة والحارية ألبستهما البرقع
بركع	2	بركع بركمة إذا قام على أربع
بركل	2	بركل الرجل إذا مشى في الطين والماء
برنت	1	rien
برنت	1	rien
برنس	2	برنس الرجل إذا أسرع وتبعثر
برهم	2	برهم الشجر إذا اجتمع ورقه وتفرده
يزرع	1	rien
يزمخ	2	يزمخ الرجل إذا تكبر
يزنت	1	rien
يسلم	2	يسلم الرجل إذا كره وجهه
يسمل	25	يسمل بسملة إذا أكثر من قول بسم الله
يعثر	5	يعثر القبر وغيره إذا بددت ترابه
يعتق	1	rien
يعذر	5	يعذرنى إذا تقضيتى
يعرض	1	rien
يعزج	5	يعزج الشيء إذا فرقه
يعكر	5	يعكره بالسيف إذا ضربه
يعتق	1	rien
يعثر	1	rien
يقط	5	يقط متاعه ويعثره
يكرك	5	يكركت الشيء إذا طرحك بعضه على بعض
يأخر	52	يأخر الرجل إذا حدا من فرع ويأخر الرجل إذا أكل
يأخص	2	يأخص الرجل إذا حدا من فرع
يأجج	1	rien
يأجم	5	يأجم البيطار الحمار إذا شد قوائمه من داء يصيبه
يأجج	2	يأجج الرجل أحيا ويأجج

الجدو

ل(1)

:

لائحة تمثل جزءا من قاعدة بيانات الفعل الثلاثي الصحيح السالم في اللغة العربية الكلاسيكية [7]

نستطيع من خلال هذه الواجهة أن نبحت عن جذر ثلاثي صحيح سالم ينتمي للغة العربية الكلاسيكية لمعرفة خصائصه الصرف - تركيبية ومعناه وتصريفاته المتنوعة. ونقدم مثالا آخرًا للدكتور عبد الرزاق التورابي:

منكلم	مخاطب	مخاطبة	غائب	غائبة
أَسْأَلُ	تَسْأَلُ	تَسْأَلِينَ	يَسْأَلُ	تَسْأَلُ
تَسْأَلُ	تَسْأَلَانِ	تَسْأَلَانِ	تَسْأَلَانِ	تَسْأَلَانِ
تَسْأَلُ	تَسْأَلُونَ	تَسْأَلْنَ	يَسْأَلُونَ	يَسْأَلْنَ

الشكل (1)

واجهة

خاصة

بمولد

صرفي

للفعل

المهموز

الثاني [8]

ح

يث يمكننا

إدخال

الجذع(سأل)أو الجذر(س،أ،ل) ليقدم الحاسوب تصريفا للفعل "سأل" في زمن المضارع.

7 - قاعدة بيانات من إنجاز الأستاذ الدكتور يوسف طاهر والطالبة الباحثة فطوم القريش، بحث لنيل شهادة الماجستير.

8 - عبد الرزاق التورابي وعبد الحميد الجيهاد، المعالجة الآلية للغة العربية، وقائع الندوة الدولية يونيو 2007، المصرف الآلي للأفعال في اللغة العربية، ص: 21.



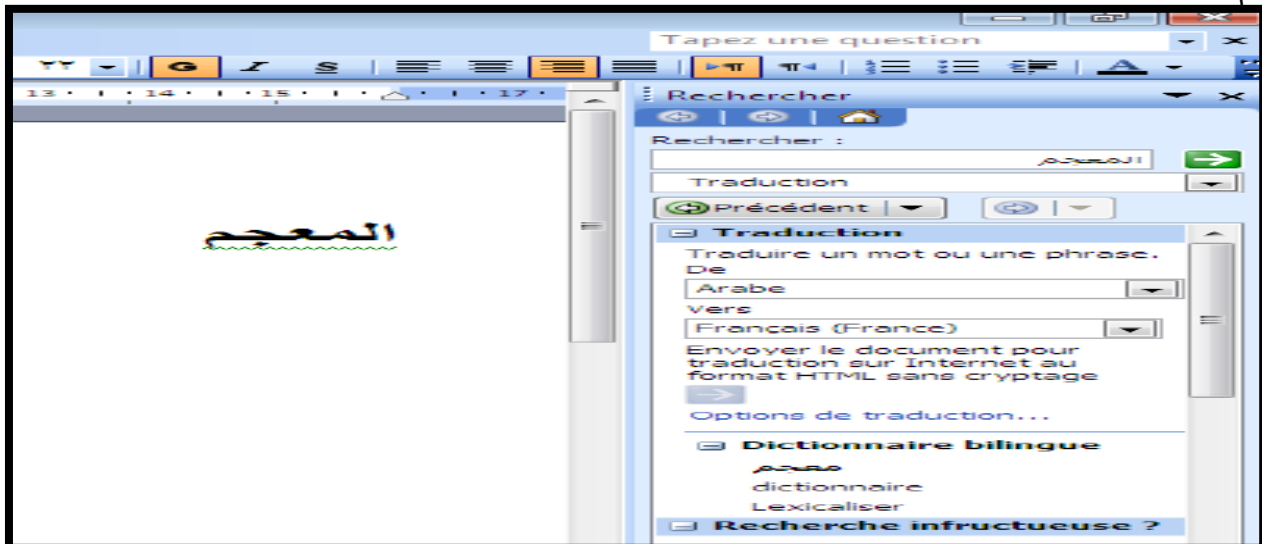
الشكل(2): واجهة تخصص معجم المعاني للغة العربية(10)

وهو معجم إلكتروني يمكن من معرفة معنى الكلمة المراد البحث عنها. وهناك العديد من المعاجم الإلكترونية في مجال القانون والطب....

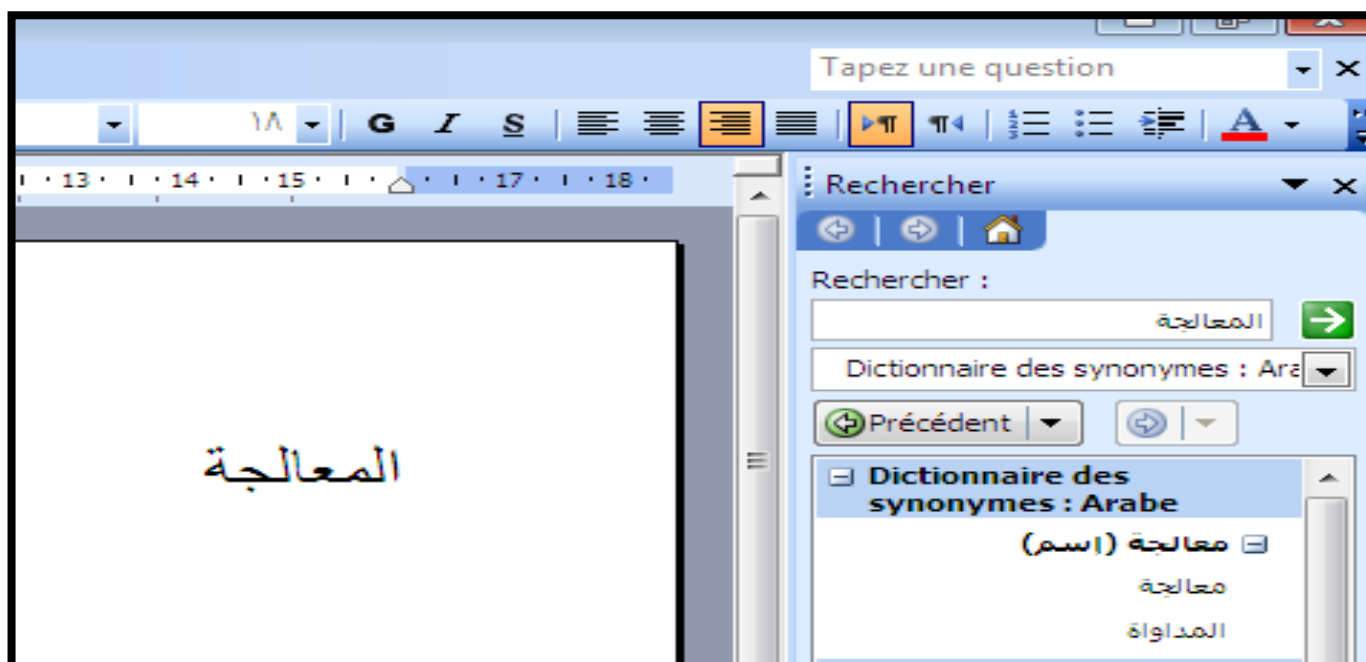
(2)- على مستوى التداول

يمكن للباحث في المعجم الآلي أن يستعين به دون أن يواجه أية صعوبة في ذلك، لأن البحث فيه يقتصر على إدخال الجذر المراد البحث عنه في واجهة المعجم الآلي ليتكلف الحاسوب بالبحث عن الجذر المخزن سابقاً، والإجابة بوجود أو عدم وجود الجذر في قاعدة البيانات ومن ثم في اللغة الطبيعية المدروسة، وذلك من خلال رموز خاصة بكل معجم آلي، ومنه يمكن لمستخدم هذا النوع من المعاجم الاستفادة منه أكثر في وقت وجيز جداً ويمكن أن يحمله معه بسهولة.

كما أن المعجم الآلي -بالإضافة إلى ذلك- يستغل في الترجمة الآلية لترجمة نص أو خطاب من لغة إلى لغة مغايرة، لأنها تقوم بالأساس على استبدال كلمات بلغة معينة إلى لغة أخرى، ويتطلب ذلك تخزين برنامج حاسوبي يضم المفردات اللغوية المراد ترجمتها ويتحقق ذلك ببرمجة معجم آلي، ولتوضيح ذلك نقدم المثال الآتي:



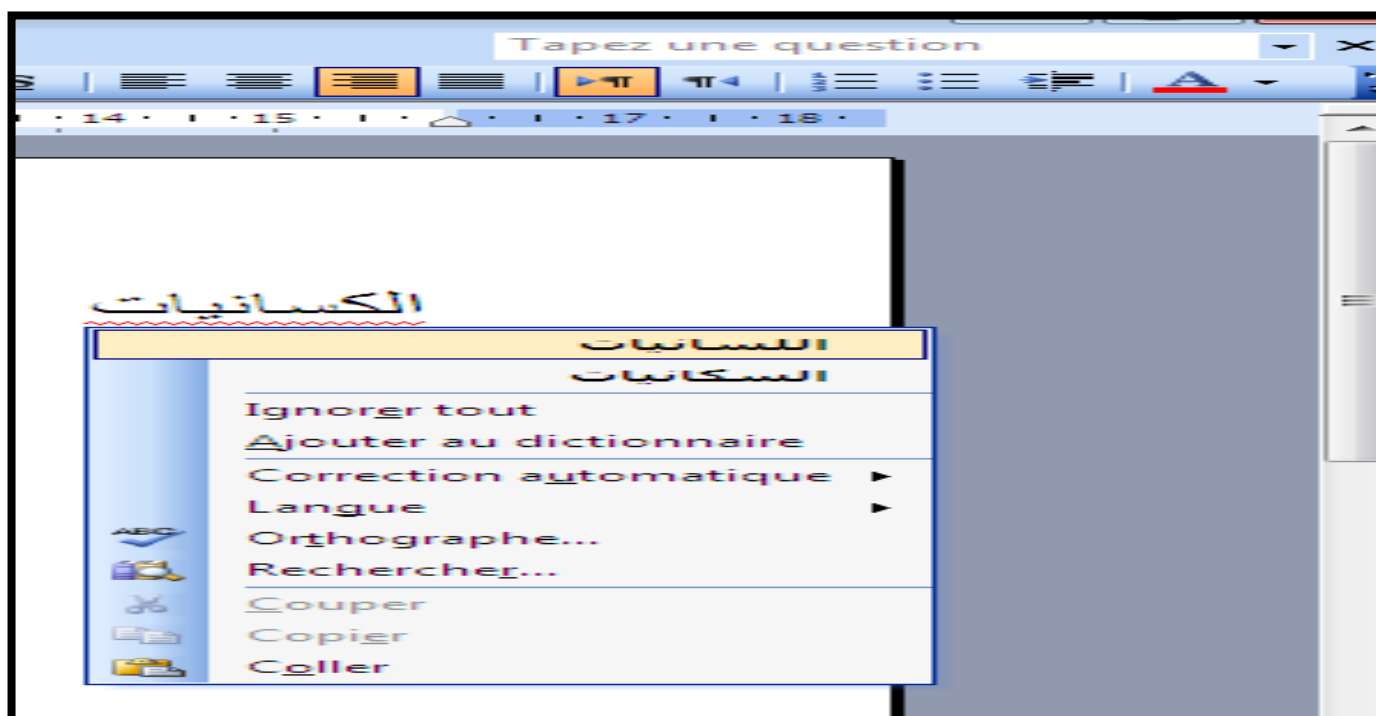
لشك
ل
(2)
واجه
ة
تخ



الشكل (4) واجهة تخص تقديم مرادف لكلمة "معالجة"

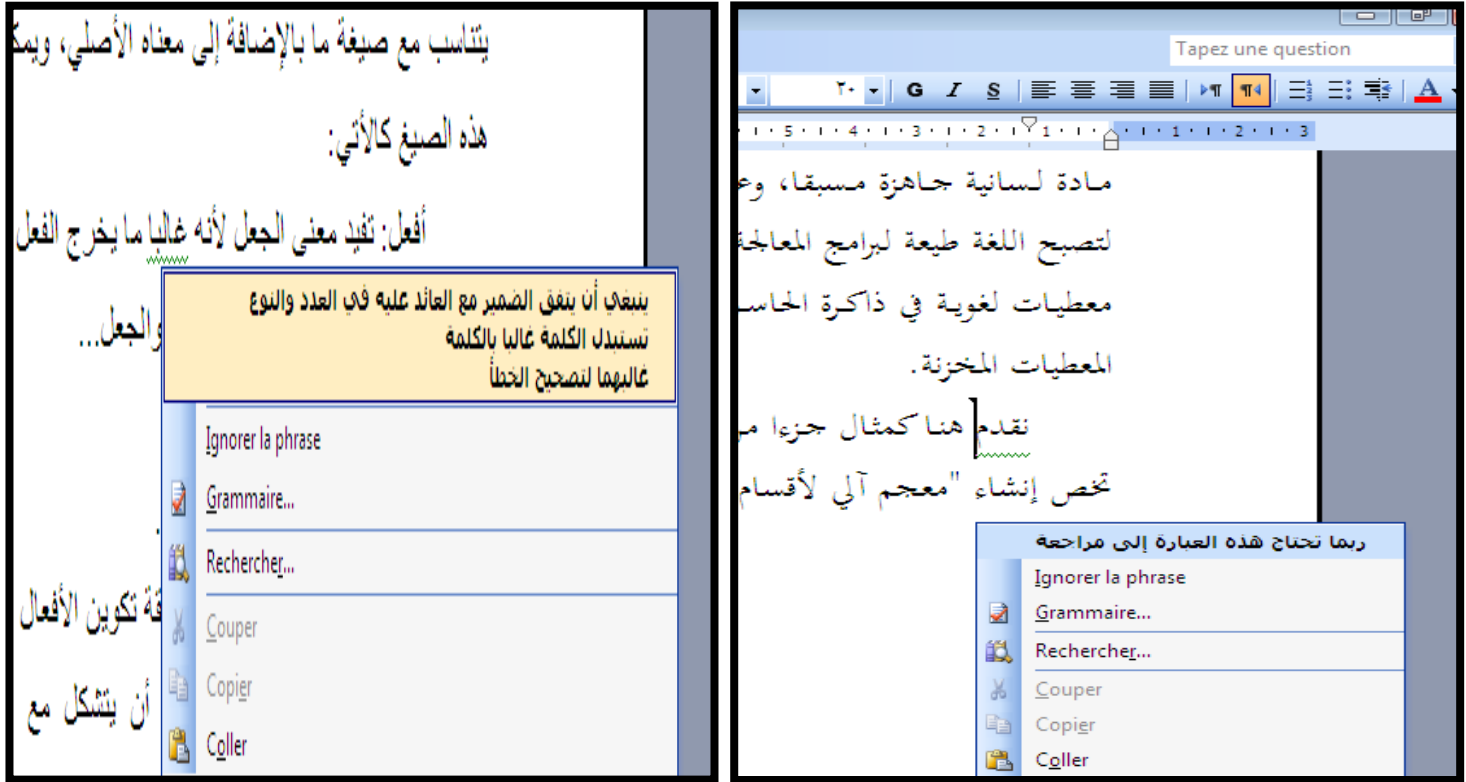
كما يستثمر المعجم الآلي أيضا في التصحيح الإملائي؛ حيث يمكن للحاسوب أن يصحح كلمة كتبت بشكل

غير صحيح بدقة فائقة، مثل:



الشكل (5) واجهة تخص التصحيح الإملائي الآلي

ويستفاد من المعجم الآلي في تصحيح عبارة صيغت بشكل خاطئ أيضا مثل:



الشكل (6) واجهتان تخصصان التصحيح الآلي لطريقة صياغة العبارات اللغوية

خاتمة:

ونستخلص من هذا المقال بأن لكل نوع من المعاجم (الورقي والآلي) سلبيات وإيجابيات؛ فعلى مستوى المادة العلمية يمكن أن يكون المعجم الورقي أوفر وأغزر لكون جل المعاجم الورقية - وخاصة القديمة- تضم كل ما يتعلق بالمفردة صوتا وصرفا وتركيبيا ودلالة وتداولاً، إلا أنه على المتصفح لهذه المعاجم أن يستخلص هذه الخصائص اللسانية بنفسه لأن المؤلف لا يبين مثلا الخصائص الصرفية (الصيغ الصرفية للفعل والاسم، ونوعية الفعل أهو جامد أم متصرف وإذا كان متصرفا أناقص التصرف أم تام التصرف...) ولا يبين أيضا الخصائص التركيبية (هل الفعل لازم أم متعد ونوعية الاسم التركيبية أعت أم حال أم مفعول...) أثناء معالجته للمفردة، عكس المعجم الآلي الذي قلما يجمع كل المستويات اللسانية لمفردة ويتوخى مؤسسه إيراد كل المعلومات اللسانية (الخصائص الصرفية والتركيبية مثلا) التي تخص مفردة ما، ومنه فإن نوع المعلومات اللغوية المخزنة في المعجم الورقي تختلف عن المعلومات اللغوية المخزنة في المعجم الآلي، فهذا الأخير يعتمد طريقة التشفير (0-1) في تخزين معلومات لغوية تخص مفردة معينة، بينما المعجم الورقي يقوم على التخطيط أو الكتابة اليدوية أو الآلية، أما على مستوى التداول؛ فكل المعجمين يستخدم من قبل الإنسان، إلا أن المعجم الآلي أسهل في البحث

وأحسن لأنه يمكن الباحث من العثور على شرح للمفردة باعتماد خطوات قليلة ومفهومة؛ فهو يقوم فقط على إدخال المفردة المراد شرحها في الشبكة الخاصة بذلك وكتابتها بشكل صحيح، ليقدّم لك الحاسوب كل المعلومات التي تخص هذه المفردة على غرار ما تمت برمجته في ذاكرة الحاسوب حول المفردة المسجلة، كما أن المعجم الآلي يستغل في البحث وفي الترجمة الآلية وفي التصحيح الإملائي للكلمة وللعبارة أيضاً، أي في المعالجة الآلية للغات الطبيعية خلافاً للمعجم الورقي الذي يستثمر في البحث فقط، ويفرض على الباحث أن يكون عالماً بطريقة ترتيب مواده.

المراجع والمصادر:

- [1] سالم الرامي، المعجم العربي العصري وإشكالاته "حوسبة المعجم العربي التقليدي". مجلة معهد الدراسات والأبحاث للتعريب، يونيو (2007).
- [2]- قاعدة بيانات من إنجاز الأستاذ الدكتور يوسف طاهر والطالبة الباحثة فطوم القريش، بحث لنيل شهادة الماستر تخصص لسانيات، بعنوان: كيفية إنشاء قاعدة معلوماتية صرفية للفعل المشتق من الجذر الثلاثي السالم في اللغة العربية، (2008-2009).
- [3] عبد الرزاق التورابي وعبد الحميد الجيهاد، المعالجة الآلية للغة العربية، وقائع الندوة الدولية يونيو (2007)، المصرف الآلي للأفعال في اللغة العربية.

[4] Aïda KHEMAKHEM, ArabicLDB : une base lexicale normalisée pour la langue arabe. Master en système d'information et nouvelles technologies : soutenue le 2 novembre(2006).

المعاجم:

- معجم العين لأبي عبد الرحمن الخليل بن أحمد الفراهيدي (100-145هـ) 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 القريش، 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 12 CITALA 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.



The Lexicon and the Automatic Dictionary: A Comparative Linguistic Study

Fettoum Krieche

Secondary school teacher qualifying

Member of Innovations Team in Mohamadia School of Engineering in Rabat

Fettoum.krieche@gmail.com

Abstract: This comparative study reveals the differences and discrepancies between the dictionaries and the mechanical dictionaries at the level of the scientific material and the level of use. As is well-known, all vocabulary of natural languages are stored in public dictionaries or private dictionaries. Since ancient times, linguists have been interested in studying the language by collecting it from the mouths of its speakers and classifying it in linguistic language books. However, with the development of technological methods and natural language processing systems, language study methods have changed and are stored in automatic dictionaries after being stored in paper dictionaries.

Keywords

Dictionaries - Automatic dictionaries - Automated processing - Database - Classification - Encryption.

Ambiguity in a Corpus Based Approach for Bilingual Ontology

Ahmed R. Elmahalawy ^{*1}, Mostafa M. Aref ^{**2}, A. A. Soliman ^{*3}

**Mathematics Department, Faculty of Science, Benha University, Benha, Egypt.*

¹aboreda92@gmail.com

³a_a_soliman@hotmail.com

*** Computer Science Department, Faculty of Computer and Information Sciences,*

Ain Shams University, Cairo, Egypt.

²aref_99@yahoo.com

Abstract— This paper presents solving ambiguity in a corpus based approach for bilingual ontology. The description of the bilingual ontology (Arabic-English) is utilizing a class definition of object oriented programming to define concepts of nouns and verbs. Four algorithms of corpus based for bilingual ontology are designed. There algorithms are: preprocessing; matching and alignment; ambiguity resolving and updating. The ambiguity algorithm resolves the ambiguity of having a word that might be noun or verb.

Keywords: Machine Translation, Ontology, Bilingual Ontology (Arabic-English), Corpus based approach, (Arabic-English).

1 INTRODUCTION

This paper presents resolving ambiguity in a corpus-based approach for bilingual ontology that can be used to describe the concepts by using classes. Ambiguity is a pervasive phenomenon in human languages. It is very hard to find words that are not at least two ways ambiguous. Sentences which are (out of context) in several ways are considered ambiguous. Ambiguity is the rule, not the exception. In the worst case, a sentence containing two words and each has two ways ambiguous may lead to four ambiguous. The word ambiguous, at least according to the Oxford English Dictionary, is ambiguous between two main types of meaning: uncertainty or dubiousness on the one hand and a sign bearing multiple meanings on the other. Philosopher's interest in ambiguity has largely stemmed from concerns regarding the regimentation of natural language informal logic: arguments that may look good in virtue of their linguistic form in fact can go very wrong if the words or phrases involved.

This paper is organized as follows: In section (2), gives a background on ontology, bilingual ontology and Machine Translation. In Section (3), gives the related works. Section (4) describes the bilingual ontology (Arabic-English) of noun and verb by using a class of object oriented programming. In section (5) presents the design of the corpus based approach for bilingual ontology using four algorithms: preprocessing, matching & alignment model, ambiguity resolving and updating. In section (6), study two cases to obtain a bilingual ontology (Arabic-English) of a noun hierarchy and similarity to verb. In section (7) starts from an Arabic-English ontology and uses a number of files (30 Arabic-English corpus files) to get a new bilingual ontology. Finally, section (8) presents conclusion and future works.

2 BACKGROUND

Machine translation is an automated translation. This translation is implemented by utilizing computer software to transform a text from a naturalistic language (such as Arabic) to another language (such as English) without any human involution. The machine translation process is shown in Fig. 1[1].

The concept is an ambiguous, if it can have one or more than meaning. For instance, the concept of "bank" which has two meaning in English the edge of a river, or a financial institution. Ontology is a debate here in the applied context of software and database engineering, yet it has a theoretical grounding as well. Ontology gives details of a range of words with which to make statements, which may be inputs or outputs of knowledge agents (such as a software program) [2].

Bilingual is the most general expressions that is utilized when people talk about people who speak two languages. For example, a bilingual person might talk Arabic and English or any other two languages. How to make ability to speak two languages, mostly depends on the person who works to find information and it make observations in the form of questions, or the policy maker and his statutory policy [3]. The word of Bilingual is divided into two parts: the first part is "Bi" which means (having two) and the second part is "lingual" which means (language). Thus bilingual means (having two languages). Bilingual is a noun, and a person can be called a bilingual, such as in the North American country like Canada, where the official languages are French and English, and where many of the citizens are bilingual [4].

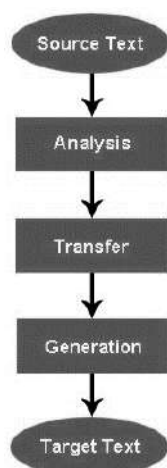


Figure 1: Machine Translation Process

3 RELATED WORK

There are many related work depend on machine translation, ontology, corpus based and bilingual ontology. In [5], the authors give a detail about the semi-automatic process of associating a Japanese word list with a semantic concept taxonomy called an ontology, utilizing an English- Japanese bilingual dictionary. This problem focuses on how to connect the Japanese lexical things with the concepts in the ontology by automatic ways, so it is also hard to know many concepts manually. They have prepared three algorithms to connect the Japanese lexical things with the concepts such as: the equivalent-word match, the argument match, and the example match. The major goal of this research improves the ratio of the open words and the closing words from the three points: semantic distance measurement, other lexicons and databases and Integration of the three algorithms.

In [6], the researcher describes an alignment system that aligns Malayalam - English texts at word level in parallel sentences. A parallel corpus is a combination of texts in two various languages, one of whom language is translated to tantamount of the second language. So, the prime objective of this method is to construct word-aligned parallel corpus to be utilized in Malayalam and English machine translation (MT). This research aims to build Myanmar-English parallel corpus can be improved by a combination of corpus-based approach and dictionary lookup approach.

In [7], the authors developed the paper in [6]. Parallel corpus assists in to create the statistical bilingual dictionary, in backing statistical machine translation and also in supporting as traineeship data for word meaning and translation disambiguation. Furthermore, the presentation of this approach can too be progressed by utilizing a listing of equations and morphological analysis. This research develops Myanmar-English parallel corpus depends on the work by [6]. This system uses the Myanmar-English corpus (1000 sentence pairs) and 250 sentence pairs for testing. The sentences were at least 4 words long. The performance report of our alignment Models by using three terms (Precision, Recall and F - measure).

In [8], the researchers describe the methodology to know the parallel Hindi-English sentences by utilizing a word alignment. This methodology helps to improve the parallel Hindi-English word dictionary after syntactically and semantic analysis of the original text from Hindi-English. This methodology develops on two ways to solve this problem. The first way is normalization of tagged Hindi-English sentences. The second way is a mapping of Hindi-English sentence by utilizing parallel Hindi-English word dictionary. The major aim of this research how to describe the alignment English-Hindi parallel corpus and obtain the type of alignment (e.g. one to one, ..., and so on). This paper starts from 555 different parallel English-Hindi sentences have been accepted.

4 DESCRIPTION OF BILINGUAL ONTOLOGY

In this section the description of bilingual ontology is going to focus on some part of speech (POS); the concept of nouns and verbs. The noun concepts are going to discuss some semantic relationships e.g. Synonyms, Hypernyms, and Hyponyms in the perspective of both English and Arabic. The verb concepts are going to discuss some semantic relationships e.g. Synonyms, Hypernyms, and Troponyms in both English and Arabic perspective.

The bilingual ontology is a set of concepts in two languages (Arabic - English), one of which is the translation equivalent of the other. The bilingual ontology described by using the concepts. The concept is defined by using a class of Object-Oriented-Programming to describe the concept of English and Arabic.

As illustrated in Fig. 2, the general description of noun concept is defined by using a class. From the class definition, the symbols and characters are defined as the following:-

- The symbol (#) means the number of
- N_S = number of Synonyms
- N_E = number of Hypernyms.
- N_O = number of Hyponyms.

N-concept		
Semantic relations	#	Concepts
# Synonyms:	(N_S)	concept 1 - - concept (N_S)
# Hypernyms:	(N_E)	concept 1 - - concept (N_E)
# Hyponyms:	(N_O)	concept 1 - - concept (N_O)
# المرادافات:	(N_S)	١ مفهوم - - (N_S) مفهوم
# الاشتمال:	(N_E)	١ مفهوم - - (N_E) مفهوم
# التضمين:	(N_O)	١ مفهوم - - (N_O) مفهوم

Figure 2: A General Description of the noun concept

As illustrated in Figure 3.2, the general description of noun concept is defined by using a class. From the class definition, the symbols and characters are defined as the following:-

- The symbol (#) means the number of
- N_S = number of Synonyms
- N_E =number of Hypernyms.
- N_T = number of Troponyms.

We discussed about the description of bilingual ontology and some examples obviously in [9].

V-concept		
Semantic relations	#	Concepts
# Synonyms:	(N_S)	concept 1 - - concept (N_S)
# Hypernyms:	(N_E)	concept 1 - - concept (N_S)
# Troponyms:	(N_T)	concept 1 - - concept (N_T)
# المرادافات:	(N_S)	١ مفهوم - - (N_S) مفهوم
# الاشتمال:	(N_E)	١ مفهوم - - (N_E) مفهوم
# المجاز:	(N_T)	١ مفهوم - - (N_T) مفهوم

Figure 3: A General Description of the verb concept

5 DESIGN OF CORPUS BASED APPROACH FOR BILINGUAL ONTOLOGY

This section describes the design of corpus based approach for bilingual ontology (Arabic-English). There are four different steps to develop this approach. These steps will be discussed in the following of the sub sections. This section refers to a work [7] to develop the three steps (Pre-processing, Matching & Alignment, and Update) and add a new step is “Ambiguity resolving” to obtain the final architecture of corpus-based approach for bilingual ontology is developed as shown in Fig.4.

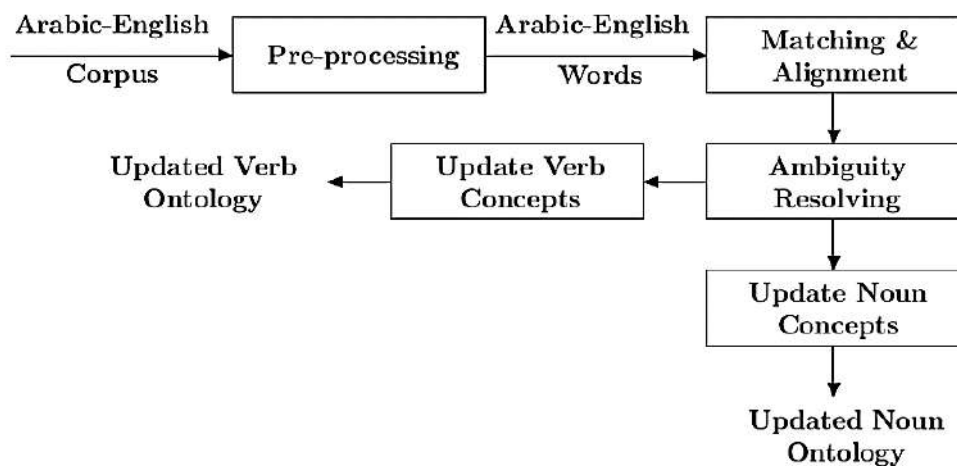


Figure 4: Architecture of Corpus-Based Approach for Bilingual Ontology

This architecture starts from an Arabic-English corpus, to get (nouns and verbs) concepts to store them in bilingual ontology. Now, let's go to explain the four steps and clarify the improvement.

