



# الحرف العربي والتقنية أبحاث في حوسبة العربية



تحرير  
يوسف العريان



# الحرف العربي والتقنية

## أبحاث في حوسبة العربية

تحرير  
يوسف سالم العريان

المشاركون

عمر السيد شعبان  
مأمون صبحي الخطاب  
محمد زكي خضر  
يحيى محمد الحاج

أمجد أبو جبارة  
إريك آتويل  
سامح محمد عويضة  
عبدالله يحيى الفيفي



## الحرف العربي والتقنية أبحاث في حوسبة العربية

يوسف العريان

الرياض ، ١٤٤٥ هـ

البريد الإلكتروني : [nashr@ksaa.gov.sa](mailto:nashr@ksaa.gov.sa)

ح / مجمع الملك سلمان العالمي للغة العربية ، ١٤٤٥ هـ  
فهرسة مكتبة الملك فهد الوطنية أثناء النشر

ص.. ص : ..سم

رقم الإيداع : ١٤٤٥/٢٢٥٦٩

ردمك: ٨-٥٢-٤٤٤-٦٠٣-٩٧٨

لا يسمح بإعادة إصدار هذا الكتاب ، أو نقله في أي شكل أو وسيلة ، سواء أكانت إلكترونية أم يدوية ، بما في ذلك جميع أنواع تصوير المستندات بالنسخ ، أو التسجيل أو التخزين ، أو أنظمة الاسترجاع ، دون إذن خطي من المجمع بذلك .

( صدر هذا الكتاب عن مركز الملك عبدالله للتخطيط والسياسات اللغوية ، والذي جرى دمجه في مجمع الملك سلمان العالمي للغة العربية ) .

هذه الطبعة إهداء من المجمع ، ولا يُسمح بنشرها ورقياً ، أو تداولها تجارياً .



أطلق مجمع الملك سلمان العالمي للغة العربية ضمن أعماله وبرامجه مشروع: (المسار البحثي العالمي المتخصص)؛ لتلبية الحاجات العلمية، وإثراء المحتوى العلمي ذي العلاقة بمجالات اهتمام المجمع، ودعم الإنتاج العلمي المتميز وتشجيعه، ويضم المشروع مجالات بحثية متنوعة، ومن أبرزها: (دراسات التراث اللغوي العربي وتحقيقه، والدراسات حول والمعجم، وقضايا الهوية اللغوية، ومكانة العربية وتعزيزها، واللسانيات، والتخطيط والسياسة اللغوية، والترجمة، والتعريب، وتعليم اللغة العربية للناطقين بها وبغيرها، والدراسات البيئية).

وصدر عن المشروع مجموعة من الإصدارات العلمية القيمة (جزء منها-ومن بينها هذا الكتاب- صدر عن مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللغوية والذي جرى دمجها في مجمع الملك سلمان العالمي للغة العربية). ويسعد المجمع بدعوة المختصين، والباحثين، والمؤسسات العلمية إلى المشاركة في مسار البحث والنشر العلمي، والمساهمة في إثرائه، ويمكن التواصل مع المجمع لمسار البحث والنشر عبر البريد الشبكي: (nashr@ksaa.gov.sa).

والله ولي التوفيق

## كلمة المركز

يجتهد مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية في العمل في مجالات متعددة تحقق تعميق الوعي اللغوي على المستويات المختلفة ( الاجتماعية والعلمية / الأهلية والرسمية ) ؛ وذلك للسمو باللغة العربية، وترسيخ منافستها للغات الحضارية في العالم، وتعميق قيادتها الدينية والتاريخية لشعوب شتى في أنحاء المعمورة.