A. Preprocessing Algorithm

Pre-processing Algorithm is an important step in the design of the corpus-based approach for bilingual ontology. It removes all Arabic and English stop words from the sentences, as well as, all character sets, digits, proper nouns (singular or plural), adverbs, adjectives, ..., and so on except a (verb or noun) as illustrated in Algorithm 1. This algorithm describes as follows: Given an Arabic language (A) and English language (E). The Arabic sentence describe as $A = A_1, A_2, \dots, A_r, \dots, A_{L_A}$ for length L_A whereas the English sentence describe as $E = E_1, E_2, \dots, E_k, \dots, E_{L_E}$ for length L_E .

The aims of this algorithm is to remove all stop words of Arabic and English Sentences (i.e. character sets, digits, nouns, verbs, proper nouns, adverbs, adjectives, ..., and so on) except nouns and verbs. This algorithm improves from the work by [7], it gets new sentence of two language (Arabic and English) as illustrated in Algorithm 1.

Algorithm: Pre-processing Algorithm

```

1: Start with a list of stop words Arabic ( $L_1$ ) and English ( $L_2$ );
2: Accept the Arabic sentence ( $A$ ) and the English sentence ( $E$ );
3: for each Arabic sentence  $A$  do
4:     Separate the sentence into words;
5:     if word is found in a list ( $L_1$ ) or word  $\in$  Character sets or word  $\in$  Digits or word  $\notin$  (noun or verb) then
6:         Removing the stop words;
7:         Removing the character sets;
8:         Removing the digits;
9:         Removing the proper nouns;
10:        Removing the adverbs, adjectives, ..., and soon;
11:    else
12:        Store the word in a new list;
13:    End if
14: end for
15: for each English sentence  $E$  do
16:     Separate the sentence into words;
17:     if word is found in a list ( $L_2$ ) or word  $\in$  Character sets or word  $\in$  Digits or word  $\notin$  (noun or verb) then
18:         Removing the stop words;
19:         Removing the character sets;
20:         Removing the digits;
21:         Removing the proper nouns;
22:         Removing the adverbs, adjectives, ..., and soon;
23:    else
24:        Store the word in a new list
25:    End if
26: end for

```

B. Matching and Alignment Algorithm

As illustrated in Algorithm 2, after preprocessing algorithm. The matching and alignment algorithm is used to make matching between two words in Arabic and English. This algorithm develops from the work by [7].

Algorithm 2: Matching and Alignment Algorithm

```

1: Accept the Arabic sentence ( $A$ ) and the English sentence ( $E$ );
2: Let  $A_E =$  set of English concepts based on Arabic word;
3: for each Arabic word do
4:     if search in  $A_E$  is found then
5:         Match the meaning with English word in  $A_E$ 
6:     else
7:         Update and make a new concept in  $A_E$ 
8:     End if
9: end for

```

C. Ambiguity Resolving Algorithm

After matching and alignment algorithm, the “ambiguity resolving” algorithm is utilized to check the concepts of (noun or verb) in (Arabic - English). This section has five cases to determine whether the concept is a noun or a verb or both as illustrated in Table I and as described in Algorithm 3.

TABLE I
CASES OF CONCEPTS

Arabic Concept or English Concept	English Concept or Arabic Concept	Final Result
noun	noun	noun
verb	verb	verb
noun	(noun and verb)	noun
verb	(noun and verb)	verb
(noun and verb)	(noun and verb)	(noun and verb)

Algorithm 3: Ambiguity resolving Algorithm

```

1: Accept list of Arabic concepts  $L_A$  and list of English concepts  $L_E$ ;
2: for each Arabic word ( $W_A$ ) and each English word ( $W_E$ ) do
3:   if  $W_A$  is a noun and  $W_E$  is a noun, conversely then
4:     Store Arabic concept and English concept as a noun
5:   Else if  $W_A$  is a verb and  $W_E$  is a verb, conversely then
6:     Store Arabic concept and English concept as a verb
7:   Else if  $W_A$  is a verb and  $W_E$  is a (verb and noun), conversely then
8:     Store Arabic concept and English concept as a verb
9:   Else if  $W_A$  is a noun and  $W_E$  is a (noun and verb), conversely then
10:    Store Arabic concept and English concept as a noun
11:  else if  $W_A$  is a (noun and verb) and  $W_E$  is a (noun and verb), conversely then
12:    Store Arabic concept and English concept as a (noun and verb).
13:  End if
14: end for

```

D. Updating Algorithm

After “ambiguity resolving” algorithm, the final step is an updating algorithm which is utilized to update the bilingual ontology (Arabic-English). It contains words to be translated into other words. The algorithm of the corpus-based approach for bilingual ontology is illustrated in Algorithm 4. We developed this algorithm from the work by [7].

Algorithm 4: Updating Algorithm

```

1: Accept the Arabic sentence  $A$  and the English sentence  $E$ ;
2: for each Arabic word do
3:   if Match is found then
4:     Corpus based Arabic word with English
5:   else
6:     Update and add in a hierarchy of the bilingual ontology
7:   End if
8: end for

```

6 CASE STUDIES

To illustrate the corpus based approach by using two case studies are presented. The first case study does not have ambiguity, and the second case study has ambiguity.

A. Case 1:

The first case study is to find a new concept in Arabic-English that is a noun or a verb in one of these language and is a (noun and verb) in the other. In this case, the results are described in two points:

1. $W_1 = \text{noun}$, $W_2 = (\text{noun and verb})$, then the result = noun.
2. $W_1 = \text{verb}$, $W_2 = (\text{noun and verb})$, then the result = verb.

This new concept is not defined in the previous (noun and verb) hierarchy of the bilingual ontology.

Example: We have two input sentences in Arabic (A) and English (E) languages as the following:

E Sentence: Thank Ahmed for going with me
A Sentence: شكرا أحمد للذهاب معي

Applying the first step to remove all English and Arabic stop words, digits, character sets, proper nouns and adverbs, adjectives, ..., and so on from the two sentences by using the preprocessing algorithm in Algorithm1. the new sentences are follows:

E Sentence: Thank go
A Sentence: شكرا ذهاب

Applying the second step to make alignment between the two new sentences by using the alignment algorithm as shown in Algorithm 2. From the alignment algorithm, this case gets a new verb concept (thank-شكرا). Now, in order to determine whether the concept is a noun or a verb or both. The “ambiguity resolving” algorithm is used as shown in Algorithm 3. After checking the two concepts, this algorithm finds the Arabic concept as a (noun and verb) and English concept as a verb. This case takes the verb concept of both of them.

The concept of ”thank” has one senses in English and two senses in Arabic and every word has one or more senses. The Synonyms, Hypernyms and Troponyms of that concept are described in Fig. 5.

V-thank		
Semantic relations	#	Concepts
# Synonyms:	1	thank
# Hypernyms:	7	convey - impart - tell - inform - communicate - interact - act
# Troponyms:	2	acknowledge - appreciate
# المرادافات:	2	حمد الله - شكر
# الاشتمال:	7	فعل - تفاعل-نقل - إعلام - يخبر - عرف - نقل
# المجاز:	2	نقدر - اعترف

Figure 5:A Description of the verb thank

The final step is applied to make an update in the hierarchy of the bilingual ontology by using an updating algorithm as shown in Algorithm 4. To get new hierarchy of the bilingual ontology, we use all the previous of verb concepts such as (eat, go, become and do) and add a new concept ”thank” as shown in Fig. 6.

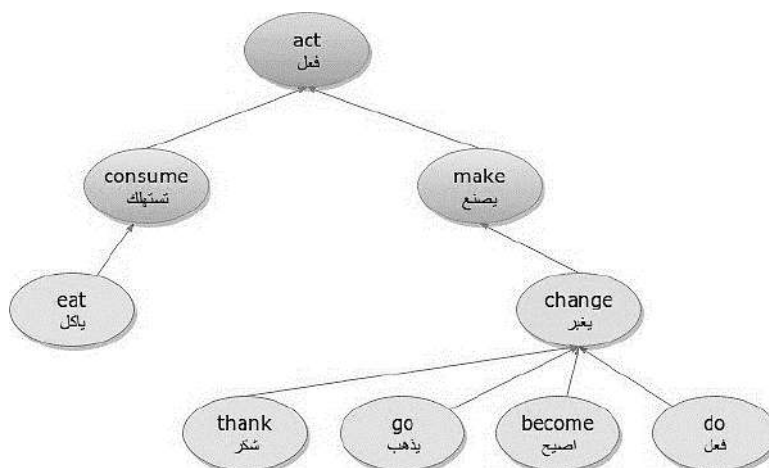


Figure 6: Description of the verb hierarchy after adding thank

B. Case 2:

The second case study is to find a new ambiguous concept in Arabic-English one of this language is a (noun and verb). This case takes all parts of speech like noun and verb. This new concept is not defined in the previous (noun or verb) hierarchy of the bilingual ontology. Finally, the researcher adds two concepts in both noun hierarchy and in verb hierarchy.

Example: We have two input sentences in Arabic (A) and English (E) languages as the following:

E Sentence: This book about the car
 A Sentence: هذا الكتاب عن السيارة

Applying the first step to remove all English and Arabic stop words, digits, character sets, proper nouns and adverbs, adjectives, ..., and so on from the two sentences by using the pre- processing algorithm as shown in Algorithm 1, the new sentences are follows:

E Sentence: book car
 A Sentence: كتاب سيارة

Applying the second step to make alignment between the two new sentences by using an alignment algorithm as shown in Algorithm 2. From the alignment algorithm, this case gets a new ambiguous concept (book – كتاب). Now, in order to

determine whether the concept is a noun or a verb or both. The “ambiguity resolving” algorithm is used as shown in Algorithm 3. After checking the two concepts, this algorithm finds the Arabic concept is a (noun and verb) and English concept is a (noun and verb). This case takes some of parts of speech, such (noun and verb) of the concept in Arabic-English.

The final step is applied to make an update in the hierarchy of the bilingual ontology by using an updating algorithm in Algorithm 4. To get new hierarchy of the bilingual ontology, we use all the previous verb concepts such as (eat, go, become, thank and do) and add a new concept ”book” as shown in Fig. 7. Also, by using all the previous noun concepts such as (person, dinner, car, college and teacher) and add a new concept “book” as shown in Fig. 8.

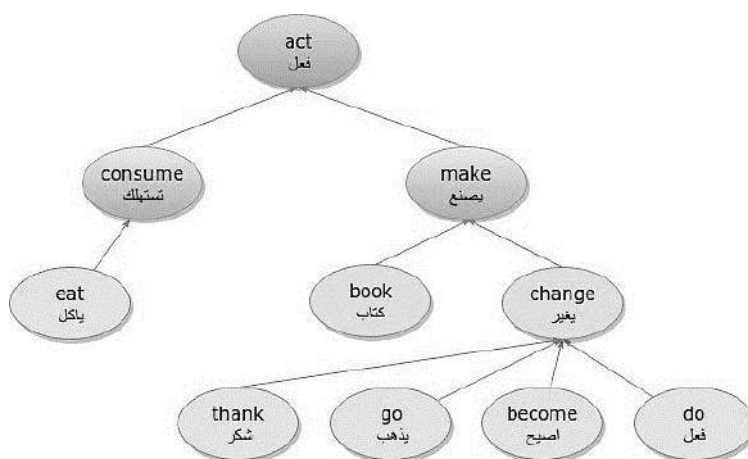


Figure 7: Description of the verb hierarchy after adding a book

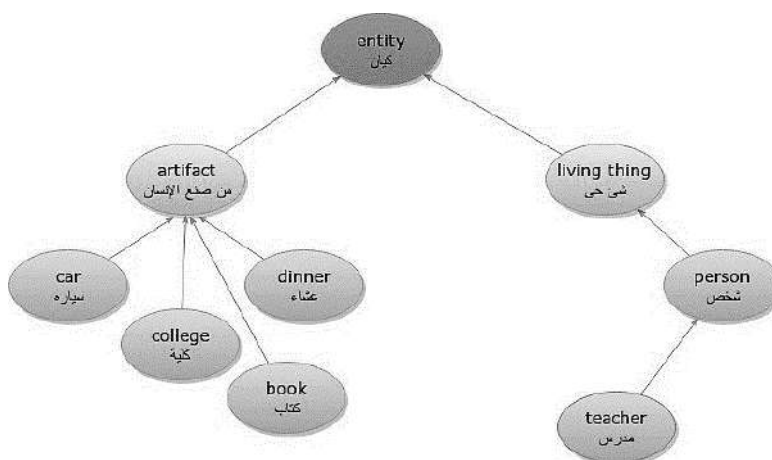


Figure 8: Description of the noun hierarchy after adding a book

7 DATA SETS

This section uses the Python programming language and some libraries to make an implementation for the Arabic-English corpus files to obtain the new concept for noun or verb, starting with an Arabic-English bilingual ontology that contains 25 concepts of noun and verb. This chapter uses 30 Arabic-English corpus files from the website [10].

The new Arabic-English bilingual ontology is stored the concept in excel sheet. The obtained results are shown in Table II. By using the comparison between the old of bilingual ontology and Arabic-English corpus files, the researcher gets 7 different nouns and eight different verb to adding in the final bilingual ontology. The research collects all nouns and verbs in a new bilingual ontology. The obtained results are shown in Table II. The executed program takes approximately 50 minutes.

TABLE II
RESULTS OF 30 ARABIC-ENGLISH CORPUS FILES

	Number of Noun	Number of Verb	Total
Initial Bilingual Ontology	16	9	25
Arabic-English Corpus Files	276	130	406
Final Bilingual Ontology	283	138	421

8 CONCLUSIONS

In this paper, we proposed a description of the bilingual (Arabic-English) ontology by using a class of object oriented programming to define concepts of nouns and verbs. The noun concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Hyponyms in the concepts of English and Arabic. The verb concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Troponyms in the concepts of English and Arabic. The paper discusses how we may resolve the ambiguity of words in both languages. Two case studies are presented.

We have applied four algorithms of corpus based for bilingual (Arabic-English) ontology such as:

1. Preprocessing
2. Matching & Alignment
3. Ambiguity resolving
4. Updating

From the description of bilingual ontology and design of corpus based approach for bilingual ontology to show the case studies. The directions of future work:

1. Applying this approach on a large number corpus files of (Arabic-English) ontology.
2. Applying this approach on multilingual ontology rather than bilingual ontology and in this case using a number of corpus files on multilingual ontology.

REFERENCES

- [1] C. Stern and A. Dufournet. What is machine translation? — systran translation technologies. Machine translation, <http://www.systransoft.com/systran/corporate-profile/translation-technology/what-is-machine-translation/>, August 2011, (accessed 22October2015).
- [2] T. Gruber. Ontology (computer science) - definition in encyclopedia of database systems. Ontology, <http://tomgruber.org/writing/ontology-definition-2007.htm>, September 2007, (accessed 14October2015).
- [3] N. Takaya. What do we mean when we say bilingual? — Psychology in action. Bilingual, <http://www.psychologyinaction.org/2012/01/17/what-do-we-mean-when-we-say-bilingual/>, January 2012, (accessed 31October2015).
- [4] I. Thinkmap. Bilingual - dictionary definition: Vocabulary.com. Bilingual, <http://www.vocabulary.com/dictionary/bilingual>, June 2013, (accessed 31October2015).
- [5] A. Okumura, E. Hovy. *Building Japanese-English Dictionary based on Ontology for Machine Translation*. In *proceedings of ARPA Workshop on Human Language Technology*, pages 236-241, 1994.
- [6] K. T. Nwet. *Building Bilingual Corpus based on Hybrid Approach for Myanmar - English Machine Translation*. *International Journal of Scientific & Engineering Research*, 2(9), 2011.
- [7] K. T. Nwet, K. M. Soe, and N. L. Thein. *Developing word-aligned Myanmar - English parallel corpus based on the ibm models*. *International Journal of Computer Applications*, 27(8), 2011.
- [8] S. Dubey, T. D. Diwan. *Supporting Large English-Hindi Parallel Corpus using Word Alignment*. *International Journal of Computer Applications*, 49, No.6, (7), 2012.
- [9] Ahmed R. Elmahalawy, Mostafa M. Aref. *Design a Corpus Based Approach for Bilingual Ontology Arabic-English*. *International Journal of Computer Applications International Journal of Computer Science Trends and Technology (IJCSST)*, pages 98-103, Volume 5 Issue 1, Dec 2017.
- [10] Arabic-English Parallel Corpus. <http://aeparallelcorpus.net/index.php/content/search>, 2015. (Accessed 10-June-2018).

BIOGRAPHY



Ahmed Reda Ahmed AbdelHamid Elmahalay Demonstrator of Mathematical Department (Computer Science) and Research, Benha University, Egypt. His research interests: Web Design, Web Developing, Ontological Engineering, Image Processing, and Artificial Intelligence. His Msc thesis title is “*Developing Corpus-based Bilingual Ontology for Machine Translation (English/Arabic)*”.



Prof. Mostafa M. Aref Professor of Computer Science and Research, Ain Shams University, Cairo Egypt. He got Ph.D. of Engineering Science, June 1988, University of Toledo, Toledo, Ohio, USA. He has more than 50 journal and conference publications. His research areas are Natural Language Processing, Knowledge Representation, Object-oriented Programming, Ontology and Real-Time Strategy Games.



Prof. Abd Elkareem Soliman Professor of Mathematics Department, Head of Department and Research, Benha University, Egypt. His research areas are Differential Equations, Partial differential equations and Integral Equations.

الغموض في المنهج القائم على النصوص لتوليد علم الموجودات ثنائية اللغة

أحمد رضا المحلاوى^{1*}، مصطفى محمود عارف^{2**}، عبدالكريم عبدالحليم سليمان^{3*}
^{*}كلية العلوم قسم الرياضيات (حاسب آلي) – جامعة بنها – مصر.
^{**}قسم علوم الحاسب الآلي- كلية الحاسبات والمعلومات – جامعة عين شمس – مصر.

aboreda92@gmail.com¹

aref_99@yahoo.com²

a_a_soliman@hotmail.com³

ملخص

يقدم هذا البحث حل الغموض في المنهج القائم على علم الوجود ثنائي اللغة. وصف الانطولوجية ثنائية اللغة (عربي – إنجليزي) باستخدام تعريف الفئة الموجود في البرمجة الشيئية لوصف المفهوم في الأسم والفعل. قمنا بعمل تصميم لأربع خوارزميات يعتمد على علم الوجود ثنائي اللغة (عربي – إنجليزي). وتستخدم هذه الخوارزميات للحصول على المفاهيم الجديدة من الأسم والفعل. اما بالنسبة لخوارزمية الغموض تعمل على حل مشكلة الغموض في بعض الكلمات التي قد تكون احيانا أسم ام فعل.

A Tool for Measuring Linguistic Variations in Machine Translation: A Corpus Based Study

Maram Elsaadany

Institute of Applied Linguistics & Translation, Faculty of Arts, Alexandria University, Alexandria, Egypt

Maram.elsaadany@gmail.com

Sameh Alansary

Bibliotheca Alexandrina, Alexandria, Egypt

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

sameh.alansary@bibalex.org

Abstract — This paper is primarily a translation analysis of the Arabic and English morpho-syntactic structures using Biber's model (1988) and Stanford program. It is a corpus based quantitative study that has used 66 features out of the 67 that has been identified by Biber and the paper is in line with Biber's model and statistical procedure. The corpus selected for this thesis is Alice Monro's collection of short stories *The Power of Love (1985)* and its translation into Arabic, *Masiret El Hob (2015)* by Mohamed Tantawi. All the English and Arabic features are counted by the aid of a computer program, *Stanford (2015)*. Stanford is a program that is used for annotating the chosen corpus and it is working on the morpho-syntactic levels of English and Arabic. The 66 features are classified into four factors for the English language and five factors for the Arabic. Only twelve features are counted manually in the Arabic analysis by the researcher herself. Finally, the findings reflect major differences between the two languages as some of the features could not be identified by the computer program.

Key words: translation, computational, linguistic variations

1 INTRODUCTION

Corpora are collections, usually electronic ones today, of texts. A 'parallel corpus' is a bilingual or multilingual corpus that contains one set of texts in two or more languages" [16]. There are altogether three types of parallel corpora, and their functions are different according to their different construction. The first type is the normal parallel corpus. This type contains only texts of language, usually source language, and their translation into another language, target language. The corpus of this study belongs to this type. The second type is the reciprocal parallel corpus. It contains not only the source texts in language A and their translation in language B, but also source texts in language B and their translation in language A. The third type contains only translations in different target languages. This type may be bilingual or multilingual.

Corpus-based translation studies are interested in how equivalence might be achieved and what kind of equivalence can be achieved, and in what context [2]. Translation unit studies are interested in the alignment of translation units and their equivalents in a given parallel corpora, and how these equivalents can be re-used by other translators in the future translations, especially by those translators who have to translate into a non-native language where their intuition is often insufficient. The unit of annotations and the choice of the annotation scheme are crucial for the quality of this research. [19] has expressed that translation units are the smallest unit in translation and they are very useful for bilingual lexicography. [22] has stated that "parallel corpora are repositories of the translation units and their equivalents". Computational linguistics (CL) combine resources from linguistics and computer science to discover how human language works. Computational linguistics is a vital field in the information age. According to [18], computational linguists create tools for important practical tasks such as machine translation, speech recognition, speech synthesis, information extraction from text, grammar checking, text mining and more. [10] has stressed the idea that contrastive Analysis (CA) is a method that is connected to Contrastive Linguistics, which is considered a branch of linguistics that focuses on illustrating the differences and similarities among two or more languages at different linguistic levels as semantics, syntax, and phonology, [24]. [9] defines contrastive

analysis as "the study of foreign language learning, the identification of points of structural similarity and difference between two languages".

The use of computerized text corpora and computer programs for the automatic identification of linguistic features made it possible to fulfill a study of this scope. The words in any text are all marked, or 'tagged', for their grammatical category, to facilitate automatic syntactic analysis. There are two main steps associated with automatic identification of the linguistic features. The first is to tag the grammatical category of each word, as a noun, verb, adjective, preposition, WH pronoun, etc. [13] has explained that this step requires a computerized dictionary so that the program can search for words in the dictionary and find their grammatical category. The tags resulting from this procedure provide the basis for the second step, which is identifying particular sequences of words as instances of a linguistic feature. For example, if a noun is followed by a "WH pronoun" and not preceded by the verb "tell or say", it can be identified as a relative clause; the sequence tell/say + noun phrase + WH pronoun might be either a relative clause or a WH clause. Working on the programs which can be used for the frequency counts of the features has spread over the years (1983- 1986).

Earlier Programs have been criticized by the lack of a dictionary; to identify linguistic features, they relied on small lists of words that were built into the program structure itself. These lists included prepositions, conjuncts, pronominal forms, auxiliary forms. Since these word lists were relatively restricted, the grammatical category of many words in texts could not be accurately identified, and therefore these programs could not identify all of the occurrences of some linguistic features. The programs have been designed to avoid skewing the frequency counts of features in one genre or another so that the relative frequencies were accurate. The main disadvantage of this earlier approach was that certain linguistic features could not be counted at all. For example, there was no way to compute a simple frequency count for the total nouns in a text, because nouns could not be identified. For these reasons, the second set of programs has been taking place.

The second stage of program development took place during the years (1985-1986). The approach used in this stage is different from that of the first stage. As a result, a general tagging program to identify the grammatical category of each word in a text was developed. The aim is to develop a program that was general enough to be used for tagging both written and spoken texts. For example, the program could not depend on upper case letters or sentence punctuation. This goal is achieved by using a large-scale dictionary together with a number of context-dependent disambiguating algorithms. The main problem that had to be solved is that many of the common words in English are ambiguous as to their grammatical category. Words like "absent" can be either adjectives or verbs; words like "acid" can be either nouns or adjectives. All past and present participial forms can function as noun (gerund), adjective, or verb. A simple word like that can function as a demonstrative, demonstrative pronoun, relative pronoun, complementizer, or adverbial subordinator.

[3]has developed algorithms to disambiguate occurrences of certain words, depending on their surrounding contexts. For example, a participial form preceded by an article, demonstrative, quantifier, numeral, adjective, or possessive pronoun is functioning as a noun or adjective. That is to say, it is not functioning as a verb in this context; given this preceding context, if the form is followed by a noun or adjective then it will be tagged as an adjective; if it is followed by a verb or preposition, then it will be tagged as a noun. Tagged texts enable automatic identification of a broad range of linguistic features that are major for differentiating between genres in English. The tagged texts are subsequently used as input to other programs that count the frequencies of certain tagged items (e.g. nouns, adjectives, adverbs) and compute the frequencies of particular syntactic constructions (e.g. relativization on subject versus non-subject position). This approach assures a higher degree of accuracy and it allows inclusion of some features that could not be accurately identified by the previous programs. The resulting analysis is thus more complete than earlier analyses. The researcher has consulted the IT team in *Stanford* who helped a lot in solving many problems that the researcher has encountered in this research. They also put Biber's algorithm into consideration, which is going to be an asset to their program.

2 BIBER'S VARIATIONS

[3] is the main model upon which this study is based. The initial step is to collect the English and Arabic texts that are used as the corpus of this study. The second step is to choose an English tagger to be implemented in the analyses of the data. A pilot study is conducted for choosing the best tagger to be used in the study which is Stanford tagger. Next, computational identification of the specified linguistic features in English and Arabic texts by the use of Stanford tagger is applied. Furthermore, an annotation of the linguistic features is manifested as the units of annotation and the choice of the annotation scheme are crucial for the quality of this research. Moreover, clustering the linguistic features into groups that occur with a high frequency is manifested. The following step is to group these features into factors and to apply a statistical analysis to interpret the features underlying each factor. Also, specifying the English and the Arabic features used in *Stanford* Corpus to

be applied in both the English texts and the Arabic equivalent translated ones. Next, annotating the English / Arabic texts by using an English tagger, Stanford is applied. Text normalization is crucial for any comparison of frequency counts across texts, because text length can vary widely. A comparison of non-normalized counts is going to give an inaccurate assessment of the frequency distribution in texts. Finally, the actual presence of the variables located in the texts and their parallel translated words are going to be checked in the SPSS program to calculate their actual number of presence.

The present study is significant for many reasons. First, the use of computer-based text corpora provides a standardized database. Second, the use of *Stanford computer program* to count the frequency of occurrence of 66 linguistic features in ten short stories and their translations and to offer a detailed analysis of the distribution of these features. It is the only computer program that can deal with the English and Arabic languages simultaneously. Next, the employment of multivariate statistical techniques, especially factor analysis, is applied to determine the co-occurrence relations among the linguistic features. Finally, the use of microscopic analysis is maintained to interpret the features underlying each factor.

While the purpose of this paper is academic, the need to accelerate the investigations in translation research is becoming a must, as translating from other languages in the modern era, with information flooding from every corner in the globe is increasingly in demand. Translation studies, have only evolved during the last decades [8]. Scientific research in this area is a very recent phenomenon, as stressed by [12]. The call for research in translation is overwhelming as "a whole range of issues seemed to be waiting for examination, and inquiry is overdue", [21]. Calls for conducting systematic comparative studies of translated and source texts [8] and those for research focusing more upon what [17] has termed "the acquisition of translation competence", have not been accomplished. In translating between English and Arabic, there is a shortage of research in translation problems that may be encountered by Arabic translators of English [15], [16].

3 THE CORPUS

The short stories that are used as the corpus of this study are ten short stories taken from Monro's collection of short stories *The Power of Love* and their translated equivalence which are translated by Mohamed Saad Tantawi and published by *Hindawi Foundation for Education and Culture* in 2015. The actual presence of the variables located in the texts and their parallel translated words are going to be checked in the SPSS program to calculate their actual number of occurrence. In order to normalize texts, in this study, the frequency counts of all linguistic features are normalized to a text length of 7,725 words so we have to delete some words to make them the same length.

4 PROCEDURE

Biber's model of textual variations (1995) is the main model upon which this paper is based. The initial step is to collect the English and Arabic texts that are used as the corpus of this study. The second step is to choose an English tagger to be implemented in the analyses of the data. A pilot study is conducted for choosing the best tagger to be used in the study which is Stanford tagger. Next, computational identification of the specified linguistic features in English and Arabic texts by the use of Stanford tagger is applied. Furthermore, an annotation of the linguistic features is manifested as the units of annotation and the choice of the annotation scheme are crucial for the quality of this research. Moreover, clustering the linguistic features into groups that occur with a high frequency is manifested. The following step is to group these features into factors and to apply a statistical analysis to interpret the features underlying each factor. Also, specifying the English and the Arabic features used in Stanford Corpus to be applied in both the English texts and the Arabic equivalent translated ones. Next, annotating the English / Arabic texts by using an English tagger, Stanford is applied. Text normalization is crucial for any comparison of frequency counts across texts, because text length can vary widely. A comparison of non-normalized counts is going to give an inaccurate assessment of the frequency distribution in texts. Finally, the actual presence of the variables located in the texts and their parallel translated words are going to be checked in the SPSS program to calculate their actual number of presence.

5 Factor Analysis: Technical Description

The first step in a factor analysis is to choose a method for extracting the factors. The use of factor analysis in linguistics is usually exploratory rather than confirmatory, a principal factor solution should be used, [13]. There are several options available, but the most widely used is known as 'common factor analysis' or 'principal factor analysis'. This procedure extracts the maximum amount of shared variance among the variables for each factor. Thus, the first factor extracts the maximum amount of shared variance, i.e., the largest grouping of co-occurrences in the data; the second factor then extracts

the maximum amount of shared variance from the tokens left over after the first factor has been extracted, and so on. In this way, each factor is extracted so that it is uncorrelated with the other factors.

Once a method of extraction has been chosen, the best number of factors in a solution must be determined, [13]. As noted above, the purpose of factor analysis is to reduce the number of observed variables to a relatively small number of underlying constructs. A factor analysis will continue extracting factors until all of the shared variances among the variables have been accounted for, but only the first few factors are likely to account for a nontrivial amount of shared variance and therefore be worth further consideration. There is no mathematically exact method for determining the number of factors to be extracted. A "scree plot", will normally show a characteristic break indicating the point at which additional factors contribute little to the overall analysis. The scree plot corresponding to eigenvalues is given in Figure (1), (2). The eigenvalues of the English and Arabic texts can be used to indicate the percentage of shared variance that is accounted for by each factor.

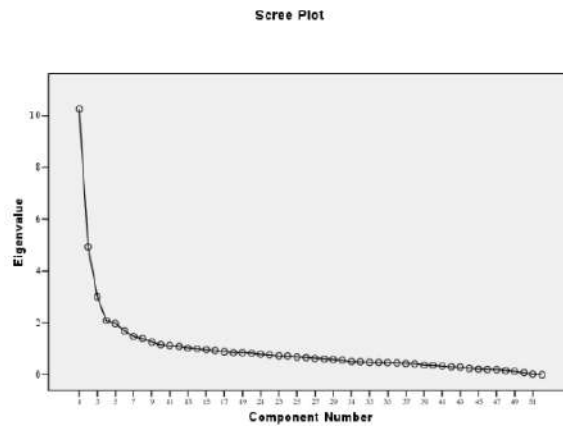


Figure 1 Scree plot of the English factors

The break in the English plot occurs between the first, second, third and fourth factors. When faced with a choice between a larger or smaller number of factors, the more conservative procedure is to extract the larger number and then discard any unnecessary factors [13]. Extracting too few factors will result in loss of information, because the constructs underlying the excluded factors will be overlooked; it might also distort the factorial structure of the remaining factors, because multiple constructs are collapsed into a single factor. The same procedure is applied to the Arabic data. The scree plot is applied to extract the number of factors needed and the features that constitute each factor.

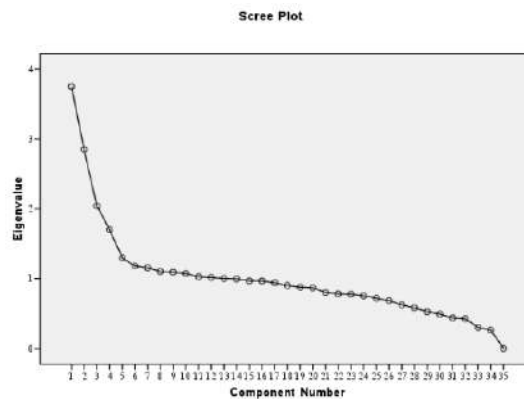


Figure 2 The Arabic scree plot.

The break in the Arabic plot occurs between the first, second, third, fourth and fifth factors. When faced with a choice between a larger or smaller number of factors, the more conservative procedure is to extract the larger number and then discard any unnecessary factors [13]. Extracting too few factors will result in loss of information, because the constructs underlying the excluded factors will be overlooked. It might also distort the factorial structure of the remaining factors, because multiple constructs are collapsed into a single factor. Factor loadings reflect the extent in which one can generalize from a given factor to an individual linguistic feature. Features with higher loadings on a factor are more representatives of the dimension underlying the factor, and when interpreting the nature of a factor, the features with large loadings are given priority. Multivariate statistical techniques such as factor analysis are not practical without the aid of computers. A factor analysis involves many computations using matrix algebra. The first point for a factor analysis is a simple correlation matrix of all variables. Factor analysis routines are usually included as part of the standard statistical packages (e.g. SPSS) available on computers at most academic institutions. SPSS computational tool makes a new range of linguistic research possible.

Factor analysis uses frequency counts of linguistic features to locate sets of features that co-occur in texts. The use of this technique to identify underlying textual dimensions is based on the assumption that frequently co-occurring linguistic features have at least one shared function [3]. It is claimed here that there are relatively few primary linguistic functions in English and Arabic, and that the frequent co-occurrence of a group of linguistic features in texts is indicative of an underlying function shared by those features. Working from this assumption, it is possible to obtain a unified dimension underlying each set of co-occurring linguistic features. In this proposal, a collection of short stories written by Monro and their translation are used as the sample of this study. The next step is to identify the linguistic features with these texts before applying factor analysis. Inadequate preparation or skewing in these theoretical prerequisites can invalidate the results of a factor analysis [13]. That is, factor analysis provides the primary analytical tool, but it is dependent on the theoretical foundation provided by an adequate database of texts and inclusion of multiple linguistic features.

6 Results

A. Interpretation of the English Results

The results of the present study reflect nine factors, four factors for the English language and five factors for the Arabic one. The nine factors identified here are general, underlying parameters of variations. Some of the features in the Arabic language cannot be identified in the five factors mentioned for the Arabic language. That is why the researcher has counted them manually by herself in a separate table. The factors do not represent all of the differences defined by the original 67 linguistic features that are identified by Biber in his model. The factors are abstractions, describing the underlying parameters of variations in relatively global terms.

Rotated component matrix in factor 1

	Factor 1
Positive	
Wh question	0.75
Type token ratio	0.72
Existential there	0.78
Amplifiers	0.47
Private verbs	0.64
Hedges	0.39
Contractions	0.72
Do as a pro verb	0.36
Word length	0.33
Past tense verbs	0.54
2 nd persons pronouns	0.47
Pronoun it	0.45
Analytic negation	0.40
1st person pronoun	0.78
Wh relative clause subject position	0.45
Split infinitives	0.39
WH relative cl. in object position	0.50
Non phrasal coordination	0.40
Demonstrative	0.63
Emphatics	0.60
Negative	
Present tense	-0.07
Attributive adjective	-0.08

t, p: t and p values for Student t-test

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 2

	Factor 2
Past participial clause	0.52
Agentless passive	0.66
By passive	0.78
Nouns	0.83
Present tense verbs	0.73
3 rd person pronoun	0.51
Perfect aspects	0.62
Public verbs	0.77
Synthetic negation	0.59
Present participial clause	0.77
Attributive adjectives	0.89
Present participial WHIZ deletion	0.42
Prepositional phrase	0.69
That deletion	0.50
Conjuncts	0.64
Adverbs	0.67
Negative	
Word length	-0.58
Prepositions	-0.54

t, p: t and p values for **Student t-test**

Rotated component matrix in factor 3

	Factor 3
Predicative adjectives	0.43
Other adverbial subordinators	0.75
Gerunds	0.81
Time adverbials	0.40
Place adverbial	0.50
Adverbs	0.67
Be as a main verb	0.82
Pied piping construction	0.63
Prediction modals	0.50
Conditional subordination	0.80
Discourse particle	0.50
Possibility modals	0.63
Necessity modals	0.73
Suasive verbs	0.48
Concessive subordination	0.67

Rotated component matrix in factor 4

	Factor 4
Positive	
That clause as a verb compliment	0.44
Past participial WHIZ deletion	0.58
That clause as adjective compliment	0.39
That clause on subject position	0.54
Sentence relatives	0.71
Demonstrative pronouns	0.63
Indefinite pronoun	0.53
Seem / appear	0.61
Down toners	0.38
That clause on object position	0.67
Nominalization	0.83
Causative subordination	0.48
Split auxiliaries	0.25
Infinitives	0.60
Phrasal coordination	0.67
Demonstrative pronoun	0.63
Negative	
Time adverbial	-0.50
Place adverbial	-0.49

t, p: t and p values for Student t-test

*: Statistically significant at $p \leq 0.05$

Overall, these results indicate that the tagging program is quite accurate. First, there are very few mis tags; the majority of 'errors' are untagged items, which do not introduce misleading analyses, and even untagged items are relatively uncommon. Secondly, there is no serious skewing of mis tags so that the results are accurate in relative terms; that is, the results enable accurate comparisons across texts because the same word types are left untagged in all texts. Last but not least, the few mistags and untagged items that do exist are of a very specialized or idiosyncratic in nature, and often these items have no bearing on the linguistic features counted for the analysis of textual dimensions. The tagged texts produced by this program thus provide a good basis for the automatic identification of the linguistic features, only the potentially important linguistic features are actually counted.

The tagging of some lexical items was so problematic that they were systematically excluded. In addition, the researcher has carried out some hand editing of the tagged texts to correct certain inaccuracies. For example, past and present participial forms were checked by hand. Although the tagging program includes elaborate algorithms to distinguish among gerunds, participial adjectives, WHIZ deletions, participial clauses, passives and perfects (in the case of past participles), and main active verbs (present or past), a high percentage of these forms was incorrectly tagged.

To a computer program without access to semantic information, however, there is no difference between these constructions, and thus at least one of the two cases will be tagged incorrectly. Similar problems were found in attempting to disambiguate the other functions of present and past participial forms. As a result, all participial forms were checked by hand. The factors reflect the fact that depictive details are important in narrative discourse. Discourse particles are generalized markers of

informational relations in a text. They help to maintain textual coherence when a text is fragmented and would otherwise be relatively incoherent. Also, subordination features occur with a variety of involved and generalized content features, and in a complementary pattern to highly informational features. Furthermore, sentence relatives are present to express attitudinal comments. Wh – clauses provide a way to “talk about” questions. Time and place adverbials depend on referential inferences by the addressee. Persuasion is one of the main techniques used in informative texts; it is a marking of the author’s own point of view or an assessment of the advisability of an event presented to persuade the reader. Narrative genre is marked by considerable reference to past time, third person animate referents, reported speech and depictive details. It has a high lexical variety. It is a discourse that reports events in the past or deals with more immediate matters but does not mix both. In conclusion, the four factors have strong factorial structures and the features grouped in each factor are functionally coherent and can be easily interpreted.

B. Interpretation of the Arabic Factors

Rotated component matrix in factor 1

	Factor1
Amplifiers	0.51
Analytic negation	0.78
Conditional subordination	0.66
Discourse particles	0.57
Emphatics	0.74
Pied piping	0.30
Prepositional phrase	0.61
Private verbs	0.66
Seem and appear	0.73
Wh clause	0.67
Type/ token ratio	0.71
Sentence relatives	0.64
Split auxiliary	0.54
Infinitives	0.50
Causative subordination	0.53

Rotated Component Matrix in factor 2

	Factor 2
Place adverbial	0.53
Present participial WHIZ deletion	0.39
Present participial clause	0.87
Public verbs	0.76
Adverbial past participle clause	0.37
Demonstrative pronouns	0.37
Indefinite pronoun	0.44
Past tense	0.87
Perfect aspect verbs	0.73
Synthetic negation	0.42
That clause on object position	0.70

Negative	
Past participle WHIZ deletion	-0.35

Rotated component matrix in factor 3

	Factor 3
Past part. Clause	0.43
Present tense	0.80
3rd person pronoun	0.71
Adverbs	0.40
Gerunds	0.40
That clause as adjective compliment	0.46
That clause an subject position	0.54
Other adverbial subordinators	0.49
Time adverbial	0.29
Attributive adjectives	0.70
Negative	
Past tense	-0.45

t, p: t and p values for **Student t-test**

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 4

	Fctor4
Phrasal coordination	0.62
Nominalization	0.81
Conjuncts	0.51
Existential there	0.48
Hedges	0.62
Wh relative clause on object position	0.47
Wh relative clause on subject position	0.45
That clause as a verb compliment	0.61
Negative	
Synthetic negation	-0.38

t, p: t and p values for **Student t-test**

*: Statistically significant at $p \leq 0.05$

Rotated component matrix in factor 5

	Factor 5
Suasive verbs	0.50
Word length	0.36
1st person	0.29
Demonstratives	0.87
Non phrasal coordination	0.47
2nd person	0.34
Nouns	0.45
Past participle WHIZ deletion	0.50
Concessive subordination	0.47
Negative	
Nominalization	-0.36
Time adverbials	-0.31
3 rd person pronoun	-0.39

B.1 Features not in the Arabic Results

Features not in the Arabic Results	No. English	No. Arabic (manual)
Pronoun it	3705	3705
Be as a main verb	14840	5687
Do as a pro verb	19400	-
Subordination that deletion	57113	-
Split infinitives	5567	1000
Possibility modals	20300	20300
Necessity modals	5950	5950
Prediction modals	20300	20300
Contractions	15100	-
By passive	20276	706
Agentless passive	8167	1700
Predicative adjective	36033	1800

The coming section deals with the features that *Stanford* program could not identify in the Arabic language .All the features encountered in this table are counted manually by the researcher herself; only 12 features are counted manually and this is done because the computer program, *Stanford*, could not identify these features due to the complex nature of the Arabic language.

The first feature is the modal verbs. Necessity and possibility modals are used either as an explicit marking of the writer’s own point of view or as an argumentative discourse designed to persuade the addressee. In addition, the necessity modals are pronouncements concerning the necessity of certain events and the possibility modals are pronouncements concerning the possibility of certain events occurring. Suasive verbs and conditional subordination act as an alternative for the prediction

modals in Arabic. They imply intensions to bring about certain events in the future while conditional subordination specifies the conditions that are required for certain events to occur.

For example:

22- But knew I shouldn't waste the milk.

كنت اعلم انني لا يجب ان اهدر اللبن

23- He should have not stayed.

كان لا ينبغي عليه البقاء

The above examples illustrate the variety of positions in which the English negation can occur. In all cases, the Arabic counterpart immediately precedes the verb. English usage will seem extremely random and complex to the Arabic-speaking student. Modals present a variety of problems to the Arabic translators of English since modals as grammatical classes do not exist in Arabic. Their meanings are conveyed by particles, prepositional phrases, and unmodified verbs.

For example:

24- She can use the telephone.

كان في استطاعتها ان تستخدم التليفون

Moreover, "can" can be rendered in Arabic as a prepositional phrase. In most cases, such a verb or prepositional phrase precedes a nominalized /?an/ clause.

Arabic translators are not familiar with vowel reduction as it occurs in English, and are likely to use the full form in all cases as the Arabic language does not allow the contraction technique in its characteristics except in some forms as:

31-

صلى الله عليه وسلم (ص)

As for the passives and other past participial clauses, they are used to emphasize abstract conceptual information over more concrete or active content. Usage of the passive form with (was/ were/ is /are) in English is totally neglected in the Arabic translation. The passive form does not exist in the Arabic language. It is not used as we drop it in the translation of the Arabic language.

For example:

32- She was cut down and taken away .

تم انزالها وحملت للداخل

Verb to be in "taken away" has been omitted by the translator and has been used only the past participle of the verb as it has been used in the first verb (was cut).

33- People are dead now.

انهما ميتان الان

The translator avoided the passive structure to change the sentence into a nominal sentence and he omitted verb to be from the sentence. He also translated /people/ as /انهما/

34- I was struck by that.

ادهشني ذلك

The translator changed the passive sentence into an active one so it can be more vivid to the reader.

35- My brothers were not bothered by any of this.

لم يكن أخواي يئزعجان

The translator changed the passive sentence into an active one. He used /lam/ as a particle to change the verb in the present to give the meaning of the past. Predicative adjectives are proceeded by a linking verb to be. Some verbs have a verb (1) and verb (2) forms. Verb (1) being an intransitive linking verb meaning seemed or appeared followed by a predicative adjective.

For example

36- Father was polite.

كان والدي غايه في التهذيب

37- That kind of life is dreary.

هذا نوع كئيب من الحياه

The verb to “be” in the present tense is neither used in the Arabic language nor verb to do. Verb to do expresses about the past simple tense /did/ or it expresses about the absence of the third person /does/

For example

38- He does not care.

لا يهتم

In this sentence “does” is used because of the word “he” and to remove the “s” of the verb to express the continuous tense with the pronoun.

39- I figured out that he didn't mind people doing new sorts.

ادركت انه كان لا يكثرث بقيام الناس بأشياء جديدة

Here “did” is used to express that the verb is in the past.

That deletion, while that can be dropped in English, /?anna/ should be always retained in Arabic.

For example:

40- Just what is it you are famous for?

كنت اود ان اسألك: بم انت مشهوره؟

“That” is dropped in the sentence but /?anna/ is not. Arabic uses relative nouns that need to agree with the head noun in case, gender, and number. The verb to “be” in the present tense is neither used in the Arabic language nor verb to do. Verb to do expresses about the past simple tense /did/ or it expresses about the absence of the third person /does/

[1] has stated that there is no neuter pronoun in Arabic, that is to say, pronoun “it”. It uses only feminine and masculine gender. The English impersonal it has no counterpart in Arabic. These independent pronouns as claimed by [20] can function as a subject of a verb, a subject or a predicate of a verb less sentence and as a copula. The English pronoun it has no counterpart in Arabic. For example:

41- They washed and combed it beautifully.

وكن يغسلن و يمشطن شعره علي نحو رانع

42- She could pay it back.

حتى يستطيع دفع ما يقابلها له

43- It was raining outside.

كانت تمطر بالخارج

As shown in these examples pronoun “it” is translated like the 3rd person pronoun because of the nature of the language itself. That is why the 3rd person pronoun is significantly larger in weight in Arabic than English. Concerning “be” as a main verb, it is typically used to modify a noun with a predicative expression instead of integrating the information into the noun phrase itself. Be as a main verb is omitted in the translation thus changing the English verbal sentence into Arabic nominal ones. That is to say into a topic and a comment. When (am, is, are) are used as main verbs, their sentences are nominal in Arabic. Therefore, they are deleted completely in Arabic. The past tense of (be, have) are translated into verbal sentence in Arabic and this is more effective in delivering the message. For example:

44- My father was not religious.

لم يكن ابي متدينا

In other cases the translator needs to change the phrase to verbal sentences and to remove verb to be.

45- At the end of the yard is a small barn.

يوجد في نهاية الفناء مخزن صغير

There is another difference in translating the following sentence:

46-They are dead now

انهما ميتان الآن

It is a passive sentence and that is why here in translating passive sentences we can use nominal sentences affirmed with "ان". In a sentence like:

47- Both of my boys were in school.

ولداي يذهبان إلى المدرسة

In this example verb to be is omitted and the word/ both/ is deleted and the duality is shown in the word "كلتا". We can conclude here that it shows that the plural in Arabic is changed into dual.

7 CONCLUSION

In view of the paper presented here, linguistic variations are considered as a field of study that requires further analyses based on the use of corpora and the refinement of parameters in register description. Obviously, register variation research has immediate applications to foreign language teaching and intercultural communication, and this type of perspective that the field offers should attract scholars and communication practitioners. Biber’s model is only dealing with the morphology and the syntax of the language. More models are needed to combine the structure and the ideology together. This sort of descriptive study is greatly facilitated by the availability of tools of corpus linguistics. The *Stanford* program used in this investigation is user-friendly and has proved very practical as an aid to human analysis of a whole text. The tagging could be grammatical (to look more closely at clause beginnings or shifts from noun to verb), functional (such as analysis of Transitivity patterns) or stylistic (the highlighting of the occurrence of particular lexical fields, an author’s favorite constructions, words with positive and negative connotations).

References

- [1]Affendi, A. (2011). A contrastive analysis between Arabic and English relative pronouns. Faculty of Education: Salatiga. Retrieved from <https://docplayer.net/39367735>.
- [2]Buránová, E., Cová, E. & Sgall, P. (2000). Tagging of very large corpora: Topic-focus articulation. In Proceedings of the 18th conference on Computational Linguistics (Coling), 1, 139–144, Saarbrücken: Germany.

- [3]Biber, D.(1988). *Variation across speech and writing*, Cambridge: Cambridge University Press.
- [4]Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*, Cambridge: Cambridge University Press.
- [5]Biber, D., Connor, U. & Upton, T. (2007). *Discourse on the move: using corpus analysis to describe discourse structure*, Amsterdam: John Benjamins.
- [6]Biber, D. & Conrad, S. (2009). *Register, genre and style*, New York: Cambridge University Press.
- [7]Biber, D. (2012). *Register as a predictor of linguistic variation. Corpus linguistics and linguistic theory*, 6 (1), 9-37.
- [8]Broeck, R. (1986). Contrastive discourse analysis as a tool for the interpretation of shifts in translated texts. In House, J & Blum Kualka, S (Eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, Tübingen: Gunter Narr P 37-47.
- [9]Crystal, D. (1992). *Introduction basic linguistics*. England: Penguin Books Ltd.
- [10]Fisiak, J. (1981). Some introductory notes concerning contrastive linguistics. In J. Fisiak (Ed.), *Contrastive linguistics and the Language Teacher* (pp. 1-11). Oxford: Pergamon.
- [11]Doorslaer, L. (1995). Quantitative and qualitative aspects of corpus selection in translation studies. *Target*, 7 (2), 245-60.
- [12]Gile, D. (1994) *Beyond testing towards a theory of educational assessment*, London: Falmer Press.
- [13]Gorsuch, R. (1983). *Factor Analysis* (2nd Ed.). Hillsdale, NJ: Erlbaum.
- [14]Hornby, A. (2005). *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.
- [15]Khalil, A. (1993). Arabic translations of English passive sentences: problems and acceptability Judgements, *Papers and Studies in Contrastive Linguistics*, 27, 169-81.
- [16]Khafaji, R. (1996). Arabic translation alternatives for the passive in English, *Papers and Studies in Contrastive Linguistics*, 31,19-37.
- [17]Krings, H. (1986). Translation problems and translation strategies of advanced German learners of French (L2). In House, J. & Blum-Kulka, S. *Interlingual and Intercultural Communication: Discourse and Cognition in Transition and Second Language Acquisition Studies*. Tübingen: Gunter Narr Verlag, 263-276.
- [18]Mitkov, R (ed.).(2015). *The Oxford hand book of computational linguistics*. 1st ed. Oxford University Press: Oxford
- [19]Olohan, M. (2004). *Introducing corpora in translation studies*. Taylor and Francis: London.
- [20]Ryding, K. (2005). *A reference grammar of modern standard Arabic*. Cambridge: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511486975>
- [21]Simon, S. (1996). *Gender in Translation: Cultural Identity and politics of Translation*. London and New York: Routledge.
- [22]Teubert, W. 2002 "The Role of Parallel Corpora in Translation and Multilingual Lexicography". In *Lexis In Contrast*, B. Altenberg and S. Granger (eds.), 189 – 214. Amsterdam: Benjamins.
- [23]Teubert, W. 2005 "My Version of Corpus Linguistics". *International Journal of Corpus Linguistics*. 10(1): 1–13.
- [24]Towell, R. & Hawkins, R. (1994). *Approaches to second language acquisition*. Clevedon: Multilingual Matters.

BIOGRAPHY



Elsaadany is a former English instructor in the Arab Academy for Science and Technology. Furthermore, she is a speaking examiner for IELTS and OET exams. She has attained her PH-D degree in Translation studies from the Institute of Applied Linguistics and Translation (2018), Faculty of Arts, Alexandria University. She has also attained her Master's degree in Applied Linguistics in (2014). Her main areas of interest are applied linguistics, computational linguistics and computational studies in the field of translation. She is also the head of an educational center for teaching the CIPP program.



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

مقترح أداء لقياس التغيرات اللغوية: دراسة مؤسسة علي مدونة لغوية

مرام السعدني

معهد الدراسات اللغوية و الترجمة - كلية الآداب - جامعة الاسكندرية - مصر

Maram.elsaadany@gmail.com

سامح الأنصاري

مكتبة الأسكندرية، أسكندرية، مصر

قسم الصوتيات و اللغويات، كلية الآداب، جامعة الأسكندرية، اسكندرية، مصر

sameh.alansary@bibalex.org

ملخص

هذا البحث هو تحليل لترجمة للهياكل المورفولوجية العربية والإنجليزية باستخدام نموذج بيير (1988) وبرنامج ستانفورد وهي عبارة عن دراسة كمية تستند إلى مجموعة من الخصائص اللغوية و تستخدم 66 ميزة من بين 67 سمة تم تحديدها من قبل بيير ، ويتمشى البحث مع نموذج بيير والإجراء الإحصائي. إن المجموعة المختارة لهذه الرسالة هي مجموعة قصص أليس مونرو القصيرة (1985) *The Power of Love* وترجمتها إلى العربية مسيره الحب (2015) لمحمد طنطاوي. يتم حساب جميع التغيرات اللغوية في اللغة الإنجليزية والعربية بمساعدة برنامج كمبيوتر ، ستانفورد (2015). ستانفورد هو برنامج يستخدم لتلخيص مجموعة النصوص المختارة ، وهو يعمل على مستويات اللغة الإنجليزية والعربية morpho-syntactic. يتم تصنيف 66 وظيفة إلى أربعة عوامل للغة الإنجليزية وخمسة عوامل للغة العربية. يتم حساب اثني عشر وظيفة يدوياً في التحليل العربي بواسطة الباحثه بنفسها. تعكس النتائج اختلافات كبيرة بين اللغتين حيث لا يمكن تحديد بعض الميزات بواسطة برنامج الكمبيوتر.

A Morphological Analyzed Corpus for Egyptian Child Language

Heba Salama, Sameh Alansary

Phonetics and linguistics Department, Faculty of Arts Alexandria University

Heba.salama.slp@gmail.com

Sameh.Alansary@bibalex.org

Abstract: The main focus of the present paper is on the construction of a MOR grammar for Egyptian children. We also introduce a morphological analyzer that was specifically developed for Egyptian children corpus. Morphological analysis is so crucial for child language studies. The corpus will be an instrument for future investigations of language development, child language acquisition theory and psycholinguistics in general. Egyptian children hand tagging of child transcript is time-consuming and error-prone, the MOR program for automatic analysis and part of speech tagging was introduced into the CHILDES system to address these problems. To date, MOR analysis programs have been constructed for 11 languages. Once a child language corpus has been automatically tagged by MOR, it is then possible to automate various systems for assessment and diagnosis of child language. The annotated corpora will support three different kinds of investigation: 1) Basic child language developmental research that examines the sequence of acquisition of grammatical morphemes and the various morphosyntactic processes of the language, 2) Computational modeling of child language acquisition, and 3) Diagnosis of language differences and disorders through methods such as automation of IPSYN or DSS scores. The corpus data contained 26,700 words, all the data are transcribed in CHAT. The paper discusses methods for preparing data for MOR analysis and developing MOR grammars. Describe the shape of rules for controlling Allomorphy and morpheme concatenation. The emergence of this new line of work represents a major step forward for Egyptian child language research. The analyzer adequately covers the entire corpus, producing detailed correct analyses for all tokens. Evaluation on a new corpus reveals high coverage as well. The result is a high-quality morphologically-annotated corpus of Egyptian children.

Key words: Morphosyntax, Part of speech tagging, Child language, MOR grammar

1 INTRODUCTION

The current paper shows the construction of a MOR system for Egyptian child language. A part-of-speech (POS) tagging system for Egyptian Arabic children. The morphological (MOR) tagging relies on the application of MOR program for automatic analysis and part of speech tagging in CHILDES database. The morphological tagging of corpora is important not only for research on morphosyntactic development but also for the development of automatic ways of evaluating children's level of grammatical development. Egyptian Arabic is rich of allomorphic patterns and complex morphology. Therefore, it demonstrates a particular interesting challenge to any system for automatic morphological analysis. Explaining child language acquisition is one of the most challenges facing cognitive science, linguistics and psycholinguistics. Acquisition of grammar is expressed through the measurement of morphosyntactic competence which is important in fields such as developmental language disorders, schooling and literacy, and second language acquisition. To test and validate theoretical predictions quantitatively, researchers have increasingly come to rely on large corpora of transcript data of verbal interactions between children and parents to examine the development of morphosyntax. CHILDES database [1] is a standard source of child language corpus data to investigate the development of morphosyntax. To date, MOR analysis programs have been constructed for 11 languages that do not include Arabic; the Arabic language do not have yet a MOR lexicons. Thus, one must create a system of part of speech using the automatic tools provided by CLAN program [2]. The CLAN software includes a language for expressing morphological grammars, implemented as a system, MOR, for the construction of morphological analyzers. There is a great need to develop such an automatic tools as a solution for the huge time required for hand annotation for child language transcript it took about 170 hours and it is a very time-consuming task. The Arabic language is characterized by much inflection; therefore, this system presents a challenge to any system for automatic MOR analysis. Part-of-speech (POS) tagging is a core natural language processing task that can benefit a wide range of processing applications. Tagging a child language has various outcomes to acquisition researchers for exploring instances, finding frequencies of particular constructions and searching specific usage. This tool serves language researchers by providing them with an automatic interface that enables quick and accurate analysis of child's language. The current paper focuses on the previously transcribed data on CHAT format from 10 children between the ages of 1;6 and 4;00 [3], cross sectional corpus. Methods for preparing data for MOR analysis, the shape of rules for controlling allomorphs, and morpheme concatenation to develop MOR grammars are generated. The project has several contributions: it presents a morphologically annotated corpus of Egyptian Arabic, with a set of tools that can be applicable to new corpora. Moreover, this is a research tool for future investigations of new perspectives in Egyptian Arabic language research, which will influence some areas such as first language acquisition, second language learning, and speech disorders.

We discuss the MORphological device and describe the shape of rules for controlling Allomorphy and morpheme concatenation. Method for lexicon building; ongoing effort on morphological disambiguation is the topic of Section 3. Finally, Section 4 outlines the ways in which the annotated database is already being used, and concludes with future work.

2 MOR DEVICE

We developed a morphological analyzer for the Egyptian child language in complying with MORphological grammars [4] for 11 languages. The CLAN software includes a language for expressing morphological grammars, implemented as MOR program by applying a MOR grammar to a corpus, a new tier below each main tier is created, %mor tier. The morphological information for each item in the main tier is listed that gives the surface representation of concatenated linear morphemes (stems + affixes) and lexical information attributed to the surface token, as shown in Fig. 1.

```

@Transcriber: Investigator
@Date: 05-MAY-2003
@Time Duration: 00:30:00
@Time Start: 01:00:00 02:00:00
@Activities: Asking Question, Naming Obejects, singing
@Media: Bilal audio
*CHI: baelloonae . •
%mor: n|baelloonae&f .
*INV: fiihae baelloonae ? •
%mor: prep|fii&f&sg-hae n|baelloonae&f ?
*CHI: ?uh . •
%mor: co|?uh .
*INV: +^ feen hijjae ?elbaelloonae ? •
%mor: pro:wh|feen pro:sub|hijjae&f ?el&def:moon:art#pfx|baelloonae&f ?
*CHI: ?aeheeh . •
%mor: pro:dem|?aeheeh&f&sg .
*INV: ?aeheeh we ?ee ?il fiihae di miin mae?ae ?elbaelloonae fessoora
henae ? •
%mor: pro:dem|?aeheeh&f&sg conj:coo|wet
pro:wh|?ee pron:rel|?il prep|fii&f&sg-hae|
24nov15[E|CHAT] * 28

```

Figure 1: Word length distribution in Arabic and English

The Egyptian analyzer (EgyMor) consists of three major components: first an a(llophone)-rules file where it is specifies various forms a root or stem can take, second a c(ontenation)-rules file which consists of higher level rules, allowing concatenation of the different categories and finally a set of lexicon files is organized by grammatical category (e.g., adverbs, function words, nouns, pronouns and verbs) and divided into several files that contain roots, stems, or whole words will described in the following sections. The use of MOR device varies in Different languages according to their requirements and needs. The Dutch grammar incorporate a short list of possible allomorphic changes and several concatenation rules; the grammar for Italian includes a more detailed A-rules file and a relatively short C-rules file; the grammars for Cantonese and Mandarin rely on lexicon files only; while Hebrew grammar as well as Egyptian Arabic extensively use all of these devices (bases and affixes) because of the changes of the stem allomorphs is handled within a set of A-Rules; and affixation possibilities are allowed (or restricted) via the C-rules [5].

A. Lexicon

The current lexicon includes over 4,000 entries, distributed across the different parts of speech as listed in table 1. Creating a lexicon is an open-ended task; we focused on adding the entries that occur in the corpora we had, and will continue to extend the lexicon as needed, given more corpora. Lexical entries are very simple. Words are entered into text files one word on each line in alphabetical order. The surface form comes first on the line, followed by the scat or syntactic category, some possible morphological features. For example, ?u^tt'a {[scat n][gen f]} "cat". The most complex entries are usually for verbs, as in this example, "?aeae" "said" {[scat v][ptn a][tense PAST][pers 3][num sg][gen m]}.