وامتدادا لذلك. ينشط المركز في مجال النشر، مستقطباً الأعمال العلمية الجادة وفق لائحة معتمدة منظّمة لذلك، كما ينشط في مجال التأليف من خلال استكتاب مجموعة كبيرة من الباحثين ؛ لتأليف عدد متنوع من الإصدارات النوعية المقروءة التي تعالج عنوانات يقتضها المركز، ويلفت الانتباه إليها، ويعلن من خلالها الفرص الممكنة لخدمة اللغة العربية في المجالات المختلفة، ملبياً بذلك الحاجات التي يلتمس المركز تطلّع المكتبة اللغوية العربية إليها، ولافتاً الأنظار إلى أهمية التعمق فيها بحثياً، واستكشاف ما يمكن عمله تنفيذياً في هذه المجالات. ويسعد المركز بأن استقطب في المرحلة الأولى من هذا المشروع ما يربو على مئتي باحث، موسّعاً دائرة المشاركة محلياً وخليجياً وعربياً وإسلامياً وعالمياً، ومنوعاً مسارات البحث الرئيسية والفرعية، ومنفتحة على كل ما من شأنه خدمة اللغة العربية بجميع الوسائل والأطر.

ويمثّل هذا الكتاب واحداً من الكتب التي صدرت ضمن سلسلة (مباحث لغوية) يحتوي عدداً من الأبحاث لأساتذة مرموقين؛ استجابوا لما رآه المركز من الحاجة إلى التأليف تحت هذا العنوان، وبأدروا إلى ذلك مشكورين.

وتوَدُّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، ومدير هذا المشروع العلمي على ما تفضلوا به من التزام علمي لا يستغرب من مثلهم، وقد ترك المركز للمحرر مساحة واسعة من الحرية في اختيار الباحثين ووضع الخطة العلمية - بالتشاور مع المركز -؛ سعياً إلى تحقيق أقصى ما يمكن تحقيقه من الإفادة العلمية، مع الأخذ بالاعتبار أن الآراء الواردة في البحوث لا تمثل رأي المركز بالضرورة، ولكنها من جملة الآراء العلمية التي يسعد المركز بإتاحتها للمجتمع العلمي وللمعنيين بالشأن اللغوي لتداول الرأي، وتعميق النظر، ونلفت انتباه القارئ الكريم إلى أن ترتيب أسماء المؤلفين على الغلاف موافق لترتيب أبحاثهم في الكتاب، وهي خاضعة للرؤية المنهجية التي تفضّل المحرر - مشكوراً - باقتراح خطتها.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي بحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة، ويمتد الشكر لمعالي نائبه، وللসادة أعضاء مجلس الأمناء نظير الدعم والتسديد لأعمال المركز.

والدعوة موجّهة لجميع المختصين والمهتمين بتكثيف الجهود نحو النهوض بلغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

## المقدمة

إِنَّ الْحَمْدَ لِلَّهِ، نَحْمَدُهُ، وَنُسْتَعِينُهُ، وَنَسْتَغْفِرُهُ، وَنَعُوذُ بِاللَّهِ مِنْ شُرُورِ أَنْفُسِنَا،  
وَمِنْ سَيِّئَاتِ أَعْمَالِنَا، مَنْ يَهْدِهِ اللَّهُ فَلَا ضَلْلَ لَهُ، وَمَنْ يَضِلَّ فَلَا هَادِيَ لَهُ، وَأَشْهَدُ  
أَنَّ لَا إِلَهَ إِلَّا اللَّهُ، وَحْدَهُ لَا شَرِيكَ لَهُ، وَأَشْهَدُ أَنَّ مُحَمَّدًا عَبْدُهُ وَرَسُولُهُ؛ أَمَّا بَعْدُ،

فَقَدْ أَيْدَ اللَّهُ -تَبَارَكَ وَتَعَالَى- رَسْلَهُ بِآيَاتٍ، ثُمَّ اصْطَفَى لَخَاتِمِهِمْ مِنْهَا كَلَامَهُ  
العربي الخالد المبين؛ فَكَانَتْ آيَتُهُ مِنْ جِنْسِ الْبَيَانِ (الرَّحْمَنُ . عَلَّمَ الْقُرْآنَ . خَلَقَ  
الْإِنْسَانَ . عَلَّمَهُ الْبَيَانَ) وَكَانَتْ (بِلِسَانٍ عَرَبِيٍّ مُبِينٍ). وَفِي هَذَا مَا يَكْفِي لِلإِشَارَةِ إِلَى  
فَضْلِ الْبَيَانِ عَمُومًا وَاللِّسَانِ الْعَرَبِيِّ خُصُوصًا.