TABLE I
NUMBER OF LEXICON ENTRIES PER PART OF SPEECH

POS	Number of entries
NOUN	1730
Verb	1550
Adjective	500
Adverb	125
Preposition	25
pronoun	78
Total	4,008

B. A-rule

The first step in building MOR grammars is to determine the basic allomorphy types of Egyptian Arabic. Egyptian Arabic is characterized by rich and complex morphology (inflectional and derivational), therefore several systematic text books of inflectional and derivational patterns used. We follow the same in languages with more complex paradigm such as Spanish and Hebrew, allomorphy types is created based on the formal segments of the nominal and verbal paradigm. For example, the stems can target past, imperative. The construction of morphological analyses depends first on the generation of a runtime lexicon combined through the operation of rules of allomorphy (arules), as they operate on the items listed in the lexicon. The function of the arules is to increase the entries in the disk lexicon into a larger number of entries in the on-line lexicon. Subsequently Words that undergo regular phonological or orthographic changes when combined with an affix just need one disk lexicon entry. The arules are used to create on-line lexicon entries for all inflectional variants. These variants are called allos. The full set of generated allomorphs is stored in a trie structure [6]. Example of the allomorphy rule for noun suffix as the following:

```

RULENAME: n-sfx
LEX-ENTRY:
LEXSURF= $X$B
LEXCAT = [scat nsfx]
ALLO:
ALLOSURF = LEXSURF
ALLOCAT = LEXCAT, ADD [allo n0]

```

An arule consists of a header statement, which contains the rule name, followed by one or more condition-action clauses. Each clause has a series of zero or more conditions on the input, and one or more sets of actions. LEXSURF matches the surface form of the word in the lexical entry to an abstract pattern. The string \$X\$B is composed of variable declarations that characterize the noun suffixes to produce form as beet-u "his home" where X and B are a variable declaration indicate any string of consonant or vowels. Next line in the rule is ALLOSURF which is used to produce an output surface. For example, a lexical entry surface form such as "beet" "Home" is converted to "beetu" "his home" to serve as the stem of the noun. where ALLOCAT determines the category of the output allos. From the above example, The first allo (morph) produced by the rule is "beet" "home" will produce words like "beetu" "his home or "beetnae" "our home". The use of this lock-and-key allomorphy pattern matching mechanism is controlled by the Crules. The application of Arules must be ordered from specific to general rule pattern to control the lock-and-key matching system for feature-value pairs through careful documentation and control of the [allo] features and other grammatical feature-value pairs such as [gen]. For example, n|beet&m&sg-u . 347 Arule is applied in Egyptian Arabic is not too complex whereas in English with a simple morphology it is 391 lines and in Spanish it is 3172 [7]. After finishing building arule it should end with a default rule to copy over all remaining lexical entries that have not yet been matched by some rule. This default rule must have this shape.

```

% default rule- copy input to output
RULENAME: default
LEX-ENTRY:
ALLO:

```

C. C-rule

The purpose of the crules is to allow stems to combine with affixes. In these rules, sets of conditions and actions are grouped together into if then clauses, performed by two input START and NEXT. The word matches are determined by the START rules that only require that a morpheme match the syntactic category [scat] of the rule. After the first rule match creates a

candidate for the first few letters, MOR continues to take in letters looking for another morpheme match. Once a new morpheme fires, there can be lock-and-key process in which the STARTCAT and the NEXTCAT must match in terms of their allo features which is called the MATCHCAT process. Crules depend on direction from RULEPACKAGES statement. START and END rule should be for each part of speech. Here is an example of crule analyzed plural noun

```
RULENAME: n-start
CTYPE: START
if
NEXTCAT = [scat nsfx]
then
RESULTCAT = NEXTCAT
RULEPACKAGE = {n-reg-plural}
```

In the above example The STARTCAT is not defined, The NEXTSURF is the noun suffix that is attached to that stem then RESULTCAT = NEXTCAT lead all category information from the start input to be copied over to the result. The RULEPACKAGE identifies which rules may be applied to rule, when that result is the input to another rule. RULEPACKAGE are not tried until after another morpheme has been found. For example, in parsing the input "baellonaet" "ballons" , the parser first finds the morpheme "ballonae" "ballon" and applies the start rules thus the rule for noun will be fired. This rule includes a RULEPACKAGES statement specifying that the rule which handles noun suffix may later be fired. When the parser has further identified the morpheme "aet" the verb conjugation rule will apply, where "ballonae" is the start input, and "aet" is the next input. Applying the rule the analyzer associates this analysis n|baellonaet&f&pl-aet. This analysis is read as the follows N is the main category (noun) and f&pl indicate feminine plural. The crules on the contrary, of arule which are strictly ordered from top to bottom, crule is determined by their CTYPE and the way in which the RULEPACKAGES statement track words from one rule to the next. 456 crule lines are used in Egyptian grammar. Generally, not huge set of rules is required for producing a relatively extensive number of outputs (above 26,500 items are analyzed by the current version of the analyzer). Currently, 95.4% of all adult word tokens and 87% of all child tokens are analyzed. Out of these analyzed tokens, approximately 29.4% of all adult forms and 15.4% of all child forms are still ambiguous.

3 LEXICON BUILDING

The first step in building a grammar is to take a tour of the analysis of Egyptian morphological system [8][9] to have a good understanding of affixes, the classes of the stem Allomorphy variation and the conditions of choosing allomorphs. After that, we create a small lexicon of the most frequent words and choose to begin with adjectives because it is better to rely on one part-of-speech at a time and contains no morpheme. Then we move to another speech parts. Therefore, a related language rules from another languages with complex morphology like Hebrew and English rules files is studied well. Consequently, we create Arules and crules files for Egyptian language. We move from build up a lexicon of uninflected stems to inflected stem so affix.cut file is created which includes categories and allos for the affixes it matched with the stem during crule fire. For example, f {[scat npfx] [pcat OR n pfx]} "f|prep" in this example the preposition attached to a noun as a prefix. In the process of building to check up the application of ar.cut and cr.cut we type mor +xi*.cha in the interactive mode to analyze the word in MOR, if it is analyzed the rules is ok, if not we changed the rules. The unrecognized words are identified by running the command mor +xl*.cha. Example of MOR analysis is shown as follows in Fig. 2 and Fig. 3.

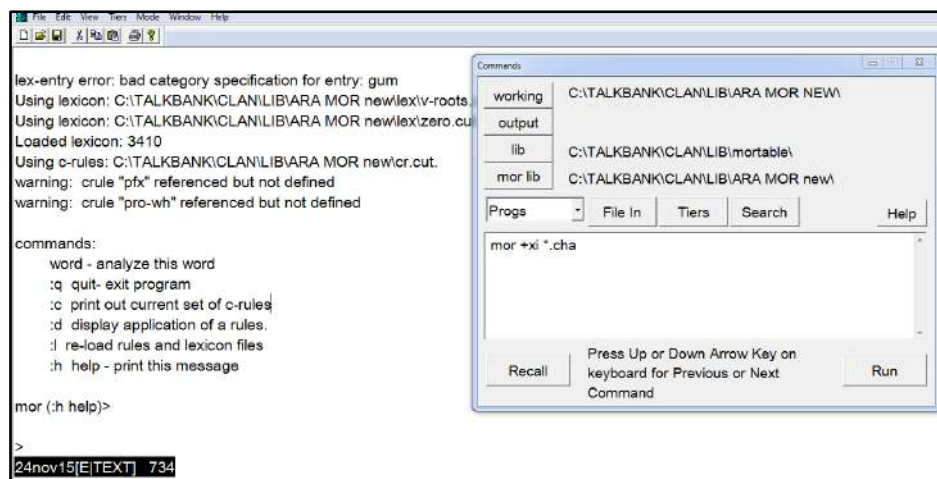


Figure 2: MOR command in interactive mode

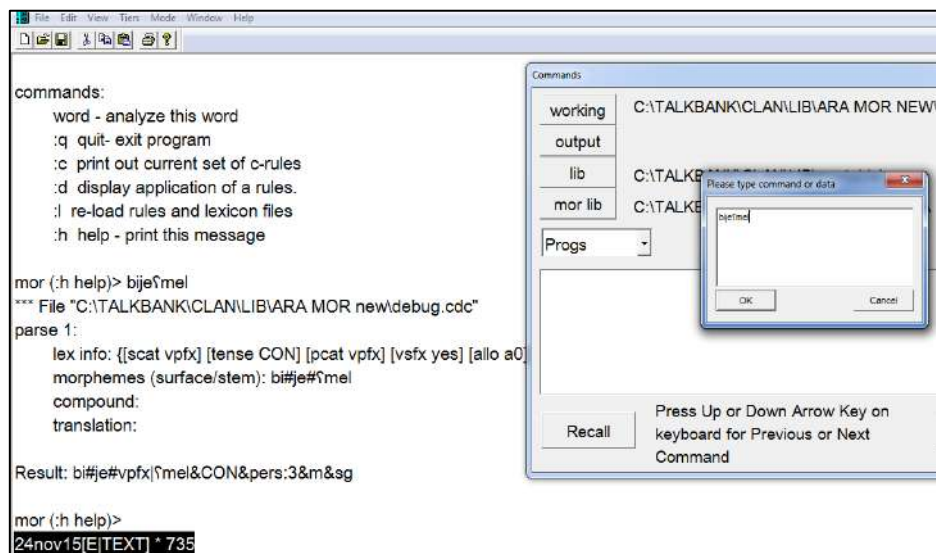


Figure 3: MOR analysis

There are 55 separate files for each of the 55 parts of speech that break out the possible words of Egyptian Arabic (and additionally, 190 affixes), distributed across the different parts of speech. After all rules are built mor*.cha command is running to insert morphological analysis layer in child transcript. The output of the MOR program is a new tier or line called the %mor line in which tags stand in one-to-one correspondence with words on the main transcript line. These files are not disambiguated. Words can receive as many as 2 different analyses or more, all concatenated with the caret (^) symbol. For example, analysis output for "ʕaemaelt "I did" is v|ʕaemaelt&PAST&pers:2&m&sg-t^v|ʕaemaelt&PAST&pers:1&sg-t "maeʕae" "with" prep|maeʕae^maeʕneg:bound|ʕa part of one word is part from another.

To achieve disambiguation of such combinations, the CLAN program offers two solutions 1) manual disambiguation tool called Disambiguator Mode in mode menu within CLAN. In this mode each ambiguous interpretation on %mor tier is produced in its alternative possibilities. The user chooses the correct option this is called manual disambiguation. This is a very time consuming process that involves many online decisions 2) Automatic disambiguation POST and POSTTRAIN programs where these automatic module requires a lot of settings such as database of disambiguation rules called post.db file that we will consider in the future work. We notice that ambiguity of our data is low for example in one file of child transcript there are 41 ambiguous words from 2620 words. Ambiguity arises especially with items that have the same orthographic like first and second person in past "kaetaebt" "wrote". A lot of morphological analysis such as frequencies of morphological categories and mean length of utterance can be performed which is valuable and import to study child language acquisition.

4 DISCUSSION

This paper describes the construction and usage of first computational systems for morphosyntactic analysis for Egyptian child language. The MOR grammar approximately covers our current corpus data. The main advantage of our building system is that it can analyze words containing 8 morphemes. For example, the result of analysis for "maebijaebulinaej]" "they did not give us" is mae#bi#je#vpx|aeb&CON&pers:3&m&sg-u-linae-f, 7 morphemes, Where the result of the word "maebijiktibuhaelhaej]" " they did not wrote to her" is 8 morphemes. mae#bi#ji#vpx|ktib&CON&pers:3&m&sg-uhae-lhae-f]. We face a lot of challenges:

- 1- Short studies deal with the description of Egyptian Arabic where this work should be guided by a good descriptive grammar of the morphology of the language therefore, we try to collect possible description from different books.
- 2- Error and replacement in child speech
- 3- Ambiguity solving

Future plans

Several issues still need to be resolved with respect to our Egyptian child corpus. First, the way to handle disambiguation is to train a part-of-speech (POS) tagger using automatic tools provided by CLAN. Such a tagger will be able to automatically select the most suitable analysis for every multiple-choice output for a given lexeme and thus saves the time. We are currently preparing training material for this task, and will use the POST program that is part of CLAN since it has worked to markedly reduce ambiguity for languages such as English, Spanish, and Chinese. The output of the POS tagger will serve as a basis for developing a statistical parser for Egyptian data of the kind that was developed for the English section of CHILDES and other languages [10].

REFERENCES

- [1]MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

- [2] (Sagae, K., Lavie, A., & MacWhinney, B. (2005, June). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 197-204). Association for Computational Linguistics. University of Michigan, USA
- [3] Salama, H., & Alansary, S (2015). Building a POS-Annotated Corpus for Egyptian Children. *The Fifteenth Conference on Language Engineering (ESOLEC'2015)* (Ain Shams University Cairo, Egypt.
- [4] Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition, 2000.
- [5] Brian MacWhinney. Enriching CHILDES for morphosyntactic analysis. In Heike Behrens, editor, *Corpora in Language Acquisition Research: History, methods, perspectives*, volume 6 of *Trends in Language Acquisition Research*. Benjamins, Amsterdam, 2008.
- [6] Fredkin, E. (1960). Trie memory. *Communications of the ACM*, 3, 490-499.
- [7] Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37, 705-729
- [8] Abdel-Massih, E. T. (2011). *An Introduction to Egyptian Arabic*. Ann Arbor, MI : Publishing
- [9] Omar, M. K. (2017). *The acquisition of Egyptian Arabic as a native language* (Vol. 160). Walter de Gruyter GmbH & Co KG.
- [10] Bracha Nir, Brian MacWhinney, and Shuly Wintner. A morphologically analyzed CHILDES corpus of Hebrew. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1487–1490. European Language Resources Association (ELRA), May 2010. ISBN 2-9517408-6-7.

BIOGRAPHY



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. She likes corpus linguistic field because it is more methodology that is powerful, scientific and open objective verification of results. The construction of database is very important in helping the researcher to manage the problem they faced and wishes to test a detailed theoretical prediction on naturalistic samples.

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.

بناء مدونة لغوية محللة صرفيا في عربية الاطفال المصريين

هبة سلامة - سامح الانصاري
كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية

ملخص

تهدف الدراسة الي بناء مدونة محللة صرفيا للأطفال المصريين،حيث أن التحليل المورفولوجي مهم للغاية في دراسة لغة الأطفال وستكون هذه الاداة في عمل أبحاث مستقبلية عن نظرية اكتساب الطفل للغة وعلم اللغة النفسي. يستغرق عمل المحلل الصرفي الكثير من الوقت والجهد والأخطاء لذلك فان انشاء محلل صرفي ألي MOR الذي سوف يساعد في التغلب علي هذه المشاكل. حتي الان تم بناء برنامج MOR ل 11 لغة لا تتضمن العربية المصرية.ويجب انشاء محلل صرفي باستخدام الادوات الموجودة ببرنامج. بمجرد أن يتم عمل محلل صرفي تقوم العديد من البرامج التحليلية التي تعمل علي تقييم وتحليل لغة الأطفال.تساعد المدونات المحللة صرفيا 3 تحليلات اساسية : (1) تحليل تطور النحو الصرفي للأطفال.(2) النمذجة الحاسوبية لإكتساب الطفل للغة. (3) تشخيص الإختلافات والإضطرابات اللغوية من خلال أدوات برنامج CLAN. تحتوي المدونة علي 26,700 كلمة تم إعدادها باستخدام CHAT. وتتناول الدراسة طرق إعداد المحلل الصرفي والقواعد المستخدمة ويمثل ظهور هذا الاتجاه من البرنامج خطوة كبيرة للأمام في بحث اللغة المصرية للأطفال.

Arabic Educational System Using Augmented Reality

Marwa Elgamal, Salwa Hamada, Reda Aboelezz, Mohamed Abou-Kreisha

AASTMT, Computer Engineering Department, Cairo, Egypt

Eng.marwa.m.m@gmail.com

Al-Azhar University, System and Computer Department at Faculty of Engineering, Cairo, Egypt

Reda2018aboelezz@gmail.com

Al-Azhar University, Mathematical Department at Faculty of Science, Cairo, Egypt.

drkresha@gmail.com

National Research Institute, Cairo, Egypt

hesalwa@hotmail.com

Abstract: Most users in Arab countries for smartphones, tablets, and laptops are children and young people. Their passion for modern technology and their rapid learning to use these smartphones and electronic devices to play games and chat with their friends and watching videos draws attention. In the meantime, less of them use their smartphones in the education or learning process to solve their homework or to learn skills and gain additional information about their scientific topics. While all over the world, digital educational programs have become like audio or videos available for all ages. Children all over the world use smartphones to access books, lessons, exercises and various scientific experiences. In recent years due to the rapid advances of wireless and mobile technologies, the AR entered educational applications such as Mathematics, Biology, Anatomy and chemistry. But research in using AR in Language learning is still limited. In this paper, we will describe how to use a marker based algorithm for Augmented Reality (AR) technology in Arabic language learning. AR will be introduced clarifying how it can be used to enhance the education systems and make it more actively to be used in the education process and to make a task that hard to understand easier. It shows how will AR can change the education process completely in the future and turn it from being a memorization and routine learning to be a journey of discovery, prediction and an active learning. The Arabic Educational System provide an educational system that gives the student a chance to learn Arabic letters and explained by simple words and presented by AR to add a virtual scene as a three-dimensional object in the front of the user. The difficulty in the Arabic language is that Arabic is a cursive language, written from right to left each character have different forms. This system used and tested with a group of young learners to teach Arabic letters and simple words for children. The goal is to use it in the future in educational books for students from age four to twelve years old. The system will act as an interactive and enjoyable learning process for students.

Keywords: Augmented reality AR, Virtual Reality VR, Arabic language learning, AR, and Edutainment.

1 INTRODUCTION

Augmented Reality is a new amazing field of computer research that combines actual scenes viewed by the user, and virtual scenes generated by the computer. This augments the scene with additional information Since the evolution of technologies and information increase kid's skills, Augmented Reality (AR) and Virtual reality (VR) are amazing technologies to be used in education. It gives the child a new experience in studying and learning with interactive way. AR may help in various ways to give students extra digital information about any subject and convert complex information to an easier one to understand. Currently, we may add some examples of Augmented Reality in education usage worldwide. Ability to connect reality and digital content has become true and opens more options for students with rapid growth in AR technology. The aim of this paper is to design a system that use AR in Arabic language learning system for kids from age four to six which is the hardest age because they start to enter a new stage in their lives which is learning how to read and write their language if they Arab students or another language if they are native students. We want to make the learning process more effective and interesting for the beginners at the same time. The system starts with designing the Arabic letters each letter will be written in a card in front of the user when the user scanned the card by AR device the three-dimensional object which describes the letter will be presented. a support video and audio will be Augmented in front of the user too. The user will see the Arabic letter and an object start with the letter and hear the sound of it and the sound of the object if the object is an animal starting with this letter. When educational or school books supported with AR technology students will become active learners and able to interact with their learning environment.

2 REVIEW USING AUGMENTED REALITY in EDUCATION

Augmented Reality enhances a user's perception of and interaction with the real world. The virtual objects display information that the user cannot directly detect with his senses. The information conveyed by the virtual objects helps a user perform real-world tasks.[1]

As Augmented Reality did not become widely used in education process especially in Arabic countries there are some computer-generated simulations of historical events, exploring and learning details of each area of the event site could come alive [1]. On higher education, there are some applications that can be used. Construct3D, a (Studierstube) system, allowed students to learn mechanical engineering concepts, math or geometry.[2] Chemistry AR apps allowed students to visualize and interact with the spatial structure of a molecule using a marker object held in a hand.[3] Anatomy students could visualize different systems of the human body in three dimensions.[4] Augmented reality technology enhanced remote collaboration, allowing students and instructors in different locales to interact by sharing a common virtual learning environment populated by virtual objects and learning materials[5]. Primary school children learn easily from interactive experiences. For instance, astronomical constellations and the movements of objects in the solar system oriented in 3D and overlaid in the direction the device was held and expanded with supplemental video information. Paper-based science book illustrations could seem to come alive as video without requiring the child to navigate to web-based materials. For teaching anatomy, teachers could use devices to superimpose hidden anatomical structures like bones and organs on any person in the classroom [4] [6].In the table(1)a preview of some projects of using AR in education from 2011 to 2016.

TABLE (1)
OVERVIEW of USING AR in EDUCATION

Author	Domain	Propose of AR Use	Year
-Chang et.al.	-Medical education (surgical training)	To provide training and guide the surgical procedures.	2011
-Yean	-Medical education (Anatomy)	To teach and test anatomy knowledge.	2011
-Singal et.al.	-Chemistry education	To provide an efficient way to represent molecules.	2012
-Fleck Simon	-Astronomy	To show Augmented views of celestial bodies.	2013
-Shoudong Wang	-Language Learning	To provide an efficient way to learn a second language	2017

3 DIFFERENCE BETWEEN VR, AR and MR

A. Virtual Reality (VR)

It is a technology that enables the user to experience access to unique virtual worlds, The Virtual Reality term used to describe a three-dimensional, computer-generated environment which can be explored and interacted with by a person. That person becomes part of this virtual world or is immersed within this environment, the user able to manipulate objects or perform a series of actions. Such as an astronaut or fighter in a battle in his small room and using his own glasses [7], like in Figure (1).



Figure 1: Virtual Reality HMD Glasses [7]

The basic elements of the virtual reality experience are hardware and software, where devices can support virtual reality where the user viewing the virtual scene through software that simulates video games. The main difference that the user

will be a part of the game events. The user can enjoy the experience via the VR headset, and load virtual reality applications into the VR app to enter an integrated experience [8].

B. *Augmented Reality (AR)*

It is an experiment where the user can see two-dimensional or three-dimensional images or videos in the user's environment through AR devices (smartphone or laptop screens, AR glasses), where these imaginary scenes are merged with the viewer reality to create the composite presentation between reality and imagination as in Figure (2). or in another term, AR termed Mixed Reality (MR) which refers to a multi-axis spectrum of areas that cover Virtual Reality VR, and AR [9].



Figure 2: Augmented Reality [9].

AR requires software that supports the operation of this technology and hardware devices that helps the user to see the augmented scene through it. The Augmented reality itself is or software application designed by developers, but it is a technology that integrates with reality and adds the imaginary object in the interviewer real environment [8], [10].

C. *Mixed Reality(MR)*

Mixed Reality (MR) brings together real world and digital elements. In mixed reality, you interact with and manipulate both physical and virtual items and environments, using next-generation sensing and imaging technologies. Mixed Reality allows you to see and immerse yourself in the world around you even as you interact with a virtual environment using your own hands—all without ever removing your headset. It provides the ability to have one foot (or hand) in the real world, and the other in an imaginary place, breaking down basic concepts between real and imaginary, offering an experience that can change the way you game and work today[11].

4 SYSTEM ANALYSIS

A. *Display*

In order to integrate the real environment with the virtual environment, it is necessary to use an (AR) application with a device that can improve the user's reception and the interaction with the application. There are three classes of devices that be used to display the (AR) scene such as:

1) *Video See-through AR*

Which the (AR) can be seen on the screen through a video camera.

2) *Monitor-Based AR*

The (AR) can be seen on the screen of portable devices such as mobile phones, tablets or laptop computers. Through the video camera of the mobile device called the projection display, the visual information is dropped directly on the physical purpose to enhance it and requires a camera to display the enhanced objects. Real world environment and the front screen to display enhancements such as information highlighted by Augmented reality markers.

3) *Optical See-through AR*

Which can we see (AR) through the glasses placed on the head dedicated to this task is also known as the screen close to the eye is a device worn by the user on the head called head-mounted display , it is an excellent tool for (AR) The screens can be placed on the head to move the view as close as possible to the user's eye where the user realizes the virtual environment in the real world and at the same time allows the user to walk in the real world, In this paper we will use the first and second classes to view the AR objects. Examples of these devices are shown in Figure (3) [10].



Figure 3: AR Display Devices

B. Markers

Augmented Reality in the field of Education can be implemented in many ways like Marker Based, Geo Location Based, Skeletal based [11]. But whatever is the way of implementing the method is too similar. All the mentioned methods before differing in the inputs like markers, voice but after getting the input and detecting the input of the user obviously projects a virtual world which interacts with the user. In this paper, we going to use the marker based augmented reality. Marker-based AR has been discovered a decade ago. There are different tools available for marker-based (AR) where the model of implementation is the same and simple. There are some markers in the dataset and each marker associated with some virtual world interaction, once the application starts camera scans for the markers once the marker is detected with the help of some image recognition techniques, the virtual world will be projected and the user can interact with it. Two types of markers will be used in this paper and they will be explained in section 8.

C. Mobile System

Modern mobile computing devices like smartphones and tablet Computers contain a Camera and MEMS sensors such as accelerometer and GPS. (AR) merges the three components (display, markers and mobile power) into a highly portable unit Research in the fields of computer vision, computer graphics, and user interfaces are actively contributing to advances in augmented reality systems[12]. A special code is incorporated within a print card and users place this card in front of their webcam. The software recognizes the code and activates a reaction which could be in the form of a 3-D modelling of a product. It turns out that a 3D object is standing on a card which users hold up to their webcams. AR allows superimposing 3D-content/products into the real world. AR allows interaction with 3D-content in real time also [13].

5 ARABIC EDUCATIONAL SYSTEM STRATEGIES USIN AR and SYSTEM REQUIERMENTS

The learner in Kindergarten is ready to receive simple, interesting, exciting information that satisfies his imagination and enthusiasm for play. Speaking about Arabic Vocabulary Concepts, Arabic language is one of the di cult languages to learn its characters are not similar to any other language. In this paper the Arabic Alphabet can be taught through pictures and letters as follows:

Each letter is related to a certain animal or bird name that starts with it. The marker has a picture to the animal or QR code with the alphabet to help kids to choose the needed one and the letter added to the marker with character pronunciation.

Adding videos which contain pictures, some details about the animals and some stories also to learn the letter and memorize the concepts without learning how to write it. This is because the purpose of this system is to teach the child the alphabet, to enrich his Arabic language with basic words addressed by Arabic books used to teach kids, and to make them learn Arabic with entertainment.

A. Software Requirement

For software, augmented reality requires a much more sophisticated artificial intelligence and 3-D modelling applications [18]. It also needs; Image processing like (Opens), Graphics like (OpenGL), Audio (Penal) and Toolkit like (OpenSpace3D) using ARToolKit library [14]. For efficient product, the augmented reality improved featured and integrated the ARToolKit library for image detection and tracking (Natural Feature Tracking) into OpenSpace3D.The ARToolKit technology is recognized as the most popular augmented reality library.

In Addition, AR can be done by using different free as well as commercial tools, OpenSpace 3D is one of such tools which is developed by I-maginer. OpenSpace3D provides different options like Face Tracking, Marker Tracker, Skeletal Tracking which is the main part of Augmented Reality and OpenSpace3D provides a lot of 3D meshes which were the part of Virtual World and supports different smart devices and sensors like Kinect, MYO, Leap Motion and many more.

B. Hardware Requirement

In terms of hardware components, we need the following requirements at any AR application:

- *Display*: In this paper, we used two ways for displaying the AR scene the first one: web camera to capture video of the real world and sends it to the computer and a computer or laptop plasma screen which works to display the virtual objects on it. The second way to display the AR scene by a portable device such as a smartphone or tablet that has its own camera to capture the Marker and display the AR scene in the mobile screen.
- *Marker*: in this paper two kinds of markers will be presented QR code and picture markers, QR code (abbreviated from Quick Response Code) is the trademark for a type of matrix barcode (or two-dimensional barcode) first designed for the automotive industry in Japan. A barcode is a machine-readable optical label that contains information about the item to which it is attached. A QR code uses four standardized encoding modes to efficiently store data; extensions may also be used. QR is two-dimensional black and white squares that have a specific pattern. It is used to display the virtual objects above it. When it moves, the virtual object should move with it and appear exactly aligned with the marker. Another way to use markers will be with pictures of objects that will be presented in augmented reality.
- *Input Hardware Devices*: devices which help us input values to the computer such as mouse and keyboard.

6 THE ARABIC EDUCATIONAL SYSTEM DEFINITION and DESIGN

Arabic educational system presented in this paper. It is a teaching system with Marker-based AR algorithm, It concerned with teaching the meaning of concepts in the Arabic language. This paper gives a case study to teach 4 concepts by their 3D image and sound. They are the rabbit Alef (Arnab), owl baa (boma), tortious seen (solahfah), dog kaf (Kalb), See the appendix figures.

A. System Operation

1) Software Environment

The system was developed using Marker-based Algorithm with an OpenSpace3D tool for creating objects scene, the object loaded from ARtoolkit library.

2) System Markers Cards

In this paper, two kinds of markers will be presented to be used in Arabic language learning. One of them is using the QR code as shown in figure (4). Another way will be used in this paper is pictures of the objects as shown in figure (5) we have the picture for the rabbit to present Alef character, Owl picture for character Baa, Dog to present Kaaf character, a Tortious picture for the character seen.



Figure 4: QR Markers used in the paper



Figure 5: Image Markers used in the paper

B. System Analyses

OpenSpace3D is able to perform this camera tracking in real time, ensuring that the virtual objects always appear overlaid on the tracking markers. Figure (6) illustrates a summarization of this process steps. This process is explained below.

1) Designing 2D marker

Design the special pattern markers for 2D objects to be placed on the marker during the video shooting and it is available in the OpenSpace3D software. We can select and design the required marker by black and white squares or by uploading images.

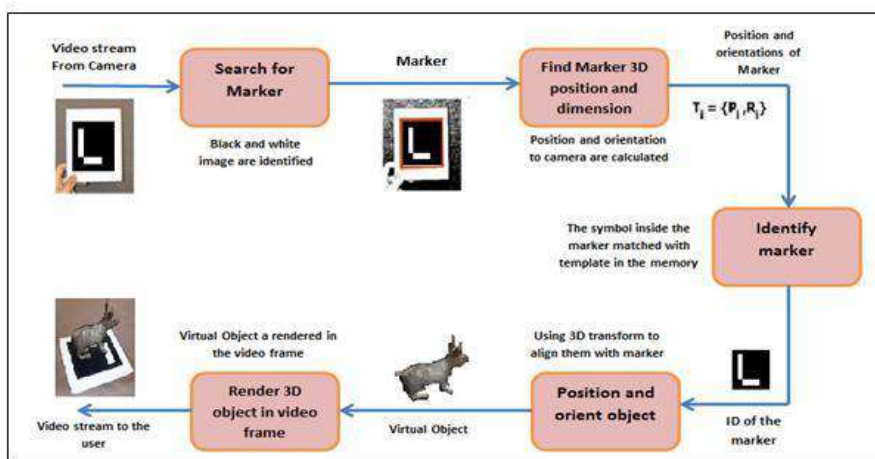


Figure 6: System Flow

2) Image Accusation and Representation (Binary Matrix)

The camera captures video of the real world and sends it to the computer developing an efficient algorithm to calculate the objects positions and orientations in the video.

3) Localization of marker (searching and matching)

Software on the computer searches for any square shapes through each video frame in the system library database. If a square or image is found, the software uses some mathematics to calculate the position of the camera relative to the black square.

4) Homogeneous transformation

Once the position of the camera is known as a computer graphics model is drawn from the same position.

5) Virtual object placement on the marker

This model is drawn on top of the video of the real world and so it appears stuck on the square marker.

6) 3D generation (virtual object generation)

The final output is a virtual object which is an animal or bird that placed on the real world (marker).

7 SYSTEM IMPLEMENTATION and TEST RESULTS

The appendix presents some snapshots during the system running to present the results of the Marker-based portable display based AR. The system contains four objects each with three phases in the case study;

A) Phase o

It represents how a kid can understand the dog, tortious, owl and rabbit words. In this phase, the kid sees these animals in front of him the animal first letter and its name and sound presented in front of the child.

B) Phase two

It illustrates how a kid can learn the letters and words and start to make a sentence with more than one word.

C) Phase three

It shows how kids can know each Arabic letter sounds and remembering the letter when seeing the animal that starts with this letter the kid can listen to animals sound and a video can be added to give a summary about the animal life.

We present our results of Arabic Educational System to twenty students from age four to twelve years old, and we asked them if they were satisfied with the lesson. The results are shown in the figure to the question: "Why did you like the lesson? ", the student like most real objects and real experiments (12), they like that the lesson different than every day (9), and they said that the learning was fun(18),Additionally the students asked if the lesson raised their interest in learning the Arabic language most of them answered positively . The next graph shows the reviews result.