وَلِعُلُومِ الْحَاسِبِ فَضْلُهَا الْمُسْتَمَدُّ مِنْ كَوْنِ الْحَاسِبِ آلَةً تَمُدُّ وَتَقْوِي مَا مِيزَ اللَّهُ  
بِهِ بَنِي آدَمَ مِنْ عَقْلٍ؛ كَمَا أَنَّ الْمُقْرَابَ امْتِدَادَ لِقُوَّةِ الْإِبْصَارِ. لِذَا، فَيُمْكِنُنَا الْقَوْلُ بِأَنَّ  
نِسْبَةَ رَقِي الْحَاسِبَاتِ عَلَى غَيْرِهَا مِنَ الْآلَاتِ كَنِسْبَةِ رَقِي الْعَقْلِ عَلَى الْجَسَدِ. ثُمَّ إِنْ  
بَرُمَجَةُ هَذِهِ الْآلَاتِ -الَّتِي ابْتَدَأَتْ بِبَعْضِ أَعْمَالِ الْإِخْوَةِ بَنِي مُوسَى- تَحْتَاجُ إِلَى لُغَاتٍ  
خَاصَّةٍ بِهَا، مِمَّا يَعُودُ عَلَيْهَا بِمَا أَسْلَفْنَا مِنْ فَضْلِ الْبَيَانِ وَاللِّسَانِ.

يَتَنَاوَلُ كِتَابُنَا هَذَا الْحَاسِبُ فِي خِدْمَةِ اللُّغَةِ الْعَرَبِيَّةِ. وَهُوَ فِي هَيْكَلِهِ يَشْبِهُ  
الشَّجَرَةَ: إِذْ يَبْدَأُ بِثَلَاثَةِ بَحُوثٍ فِي الْبُنَى التَّحْتِيَّةِ: أَحَدُهَا يَنَاقِشُ الْأَسَاسَ التَّقْلِيدِيَّ  
لِبَرَامِجِ حَوْسِبَةِ اللُّغَةِ، وَهُوَ تَمَثِيلُ اللُّغَةِ الْعَرَبِيَّةِ فِي الْحَاسِبِ، وَالْآخَرَانِ لِلْمَدُونَاتِ  
اللُّغَوِيَّةِ وَقَوَاعِدِ الْبَيَانَاتِ الَّتِي تَبْنِي عَلَيْهَا النِّظْمُ الْمُتَعَلِّمُ. ثُمَّ يَنْتَقِلُ إِلَى مَا يَسَامِي  
الْجَذْعَ -وَهِيَ الْأَدَوَاتُ الَّتِي تَحْمِلُ غَيْرَهَا مِنَ التَّطْبِيقَاتِ الْمُثْمِرَةِ- فَجَاءَ بِبَحْثَيْنِ  
اسْتَقْصَى أَحَدُهُمَا التَّقْنِيَّاتِ التَّقْلِيدِيَّةِ وَالْإِحْصَائِيَّةِ لِمُعَالَجَةِ اللُّغَاتِ الطَّبِيعِيَّةِ  
وَتَخْصُصُ الْآخَرِ فِي طَرِيقَةٍ لَتَعْلَمَ التَّشْكِيلَ آليًا. ثُمَّ خَتَمَ الْكِتَابَ بِفُرْعَيْنِ مُثْمَرَيْنِ



استقصيا تطبيقات الترجمة الآلية من العربية وإليها وتطبيقات التعرف الآلي على قراءة القرآن الكريم.