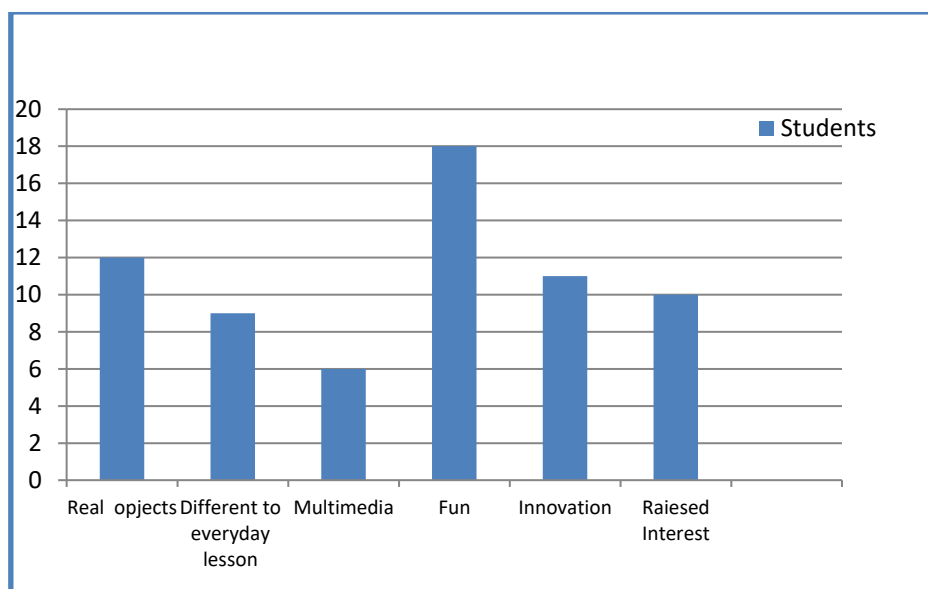





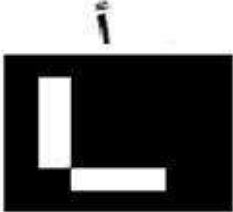




Figure 7: Students Evaluation Results

8 CONCLUSIONS

The process of education is interactive, entertaining and amazing by using Augmented Reality technology. The learner, in any stage, spends a lot of time studying new lessons without getting bored. This paper present marker based algorithm in AR technology, its applications and how to use it in the educational process. This is done through an AR education system that teaches letters and word concepts. We can design all Arabic alphabets using OpenSpace3D to design and implement Arabic characters and words to give kids from age four to twelve years an interactive learning using AR in Arabic language learning. The Arabic educational system is tested on twenty students from age four to twelve years in kindergarten and primary school; they are impressed, interacted with it and also enjoyed. The student like most real objects and real experiments, some of them like that the lesson different than every day but most of them said that the learning was fun.

Our aim in future work is to enhance the Arabic scripts recognition then designing a system that translates Arabic educational and storybooks to enjoyable books using AR technology which give the students to "visualizing the invisible" to explain tasks that might be hard to understand by reading from a book.

APPENDIX

Object Augmented above the marker in reality	Image Marker	Object Augmented above the marker in reality	QR Marker
			
			

REFERENCES

- [1] Ronald T. Azuma, "A Survey of Augmented Reality", *Teleoperators and Virtual Environments Journal*, volume 6, 1997.
- [2] Steven Wiseley, "Augmented and Virtual Reality: The Future of Learning Experiences ", *Virtual speech Journal*, <https://virtuallspeech.com/blog/augmented-virtual-reality-future-of-learning-experienc>, 2018.
- [3] Hannes Kaufmann, "Virtual Environments for Mathematics and Geometry Education, Themes in science and technology education", Special Issue, Pages 131-152, Klidarithmos Computer Books, 2011.
- [4] L. N. Birgitte, B. Harald, S.Hakon, "Augmented Reality in science education a ordnances for student learning ", *Norina journal*, pp. 12-24, 2016.
- [5] Lizana, P. and M. Rubilar, Cristian and Bassaber, Arlette and H. Flores, Ricardo, V.Fernndez and B. Octavio, "Learning Human Anatomy Using Three-Dimensional Models Made from Real-Scale Bone Pieces", *International Journal of Morphology*, 2015.
- [6] Shelton and Hedley, "Using Augmented Reality for Teaching Earth-Sun Relation-ships to Undergraduate Geography Students", 2002.
- [7] A. Ellinogermaniki, "Augmented Reality in Education", *Open Classroom Conference, Science Centre To Go Workshops*, 2011.
- [8] "What is Virtual Reality", <https://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html>
- [9] R. Silva, J. C. Oliveira, G. A. Giraldi, "Introduction to Augmented Reality", *National Laboratory for Scientific Computation*, 2012.

- [10] Salwa Hamad, "Education and Knowledge-Based Augmented Reality (AR)", *Intelligent Natural Language Processing*, Vol 740, Springer, 2017.
- [11] Jack C. P. Cheng, Wei Chen, K. Chen, "Comparison of marker-based AR and markerless AR: A case study on indoor decoration system", *Conference: Lean Computing in Construction*, 2017.
- [12] Y. Dingtian, H. Huosheng, "Application of Augmented Reality and Robotic Technology in Broadcasting: A Survey", *MDPI*, 2017.
- [13] http://en.wikipedia.org/wiki/Augmented_reality.html
- [14] ARToolworks support library, "How ARToolKit works", last modification, 2012, <http://www.artoolworks.com/support/library/index.php>

BIOGRAPHY



Dr Salwa Hamada, Winner of the Canadian Best Arab Researcher Award. Nominated by Ain Shams University for the King Faisal Prize for Arabic Language and Computer by Ain Shams University. She is the writer of a series of books like Automatic processing of the Arabic language Issues and solutions, 2009. Theory and Application. Had been invited to many seminars and conferences in Arab and foreign countries. She has many articles about Arabic. It has more than sixty published papers in national and international journals.



Dr Reda Abo Elezz, A scientist in Intelligent Software Systems, Computer and Natural Language Engineering. Was the Head of Department of Systems and Computer Department at Faculty of Engineering, Al-Azhar University have been invited to many seminars and conferences in Arab and foreign countries. Supervised and judged many master and doctoral research. Published more than hundreds of papers in national and international journals.



Dr Mohamed Taha Abou-Kreisha, Executive Manager of E-Learning Program for Islamic and Arabic Sciences, Al-Azhar University. PhD in Computational Mathematics, Faculty of Science, Cairo University. M.Sc in Computational Mathematics, Faculty of Science, Cairo University. Published papers in many national and international journals, supervised many doctoral and master thesis.



Eng Marwa Elgamal Had a Bachelor of Science degree and a Master of Science degree from in Computer Engineering at Arab Academy of Science Technology and Maritime Transportation. Worked as a lecturer in AAST, Currently, working in PhD thesis in Al-Azhar University, concern about researching in Augmented Reality and Machine learning.

نظام تعليم اللغة العربية باستخدام تقنية الواقع المعزز

مروه محمد الجمل , أ.د. سلوى حماده , أ.د. رضا أبو العز , د. محمد طه أبو كريشه

الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحري , كلية هندسة الحاسب الآلي , القاهرة , جمهورية مصر العربية

Eng.marwa.m.m@gmail.com

جامعة الأزهر , قسم هندسة النظم والحاسبات , كلية الهندسة بنين , القاهرة , جمهورية مصر العربية

Reda2018aboelezz@gmail.com

جامعة الأزهر , قسم الرياضيات كلية العلوم , القاهرة , جمهورية مصر العربية

drkresha@gmail.com

المعهد القومي للبحوث , القاهرة , جمهورية مصر العربية

hesalwa@hotmail.com

ملخص

إن أكثر المستخدمين في الدول العربية للهواتف الذكية والأجهزة اللوحية وأجهزتهم الكمبيوتر من الصغار لشغفهم بالتكنولوجيا الحديثة وتعلمهم السريع لكيفية استخدام هذه الأجهزة وإستيعابهم لإستخدامها أسرع من الكبار في السن ويقوم الصغار باستخدام هذه الأجهزة الذكية للوصول إلى مواقع المقاطع الصوتية والصورية ومواقع التواصل الإجتماعي والتواصل مع أصدقائهم وأقاربهم و اللعب بالألعاب الإلكترونية و الدردشة مع أصدقائهم وأقاربهم.

وفي المقابل يستخدم القليل منهم هواتفهم الذكية في عملية التعلم لحل واجباتهم المدرسية ،أو تعلم مهارات والحصول على معلومات إضافية حول المواضيع العلمية. بينما في جميع أنحاء العالم أصبحت البرامج التعليمية الرقمية من مقاطع صوتية أو بصرية متوفرة لجميع المراحل العمرية و يستخدم الأطفال الهواتف الذكية للوصول إلى الكتب و الدروس والتمارين والتجارب العلمية المختلفة.

وفي بعض الدول أصبح التعليم الرقمي جزءا معتمدا من أنظمة التعليم. إن تعلم اللغات مثلا هو دائما مهمة صعبة للجميع ولكن مع التكنولوجيا الحديثه أصبح بإمكان الطالب تعلم أكثر من لغة من الناطقين بها بطريقة سلسة وسهلة حيث يمكنه التعلم في مكانه باستخدام الأجهزة الحديثه دون أن يضطر للسفر لتعلم لغة معينه.

في هذه المقالة العلمية سوف نوضح كيفية استخدام تقنية الواقع المعزز (Augmented Reality) باستخدام Marker Based Algorithm في تعلم اللغة العربية لصغار الدارسين في مرحلة رياض الأطفال و المراحل الإبتدائية ولغير المتكلمين باللغة العربية.

كما سنقوم بتوضيح كيف يمكن استخدام الواقع المعزز لتعزيز أنظمة التعليم المختلفه وجعلها أكثر فعالية لاستخدامها في العملية التعليمية. وكيف يمكن للعملية التعليمية التغير بالكامل في المستقبل وتغييرها من مجرد عملية تلقي وتعليم روتيني للطلاب وتحويلها إلى رحلة إستكشاف مليئة بالتجارب والتفاعل مع المواد العلمية المختلفه ليكون التعلم رحلة إستكشاف وتوقع وتفاعل مع العلوم المختلفه بدلا من التلقي والحفظ بدون الفهم. سيتم إستخدام هذا النظام وتجربته مع مجموعه من عشرين من صغار الدارسين لتدريس الحروف العربية والكلمات البسيطة للأطفال والهدف لإستخدامه في المستقبل في مرحلة رياض الأطفال من أربع سنوات إلى إنتى عشر سنه وسوف يعمل النظام مثل عملية تعلم تفاعلي ممتع وسهل للطلاب.

الكلمات المفتاحية: الواقع المعزز، التعلم الإلكتروني، تعلم اللغة العربية، الواقع الافتراضي

فهرسة المدونات النصية والإستفادة منها في تعليم اللغات اللغة العربية نموذجاً

سلوى حمادة

قسم بحوث المعلوماتية بمعهد بحوث الإلكترونيات

ع.م.ج

hesalwa@hotmail.com

ملخص

فرضت الوقائع والأحداث اليومية في المنطقة العربية أصبحت مؤثرة في الأحداث والسياسات العالمية، على الآخرين من الجاهلين بالعربية أو المتجاهلين لها وجوب تعلمها وفهمها حتي يستطيعوا التواصل مع المنطقة العربية. وأصبحت أغلب جامعات العالم تهتم باللغة العربية وتدرسيها. واللغة العربية لغة ثرية بمفرداتها غنية بمشتقاتها وتتميز عن غيرها بقواعد صرفية كثيرة قد تجعل من تعلمها أمراً يبدو صعباً. والغرض من هذا البحث هو التعرض لفرع مهم من الدراسات اللغوية وهو المدونات اللغوية بهدف المساهمة في تعليم اللغة الثانية. وتقر أغلب الدول المتقدمة الآن استخدام المدونات في تدريس اللغات؛ اللغة الثانية وحتى اللغة الأم.

ولأن لا توجد هناك طرق منهجية لتعلم اللغة الثانية من خلال المفهرسات رغم توفر المدونات العربية للبحث والتنقيب وإن كانت هناك أبحاث كثيرة لاستخدام هذا المنهج مع اللغة الإنجليزية وغيرها من اللغات إلا أنه لا توجد أية أبحاث لاستخدام أية منهج محدد مع اللغة العربية.

وتعد المفهرسات من أهم الأدوات في استكشاف المدونات والاستفادة منها. ويوضح البحث أن استخدام المدونات في تعلم اللغة العربية كلغة ثانية لم يعد خياراً بل ضرورة لكنه ينطلق من مستويات مختلفة تختلف بالغرض والسياق وغيرها. ويعرض البحث تجارب ساهمت فيها الباحثة في المدونات تستخدم في تعليم اللغة العربية.

الكلمات المفتاحية: لغة ثانية – المدونات / الذخائر – المتون – التعليم – المفهرسات – الإحصاء اللغوي – التذييل / الترميز.

1. مقدمة

مفهوم المدونة موجود منذ خلق البشر حتي لو يحدد أو يطلق عليه الاسم. وقد أصبحت لفظة "المدونة" حالياً أحد الكلمات الشائعة المستخدمة في تدريس اللغة عالمياً - وإن قل في التوجه العربي في استخدام المدونات- ويتزايد باستمرار استخدام كل من المعلم والدارس للمنتجات التعليمية التي بنيت على فكرة "المدونة" مثل القواميس وكتب القواعد بل أيضاً لكل مستخدم للغة. ناهيك عن استخدام المدونات نفسها في أغراض كثيرة تبدأ بالأغراض البحثية وتنتهي بالأغراض التعليمية.

لقد أصبحت المدونات اللغوية الآن آلية ضرورية ومهمة في البحث اللغوي بصفة خاصة والتعليم بصفة عامة. ويطلق عليها "المدونات الحاسوبية" أو "المدونات الإلكترونية" أو المتون النصية أو الذخائر النصية. والأبحاث التي لا تعتمد على شواهد من مدونات شاملة وممثلة للغة تعتبر ناقصة وتفتقر إلى الدقة والبرهان.

وتعد المدونة منذ زمن بعيد أحد المصطلحات الشائعة في تعلم اللغة في بعض جامعات أوروبا، والدول المتقدمة. وحتى لو لم يتم التدريس باستخدام المدونة إلا أنه يتم باستخدام منتجات من المدونة كالقواميس وقوائم الكلمات الأكثر

شيوعا. أما بالنسبة للغة العربية بالمنطقة العربية فإن العاملين في مجال تعليم اللغة ومعلموها ليس لديهم أدنى فكرة عن المدونات، أو فكرة بسيطة ولا يقدروا أهميتها في تعليم اللغة، وأكثرهم لا يعرفون حتى: ما هي المدونة؟ وبالطبع ليس هناك استعمال مباشر للمدونات في التعليم لجهل أغلبهم بأهمية دور المدونة في تعلم اللغة؟ وكيف تطوع لهذا الغرض؟

وهذا البحث دراسة شاملة لأغلب أنواع المدونات المتاحة. وتشرح دور المدونات في البحث العلمي وفي تعليم اللغة. وتشير إلى عدم استخدام هذه التقنيات العلمية بالصورة المناسبة وخاصة في تعليم اللغة العربية لغير الناطقين بها. وقد حاولت عرض الموضوع باستفاضة لولا التقيد بعدد صفحات البحث. وللبحث أهمية علمية؛ تشمل الدراسة مدونات مستخدمة فعليا بالإضافة لعرض مدونات شاركت فيها وشرح كيفية استخدامها وأهمية ذلك في قضية ينبغي التوجه إليها والاهتمام بها. يعرف البحث المدونة وبعضها من أنواعها. كذلك يناقش مزايا استخدامها في تدريس اللغة الثانية، لاسيما تدريس اللغة العربية وكيفية تزويد المعلمين والطلبة بأمثلة الاستخدام للغة الحياة اليومية في الفصول الدراسية أو حتى في التعليم الذاتي. وشرح وضع المدونات ضمن عملية تعليم اللغة ككل. ولها أهمية عملية؛ حيث يطرح البحث فكرة عمل مدونة شاملة تساعد متعلم اللغة العربية على استيعابها من خلال ما تمده به من الأمثلة والسياقات المختلفة للمفردات والقواعد التي يتعلمها ويمكن توفيرها مفتوحة المصدر على شبكات الاتصال. ويشرح تكييف الفهارس والمعجم اللغوي لعملية تعلم اللغة الثانية وقد زود البحث بصور إحصائيات وبعض نتائج التحليل اللغوي وفي سياق البحث نتعرض لبعض المتطلبات لأجل المضي قدما في هذا الفرع المهم من الدراسات اللغوية. وينتهي بأمثلة لمدونات تستخدم في العملية التعليمية سواء مكتوبة أو صوتية. ونتائج هذه الدراسة تقترح وتضع الحلول لعمل مدونة تخدم المجتمع العربي، سواء كان ذلك في الجانب الثقافي، أو التعليمي، أو القرابي، أو الإعلامي، أو أي جانب آخر، من خلال نصوص مختارة تغطي جميع المجالات وتبني الحلول المقترحة.

– وتنقسم أهمية هذه الدراسة إلى قسمين.

- المدونات النصية وبناءها.
- القسم الأول المباحث التي تتعلق بالمشاكل والقضايا التي تشير إليها الدراسة

2. المدونات النصية /Textual Corpus /الذخائر النصية

(1) ما المقصود بالمدونة؟ Corpus definition

المدونة Corpus هي جسم غير منتظم من النصوص المكتوبة أو المنطوقة التي تُستخدم لدراسة جوانب اللغة، يمكن قراءتها والتعامل معها إلیًا بعد إدخالها على الحاسب الآلي، كما يمكن التحكم في بياناتها ومُدخلاتها، بالإضافة أو الحذف أو التعديل من خلال قواعد بيانات (Databases) صُممت خصُوصًا للتعامل مع هذه النصوص. وتُعتبر قاعدة البيانات الحاوية لنصوص المدونة اللغوية مخزنًا كبيرًا للغة، يُرجع إليه وقت الحاجة، ويتحمّل أيّ قدرٍ من النصوص التي تُضاف إلى المادة الأساسية مُستقبلًا [1].

(2) أنواع المدونات

أ. المدونة الذهنية:

هي كم من النصوص والشواهد والمفردات التي نتعرف عن طريقها على الأشياء ومعانيها من خلال حصيلة الخبرة البشرية المتراكمة في شتى المجالات والميادين. أما أقسام هذه المدونة فيمكن أن تكون صوتية أو تصويرية أو حسية أو لفظية بخلاف المعاجم الآلية التي يفترض في مداخلها أن تكون لفظية فقط. ولكل منا مدونته الذهنية الخاصة به التي ترتبط بثقافته وخبرته وعن طريقها يستطيع أن يتعامل مع العالم من حوله. والتعليم ما هو إلا توسع جديد أو تبديل أو حذف أو تطوير في تلك المدونة وكلما أخذ التعليم شكل المدونة كلما كان أسهل استيعابا وفهما وهذا ما نهدف في تطبيقه على المدونات الحاسوبية.

ب. المدونات الحاسوبية:

يجب أن تخضع مادة المُدَوَّنة اللُّغَوِيَّة لمجموعة من الأسس والمعايير لنطلق على مجموعة من النصوص الضخمة "مُدَوَّنة لغوية" فهي ليست نصوصاً مجموعة بطريقة عشوائية؛ على الرغم من أنها كتلة غير منتظمة من النصوص. وهذه الأسس والمعايير، يُحدِّدها الهدف المنشود من المُدَوَّنة اللُّغَوِيَّة؛ فالمُدَوَّنة التي يُعتمد عليها في صناعة مُعجَم لُغَوِيٍّ، ستختلف مادتها عن تلك المُستخدمة في حصر مجموعة من الأنماط التركيبية أو البِنَوِيَّة للغة، كما تختلف مادة المُدَوَّنة المُستخدمة في صناعة مُعجَم تَكَرَّريٍّ عن تلك التي يُعتمد عليها في صناعة المَعاجِم التاريخية. كذلك.. فإنَّ المُعالِجة الآليَّة للنصوص تتفوق وطبيعة المُدَوَّنة؛ فالبرامج الحاسوبية المُستخدمة، وطريقة مُعالِجة النصوص، وطرائق إدارة قواعد البيانات، كلُّ هذا يخضع لتلك الأسس والمعايير التي يُحدِّدها هدف المُدَوَّنة اللُّغَوِيَّة [1]. ولمزيد من التفاصيل راجع [2].

وعلى الرغم من هذا كله فأنا أؤيد عمل مُدَوَّنة تغطي جميع الأغراض السابقة أو بعبارة أخرى فإن المُدَوَّنة الممثلة للغة يجب أن تصلح لجميع الأغراض. وإن ظل الحديث هنا قاصراً على المدونات النصية باللغة العربية التي تعاني تقلصاً هائلاً نسبة لنظائرها في اللغة الإنجليزية. ولا ننكر جهود الباحثين المخلصين في هذا المجال مثل لطيفة السليطي (1) التي سعت لجمع مدونه نصية هائلة ومتنوعة لسد حاجة اللغويين في هذا المجال إلا أن عملها ظل محدوداً بالوقت المخصص للبحث وبالجهود الفردي أيضاً.

ونسوق هنا صوراً لتصنيف المدونات الحاسوبية من [2]

- (1) من حيث التحليل اللغوي:
- (2) من حيث اللغة: (وحيدة اللغة وثنائية اللغة ومتعددة اللغات).
- (3) من حيث التطبيق اللغوي والعلاقة بين اللغات في المُدَوَّنة متعددة اللغات
- (4) من حيث العموم : عامة أم خاصة.
- (5) من حيث نوعية النص/ المادة النصية .
- (6) مدونات اللهجات Dialect Corpora مثل المدونات الخاصة بلهجة واحدة كمدونات اللهجة البورسعيدية ومدونة اللهجة الصعيدية.
- (7) من حيث نوعية مستخدم المُدَوَّنة: هل هي لاستخدام ابن اللغة native speaker أو لمتعلم اللغة learner. وهنا يرتبط تصميم المُدَوَّنة بهذا الفرق.
- (8) من أصل اللغة: هل لغة المُدَوَّنة أصلية original أو تمت ترجمتها translations. ويجب التمييز بين النوعين هنا لتأثر لغة المُدَوَّنة المترجمة بأسلوب ودقة وطريقة الترجمة.
- (9) من حيث زمن اللغة: مُدَوَّنة محددة بفترة زمنية محددة Synchronic Corpora كمدونة العربية الفصحى (2004) بجامعة مانشستر، وهي محددة من الفترة قبل الإسلام حتى القرن الحادي عشر.
- (10) مُدَوَّنة شاملة مختلفة الفترات Diachronic Corpora.
- (11) من حيث الإتاحة: وقد أدى ظهور شبكات الاتصال إلى التمييز بين المدونات الرسمية – Standard Corpora وهي مدونات مصممة وفقاً لمبادئ خاصة - ومُدَوَّنة حرة Free Corpora مثل المتاحة على شبكات الاتصال والتي تضم نصوصاً لا نهاية لها.

(1) Designing and Developing a Corpus of Contemporary Arabic, Latifah Al-Sulaiti, (2004).. Leeds University. P.1.[3]

(12) من حيث تحشية/ تذييل المدونات: سبق وأوضحنا أن هناك نوعين من المدونات:

- مدونات خام Raw Corpora / متون: وهي نصوص توجد على صورتها الخام ولم تتم أي إضافة لها.

- مدونات مزودة بمعلومات لغوية Annotated Corpora: ورغم شيوع استخدام المدونات الخام في الدراسات اللغوية، فإن هناك بعض القضايا اللغوية التي تتطلب استخدام النوع الآخر من المدونات والمزودة بمعلومات لغوية مختلفة ومتنوعة. وأحد أهم المزايا الأساسية لهذه المدونات أنها تحولت من صورتها الخام إلى صورة يتضح فيها شتى المعلومات اللغوية. وتسمى هذه المعلومات "تحشية". وهناك العديد من الأبحاث حول عملية التحشية ومبادئها وقواعدها .

(13) من حيث المنتج:

- مؤسسات أو هيئات Organization: وهي خاصة بمشروعات تجارية وتمتاز بتوفر الدعم المادي والبحثي والحكومي أحياناً مثل مدونات [3].

- أفراد أو مجموعات بحثية: وهي التي ينتجها أفراد للمشاركة في مشروعات بحثية أو للحصول على درجات علمية مثل [3].

(3) تخزين المدونة:

يوجد كثير من الصور الإلكترونية للنصوص لا تمكننا من التعامل معها تعاملًا مباشرًا سواء باستخدام برامج التحليل اللغوي أو المفهرسات الآلية، لذلك يجب تحويل هذه الصور جميعها إلى الصورة النصية **plan text format** مثل (Doc, Text, Rtf,..) حتى يسهل التعامل معها بأي محرر نصي text editor. من هذه الصور:

صورة لغة الربط بين النصوص التشعبية (HTML)، صورة الوثيقة المحمولة، والصورة النصية الصورة الصوتية والصورية (وضع الوسائط المتعددة في المدونة) : فهي تحتاج نوعية خاصة من التعامل، وكذلك اللغات التي لا يمكن لجميع الحواسيب التعامل معها مثل الصينية واليابانية. ويفضل تخزين المدونات في صورة قواعد بيانات يتم ربطها بالمعلومات الخاصة بها.

(4) مزايا إخضاع المدونات للحوسبة

- 1) حصر خبرات جميع الكتاب واللغويين.
- 2) مكن من تغطية كل الظواهر اللغوية المعروفة وإيجاد ظواهر جديدة.
- 3) مكن من تسريع معالجة البيانات مع الدقة العالية لو توافرت شروط بناء المدونة الممثلة للغة.
- 4) مكن من تفادي التحيز البشري human bias واختلاف وجهات النظر في النتائج.
- 5) مكن من إثراء قاعدة البيانات بالبيانات المنطوقة metadata.
- 6) مكن من عمل إجراءات آلية لعمليات مثل التحليل الصرفي والنحوي والدلالي والتذييل وغيره.
- 7) مكن من البحث في حجوم هائلة من النصوص والخروج بنتائج شاملة.
- 8) رفع من دقة وتجانس النتائج.
- 9) اثبات النظريات اللغوية القديمة أو دحضها.
- 10) ومن الطريف في العمل الحاسوبي أنه عند عمل برمجية لتطبيق معين على المدونة نكتشف إمكانية استخدامها في تطبيقات جديدة لم تكن في الحسبان عند عمل هذه البرمجية.

(5) الحس والخبرة اللغوية والتدوين اللغوي

كانت القواعد والاستنتاجات فيما سبق تعتمد كثيراً على حس الباحث وخبرته. أما الآن وباستخدام المدونات فتم حسم كثير من المشكلات فمثلاً:

- (1) تأثير اللهجة والبيئة والعوامل الإجتماعية والتي كان يصعب حصرها.
- (2) التطور اللغوي لا يمكن أن يدرك دون مدونات من مراحل تاريخية مختلفة تمثل تطور اللغة. ومن أهم الأمثلة على ذلك المعجم التاريخي الذي تم عمله في قطر.
- (3) المدونة تمثل اللغة وتبنى على النصوص الأصيلة أو الحقيقية (المستخدمة فعليا لا المؤلفة).
- (4) التحليل المعتمد على الحاسوب يُمكن أن يَسترجع اختلافات لا يَستطيعُ الحدس أو الخبرة إدراكها.
- (5) التعامل مع كل الظواهر اللغوية يمكن أن يتم في أن واحد.
- (6) الألفاظ ذات المعنى الواحد على مستوى اللهجات المحلية أو العربية.
- (7) تنفيذ الكثير من الدراسات اللغوية القديمة والقائمة اليوم والتي تستنزف جهد الباحث وماله ووقته. هذه الدراسات يمكن أن توفر عليه جميع أنواع المشقة من خلال المدونات الإلكترونية.

ولا يعني ذلك إهمال الحدس والخبرة بل إنهما المفتاحان للوصول إلى المعلومات باستعمال المدونات. فالمدونة في حد ذاتها غير مفيدة دون استخدام برمجيات تمكننا من استخلاص ما يمكننا الحصول عليه من معلومات منها كما ذكرنا من قبل. وهذه المعلومات يوجهها الحدس والخبرة اللغوية.

(6) اللغويات التطبيقية Applied Linguistic

- (1) تستكشف أنماط فعلية لإستخدام اللغة.
- (2) أداة لتطوير موارد للدروس اللغوية.
- (3) تستكشف أسئلة مختلفة لاستعمال اللغة.
- (4) للحصول على أداة قوية لتحليل اللغات الطبيعية.

(7) المجالات ذات العلاقة:

وهناك العديد من الأبحاث في هذا المجال، ولمزيد من المعلومات ارجع للمرجع الذي يعتبر الأكثر تغطية لهذه الناحية:

http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/

(8) معايير بناء المدونات

قبل البدء في عمل المدونة نسرد معايير Criteria لتجميع المدونة حيث يجب أن يقوم أيّ اختيار أو تطبيق بحثي على بعض المعايير والخطوة الرئيسية الأولى في بناء المدونات ومن ثم تصميم المعايير التي تُختار بها النصوص التي تُشكّل المدونة. وإلى الآن لا توجد معايير قياسية لبناء المدونات [4] ورغم قدم هذا البحث إلا أن هذا الوضع مازال قائماً. وإن كان هناك اتفاق عرفي بين بناء المدونات على بعض المعايير وتضمن المعايير العامة:

- (أ) حجم المدونة: ويقاس بعدد الكلمات غالباً،
- (ب) نمط الكتابة؛ سواء اللغة في الوقت الحاضر النمط الإلكتروني،
- (ت) نوع النص؛ على سبيل المثال إذا مكتوباً، سواء في كتاب، أو مجلة، أو ملاحظة، أو رسالة،

- (ث) طبيعة النص؛ منطوق ، مكتوب، لغة إشارة.
 (ج) مجال النص؛ على سبيل المثال أكاديميا كان أو عاما،
 (ح) اللغة أو تنويعات اللغة في المدونة؛ على سبيل المثال لغة عامية أو فصحي أو عامية متفصحة،
 (خ) موقع النصوص الجغرافي؛ على سبيل المثال سوريا كان أو مصريا أو مغربيا،
 (د) البعد التاريخي؛ تاريخ كل نص.

(9) التخطيط لإنشاء المدونة **Planning the construction of a corpus** للتفاصيل والشرح راجع [2].

(أ) جمع وحوسبة البيانات **Collecting and computerizing data** للتفاصيل والشرح راجع [2].

تنقسم إلى مرحلتين

(أ) مرحلة حوسبة البيانات للتفاصيل والشرح راجع [2].

(ب) مرحلة مراجعة المدونة

بعد الحصول على المدونة من مصادرها المختلفة تجهز المدونة للتذليل في عدة خطوات- لمزيد من التفاصيل انظر [1] و [2] ويضاف إليها:

(ت) مراجعة المدونة مراجعة لغوية إملائية [1].

(ث) تنسيق المادة المدخلة تمهيدا لمعالجتها آليا [1].

(ج) مرحلة تذليل / ترميز المدونات **corpus annotation** [2].

(10) تذليل / ترميز المدونات [2].

(أ) خطوات التذليل : التقطيع **tokenization** ، **lemmatization** تحليل الكلمة صرفيا / تحليلا تكوينيا ، تعليم

أجزاء الكلام، الإعراب الجزئي، الإعراب النحوي الكامل، معالجة معاني الكلمات والحديث

(ب) أنواع التذليل المختلفة **Different kinds of annotation**: وبعيدا عن تذليل أجزاء الكلام (POS)

التذليل صرفي يوجد أنواع أخرى من تذليل المدونات تبعا لمستويات التحليل المختلفة **different levels of linguistic analysis** ومنها:

(أ) التذليل الصوتي **Phonetic annotation**:

(ب) مثل إضافة معلومات عن كيف تنطق الكلمة في المدونة المنطوقة " **spoken corpus** " من أنواع التذليل

الصوتي: إضافة معلومات عن السمات **prosodic** مثل التشديد والترنيم **intonation** والوقوف **pauses** و

الإجهاد **stress** والترنيم **intonation** والإيقاع **rhythm**.

(ت) التذليل المعجمي **Lexical annotation**: من أمثله رموز أجزاء الكلام .

(ث) التذليل النحوي **Syntactic annotation**: ومثال ذلك: - معلومات مضافة عن كيفية تحليل الجملة وتساهم

في إزالة غموض حد الجملة **Sentence boundary**. فنهايات الجمل عادة ما تكون (نقطة، علامات تعجب-

استفهام-..). ويمكن أن تكون النقطة نهاية الجملة أو جزءا من مختصر (ج.ع.م) أو تكون تكلمة (إلخ ...) وقد

تقوم بدور مزدوج كما في الحالة الأخيرة. ويضيف من ناحية التحليل النحوي معلومات عن الوحدات من

العبارات **phrases** وشبه الجمل **clauses**.

(ج) التذليل الدلالي **Semantic annotation**: ومثال على ذلك: - معلومات مضافة حول التصنيف الدلالي

لل كلمات مثلا لكلمة "عين" جزء من الجهاز البصري للإنسان أم كمصطلح سياسي -بمعنى جاسوس-

ويرجع ذلك إلى التصنيفات الدلالية المختلفة، بالرغم من أنه ليس هناك إختلاف في تهجئ أو التلفظ. كما يفيد

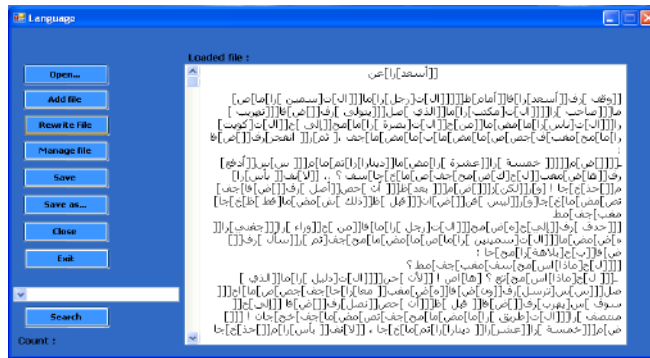
الترميز بكلمة: معنى **senses** في (حل لبس المعنى **sense resolution**).

- ح) التذييل الواقعي Pragmatic annotation: ومثال على ذلك: - معلومات مضافة حول أنواع العمل الخطابي (دور حوار) وكلمات الحوار - وهذا يحدث في لغة الحوار - الحديث في المناسبات المختلفة قد تكون اعترافاً أو طلباً للتعليقات، أو مرحلة جديدة من المناقشة.
- خ) تذييل الخطاب Discourse annotation: ومثال على ذلك: - المعلومات المضافة حول مرجعية anaphoric الضمائر في النصّ. على سبيل المثال مرجعية الضمير "هم" على سابق antecedent في الجملة عاقب المدرسون الطلاب وضربوهم. وحين ينفذ هذا التذييل يدويا يعطي على دقة عالية. وترى سلوى بعض الحلول الآلية لهذه المشكلة والتي تصل بالدقة لدرجة معقولة [2].
- د) التذييل الأسلوبي stylistic : ومثال على ذلك: - معلومات مضافة حول عرض presentation الفكر والخطاب حديث مباشر، أو غير مباشر إلخ..

(11) مثال على رموز التذييلات الصرفية والنحوية

نَعْنِي بها مجموعة الرموز المتفق عليها والتي وظفت لتمثيل التذييلات، كُمَيِّز عن المدونة الأصلية. مثال لبعض الرموز الصرفية والنحوية التي استخدمت في تذييل المدونة [5]:

جمله اسمية:جا	مركب حرفي:مح
جمله فعلية:جف	مركب وصفي:مو
جمله فعلية ناقصة:جفن	مركب حالي:محا
جمله اسمية دخل عليها حرف	مركب خبري:مخ
ناسخ:جان	
صلة الموصول:جص	مبتدأ:م
أسلوب الشرط:جش	خبر:خ
أسلوب التعجب:تج	فاعل:فا
أسلوب القسم:قس	فعل ناسخ:فن



شكل(1) المدونة المذيلة مثلا مدونة (مساك) [5]

(12) شروط التذييل الجيد

أ) تحديد طريفته وزمنه ومن قام به وما هي مراحل المراجعة التي انتجت من ووترتيب الإصدارات الجديدة، إلخ.

- (ب) يَجِبُ أَنْ يُنَبَّهَ المستخدم لاحتتمال وجود أخطاء بالتذييل وعلى الرغم من ذلك فالمدونة المذييلة مفيدة فعلاً.
- (ت) مخططات التذييل يَجِبُ أَنْ تستندَ بقدر الإمكان على معايير متفق عليها وبطرق محايدة.
- (ث) يجب أن يسهل فصل التذييلات عن المدونة المذييلة. Annotations should be separable لاسترجاع المدونة الخام/ الأصل. فأحياناً لا يفضل بعض المستخدمين وجود التذييلات لذلك من الأفضل أن يكون من السهل فصلها عن المدونة. وفصل التذييل لا يَجِبُ أَنْ يُؤدِّي إلى ضياع أية معلومات حول بيانات المدونة الأصلية original corpus أو أجزاء من النص الأصلي.



شكل (2) المدونة مساك بعد انتزاع التذييل [5]

(13) مثال على المدونات المذييلة/ المرزمة بأجزاء الكلام

دمج المفهرس مع التحليل الصرفي: مثال: مدونة معجم عربي معاصر:

وهي مدونة للغة العربية المعاصرة زاد حجمها عن ستة ملايين كلمة وتنوعت موضوعاتها لأكثر من موضوع وتحتوي مليون كلمة منها (6127718) كلمة مكتوبة في ثمانية مجالات و(432063) كلمة منطوقة في ثلاثة مجالات. وزعت على عشرين ملفاً نصياً بواقع ألف صفحة للملف الواحد. وقد تم بنائها تحت إشرافي [1]. كما تم تصميم نموذج لدمج المفهرس مع التحليل الصرفي² وقد صمم المحلل على عينات بسيطة من البيانات نظراً لتكلفة العمل المطلوبة والوقت المحدد.

وكما يوضح المخرج المفردة (وادي) مثالا:

المفردة اللغوية: وادي	جذرها: ودي
عدد مرات تكرارها (ترددها في النص):	بنيتها: اسم فاعل/ فاعل.
دلالتها: مصطلح جغرافي.	تركيبها: مشتق ودي

² بالتعاون مع د. المعزز بالله السعيد بكلية دار العلوم - جامعة القاهرة حيث قام بعمل كود البرنامج أيضا

3. أمثلة للمدونات

(1) مدونات البحث العلمي والتعلم

(أ) <https://www.sketchengine.co.uk/>

(ب) Antconc

(ت) <http://www.laurenceanthony.net/software/antconc/>(ث) <http://wordsmith.org/>

(ج) المدونة العربية: [7]

وهي مدونة تمد الباحثين في مجالات اللغة بمبتغاهم بطريقة سهلة ولكن فقط من خلال أصحاب المدونة حيث أنها ليست متاحة.

وهي من إنتاج مدينة الملك عبدالعزيز للعلوم والتقنية ، وستكون أكبر وأضخم مدونة لغوية للعربية. وهي في مرحلتها الأولى تسعى لجمع سبعمائة مليون كلمة وسوف يزداد حجمها الى ان تصل الى بليون كلمة في مرحلة لاحقة ان شاء الله. راعي تصميم المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، عدة معايير خارجية لاختيار نصوص المدونة تعتمد على خمس ركائز أساسية هي البعد الزمني، والبعد الجغرافي، والوعاء المعلوماتي، و المجال المعرفي والتصنيف الموضوعي. إضافة الى هذا فإن المدونة في مرحلتها الحالية هي للنصوص المكتوبة والكاملة فقط ولا تحوي أي نصوص منطوقة مثل الحوارات التلفزيونية أو الخطابات السياسية أو أي نصوص غير مكتملة مثل فصل من كتاب او جزء من مقال.

رابط المدونة: يمكن الوصول للمدونة العربية على الرابط www.kactac.org.sa

(ح) مدونة براون Brown Corpus للتفصيلات راجع

Francis, N.W. and Kucera, H. (1979). *Brown Corpus Manual: Manual of Information to accompany A Standard Corpus of Present Edited American English for use with Digital Computers, Revised and Amplified*. Available at:

<http://icame.uib./brown/bcm.html#bc2>

(خ) المدونة اللغوية العربية العالمية (3) ICA The International Corpus of Arabic

الجهة العاملة عليها: مكتبة الإسكندرية.

نبذة عن المدونة: مكتبة الإسكندرية هي إحدى المؤسسات المصرية العالمية التي تسهم في نشر الثقافة والمعرفة ودعم الأبحاث العلمية، وقد قامت بدعم بناء المدونة اللغوية العربية العالمية التي هي أحد المحاولات الحقيقية الطموحة لبناء مدونة لغوية للعربية المعاصرة تحوي (100 مليون) كلمة مُحلَّلة صرفياً ونحوياً ودلاليًا، وقد روعي فيها أن تكون ممثلةً لقطاع إقليمي كبير من الدول الناطقة باللغة العربية المعاصرة وعاكسةً بشكل حقيقي وواقعي لأنماط استخدام اللغة العربية المعاصرة في أنحاء العالم العربي. وبمجرد الانتهاء من بناء المدونة ستكون أول مدونة محللة ومتاحة بوصفها موردًا لغويًا للباحثين بصفة عامة، والباحثين اللغويين بصفة خاصة؛ إذ سَتُفيد في وصف نظريات اللغة من خلال الاستخدام الواقعي للكلمات.

(2) مشاكل مدونات تعليم اللغة الثانية

(أ) المفردات اللغوية

(3) ICA 2016, <http://www.bibalex.org/ica/ar/login.aspx>

من أهم مشاكل تعلم اللغة هي اكتساب المفردات وهو من أهم التحديات التي تواجه معلمي اللغة الثانية بل وحتى اللغة الأم. ومن الملفت للنظر أنه يمكن أن تكون المفردات الشائعة الكثيرة الإستخدام في نص ما تمثل أكثر من ثلثي كلمات النص. لذل اضطر بعض المتخصصين لعمل قوائم للكلمات الأكثر شيوعاً لتبني عليها المناهج الدراسية ولكن بالرغم من هذا وقد تضر هذه القوائم الطلاب وتجعلهم يركزوا عليها دون غيرها وتصبح حصيلتهم اللغوية هشة.

(ب) المشكلات الدلالية و المعجمية

اللغة العربية لغة ثرية بمفرداتها وهي لغة اشتقاقية، تختلف في بنيتها ونظامها اللغوي عن اللغات اللصقيّة كالإنجليزية والألمانية، وفصيلة اللغات الجرمانية. وتتميز بإمكانية استغلال ظاهرة الاشتقاق في التعبير عن كثير من المعاني لكن ذلك لا يعني ضرورة اختلاف مناهج دراستها عن اللغات الأخرى، فاللغة نظام مُعقّد يمكن التعامل معه في بعض جوانبه من خلال مناهج ونظريات عديدة، بغض النظر عن طبيعة النظام اللغوي للغة المُعَيّنة. واللغة العربية لبسية كغيرها من اللغات ويرجع السبب لوجود اللبس بهذا المعدل الكبير للانتقاص من أجزاء الكلمات بعدم كتابة التشكيل؛ حيث يمثل التشكيل جزءاً هاماً من أجزاء الكلام ومميزاً قوياً لانتقاء المعنى المطلوب. كذلك الحال مع علامات الترقيم. ويعد كل هذا من أهم مشاكل فهم للنصوص. وما سبق يرتبط أساساً باللبس المعجمي، أما عن اللبس التركيبي فالوضع أكثر تعقيداً؛ وذلك لمرونة اللغة العربية فهي تسمح بالابتداء باسم أو فعل أو حرف، وبها مرونة تسمح بتغيير ترتيب الكلمات وذلك بخلاف غيرها من اللغات كالإنجليزية مثلاً.

والدارس عندما يبدأ بتعلم لغة أجنبية فإنه بالطبع لا يتقنها في المرحلة الأولى، وبالتالي فإننا إذا لاحظنا لغة الدارس في هذه المرحلة نلاحظ عجباً لأنه يتكلم لغة غريبة لا هي اللغة الهدف التي تعلمها ولا هي اللغة الأصلية له، ويطلق عليها اللغة الانتقالية. ولهذه اللغة صفات أهمها: أنها تجمع خصائص لغة الدارس الأم وبعض خصائص اللغة المنشودة، ولكن لماذا تجمع بعض خصائص اللغة الأصلية؟ لأنه يحاول أن ينقل إلى لغته من اللغة الهدف، هذا في المرحلة الأولى، وعملية التأثير باللغة الأم تتأثر في جميع الجوانب اللغوية من أصوات ينطقها بلغته الأم وتراكيب يحاول استخدامها بتراكيبه المعروفة في لغته، كأن يجمع بعض الكلمات على أوزان لغته أو غير ذلك فهو يحاول أن يعمم قاعدة لنفسه [8].

(ت) من أهم المشكلات المعجمية و الدلالية التي يعاني منها متعلمو اللغة العربية من غير الناطقين بها ما يلي [9]:

- أ) كثرة مفردات اللغة مما يجعل من العسير على متعلميها الإلمام بها وفهماها.
- ب) تعدد دلالات الوحدات اللغوية (كلمة – مركب- شبه جملة- جملة) وانتقال الوحدة اللغوية من المعنى الحقيقي إلى المعنى المجازي، مما يسبب صعوبة في فهم المعنى المقصود من النص المقروء و تظهر هذه المشكلة إذا تم اختيار المواد اللغوية و تقديمها للمتعلم على أسس غير علمية من حيث الشيوع و الأهمية والتدرج و غيرها من المعايير التي يجب أن تؤخذ بعين الإعتبار في إعداد المناهج.
- ت) ارتباط الكلمات العربية بالتصريف وخضوعها للقواعد التصريفية من حيث الشكل أو البنية والميزان الصرفي والتوزيع و يشكل صعوبة على المتعلم، فالكثير من المتعلمين الذين لم يتعودوا على هذا النوع في لغاتهم يعتقدون أن تعلم الكلمة في اللغة الهدف لا يتعدى حفظها و فهم معناها و لهذا يلجؤون إلى وضع الكلمات في قوائم و حفظ معانيها معزولة عن سياقها. يواجه متعلمو اللغة العربية مشكلات في فهم بعض الكلمات و استعمالاتها و يخطئون في ذلك نتيجة تعميم القاعدة التي تعلموها في بنية الكلمة و دلالاتها و يكون الخطأ بسبب طريقة التدريس أو سوء تنظيم المنهج عندما تقدم الكلمات في قوائم مفصولة عن سياقاتها الطبيعية أو بشكل قواعد جافة لا تراعي فيها الجوانب الاتصالية و الوظيفية، أو بسبب تأثير اللغة الأم في التعلم لدى الناطقين بلغات تكتب بالحروف اللاتينية أيضاً عدم التفريق بين كلمات الاعجمية و أسماء الأعلام

- فقد يبحث طويلا عن معنى كلمة تدل على شخص ففي بعض اللغات كالإنجليزية تميز الأعلام بحرف كبير تتصدر الكلمة .
- (ث) تصور متعلمي اللغات الأجنبية أن جميع المعاني في اللغات واحدة وأن الإختلاف فقط في الكلمات الدالة عليها و يعتقدون أن كلمة في اللغة الهدف ما يقابلها في لغته الأم و هذا غير صحيح في كثير من الحالات مثال (عم و خال) كلمتان تقابلهما في الإنجليزية كلمة واحدة هي (Uncle).
- (ج) تصور المتعلم أن المعاني التي تدور في ذهنه يمكن استعمالها بالطريقة التي كان يستعملها في لغته الأم مع اختلاف اللفظ فقط والسبب في هذه المشكلة الاعتماد على الترجمة من تأثير لغته الأم وصعوبة التعبير عن المعاني العربية بالكلمات والأساليب العربية .
- (ح) إغفال المتعلمين الجوانب الثقافية و الدلالات الثانوية لبعض الكلمات فلا يدرك بعضهم أن المعنى المعجمي لا يكفي لبيان معنى الكلمة ما لم تشرح في السياق الذي وردت فيه .
- (خ) صعوبة البحث في المعاجم العربية عن معنى الكلمة التي يصعب على المتعلم فهمها لأن ذلك يستلزم أن يحدد مادة الكلمة و جذرها و هذا الأمر صعب خاصة في المراحل الأولى من التعلم.

(ث) مشكلات خاصة بالنظام اللغوي [10]

- أ) المشكلات التي يواجهها الطلاب, عند تعلم النظام الصوتي و النظام النحوي والدلالي للغة العربية.
- ب) المشكلات التي يواجهها الطلاب, في فهم ثقافة اللغة العربية.
- ت) المشكلات التي يواجهها الطلاب, وهم يتعلمون مهارات القراءة والحديث والاستماع الكتابة باللغة العربية.
- ث) مشكلات خاصة بالجانب التربوي والتعليمي النفسي. لتفاصيل أكثر راجع [10].

(3) مدونات تعليم اللغة العربية لغير الناطقين بها 1) المدونات العربية التعليمية الصوتية:

- (أ) مدونة معموري: يقول معموري في بحث مرجعي عن المدونات الصوتية وما تم إنجازه فيها [11]:
- أ) في عام 1995 تم جمع 18 تنوعاً لغوياً 18 linguistic varieties لدعم بحث تحويل النص الصوتي لنص مكتوب.
- ب) تحتوي على أكثر من 450 محادثة تليفونية بالعامية المصرية.
- ت) تم تحويل 10 دقائق من كل محادثة – 200 محادثة – بينما أهملت 120 محادثة.
- ث) المنشورات تضم استقبال الصوت الاعتيادي plain audio و رصف الرمز الصوتي بالوقت -time aligned transcripts ومعجم النطق pronouncing lexicon
- ج) المعجم يحتوي : التركيب الظاهري surface form ، النطق pronunciation التحليل الصرفي morphological analysis وتردد في ثلاث مجاميع من عينات البيانات frequency in 3 .data sets

(ب) مدونة الصوامت الصعبة في اللغة العربية (ACOCO) The Arabic Consonant Corpus

- أ) الجهة العاملة عليها: المدونة جزء من مشروع بحثي لدرجة الماجستير في الأرصاد الإلكترونية للغة العربية من جامعة عين شمس، مقدمة من الباحث محمد الغليص، تحت إشراف أ.د. علي هنداوي و أ.د. محسن رشوان [12].

(ب) نبذة عن المدونة: تطبيق إلكتروني مستمد من مدونة صغيرة (مركزة) مكونة مبدئياً من (400) كلمة، تراعي – التركيز على- الوحدات الصوتية العربية التي لا يوجد لها مكافئ في اللغات المختلفة صاحبة الأعداد الكبيرة من البشر، التي تمثل صعوبات لدارسي العربية من أصحاب تلك الألسن، ومن ثم تذييل تلك المشاكل من شأنه نشر اللغة العربية لدى الأجانب.

(ت) مدى إتاحتها: متاحة بشكل كامل بدون تسجيل على موقع المدونة على شبكة المعلومات الدولية (الإنترنت)

<http://daadacademy.org/arabic-consonants-corpus-ACOCO/>

2) مدونات متعلمي اللغة العربية الموسومة/ المذيبة/ المرزمة بالأخطاء/ مدونات الطلاب (مدونات الأخطاء):

هي مدونات تعرف بمدونات الأخطاء اللغوية الموسومة. وذلك لدراسة وتحليل الأخطاء الطلاب. هذه المدونات غالباً ما تكون مأخوذة من واجبات و انتاج الطلاب في مامهم الدراسية من مدونات متعلمي اللغة. وتهدف دراسة الأخطاء إلى تزويد من ينشئ المدونات بأدلة عن كيفية تعلم اللغة ، والأساليب المستخدمة لاكتسابها ووضع المادة التعليمية والمناهج الملائمة ووضع مخطط للتقييم العملية التعليمية. وهكذا يمكن إنشاء مدونة لمتعلمي اللغة العربية مصممة وفق معايير دقيقة لدعم الدراسات التطبيقية المعتمدة على هذه المدونات في جميع المجالات البحثية. هذه المدونات تراعي في مناهجها الأخطاء السابقة لمتعلمي اللغة وكيفية تجنب الوقوع فيها ثانية [6]. وذلك كله سيقفنا على إعادة النظر في كثير من المقولات الشائعة والراسخة في تحليل الأخطاء وتعليم اللغة الثانية، من أهمها إعادة النظر وبطريقة أدق وأعمق في أثر اللغة الأم في تعلم اللغة الثانية ومدى سلبه أو إيجابه.

ولا شك أن تحصيل هذه النتائج من تحليل الأخطاء سيترتب عليه نتائج مهمة وكبيرة في بناء مناهج تعليم اللغة العربية للناطقين بلغات محددة، وستتميز هذه المناهج باعتمادها على نتائج دقيقة إلى حد بعيد بالنظر إلى اعتمادها على مئات الدراسات والبحوث والحالات.

وأهم هذه المدونات : المدونة اللغوية لتعلم اللغة العربية Arabic Learner Corpus الهدف من هذا المشروع توفير مدونة لغوية لنصوص حررها متعلمو اللغة العربية في المملكة العربية السعودية. وتتألف من مجموعة من المواد المكتوبة والمنطوقة. قام ببناء المدونة الباحث: عبدالله الفيقي تحت إشراف إرك أتول. المدونة مشروع مفتوح المصدر، ومرخص ضمن Creative Commons Attribution. تتألف المدونة اللغوية لمتعلمي اللغة العربية من مجموعة من المواد المكتوبة والمنطوقة التي حررها متعلمو اللغة العربية في المملكة العربية السعودية. تم جمع بيانات المدونة في ٢٠١٢ و ٢٠١٣، وهي تضم ١٥٨٥ نصاً (٢٨٢,٧٣٢ كلمة)، شارك في تحريرها ٩٤٢ طالباً من ٦٧ جنسية، و ٦٦ لغة أم مختلفة. متوسط طول النصوص ١٧٨ كلمة.

وصلة المدونة: <https://www.arabiclearnercorpus.com/home-ar>

(أ) أنواع الأخطاء [6]

ثمة أنواع الأخطاء التي يقع فيها الطالب المتعلم لغة غير لغته الأم، "فالأخطاء منها ما يعد كسراً للنظام وبنية اللغة"، ومنها يعد ما يعد سوء استعمال لها. فمن الأول ما يتصل بالتركيب والسياق ، ومن الثاني ما يتصل ما يتصل بسوء ما يتصل بسوء بين اللفظ واللفظ ، أو بين اللفظ والمعنى ، ومن ما يتصل بعدم مراعاة المقام.

(ب) مصادر الأخطاء

ويمكن تقسيم الأخطاء إلى ما يلي:

- أ) الأخطاء التي تمثل التداخل اللغوي أو نقل الخبرة.
 ب) الأخطاء التي تمثل التداخل اللغوي اللغة نفسها.
 ت) الأخطاء التي تمثل التداخل اللغوي التطور اللغوي.

(ت) تحليل الخطأ

- أ) تحليل الأخطاء الإملائية والصوتية. -تحليل الأخطاء الصرفية.
 ب) تحليل الأخطاء النحوية. تحليل الأخطاء المعجمية والدلالية.
 ت) ويحدد منها [6]
 ث) الأخطاء المشتركة التي يرتكبها الناطقون بلغة أم معينة (التركية مثلاً).
 ج) الأخطاء المشتركة التي يرتكبها الناطقون بمجموعة لغات أم بينها صلات محددة (الإنجليزية والإسبانية والإيطالية).
 ح) الأخطاء المشتركة التي يرتكبها جميع متعلمي اللغة العربية على اختلاف لغاتهم الأم.
 خ) الأخطاء النادرة جداً التي يمكن ردها إلى أخطاء فردية مثلاً.

(ث) ترميز الأخطاء

انظر فهرس (2) [13]

(ج) أدوات تقدمها المدونة

أداة لوسم الأخطاء اللغوية بمساعدة الحاسب Computer-Assisted Error Annotation ، والتي توفر مجموعة من الوظائف العملية للمساعدة على وسم الأخطاء، مثل وظيفة التحديد الذكي Smart Selection ، ووظيفة الوسم الآلي Auto Tagging ؛
 أداة أخرى للبحث في نصوص المدونة على شبكة المعلومات الدولية ("الإنترنت www.alcsearch.com") ، وهذه الأداة تمكّن المستخدم من البحث في نصوص المدونة وفق مجموعة من المحددات، مع السماح له بتنزيل هذه النصوص على أكثر من هيئة (TXT، و XML ، و PDF، و MP3).
 وقد تم استخدام واسع لنصوص المدونة منذ إنشائها في عدد من المشاريع البحثية النظرية والتطبيقية، ومنها مجموعة من الدراسات في مجال اكتشاف الأخطاء وتصحيحها آلياً، وتقييم المحللات الصرفية للغة العربية، واكتشاف اللغة الأم لكاتب النص، وكذلك مجموعة من الدراسات في مجال اللغويات التطبيقية، والتعليم الموجه بالبيانات Data-Driven Learning ، وغيرها.

4. من تطبيقات دمج التدوين اللغوي والعمل الإحصائي: المفهرس Concordance كأداة تعليمية

تلعب المُدَوَّنَاتُ اللُّغَوِيَّةُ دورًا مهمًّا في دعم الإحصاء اللُّغَوِيِّ، أو ما يُعرَفُ بعلم اللُّغة الإحصائيِّ Statistical Linguistics. واعتماداً على نُصُوصِ المُدَوَّنَاتِ اللُّغَوِيَّةِ يُمكن إحصاء الكثير من خلال برنامج يطبق على مدونات مرمزة صرفياً ووتركيبيًا ونحوياً ومن أمثلتها بنوك الأشجار العربية السابق الإشارة إليها في [5]. والمدونة التي قامت الباحثة بالإشراف وتحليلها [5] وأخرى بالتعاون مع [1]. ويتصل بقاعدة بيانات تنقسم إلى ثلاثة أقسام لقواعد البيانات:

(1) استكشاف النصوص من خلال المفهرس Concordancer

يمكن تعريف المفهرس الآلي Electronic Concordance بأنه برنامج إداري يُستخدم في تنظيم النصوص وفهرستها وترتيب مفرداتها حسبما تقتضي طبيعة الدراسة المُعدّة؛ ويُعتبر أحد الأدوات الأساسية المُستخدمة في تحليل نصوص المُدونات اللغوية، لاسيما المُدونات المُصنوعة لأغراض مُعجمية. ويُوفّر المفهرس الآلي الكثير من الوقت والجهد، إذ يُعيد تشكيل النصوص المُدرجة لتظهر في صورة مُنظمة، يسهل التعامل معها آلياً، سواء على مستوى المُفردات، أو على مستوى الجمل والتراكيب.

وأهم ما يميّز المفهرس الآلي أنه قادر على:

- (1) ترميز حروف اللغة المُعينة ضمن مفردات النص.
- (2) إعادة ترتيب حروف اللغة المُعينة وفقاً لنظام الحروف الدولي الموحد Unicode.
- (3) إعادة تعيين النص المُدرج بعد حصر المفردات وترتيبها.
- (4) تعيين كلمات النص المُدرج ضمن سياقاتها.
- (5) تجميع المفردات المُتمثلة في حُقول وإعادة ترتيبها.
- (6) كذلك.. يُعطي البرنامج عدداً من خيارات الترتيب للنتائج من المفهرس:

- (أ) ترتيب المفردات ألفبائياً ترتيباً تصاعدياً.
- (ب) ترتيب المفردات ألفبائياً ترتيباً تنازلياً.
- (ت) ترتيب المفردات بحسب أكثرها شيوعاً.
- (ث) ترتيب المفردات بحسب أقلها شيوعاً.
- (ج) ترتيب المفردات تبعاً لموضعها في النص المُدرج.

(2) أهم مشكلات المفهرسات الآلية العربية سنورها بعينها ولمزيد من التفصيل راجع [1]:

لا تُراعي النظام البنوي للغة العربية، تحتاج إلى ذاكرة حاسوبية كبيرة لمعالجة الجمل الطويلة، تحتاج إلى وقت طويل للقيام بتحليل وترتيب النصوص المُدخلة. ، بعض من المفهرسات الآلية المُتاحة إلكترونياً والتي تتعامل مع اللغة العربية.

(3) استخدام المفهرس

الفهرسة Concordance: خطوط في النصّ توضّح كلمة البحث، والتي تسمى العقدة the node. البحث بالمفهرس: في المفهرس، الخطوط الظاهرة تسمى: الاكتشافات finds. والكلمات المحددة بلون أو خط أسفلها تسمى العقدة. وتسمى الكلمة الدليلية في السياق Key KWIC Word In Context. إذا قارنت بين استعمال كلمتين ومعرفة تردداتهن النسبية التي قد تختلف تبعاً لحجم المدونة مما يؤثر على النتائج.

(4) نماذج من مفهرسات اللغة العربية المُتاحة على الشبكات:

المفهرس الآلي Concordance.

الجهة المُنتجة: مجموعة R. J. C. WATT. ، اللغات التي يدعمها: جميع اللغات المُعتمَدة لدى Microsoft. برنامج تنفيذي بامتداد EXE، يُمكن الحصول على البرنامج من الرابطة:

<http://www.findmysoft.com/download-Concordance/conc320.exe>

Headw...	No.	Context	W...	Context	Line
وحدة	1704	...فصلاً كما لا أحد يقص...	الندرية	...على التمييز بين من م...	14228
كلمات	1712	...تصانيف من التي ت...	الندرية	...تسببت تغيراً في عقد...	14481
كثير	1714	...حصولاً على نصيحة من...	الندرية	...أن لم تعلم بوظيفة لا...	14599
كثير	1715	...وكلمات العودة برأس ح...	الندرية	...بشركهم كانوا من لروؤ...	14955
النص	1718	...وكانت ذلك طاهره أن...	الندرية	...لشعروها لثباتها أ...	14956
النص	1718	...بأنه ذلك صفة تسمى...	الندرية	...بنتابون على قلبها و...	15240
النص	1718	...بأنه ذلك من فرانسوا ز...	الندرية	...التي تحدث برونه من...	17064
النص	1719	...أن يفهم جميع المدن...	الندرية	...التي لا تفت جوتها و...	18011
النص	1720	...الجملة	الندرية	...التي لا تكاد تجد أ...	20705
النص	1722	...الإكتشاف	الندرية	...التي لا تكاد تجد أ...	21768
النص	1723	...التي تأتي وجوده وسط...	الندرية	...التي لا تكاد تجد أ...	22107
النص	1725	...وكان هذا أن تكون أكثر...	الندرية	...التي لا تكاد تجد أ...	29993
النص	1725	...لكن لم يسمح لها على...	الندرية	...التي لا تكاد تجد أ...	30471
النص	1726	...أن تترك الجمل كذا أو...	الندرية	...التي لا تكاد تجد أ...	31400
النص	1733	...بجهد الجهد من ج...	الندرية	...التي لا تكاد تجد أ...	32158
النص	1737	...بالفعل فاصلة وحدة و...	الندرية	...التي لا تكاد تجد أ...	32779

شكل(4) نموذج من ناتج المفهرس من البحث عن كلمة "البشرية"

(5) تهيئة المدونة بخصائص لعلاج المشكلات الدلالية والمعجمية لمتعلمي اللغة من غير الناطقين بها:

يقول فرازي تُجمع المدونة التعليمية من كتب لغويين وكتاب ذوي إلمام واسع بالمفردات والقواعد اللغوية. واقترح البحث أن تمثل المدونة مجالات اللغة جميعها. ولكن يمكن أن نستقطع جزء منوع من المدونة نهيئه لحلول المشكلات الدلالية و المعجمية التي تعرض لها [14] حيث يتم:

(1) يجب الوقوف على الفروق الأجناسية للمتعلمين، وكذا الفروق المتعلقة بأعمارهم وثقافتهم. وكلها عوامل تؤثر في قدرة كل واحد منهم على حدود استعمال المفردات المتداولة، بل يجب أيضا مراعاة التراكيب والجمل المختلفة والمتنوعة التي يستوعبها طرف دون الطرف الآخر.