وكان حرصنا منصبا في هذا الكتاب العربي على المقدمات الاستقصائية لأنه باكورة في سلسلة، فناسبه التمهيد لما بعده. كما حرصنا أن يكتب موضوعاته نخبة من أساتذة الجامعات ومدراء الشركات المتخصصة في المملكة العربية السعودية والمملكة الأردنية الهاشمية والمملكة المتحدة والولايات المتحدة الأمريكية. وإننا لندرجو أن تستمر هذه السلسلة المباركة وأن يفيد من هذا الكتاب وما بعده باحثو اللسانيات وباحثو الحاسوب وغيرهم، والله ولي ذلك والمبارك فيه.

**المحرر :**

**د. يوسف بن سالم العريان**

## حول نظام تمثيل الحرف العربي

أ. مأمون صبحي الحطاب (\*)  
m.hattab@arabtext.ws

---

(\*)

دار حوسبة النص العربي

باحث في حوسبة اللغة العربية، قام بتطوير عدد من الأنظمة المحوسبة في مجالات: الصرف والنحو، ومحركات البحث العربية، وتطوير التقنية لخدمة ذوي الاحتياجات الخاصة العرب. وهو عضو مؤسس لجمعية حوسبة اللغة العربية وإثراء المحتوى العربي على شبكة الإنترنت.



## ١. المقدمة

تعتمد برامج التحليل اللغوي على بنية الكلمة ورسمها كمصدر لمعلومات التحليل اللغوي. فعمل الحاسوب يتكون من ثلاث مراحل هي مرحلة الإدخال، ثم مرحلة المعالجة، تتبعهما مرحلة إخراج النتائج. ولم تحقق برامج التحليل اللغوي العربية -حتى الآن- نتائج تؤدي إلى بناء تطبيقات لغوية لها أثر مرض في الواقع. ويعرض البحث خصائص نظم تمثيل العربية وبعض مشاكلها في سياق مقارنتها بنظم تمثيل اللغة الإنجليزية، وهي اللغة العالمية الحالية.

لقد أدت اللغة العربية دورا بارزا في نقل الحضارة الإنسانية واستيعابها لعدة قرون دونما انقطاع. ويعبر عن هذا الدور الحضاري مؤسس علم اللغة الحديث إدوارد ساپير Edward Sapir ١٩٤ : ١٩٢١، حيث يقول:

"إن اللغات الخمس التي كان لها دور رئيس في حمل الحضارة والإنسانية هي: الصينية القديمة والسنسكريتية والعربية واليونانية واللاتينية".

والملاحظ أن اللغة العربية هي الوحيدة من بين هذه اللغات الخمس التي ما زالت تؤدي دورها دون انقطاع. وقد شهد للعربية بأنها لغة علمية -العالم اللغوي رونالد لانغاكركر Ronald Langacker (١٨٢ : ١٩٧٣) الذي يرى أن طبيعة الكلمات المستعارة من لغة ما تعكس مدى تأثيرها في اللغة المستعيرة، وأن نسبة كبيرة من الكلمات العربية الداخلة في الإنجليزية هي كلمات علمية مثل Zero و Alchemy و Algebra وغيرها كثير. وهي كلمات دخلت الإنجليزية عن طريق الإسبانية. ويذكر أن الدكتور وجيه حمد عبد الرحمن قد أثبت وجود ١٠٠٠ جذر عربي في معجم أوكسفورد الإنجليزي (١).

أما المستشرقة الألمانية ريجارد هونيكه فتقول في كتابها الشهير "شمس العرب تسطع على الغرب" (٢٤٣: ١٩٦٣) : إنه قبل ستمئة عام كان لكلية الطب الباريسية أصغر مكتبة في العالم، لا تحتوي إلا على مؤلف واحد، لعربي عظيم . وكان هذا الأثر العظيم هو المرجع الأساسي لمدة تزيد عن أربعمئة عام بعد ذلك التاريخ دون أن يزاخمه مزاحم، أو تؤثر فيه أو في مكانته مخطوطة من المخطوطات الهزيلة التي دأب في صياغتها كهنة الأديرة قاطبة. وهذا العمل الجبار خطته يد الرازي (أبو بكر محمد بن زكريا). وقد اعترف الباريسيون بقيمة هذا الكنز العظيم وبفضل صاحبه عليهم وعلى الطب إجمالاً فأقاموا له نصبا في باحة القاعة الكبيرة في مدرسة الطب لديهم، وعلقوا صورته في شارع سان جيرمان.