(2) ينبغي أن تراعي حاجة المتعلمين إلى كلمات معينة في كل مرحلة من مراحل التعلم حتى لو كانت هذه الكلمات معقدة خاصة المصطلحات التي يحتاجها المتعلمون في مجال قراءة النصوص الشرعية أو الأدبية.

(3) العمل على تنوع الأجناس الأدبية التي تجعله في منأى عن الرتابة، وبالتالي الوصول إلى أهداف ثقافية زاخرة.

(4) استبدال المجازات والألغاز التي لا يدركها إلا الراسخون في ميدان اللغة بألفاظ سهلة وواضحة لو لم تتوفر أجزاء من المدونة بعيدة عن هذه المجازات والألغاز.

ملاحظة الباحثة: لا أحبذ الإستبدال ولكن نوفر للمستخدم بدائل أبسط وأسهل معها.

(5) التدخل في كثير من الأحيان في طبيعة النص، وجعله يستجيب للهدف المنشود. وغالبا ما ما يفرض هذا التغيير والاستثناء، والتبسيط، انطلاقا من الكلمات المستخدمة في شتى الموضوعات المدروسة.

(6) اختيار الكلمات اختيارا علميا دقيقا وتقديم للمتعلمين تقديمًا جيدا يراعي فيها الأساليب العلمية التربوية فلا بد من العمل على تبسيط اللفظ لجعله مألوفا ومتداولًا لدى العامة من الناس، ما دامت الغاية من تعلم اللغة العربية، التواصل مع الناس، والعمل على أن تكون الكلمات شائعة الاستعمال وتقدم بشكل تدريجي من السهل إلى الصعب، وهذا عائق منهجي يجب مراعاته.

(7) الاحتياط من الوقوع في نص يزخر بالضمائر الغيبية، وكثرة الجمل المعتمدة والمبنية على الصفات وغيرها.

(8) الاعتماد على تكوين الجمل والتراكيب القصيرة والسهلة التي تخول للدارس المبتدئ، التأقلم مع العربية وبعشق استكشافي مثير.

ملاحظة الباحثة: يشترط أن تكون مأخوذة من سياق المدونة لا مؤلفة.

(9) تشجيع الطلاب على فهم الكلمة في سياقها الذي وردت فيه و عدم حفظها في قوائم معزولة عن سياقها .

(10) تقديم الكلمات بنسب مقاربة من خلال أنماط شائعة الاستعمال و متدرجة من حيث الصعوبة التعقيد بما يتناسب مع مستوى المتعلمين بحيث تكون فوق مستوى المتعلمين قليلا فلا تكون سهلة جدا و لا صعبة جدا.

- (11) يكون اختيار الكلمات و ترتيبها و تقديمها مبنيًا على الدراسات اللغوية النفسية في اللغة العربية التي تبين التدرج الطبيعي لاكتساب مورفيمات اللغة العربية وصيغها الصرفية ووظائفها النحوية .
- (12) يجب أن يكون محتوى النصوص معروفا و مفهوما لدى المتعلمين فلا يجمع النص بين صعوبة الكلمات و غرابة المعنى . و تقدم الكلمات الجديدة ذات المعاني غير المألوفة لدى المتعلمين من خلال أنماط مألوفة و تراكيب قصيرة و أساليب سهلة ليتمكن الطالب من معرفة معنى الكلمة الجديدة من غير حاجة للبحث عنها في المعجم .
- (13) الإهتمام بالجزئية الصوتية في المدونة من أجل عملية الإنصات: وهي عملية متعلقة بالقراءة لكتب و قصص يستفيد منها المتقدمون باللغة لتزداد ثروتهم اللغوية ، و لا تخلو بدورها من مهارة، علما و إيماننا أننا كثيرا ما نتجاهل عمق الإنصات و ما يترتب عنه من تواصل عميق و دقيق بغية الوصول إلى الفهم. و الهدف منه تعزيز المفردات و الجمل التي سبق أن تعلمها الطلاب في سياقات مسموعة بعد تعرضهم لها في سياقات مقروءة.
- (14) تزويد الشروح بالصور التوضيحية اللازمة بشرط أن تكون واضحة و ضرورية و أن توضع في مكانها المناسب.
- (15) تحذير الطلاب من استخدام معاجم ثنائية اللغة و حثهم على استخدام معاجم أحادية اللغة لأنها تثري حصيلتهم اللغوية و تزودهم بالأمثلة و الإستعمالات الحقيقية للكلمة من خلال شواهد المدونات. إذن يجب ربط المعجم بشواهد و أمثلة لإستخدام الوحدات اللغوية من خلال المدونة .

5. مدونات تعليمية مرزمة للتعلم من مساهمات الباحثة

(1) مدونة مساك [5]

تم عمل مدونة تمثل أغلب النصوص العربية و تحتوي أكثر من نصف مليون كلمة. و تم عمل تذييل لهذه المدونة على المستوى الصرفي و النحوي و الأسلوبي من قبل لغويين بشريين مما كان مكلفا جدا كبحث فردي. و قد تم حصر التذييلات و عمل رسوم بيانية للتوضيح كما في فهرس (3).

مرحلة عمل و بناء مساك

برنامج التحليل الإحصائي انظر ملحق(3):

البيئة البرمجية للبرنامج 2008 / 2005 Microsoft Visual Studio .net حيث تمد المبرمج بالقدرة على استخدام العديد من اللغات و قد استخدم هنا السي شارب #c. استخدمنا لقواعد البيانات Access 2007 على الرغم من قدرة Microsoft SQL server 2005/2008 إلا أنها تحتاج رخصة حق استخدام غالية الثمن. و يحتاج البرنامج لحاسوب ذو مشغل سريع high processor و رمات RAM عالية ليعمل بسرعة مناسبة.

خواص البرنامج:

القدرة على رفع الرموز من الملف المرمز. ، القدرة على إضافة ملف للملف المفتوح في صندوق النص text box . ، القدرة على كتابة ملف جديد. ، القدرة على حصر عدد الوحدات اللغوية مثل: الجمل و التراكيب و الساليب و الكلمات بكافة صورها.

القدرة على عرض كل تراكيب لكل الوحدات اللغوية السابقة على حدة.

(2) المدونة (هوية) و المعجم كمنتج تعليمي

تم عمل مقترح لمعجم عصر المعلوماتية من خلال المدونة اللغوية أسميتها هوية [2] و وضع منهجية لوضع السمات الصرفية و النحوية و الدلالية و العلاقات بين الكلمات بطريقة توضح المعاني و تجلي الغموض و من ثم يمكن الاستفادة منها في أغلب التطبيقات على اللغة العربية. كذلك في حالات تحديد المجازي و المصطلحات حيث يقوم المعجم المقترح على أسس علمية منطقية سواء في التصنيف أو في تحديد أشكال العلاقات داخل الحقل المعجمي الواحد مع

الاهتمام ببيان العلاقات الموجودة بين كلمات الحقل الواحد ووضعها في صورة خصائص أو ملامح دلالية تتلاقى و تتقابل داخل الحقل الواحد.

6. مساهمات مقترحة للبحث الإحصائي في المدونات لتعلم اللغة:

- (1) إيجاد مداخل معجمية جديدة
- (2) تدعيم المعجم العربي بسمات وخصائص ودلالات جديدة. اثرء المعاجم العربية بالعلاقات والسمات [1]. وإثراء المداخل معجمية الموجودة بالمعلومات الإضافية التي اكتشفت أثناء تحليل المدونة (ومثال على ذلك:- أشكال الأكثر شيوعاً، التضمين، الخ).
- (3) ملاحظة سمات مهمة في المعنى وقواعد الكلمة بالعمل على بيانات العمل مع واضحة وميسرة.
- (4) استعمال تحليل تردد الكلمة لتدليل المداخل المعجمية به.
- (5) تحديد المصاحبات اللغوية وفك اللبس عن طريق توظيفها في فكه. وتجمعات الكلمات تجمع وتنظم وتعرض كالتعبيرات الاصطلاحية مثلاً.
- (6) عرض "مواد معجمية من غير المحتمل أن تُوجَد في مصادر القاموس (ومثال على ذلك: - أسماء العلم)
- (7) عرض أمثلة وشواهد للغة توضح طريقة استعمالها والتعامل معها.
- (8) تحديد المفردات الأكثر شيوعاً.
- (9) تحديد المترادفات الأكثر شيوعاً، وكذلك الأضداد والمتضادات، وجميع الظواهر اللغوية الأخرى.
- (10) دراسة مدى تأثير غياب حركات التشكل، وعلامات الترقيم على معني الكلمات.
- (11) تحديد مدى استخدام وتأثير المفردات والتراكيب العامية والأجنبية في اللغة.
- (12) تحديد الأخطاء اللغوية الشائعة في الكتابات المعاصرة.
- (13) انشاء معجم آلي يضم السمات اللغوية على جميع المستويات .
- (14) المساهمة في تصحيح الأخطاء (الصرفية -النحوية-الدالية).
- (15) تحدي التعبيرات الاصطلاحية واستخدامها ومعانيها من خلال السياق. وقد اكتشفنا تنوع المعاني واكتسابها معاني جديدة من خلال السياق - بحث -.
- (16) دعم التصنيف الدلالي للمجالات.
- (17) عمل معاجم شاملة ومتكاملة .
- (18) مد المعجم بأمثلة من اللغة الحية. ودعم المعاني وشرحها - شواهد لغوية hard measurable evidence - ويمكن استخدامها في تعليم اللغة لأبن اللغة والأجنبي عنها.

7. التعليق النهائي

حاولنا في هذا البحث أن نمد مسؤول تعليم اللغة العربية والباحثين الراغبين في العمل بتعليم اللغة العربية أو في صناعة المدونات بكل ما يلزمهم من معلومات. فقد لمسنا ندرة هذه النوعية من الأبحاث عن اللغة العربية وبها. والإشارة لأهمية المدونة الخام والمذيلة في تعليم اللغة من خلال ما تمدنا به من معلومات. وقد قمت بعرض بعض التجارب لبعض في كل الجزئيات متى أتيت لي الفرصة حتى يشعر القارئ بإمكانية التطبيق العملي وخاصة المدونة مساك وضح البحث المنهجية التي اتبعت لخصر صور الجملة العربية من خلال مدونة نصية مكونة من نصوص متنوعة مكتوبة في العصر الحديث أدرجت تحت مسمى اللغة العربية المعاصرة. وأمكن تطبيق نفس المنهجية مع النص القرآني الشريف.

إن المنهج المتبع يمكننا من عدة أهداف؛ منها رسم صورة واضحة للجمل والمركبات في العربية المعاصرة ونسبة شيوع وانتشار هذه الجمل والمركبات، فمثلاً يمكننا من معرفة نسبة شيوع الجمل الاسمية بالنسبة للجمل الفعلية كما يمكننا من معرفة النسب المختلفة للمركبات التي تقع في موقع وظيفي معين؛ فمثلاً يمكننا من معرفة


نسبة وقوع المبتدأ ضميراً ظاهراً، ونسبة وقوعه اسماً ظاهراً أو أحد المركبات الاسمية أو مصدراً مئولاً، ويمكننا من معرفة نسبة شيوع الجمل المثبتة بالنسبة للجمل المنفية. أن لهذا المنهج جدارته في تعلم اللُّغة العربية وتعليمها وذلك لمراعاته اطراد القواعد، والسعي إلى الفهم والتفسير، واستخدامه لرموز عربية ذات دلالة لمتعلم اللغة العربية. وقد شرح البحث العوامل التي تحتاجها مثل هذه الأعمال.

ثم عرض بعض نتائج البرنامج الإحصائي لخصر صور الجمل والتراكيب والأساليب في نصوص المدونة والتي تشمل النصوص الحقيقية المستعملة في الحياة اليومية لا تحليل مجموعات مختارة من الجمل كما كان يحدث في أغلب الأعمال من قبل. ومن نتائج هذا الخصر وضع تصور لبرنامج آلي لتحليل النصوص وهذا نتركه لبحث جديد.

8. المراجع

- [1] المعزز بالله السعيد ، رسالة ماجستير بعنوان: "مدونة معجم عربي معاصر، معالجة حاسوبية"، جامعة القاهرة، كلية دار العلوم، 2009.
- [2] سلوى حمادة، كتاب " المدونات النصية و دور اللغة العربية في التعامل معها"، نور للطباعة والنشر ، ألمانيا، 2017. كما تتوفر نسخة على موقع الأمازون للبيع.
- [3] Latifah Al-Sulaiti, Designing and Developing a Corpus of Contemporary Arabic, Leeds University. P.1. ,2004.
- [4] S. Kulick and others, Kulick
<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49>, 2006
- [5] سلوى حمادة، " تحليل للجمل العربية مبني على مدونة نصية من النصوص المعاصرة"، المؤتمر العاشر لجمعية هندسة اللغة، 15-16 ديسمبر، 2010، القاهرة.
- [6] وليد العناتي، "تحليل الخطاب وتعليم اللغة العربية لغير الناطقين بها"، البصائر – العدد رقم 2، 2010.
<http://archive.sakhr.it.co/newPreview.aspx?PID=3006500&ISSUEID=18308&AID=410765>
- [7] عبد الله الفيقي، "المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية ArabicLearner Corpora", "المؤتمر الدولي لعلوم وهندسة الحاسوب، كلية الحاسبات، القاهرة، جمهورية مصر العربية، 2012.
- [8] سلوى حمادة، "نحو وضع تصور لتحليل النصوص العربية المعاصرة"، المجلة العربية لعلوم وهندسة الحاسوب ، المجلد الأول ، العدد الثاني ، 2007، فليبس للنشر، الولايات المتحدة الأمريكية
<http://www.phillips-publishing.com/jcsea/publications.html>
- [9] من مناقشات مع متخصص في تدريس اللغة العربية لغير الناطقين به الأستاذ شادي مجلي سكر من الأردن.
- [10] راضي فوزي، "من مشكلات تعليم اللغة العربية لغير الناطقين بها"، 2009
<http://www.al-maqha.com/t9096.html>
- [11] Maamouri, M. & Cieri, C. (2002). Resources for Arabic Natural Language Processing at the Linguistic Data Consortium. Proceedings of the International Symposium on Processing of Arabic. Faculté des Lettres, University of Manouba, Tunisia.
- (PDF) *The penn arabic treebank: Building a large-scale annotated arabic corpus*. Available from: https://www.researchgate.net/publication/228693973_The_penn_arabic_treebank_Building_a_large-scale_annotated_arabic_corpus [accessed Oct 17 2018].
- [12] محمد رمضان الغليظ، " المدونات الحاسوبية لأرصدة اللغة العربية" رسالة ماجستير من كلية الآداب، جامعة عين شمس، 2018م.
- [13] Abdulla Alaifi, Eric Atwell, " Second Workshop on Arabic Corpus Linguistics (WACL-2), 22nd July 2013, Lancaster University, UK
- [14] عبد السلام فزاري، " أول مدونة لتعليم اللغة العربية لغير الناطقين بها"، Access on October 2018

<http://fizabiabdeslam.ahlablog.com/Aaa-aIaaE-b1/EUaia-CaaUE-CaUNEiE-aUiN-CaaCOia-EaC-b1-p15.htm>

	<p>نبذة عن الباحثة: أ.د سلوى حمادة الحائزة على جائزة الكندي لأفضل باحث معلوماتي عربي. ورشحت من قبل جامعة عين شمس لجائزة الملك فيصل اللغة العربية والحاسب من قبل جامعة عين شمس. وهي كاتبة سلسلة كتب : المعالجة الآلية للغة العربية وصدر منها: 1- المشاكل والحلول طباعة القاهرة دار غريب 2009 . 2- النظرية والتطبيق طباعة وزارة الثقافة 2012. 3- المدونات النصية طباعة دار نور للنشر بالمانيا. دُعيت للعديد من الندوات والمؤتمرات في الدول العربية والأجنبية. تم كتابة الكثير من الموضوعات الصحفية عنها. وتم عمل لقاءات إذاعية وتلفزيونية معها. لها العديد من المقالات عن اللغة العربية. لها أكثر من ستين بحثا منشورا.</p>
---	---

Concordance of Linguistic Corpora and Its Use in Language Teaching Arabic as an Example

Salwa Hamada

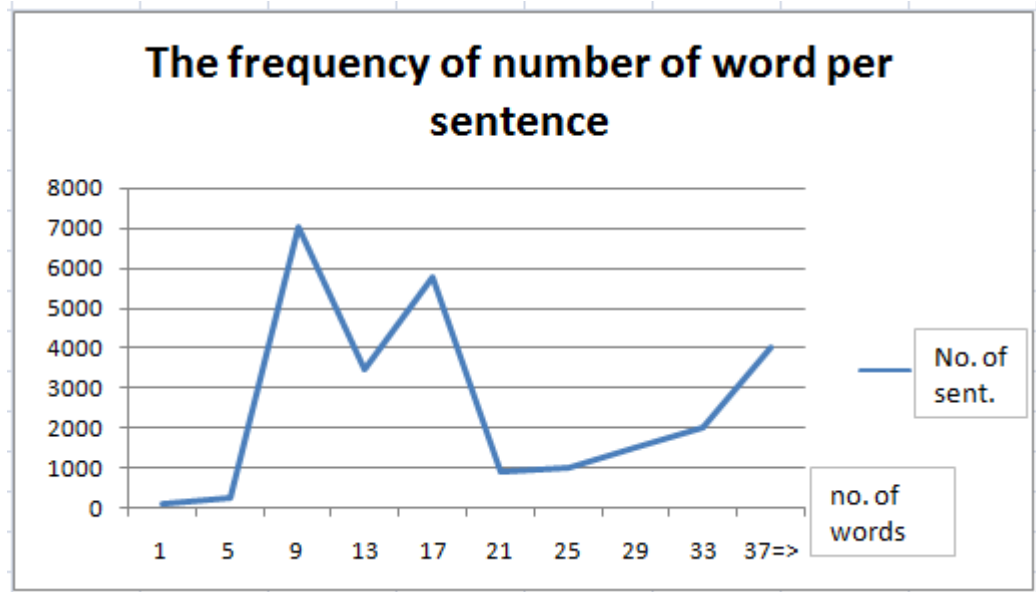
Department of Informatics, Electronics Research Institute

Egypt

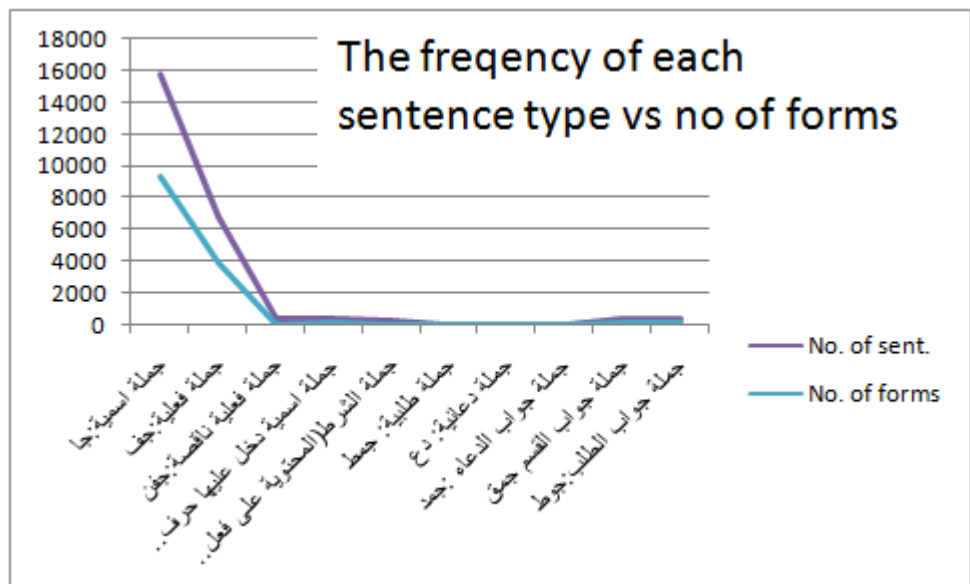
hesalwa@hotmail.com

Abstract: Arabic Language has now become a universal language and with its influential on the world events, a lot of attention is being directed to it in all countries of the world. One of the most important methods of teaching methods of language all over the world is through building Corpus. Despite of importance of educational Corpus, the Arab region is still not paying attention to it which forced us to write this research. The paper aims to clarify Corpus typed and criteria of building them. It also explained the Corpus Linguistic and methods of using them. The importance of the learning Corpus has been shown with presenting some of the Learning Corpus, especially which I contributed in their construction.

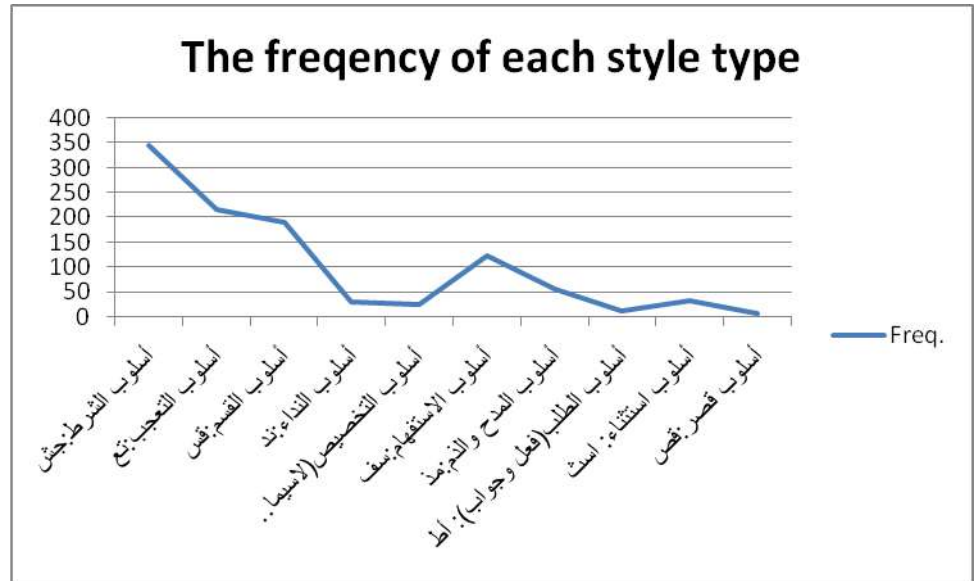
من نتائج البحث في المدونات مساك



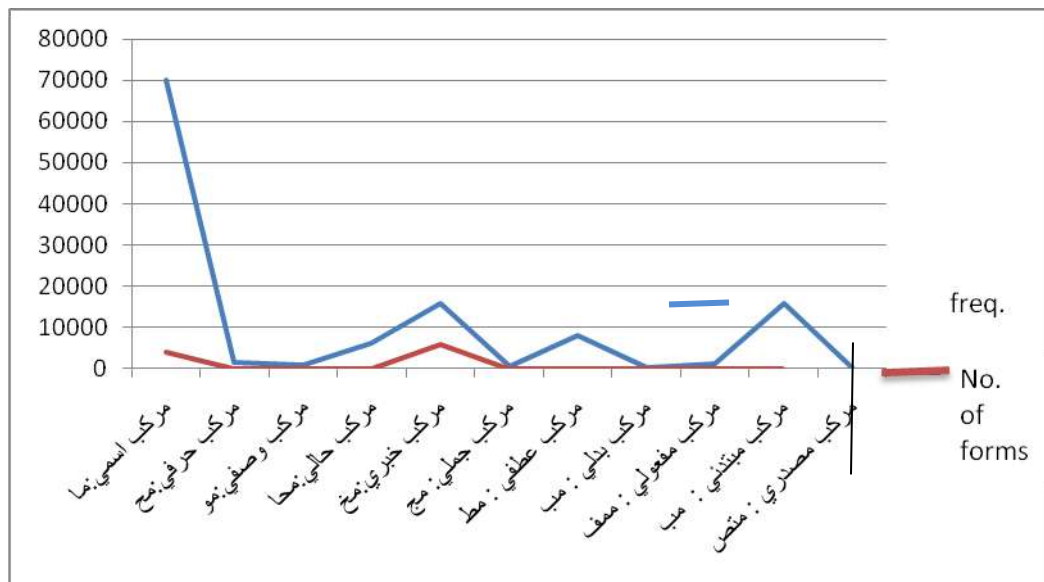
شكل (5). عدد الجمل بالنسبة لعدد الكلمات في مساك [5]



شكل (6). تردد الجمل بالنسبة لأنواعها. types of sentences number versus their forms.



شكل (7). تردد جميع أنواع الأساليب. The frequency of each style type.



شكل (8). تردد جميع أنواع المركبات النحوية
The frequency of each syntactic structure type

فهرس (2)
مدونة متعلمي العربية وترميز الأخطاء [11]

Error Category مجال الخطأ	Error Type نوع الخطأ	Arabic tag الرمز العربي	English tag الرمز الإنجليزي	Error freq. in the test data
Orthography الإملاء 'l'imlā'	1. Hamza (ء، آ، إ، ؤ، ي، ذ)	<إء>	<OH>	5
	2. Tā' Mutadarrifa (ت، ث)	<إث>	<OT>	0
	3. 'alif Mutadarrifa (أ، ي)	<إي>	<OA>	1
	4. 'alif Fāriqa (كثيرا)	<يث>	<OW>	5
	5. Lām Samīya (الطلب)	<إل>		0
	6. Tanwīn (ن، و، ي)	<إل>	<ON>	0
	7. Fasl wa Wasl (Conjunction)	<إو>	<OF>	9
	8. Shortening the long vowels تقصير الصوائت (أوي → أوي)	<إف>	<OS>	2
	9. Lengthening the short vowels تطويل الصوائت (أوي → أوي)	<إق>	<OG>	0
	10. Wrong order of word characters الخطأ في ترتيب الحروف داخل الكلمة	<إط>	<OC>	1
	11. Replacement in word character(s) استبدال حرف أو أحرف من الكلمة	<إس>	<OR>	6
	12. Character(s) redundant وجود حرف أو أحرف زائدة	<إز>	<OD>	6
	13. Character(s) missing وجود حرف أو أحرف ناقصة	<إن>	<OM>	3
	14. Other orthographical errors أخطاء إملائية أخرى	<إخ>	<OO>	0
Morphology الصرف 'ssarf'	15. Word inflection صيغة الكلمة	<صص>	<MI>	2
	16. Verb tense زمن الفعل	<صز>	<MT>	1
	17. Other morphological errors أخطاء صرفية أخرى	<صص>	<MO>	0
Syntax النحو 'nnaḥw'	18. Case/Mood Mark الوضع الإعرابي أو علامة الإعراب	<خب>	<XC>	1
	19. Definiteness التعريف والتذكير	<خغ>	<XF>	11
	20. Gender التذكير والتأنيث	<خذ>	<XG>	3
	21. Number (Singular, Dual and plural) (الإفراد والتثنية والجمع)	<خف>	<XN>	0
	22. Word(s) order ترتيب المفردات داخل الجملة	<خت>	<XR>	1
	23. Word(s) redundant وجود كلمة أو كلمات زائدة	<خز>	<XT>	4
	24. Word(s) missing وجود كلمة أو كلمات ناقصة	<خن>	<XM>	9
25. Other syntactic errors أخطاء نحوية أخرى	<خج>	<XO>	0	
Semantics الدلالة 'ddalāla'	26. Word selection اختيار الكلمة المناسبة	<خدب>	<SW>	17
	27. Phrase selection اختيار العبارة المناسبة	<خدق>	<SP>	1
	28. Failure of expression to indicate the intended meaning فشل التعبير عن أداء المعنى المقصود	<خدد>	<SM>	2
	29. Wrong context of citation from Quran or Hadith الاستشهاد بالكتاب والسنة في سياق خاطئ	<خدس>	<SC>	0
	30. Other semantic errors أخطاء دلالية أخرى	<خدج>	<SO>	0
Style الأسلوب 'l'uslūb'	31. Unclear style أسلوب غامض	<خدع>	<TU>	2
	32. Prosaic style أسلوب ركيك	<خدض>	<TP>	7
	33. Other stylistic errors أخطاء أسلوبية أخرى	<خدج>	<TO>	0
Punctuation علامات الترقيم 'alāmāt 't-tarqīm'	34. Punctuation confusion الخلط في علامات الترقيم	<خط>	<PC>	21
	35. Punctuation redundant علامة ترقيم زائدة	<خز>	<PT>	1
	36. Punctuation missing علامة ترقيم مفقودة	<خن>	<PM>	35
	37. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<خج>	<PO>	0

Syllables Classification for ASR using Variable State Hidden Markov Model

Doaa N. Senousy^{1*}, Amr M. Gody^{2*}, Sameh F. Saad^{3**}

*Electrical Engineering Department, Fayoum University, Fayoum, EGYPT

¹dn1144@fayoum.edu.eg

²amg00@fayoum.edu.eg

** Modern Sciences and Arts University, 6 October City, Giza, Egypt

³dr.sam.far@gmail.com

Abstract: This research is intending to introduce preprocessing classification approach for Automatic Speech Recognition (ASR). Four hybrid models are provided to emphasize the principle idea of this research. The first hybrid model is constructed of fixed state, structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Encoding (FS-HMM-GM-MBT). The second hybrid model is constructed of fixed state, structured Hidden Markov Model, Gaussian Mixture, Mel scaled Frequency Cepstral Coefficients (FS-HMM-GM-MFCC). The third hybrid model is constructed of variable state, dynamically structured Hidden Markov Model, Gaussian Mixture, Mel scaled Best Tree Encoding (VS-HMM-GM-MBT). The fourth hybrid model is constructed of variable state, dynamically structured Hidden Markov Model, Gaussian Mixture, Mel scaled Frequency Cepstral Coefficients (VS-HMM-GM-MFCC). A subset of TIMIT database is used in this research. The involved classes in this research are vowel, consonant, liquid, nasals, stops, and plosives. VS-HMM-GM-MBT achieves the highest overall recognition rate (81.01%). It succeeds in recognizing stops, consonants, and liquids at a higher rate than MFCC in all the experiments. This research provides scientific base of the included syllables characteristics in terms of spectrum analysis. This study is useful for anyone intending to go this way of syllables classification.

Key words: Automatic Speech Recognition, Classification technique, Hidden Markov Model, MFCC, Wavelet Packets.

1 INTRODUCTION

Automatic speech recognition (ASR) is used now in many applications. ASR tends to help people to make life easier. This research is providing a novel classification method to increase the success rate of ASR.