بالرغم مما تقدم من حقائق علمية ثابتة فقد تعرضت اللغة العربية لحملات مغرضة تفتقر إلى الموضوعية وتتناقض مع حقائق علم اللغة الحديث. وقد انصب الهجوم على عدة محاور منها:

١. الادعاء بعدم صلاحية الخط العربي والدعوة لتلتيه.
٢. الادعاء بأن العربية لغة دينية - بالمفهوم الكهنوتي للدين - وليست لغة علمية.
٣. الادعاء بجمود العربية الفصحى وضرورة السماح بالتغيير اللغوي المتمثل في العاميات والدعوة لإحلالها محل الفصحى.
٤. القول بافتقار العربية للعدد الكافي من اللواحق (السوابق واللواحق) لترجمة ما يعادلها في اللغات الأوروبية الحديثة التي تنقل العربية عنها في القرن العشرين.

## ٢. تاريخ تطور نظام كتابة العربية

عرفت العرب الكتابة في جاهليتها - قبل الإسلام - واعتبرتها شرطا في كمال الرجل العربي، وتعود معرفتهم بالكتابة إلى اتصالهم بالأمم المتحضرة في بلاد

اليمن وتخوم الشام، فأنشأ الأنباط - مثلاً - ممالكهم على أطراف بلاد الشام في الناحية الشمالية الغربية من شبه الجزيرة العربية (١٦٩ ق م - ١٠٦ م)، واتخذت البتراء "سُلع" عاصمة لها، وكانت لهم صلات بالأراميين، فتأثروا بهم وتحدثوا لغتهم، واستنبطوا لأنفسهم خطاً خاصاً بهم عرف بالخط النبطي، اشتق منه عرب الشمال الخط الأنباري والخط الحيري، أو الخط المدور والخط المثلث.

ونجد عند العودة إلى تاريخ الكتابة العربية أنها مرت بثلاث مراحل (٢) هي: مرحلة النشأة التي تمتد حتى ظهور الإسلام، ومرحلة الاستخدام الواسع وتمتد حتى بعيد منتصف القرن الثاني الهجري ثم مرحلة تقعيد الإملاء.

فقد شجع الإسلام على تعلم الكتابة، وسلك في ذلك وسائل مختلفة، حتى أن رسول الله صلى الله عليه وسلم اشترط لفكاك الأسير من قريش في بدر تعليم عشرة من صبيان المدينة الكتابة، فراجت الكتابة في عصره صلى الله عليه وسلم، حتى بلغ عدد كتاب الوحي أكثر من أربعين كاتباً.

وتعد الحجاز أول بلاد العرب معرفة للكتابة، وكانت قريش في مكة، وثقيف في الطائف أكثر القبائل شهرة بها، ومن أبنائها اختير كتاب صحف أبي بكر الصديق رضي الله عنه، وكان عمر بن الخطاب رضي الله عنه يقول كما روى جابر بن سمرة: لا يملين في مصاحفنا هذه إلا غلمان ثقيف. وعندما جمع عثمان بن عفان رضي الله عنه مصاحفه قال: اجعلوا المملي من هذيل، والكاتب من ثقيف. (٢)