The research goals in this research paper are:

- 1- To evaluate the proposed syllables classification method.
- 2- To evaluate Mel scaled Best Tree Encoded (MBT) features against the most popular features in the market Mel Frequency Cepstral (MFCC) features.
- 3- To provide a comparative study about syllables characteristics in terms of spectrum analysis using two different hybrid approaches.

A. Literature Review

Researchers are trying to invent a speech recognition method that works as the same as our brain or to improve the recognition rate of techniques used now. The rate of correctness and accuracy of the speech recognizer is affected by many reasons such as noise, feature extraction technique, the pronunciation model, the acoustic model, and the language model. Automatic speech recognition (ASR) is used in a huge number of applications such as dictation, Siri application, robots, mobile devices and others.

Feature vectors are extracted from the speech in the process of speech parameterization techniques. There are many speech feature methods: Real Cepstral Coefficients (RCC) [1], Linear Prediction Coefficients (LPC) [2], Subband Based Cepstral

(SBC) [3], Wavelet Packet Features (WPF) [4], Wavelet Packet parameters for Speaker Recognition (WPSR) [5], but the most popular one of them is Mel-frequency cepstrum coefficients (MFCC) [6]. Gody in [7] invented a new speech parameterization technique called Best Tree Encoding (BTE).

Researchers use many classification methods to get a higher rate of recognition. Speech classification has many applications in life, such as prosody.

Reynolds and Antoniou in [8] introduced research into the classification of the speech signal into seven classes. These classes are fricatives, semi-vowels, diphthongs, plosives, nasals, closures, and vowels. A set of 39 TIMIT phone set was used. This study was implemented by four feature extraction techniques which were: MFCC, perceptual linear prediction (PLP), LPC and posterior net combination between them. This survey was performed using HMM and Multi-Layer Perceptron (MLP). The highest rate of recognition was achieved using a combination of MFCC, PLP, and LPC. The achieved phone classification rate was 84.1%.

Scanlon et al. in [9] suggested a framework to classify the speech signal into vowels, stops, fricatives, and nasals. TIMIT database was used. PLP with first and second derivatives was used in this system. The classification was performed using MLP and HMM. The highest obtained phone accuracy was 74.2%.

Kiss et al. in [10] introduced a research on the segmentation and classification of the speech signal into the unvoiced closure period, voiced spirant noise, unvoiced spirant noise, voiced burst, unvoiced burst, voiced closure period, nasal, high vowel and low-middle vowel. MRBA, KIEL, and TIMIT were used as the corpus in this study. Feature vectors were determined using Bark-scale spectral resolution. TIMIT corpus achieved the highest average of classification accuracy with 80%. The confusion matrix showed 90% success in low-middle and high vowel recognition.

Nasereddin and Omari in [11] introduced a research into the classification of the speech signal into 4 classes. MFCC was used as a feature extraction technique. The classification was made using HMM, Dynamic time warping (DTW), and Dynamic Bayesian Network (DBN). Results showed that DBN outperformed on the other techniques in recognizing one class, but HMM succeeded in achieving the highest recognition rate for the other three classes.

The current research presents new classification techniques with two methods of feature extraction. Various Gaussian mixture numbers are used to get a higher rate of recognition. Comparison between each classification technique is made by applying different feature extraction techniques with different Gaussian mixture numbers.

The following sections explain the stages of the experiment: Section 2 illustrates the steps of best tree encoding and the steps of MBT feature extraction technique. Section 3 illustrates the experiment procedure while section 4 indicates HMM models that are used. Section 5 presents the results of the experiment and section 6 concludes the paper.

2 BEST TREE ENCODING STAGES

A. Framing and windowing

The signal is divided into a number of frames to be stationary in the frame period as the speech is a non-stationary signal. The length of the frame is 20 ms. Then the hamming window is applied to make a smooth transition to the signal to be continuous.

B. Wavelet packet decomposition

The signal can be represented using wavelet packets. Applying low pass filter (LPF) and high pass filter (HPF) on the base signal. The mother signal is divided into two signals and then applying LPF and HPF again on the two signals. This process continues until level 4 is reached. The output of this stage is a tree structure. Example of the tree structure is as shown in Fig.1.

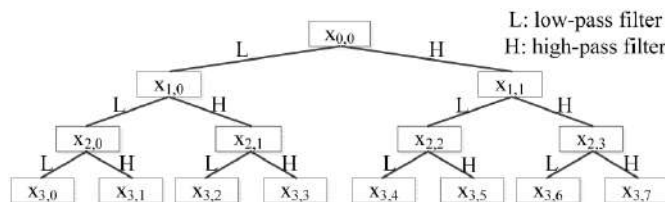


Figure 1: Wavelet packets for 3 levels decomposition. L represents low pass filter and H represents high pass filter. [12]

First, BTE makes an adjustment to the neighboring groups as in Fig.2. Second; the concept of BTE is applied. This concept is based on that the node indices are ordered in such that the nearest node has the nearest frequency as in Fig.3.

A. Entropy and Encoding

We use Shannon entropy to obtain the nodes that contain useful information. The output of this stage is the best tree. The bandwidth is divided into four regions. Each region has the quarter of the bandwidth. The 4 point encoding is indicated as shown in Fig. 4. The best tree is encoded into 7 bits to get feature vectors. Feature vectors are represented as a decimal number. Fig.5 gives an example of feature encoding. Circles point to leaf nodes in BTE that contain information. The feature vector of BTE has 4 components. Table 1 indicates the Best tree encoding evaluation of Fig. 5.

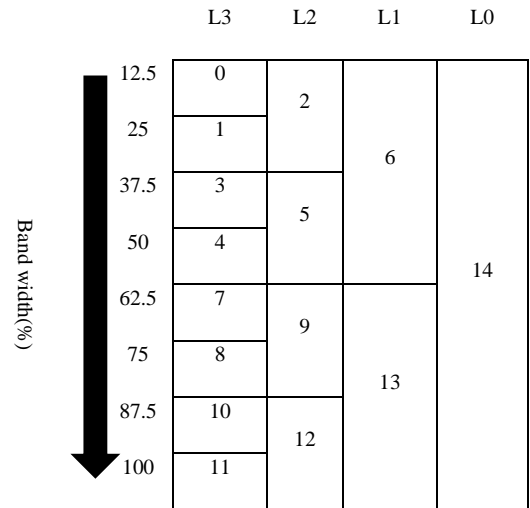
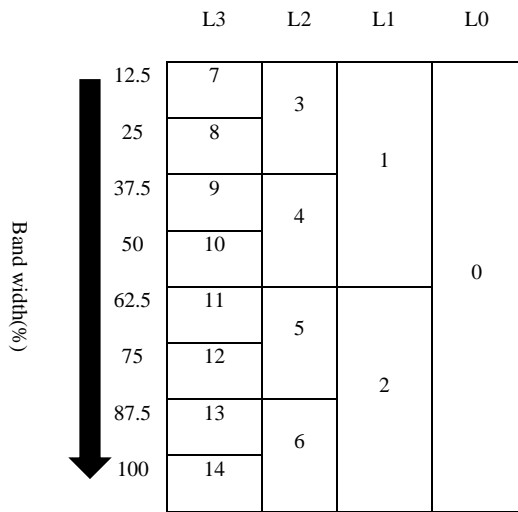
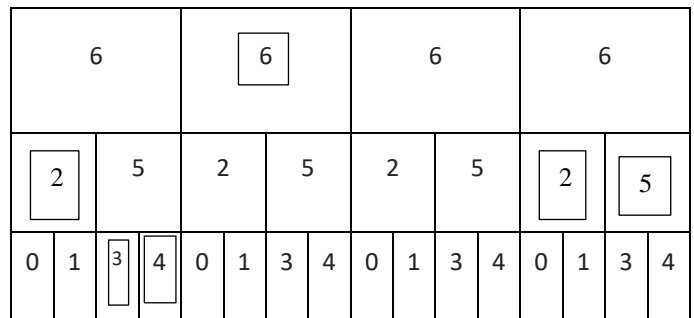


Figure 2: Wavelet packet tree analysis chart to figure out adjacent bands.L0 represents the base signal and from L1 to L3 represents applying LPF and HPF on the signal. [7]

Figure 3: Proposed indexing to solve the adjacency problem due to wavelet packets indexing system. Numbers from 0 to 14 indicate new indexing system.[7]

Band Width	Level 4	Level 3	Level 2	Level 1	Level 0
0-25%	0	2	6	Group 1	Base Signal
25-50%	1				
	3	5			
	4				
	0	2	6		
1					



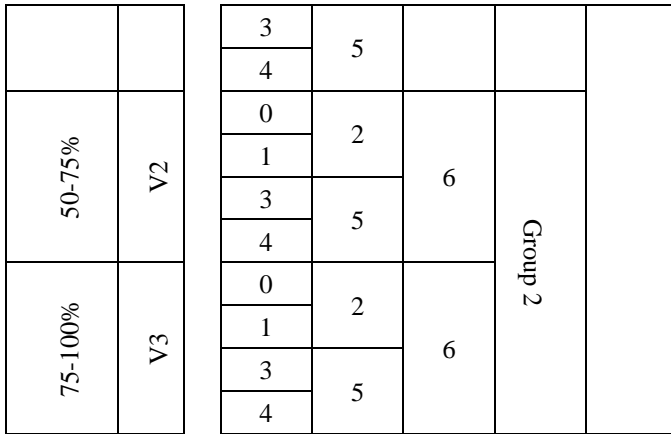


Figure 4: Chart to illustrate the 4 point encoding. V0 to V3 represent the feature vectors. [7]

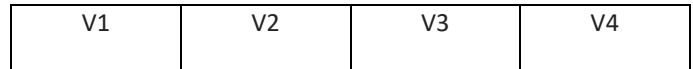


Figure 5: Best tree encoding example. Numbers in rectangular shape represent nodes that contain information.

TABLE 1
BEST TREE ENCODING EVALUATION OF FIGURE 5.

Element	Binary Value	Decimal Value	Frequency Band
V1	0011100	28	0 - 25 %
V2	1000000	64	25% - 50%
V3	0000000	0	50%-75%
V4	0100100	36	75% - 100%

The feature vector can be represented as F.

$$F = \begin{bmatrix} 28 \\ 64 \\ 0 \\ 36 \end{bmatrix}$$

The disturbance measurement or cost function measures the distance between training classes and testing classes. To obtain the best cost function, the results should indicate the lowest distance between the same classes and the largest distance between the different classes.

Enhanced Wavelet Packets Best Tree Encoding (EWPBTE) was introduced in [13]. It was an enhancement to the encoding process in BTE using the genetic algorithm. It used the accumulated Euclidean distance as a cost function and applied the genetic algorithm to adjust the pre-encoding of BTE [13]. The result was as indicated in Fig. 6.



Figure 6: Wavelet Packet index pattern using EWPBTE. Numbers from 0 to 6 represent the enhanced indexing system. [13]

B. MBT stages

MBT is introduced in [14]. It is a new version of BTE. The wave is converted to MBT by the following stages:

- 1- The input speech samples are resampled at 10 kHz.
- 2- Converting the frequencies to Mel frequencies using the Mel scale curve.
- 3- Generating EWPBTE matrix from Mel scaled data vectors using a filter bank matrix with 50 linearly spaced filter banks and they are overlapped by 50%. The filters are shaped as rectangle window. The outputs are the feature vectors of MBT.

3 EXPERIMENT FRAMEWORK

TIMIT speech database is used. It is split into training and testing sets. The database consists of 119 speech samples in WAV format. Hidden Markov toolkit (HTK) is used as a speech recognizer and Matlab [15]. The database is processed to modify the transcription files for the syllables detection goal of this research. The syllables are liquid, vowels, stops, plosives, nasals, and consonant. The following table shows each classifier with phones assigned to it.

TABLE 2
Phones of classifiers.

Classifiers	Number of labels	TIMIT labels
Vowel (V)	20	iy ih ey aa aw ay ao oy ow uh uw ux er ax axr ax-h ah ae eh ix
Consonant (C)	18	dh dx q jh ch sh dcl kcl s gcl tcl z zh pcl th bcl f v
Liquid (L)	8	l r w j y hh hv el
Nasals (N)	7	n m ng em en eng nx
Stops (S)	3	h# pau epi
Plosives (P)	6	b p t d k g

TABLE 3

EXPERIMENT RESULTS FOR THE FIXED STATE HMM DESIGN.

Mixture Count	MFCC SR%	MBT SR%
1	61.77	54.18
2	64.05	34.43
4	65.82	58.73
8	70.38	66.58
16	74.18	60.25

4 VARIABLE STATE HMM DESIGN

Models are designed using two different topologies. The first topology is shown in Fig.7; it points out that all classifiers are trained using the same HMM fixed number of states Model which has three emitting states and two non-emitting states (The non-emitting states are needed by the software to identify the entry and the exit state in HMM model). The second topology is shown in Fig.9; it illustrates that stops and plosives are modeled using one emitting state due to their short time duration almost lower than 20 ms. Fig.8 represents consonant design in the second topology.

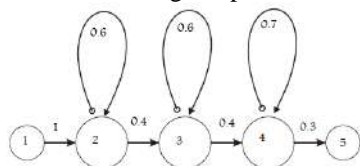


Figure 7: Fixed states model.

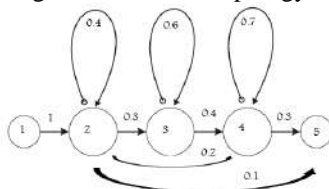


Figure 8: Consonant design in the Variable State model.

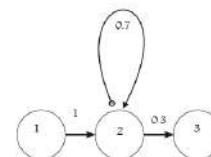


Figure 9: Stops and plosives design in the Variable State model.

GMM is used to act as spatial distribution probability density functions of attribute vectors of N-Dimensions. The mixture count is variable in this research. The system is tested against different Gaussian mixture counts (1, 2, 4, 8 and 16).

5 RESULTS AND DISCUSSION

Two sets of features vectors are used. First, MFCC technique (MFCC_0_D_N_Z) is used which has 25 coefficients. C0 represents energy component, D represents delta coefficients, N represents that absolute energy was suppressed and Z means that it has zero mean static coefficients. Second, MBT is used. MBT is a vector of 4 components as indicated in section 3. Success Rate (SR) of each syllable is used for evaluating the proposed model. The comparative study is provided to illustrate the details and the key power in each feature set and in the proposed model for each classifier. Evaluation of MBT is made by comparing its success rate with MFCC_0_D_N_Z success rate. Success rate (SR) can be calculated by the following equation. D represents Deletions (phones not found in the output). S represents Substitution (phones replaced by other phones). N represents the number of phones in the expected transcription.

$$SR = \frac{N - D - S}{N} \tag{1}$$

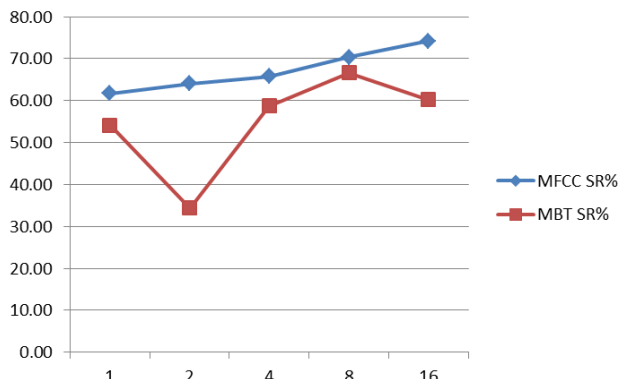


Figure 10: Success rate results in the Fixed State HMM Design.

The success rate of the Fixed State HMM Design model using the two techniques of feature extraction is shown in Table 3 and Fig. 10. MFCC_0_D_N_Z is taken as a reference. The relative success rate of MBT as a percentage of MFCC can be calculated by the following equation.

$$successrate = \frac{SRofMBT}{SRofMFCC} \tag{2}$$

According to this relative equation, MBT can achieve 95% from MFCC_0_D_N_Z in the Fixed State HMM Design.

TABLE 4

EXPERIMENT RESULTS FOR THE VARIABLE STATE HMM DESIGN.

Mixture Count	MFCC SR%	MBT SR%
1	66.58	55.95
2	67.09	64.56
4	69.11	81.01
8	70.89	73.67
16	72.66	72.15

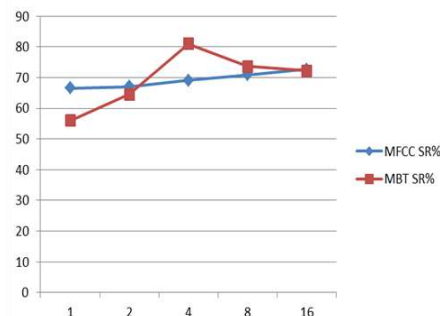


Figure 11: Success rate results in the variable state HMM design.

The results in Table 4 and Figure 11 show that in the variable state HMM design, MBT outperformed MFCC_0_D_N_Z using Gaussian mixture number 4 and 8. MBT achieves 81.01% success rate. This success rate is the highest rate obtained using the two techniques of feature extraction by the fixed and variable states of HMMS.

A. List of different cases of suggested hybrid acoustical models

Case 1:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 1 in all states.
- Features type: MFCC.
- Features vector size: 25 components.

Case 2:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 1 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Case 3:

Case 4:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 2 in all states.
- Features type: MFCC.
- Features vector size: 25 components.

Case 5:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 4 in all states.
- Features type: MFCC
- Features vector size: 25 components.

Case 7:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 8 in all states.
- Features type: MFCC
- Features vector size: 25 components.

Case 9:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 16 in all states.
- Features type: MFCC
- Features vector size: 25 components.

Case 11:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 1 in all states.
- Features type: MFCC
- Features vector size: 25 components.

Case 13:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C),

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 2 in all states.
- Features type: MBT
- Features vector size: 4 components.

Case 6:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 4 in all states.
- Features type: MBT
- Features vector size: 4 components.

Case 8:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 8 in all states.
- Features type: MBT
- Features vector size: 4 components.

Case 10:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units.
- Gaussian mixture (GM): 16 in all states.
- Features type: MBT
- Features vector size: 4 components.

Case 12:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 1 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Case 14:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C),

Liquid (L), Nasals (N), Stops (S), Plosives (P).

- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 2 in all states.
- Features type: MFCC
- Features vector size: 25 components.

Case 15:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 4 in all states.
- Features type: MFCC.
- Features vector size: 25 components.

Case 17:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 8 in all states.
- Features type: MFCC.
- Features vector size: 25 components.

Case 19:

- Syllables classification using MFCC, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 16 in all states.
- Features type: MFCC.
- Features vector size: 25 components.

Liquid (L), Nasals (N), Stops (S), Plosives (P).

- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 2 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Case 16:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 4 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Case 18:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 8 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Case 20:

- Syllables classification using MBT, HMM and Gaussian mixture.
- Classified syllables: Vowel (V), Consonant (C), Liquid (L), Nasals (N), Stops (S), Plosives (P).
- HMM design: 5 states for all classified units except P and S unit has 3 states.
- Gaussian mixture (GM): 16 in all states.
- Features type: MBT.
- Features vector size: 4 components.

Confusion matrix informs us about the classes that have a higher rate of correctness and the other classes that have a higher rate of error. Table 5 indicates that MBT outperformed MFCC in the Fixed State HMM Design at (GM= 1, 4, 8) in recognizing the liquid class and recognizing the stops class at (GM=8). Also, the results indicate that MBT outperformed MFCC in the variable state HMM design at (GM= 1, 4, 8, 16) in recognizing the liquid class, also at (GM= 4, 8) in recognizing the Consonant class and at (GM=1, 4) in recognizing the stops class.

Table 5 represents different cases in the experiment. Cases (1, 3, 5, 7 and 9) indicate results of MFCC technique and Cases (2, 4, 6, 8 and 10) indicate results of MBT technique in the Fixed State HMM model using a number of the Gaussian mixture (1, 2, 4, 8 and 16) respectively. Cases (11, 13, 15, 17 and 19) indicate results of MFCC technique and Cases (12, 14, 16, 18 and 20) indicate results of MBT technique in the variable state HMM model using a number of the Gaussian mixture (1, 2, 4, 8 and 16) respectively.

The chart in Fig.12 shows that the plosives and nasals classes reach to (100%) success rate in most of the cases. While the stops reach to (100%) success rate only in case 10 which represents the case using MBT at GM=16 in the fixed HMM model.

TABLE 5
CONFUSION MATRIX OF DIFFERENT CASES FOR (GM=1, 2, 4, 8, 16).

	Case1	Case2	Case 3	Case4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10	Cas e11

V	84.5	77.7	90.6	40	90.7	66.9	90.2	82.6	94.1	80.5	84.4
C	87.6	8.8	81.7	12.6	85.6	50.5	91.3	56.3	89.6	52.5	87.9
L	72	78.1	73.8	28.8	76.1	81.5	80.4	81.8	86.5	73.8	73.5
N	88.2	73.3	100	87.5	100	60	100	60	100	56.3	93.3
S	20	0	33.3	14.3	50	37.5	40	87.5	20	100	33.3
P	100	97.4	100	100	100	86.1	100	69.4	100	42.4	100

	Case12	Case13	Case14	Case15	Case16	Case17	Case18	Case19	Case20	Best SR
V	72.9	87.5	67.4	91.1	75.2	92.1	53	93	56.7	94.1
C	37.1	89.8	77.5	90.6	92.4	89.5	95	91.8	89.7	95
L	76.7	64.7	40	72.3	91	77.6	88.2	83.7	87	91
N	35.3	100	100	100	29.4	100	47.1	100	70.6	100
S	66.7	33.3	0	42.9	44.4	60	50	33.3	22.2	100
P	100	100	94.7	100	97.4	100	94.7	100	85.7	100

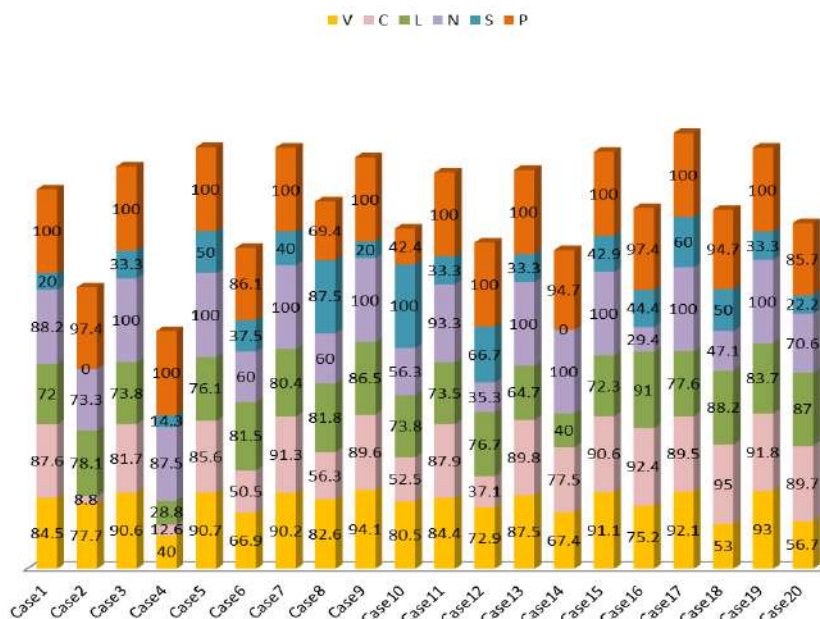


Figure 12: Chart of different SR of classes at different cases. Each classifier is represented by the first capital letter of it.

6 CONCLUSION

Two models are presented in this paper to manipulate the preprocessing classification task in order to enhance the overall performance of ASR system. The scope of this research is to propose the classification process. The first hybrid model is called variable state, dynamically structured Hidden Markov Model, Gaussian Mixture (VS-HMM-GM). The second hybrid model is called fixed state, structured Hidden Markov Model, Gaussian Mixture (FS-HMM-GM). The first model gives a higher rate of correctness than the second model. Adapting both models for best overall success rate, by changing the Gaussian mixture counts till 8 mixture is considered. Results indicate that improvement of overall success rate is noticeable using MBT features into the hybrid model (VS-HMM-GM). This is indicating that the variable state structure can be used to improve the overall success rate of the recognition engine due to the difference in time period behavior of each speech class.

To be specified in terms of specific class classification performance, the highest success rates are achieved, using (FS-HMM-GM-MBT) at (GM=16), as of almost 100% for stops class and of 94.1% for vowel's class. Using (VS-HMM-GM-MBT), the highest success rates are achieved as of 95% at (GM=8) for Consonant class and as of 91% at (GM=4) for liquid class. This is implies that using the dynamic structure HMM engine is more efficient in case of consonants but fixed structure HMM engine is more efficient in case of vowels detection.

The highest overall success rate (81.01%) is achieved using (VS-HMM-GM-MBT). As of both MBT and MFCC are used equally in all models, then it is concluded that MBT is indicating more efficiency than MFCC. Using MBT as features in both

hybrid model families achieves a higher performance using only 4 components than of MFCC which has 25 components. Although MBT still in the development phase, but this preliminary results indicating that this direction is very promising.

REFERENCES

- [1] Alan V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering, The Journal of the Acoustical Society of America, vol. 45, no. 2, pp. 458-465, 1969.
- [2] B. S. Atal, Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", The Journal of the Acoustical Society of America, vol. 50, no. 2, pp. 637-655, 1971.
- [3] R. Sarikaya, J.H.L. Hansen, "High resolution speech feature parametrization for monophone-based stressed speech recognition", IEEE Signal Processing Letters, vol. 7, no. 7, pp. 182-185, 2000.
- [4] O. Farooq, , S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition", IEEE Signal Processing Letters, vol. 8, no. 7, pp. 196-198, 2001.
- [5] Mihalis Sifarikas, Todor Ganchev, Nikos Fakotakis, Kokkinakis George, "Wavelet packet approximation of critical bands for speaker verification", International Journal of Speech Technology, vol. 10, no. 4, pp. 197-218, 2007.
- [6] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.
- [7] Amr M. Gody, "Wavelet Packets Best Tree 4 Points Encoded (BTE) Features", The Eighth Conference on Language Engineering, Ain-Shams University, Cairo, Egypt, pp. 189-198, 17-18 December 2008.
- [8] T. Jeff Reynolds, Christos A. Antoniou, "Experiments in speech recognition using a modular MLP architecture for acoustic modelling", Information Sciences, vol. 156, no. 1-2, pp. 39-54, 2003.
- [9] Patricia Scanlon, Daniel P. W. Ellis, Richard B. Reilly, "Using Broad Phonetic Group Experts for Improved Speech Recognition", IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, pp. 803-812, 2007.
- [10] Gábor Kiss, David Sztahó, Klára Vicsi, "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features", The 4th IEEE International Conference on Cognitive Infocommunications, Budapest, Hungary, pp. 579-582, 2-5 Dec 2013.
- [11] Hebah H. O. Nasereddin, Ayoub Abdel Rahman Omari, "Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation", IEEE Computing Conference 2017, London, UK, pp. 200-207, 18-20 July 2017.
- [12] L. Zhao, , L. Wang, , R. Yan, , "Rolling Bearing Fault Diagnosis Based on Wavelet Packet Decomposition and Multi-Scale Permutation Entropy", Entropy, vol. 17, no. 9, pp. 6447-6461, 2015.
- [13] Amr M. Gody, Rania AbulSeoud, Mohamed Hassan, "Automatic Speech Annotation Using HMM based on Enhanced Wavelet Packets Best Tree Encoding (EWPBTE) Feature", PESCT, Faculty of Engineering, Fayoum University, ElFayoum, Egypt, December 2011.
- [14] Amr M. Gody, Amr AbdAllah Emam Saleh, Manal Shabaan Mohammed, Automatic Speech Segmentation Using Hybrid Model, The Conference on Language Engineering, Ain-Shams University, Cairo, Egypt, December 2017
- [15] Michel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggi, Wavelet Toolbox™ 4 User's, 2009.

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 the Acting chief of Electrical Engineering department, Fayoum University in 2010, 2012, 2013, 2014 and 2016. His current research areas of interest include speech processing, speech recognition and speech compression. He is author and co-author of many papers in national and international conference proceedings and journals such as Springer(International Journal of Speech Technology), the Egyptian Society of Language Engineering (ESOLE) journal and conferences, International Journal of Engineering Trends and Technology (IJETT), Institute of Electrical and Electronics Engineers (IEEE), International Conference of Signal Processing And Technology (ICSPAT), National Radio Science Conference(NRSC), International Conference on Computer Engineering &System (ICCES) & Conference of Language Engineering(CLE).



Doaa N. Senousy received the B.Sc. degree in electrical engineering – communications and electronics Department from Akhbar El-Yom Academy, Giza, Egypt, in 2013. She joined the M.S program in the communications department in Fayoum University, Fayoum, Egypt, in 2015. Her current research areas of interest include automatic speech classification.



Sameh F. Saad received the B.Sc. degree in mechatronics engineering from High Institute of Engineering, Giza, Egypt, in 2000, the M.S. degree in mechatronics engineering from Mechatronics Laboratory, Paderborn University,

NRW, Germany, in 2004 and the Ph.D. degree in electrical engineering at Cairo University, Giza, Egypt in 2012. From 2016 to 2018, he was a Researcher and a Lecturer with the October University for Modern Sciences and Arts, Giza, Egypt. His research interest includes the development of mobile robots and autonomous vehicle, automation of aquaponics ecological system, and automation of mechatronics systems.

التجهيز الأول للتعرف التلقائي على الكلام باستخدام تصنيف المقاطع الصوتية بنموذج ماركوف الخفي ذو الهيكلية المرنة

*¹دعاء نبيل سنوسي, *² عمرو محمد جودى , **³سامح فريد

*قسم الهندسة الكهربيه , جامعة الفيوم , مصر

¹dn1144@ fayoum.edu.eg

²amg00@fayoum.edu.eg

**جامعة العلوم و الاداب، 6 اكتوبر ، الجيزة، مصر

³ dr.sam.far@gmail.com

ملخص

يقدم هذا البحث طرقا مختلفه لعملية التصنيف المستخدمة في عملية التعرف التلقائي على الكلام (ASR). تم استخدام بنيات مختلفة لنموذج ماركوف الخفي ذو الهيكلية المرنة (HMM). تم تصنيف المقاطع الصوتية إلى ستة مقاطع هي: حروف ساكنة (S)، حروف متحركة (V)، حروف لا تحتوي كلام (P) و حروف ذات طبيعة انفجارية (C) و حروف أنفيه (N) و حروف (liquid (l). تم استخدام جزء من قاعدة البيانات TIMIT في هذا البحث. تم استخدام عدد مختلف من GM في هذا البحث لتكوين نموذج هجين. وحيث أن نجاح عملية التعرف التلقائي على الكلام يعتمد بشكل كبير على إجراء عملية التصنيف بطريقة صحيحة. لذا فإن هذا البحث يركز على الوصول لعملية التصنيف الصحيحة لرفع الكفاءة لنظام التعرف التلقائي على الكلام. تم استخدام نوع جديد من الطرق لإستخراج المميزات الخاصة للمقاطع يسمى (MBT). تم إجراء تقييم للطريقة الجديدة المستخدمة (MBT) بعمل دراسة مقارنة و ذلك بمقارنة النتائج المستنبطة من النموذج الهجين المقترح بنموذج هجين مثل و لكن يعتمد على المميزات (MFCC) و باستخدام نفس قاعدة البيانات. النموذج الهجين المقترح المعتمد على المميزات المقترحة (MBT) حققت اعلى نسبة تعرف على المقاطع بنسبة تصل إلى (81.01%). ومن ذلك نستنتج تفوق (MBT) على نظيره (MFCC) في التعرف على المقاطع الاتيه: S, C, L. وتعد هذه النتائج واعدة في مجال التعرف التلقائي على الكلام.