ثم أتت مرحلة التقعيد، وهي المرحلة التي تأثرت بالقواعد الصرفية والنحوية، مما جعل الإملاء العربي، في مواضع كثيرة، يشير إلى قواعد النحويين أكثر مما يصف واقعاً كتابياً ملموساً. (٣) حتى رأى بعض الباحثين المعاصرين (٥) أن اللغة العربية مكونة من ثلاثة أنظمة، هي: النظام الصوتي والنظام الصرفي والنظام النحوي، وقائمة من الكلمات التي لا تنظم في جهاز واحد. وأن هذه الأنظمة تكون

معينا صامتا، فإذا أردنا أن نتكلم أو أن نكتب؛ نظرنا في هذا المعين الصامت فوضعنا محتوياته في حالة عمل وحركة، فأخذنا منه الكلمات ورفضناها على شروط الأنظمة، أي بحسب قواعد اللغة.

كان الإملائيون العرب أنفسهم نحويين وصرفيين، لذلك لم يفصلوا -عند وضعهم لقواعد الإملاء- الكتابة عن اللغة، ولم يفصلوا قواعد الإملاء عن قواعد اللغة، بل إنهم أقحموا قواعد اللغة في الكتابة والإملاء، مما جعل الإملاء العربي -في مواضع كثيرة- يشير إلى قواعد النحويين أكثر مما يصف واقعاً كتابياً ملموساً. (٥)

### ٣. تمثيل الكتابة لصوت الكلمة

حاول الرسم الكتابي منذ نشأته أن يجاري المنطوق في وظيفته وأن يكون مطابقاً لأصوات الكلام، فتطور في سبيل ذلك من الصوري إلى المقطعي ثم إلى الأبجدي، واتخذ رموزاً مميزة للحروف والحركات، وابتدع علامات خاصة لتمثيل الظواهر الصوتية الأخرى وتقريب المكتوب من المنطوق. ومما ابتدعته العربية في هذا الشأن الشدة والمدة وعلامات الحركات والسكون والنقط الذي يختلف عدداً ومكاناً وأدخل - لما شاع اللحن إثر دخول غير العرب في الإسلام- على الكتابة العربية ما يطورها مثل:

#### أ. النقط

وتجمع معظم المصادر على أن أول من أدخل النقط هو عالم اللغة المشهور أبو الأسود الدؤلي (توفي عام ٦٨٨ م) وذلك لما لوحظ من لحن في تلاوة كتاب الله العزيز.

#### ب. الإعجام

ويعرف ابن جني الإعجام بأنه إزالة العجمة. وهو مشتق من "أعجم" أي أزال

العجمة ومنها كلمة معجم الذي يزيل عجمة الكلمات بإيضاح معناها.

## ت. الشكل

وأول من أدخل الحركات القصيرة الضمة والفتحة والكسرة هو الخليل بن أحمد الفراهيدي (المتوفي عام ٧٨٦م). كما أدخل السكون (°) الذي يعني غياب الحركة.

وبإدخال هذه الاصلاحات الضرورية على نظام كتابة العربية أصبح ذلك النظام أقرب ما يكون إلى نظام الكتابة الصوتي (Phonetic Alphabet) الذي يحقق تناظرا بين الصوت والحرف وذلك على النحو الآتي:

أولا : التناظر بين الحروف الصائتة والأصوات الصائتة (وبلاحظ هنا أن كل حرف يمثل صوتا واحدا لا غير والعكس صحيح):

/bæb/	باب	/b/	ب/
/tæb/	تاب	/t/	ت/
/ʔaθæθ/	أثاث	/ʔ/	ث/
/dʒær/	جار	/dʒ/	ج/
/ħamal/	حَمَل	/ħ/	ح/
/xulu:d/	خلود	/x/	خ/
/dær/	دار	/d/	د/
/ðæk/	ذاك	/ð/	ذ/
/rabb/	رب	/r/	ر/
/zær/	زار	/z/	ز/
/ʔasæs/	أساس	/ʔ/	س/
/ʃams/	شمس	/ʃ/	ش/
/ʂaff/	صَفّ	/S/	ص/
/Dəbæb/	ضباب	/D/	ض/
/Tibb/	طبّ	/T/	ط/



ظ/	/Z	ظَلَمَ	/Zulm/
ع/	/ɕ	عَلِمَ	/ɕilm/
غ/	/ɣ	غَيَّمَ	/ɣaim/
ف/	/f	فَشَلَ	/fafal/
ق/	/q	قَلَبَ	/qalb/
ك/	/k	كَتَابَ	/kitæb/
ل/	/l	لَيْلَ	/lail/
م/	/m	مَلَكَ	/malak/
ن/	/n	نَوَّرَ	/nu:r/
ه/	/h	هَرَبَ	/haraba/
س/	/ʔ	أَبَ	/ʔabb/

ثانيا : الحروف الصائتة والحركات

و/	/:w,u	وَرَاءَ	/ʔ:wara/
		نور	/nu:r/
ي/	/:j,i	يَوْمَ	/jaum/
		فِيلَ	/fi:l/
ا/	/æ	سَمَاءَ	/ʔsamæ/
ا/	/:a	غَارَ	/ɣa:r/
ـَ/	/a	كُتِبَ	/kataba/
ـِ/	/i	كُتِبَ	/kutiba/
ـُ/	/u	كُتِبَ	/kutub/

وهكذا يتضح أن درجة كفاءة الكتابة العربية تضاهي أحدث وأدق ما توصل إليه علماء اللغة والأصوات في الأوساط الغربية ألا وهو "نظام الكتابة الصوتية" الذي وضع للتغلب على الصعوبات المستعصية التي تواجهها أنظمة الكتابة لمعظم اللغات الحديثة التي تعرضت لحدوث بون شاسع بين نظامي الكتابة والنطق.

في المقابل نجد اللغة الإنجليزية أبعد كثيرا عن نظام الكتابة الأمثل وذلك للأسباب الآتية:

١- تمثيل بعض الحروف لأكثر من صوت. ففي اللغة الإنجليزية -وهي اللغة العالمية الأولى في عالمنا المعاصر- نجد ما يأتي:

أ. يرمز الحرف "a" إلى الأصوات التالية: \æ\ ، \c: \ ، \a: \ ، \i \ ، \i: \ كما في الكلمات fat ، ball ، father ، village و Caesar على التوالي.

ب. يرمز الحرف "s" إلى \z \ ، \s \ ، \iz \ ، \3 \ ، \f \ ، \ϕ \ كما في الكلمات bags ، cats ، dishes ، measure ، pressure و island على التوالي.

وينطبق هذا على كثير من الصوامت والصوائت الإنجليزية.

٢- تمثل بعض الأصوات بأكثر من حرف واحد. فالصوت \i: \ وهو صوت مد في الإنجليزية يمثل على النحو التالي وبالحروف التالية :

ee ، ea ، ei ، ie ، ey- ، e ، eo ، æ

وذلك كما في الكلمات see ، tea ، receive ، thief ، key ، crises ، people ، Caesar على التوالي.

٣- توجد حروف صامتة في معظم الكلمات وهو ما يضاعف صعوبة القراءة والكتابة في الإنجليزية للناطقين بها من ابنائها وغيرهم. وفيما يلي أمثلة على ذلك:

المثال	الصامت
Thumb	B
Wednesday	D
Ate	E
Psychology	P
Apostle	T
River	r.
Damn	N
Walk	L
Wrong	W
Knife	K, E
Sign	G
Honest	H

وقد نجم عن هذه العوامل وأمثالها بون شاسع بين نظامي الكتابة والنطق في لغة العلم والحضارة في يومنا هذا مما حدا بالبرلمان البريطاني لطرح هذه المسألة المؤرقة على بساط البحث لإعادة التوازن بين النظامين.

إن الذي أدى إلى حدوث مثل هذا الخلل الجسيم في النظام اللغوي الإنجليزي هو تغير النطق دون أن يواكبه تغير في نظام الكتابة. حتى إن الكاتب الإيرلندي الشهير جورج برناردشو George Bernard Shaw يوضح حجم المفارقة إذ يقول إن كلمة "ghoti" يمكن أن تنطق "fish" إن استعصنا عن حروف الأولى بالأصوات المأخوذة من الكلمات أدناه:

laugh = f \laif\

women = i \wimin\

nation = ŋ \neiŋan\

هذا القصور وما يمثله من تحد حاسوبي لا يقارن بذلك الموجود في نظام كتابة اللغة العربية، ومع ذلك فإن التقدم في نظم معالجة اللغة الإنجليزية بالحاسوب لم يتوقف طويلاً أمام تلك التحديات.

#### ٤. بعض مشاكل تمثيل الكتابة العربية

كان الإملائيون العرب أنفسهم نحويين وصرفيين - كما ذكرنا سابقاً - فلم يفصلوا، عند وضعهم لقواعد الإملاء، الكتابة عن اللغة، ولم يفصلوا قواعد الإملاء عن قواعد اللغة، بل أقحموا قواعد اللغة في الكتابة والإملاء، مما جعل الإملاء العربي، في مواضع كثيرة، يشير إلى قواعد النحويين أكثر مما يصف واقعاً كتابياً ملموساً. (٥) وجاء هذا التعقيد على حساب دقة التمثيل الصوتي للمنطوق وأدى إلى قصور فيه، كما جعل قواعد الإملاء عرضة لاختلاف وجهات نظر العلماء تبعاً لاختلاف وجهات نظرهم في الصرف والنحو.

ولم يكن واقع التطبيق بهذه المعيارية الصارمة، فاعتري الرسم الكتابي العربي مشكلات متعددة، وأخطاء شائعة عند الاستعمال. فمن المشكلات البارزة غياب الحركات وعدم اندغامها في بنية الكلمة العربية المكتوبة. وهي تزود من الذاكرة بناءً على الخبرة اللغوية المستكنة في الذهن، وعلى السياق اللغوي الذي يحصر المعنى. وتُلفظ الكلمة في النصوص غير المشكولة كما تفهم من السياق وكما يحددها الحس اللغوي مما يصعب للحاسوب معها.

إن النظام الكتابي الحالي للغة العربية هو نظام صامت Consonantal عموماً لا يعترف إلا بالحروف الصحيحة وحروف المد واللين، ولا يعترف بالحركات

حروفاً. فكل ما ليس له رمز في الألفباء العربية لا يعد حرفاً، على الرغم من قول علماء العربية (٣) إن الحركات أبعاد حروف المد واللين، وهي تماثلها من حيث النوع ولكن تختلف عنها من حيث الكم، وإدراكهم قيمة الحركات في التمييز بين الألفاظ المعجمية والصيغ الصرفية والحالات الإعرابية.

وقد استعاض عن دمجها في بنية الكلمة المكتوبة بوضع علامات لها -Diacritics تكتب خارج الكلمة فوق الحروف أو تحتها.

ومن المشكلات رسم الهمزة الذي يختلف حسب موضعها من الكلمة (أي مجيئها أولاً أو وسطاً أو آخرًا) وحركة ما قبلها أو سكونه، وحركتها هي أو سكونها. ويتصل بهذا الجانب التباس همزة القطع بهمزة الوصل التي يؤتى بها توسلاً إلى النطق بالسكان في أول المجموعة الكلامية Utterance، وهي تختفي إذا سبق الساكن بحركة في درج الكلام. ومع ذلك نظل نحفظ برسمها في ذلك الموضع، تنبيهاً على أن هذا الموضع هو مما يستوجب إدخال الهمزة عندما لا تكون هناك حركة. وواضح ما تدخله همزة الوصل من تغيير في النمط المقطعي للكلمة.

ومن المشكلات كذلك كتابة الألف في نهاية الكلمات من أفعال وأسماء قائمة أو غير قائمة (أي مقصورة في الاصطلاح الشائع)، والتفريق بين هاء الضمير (كما في له وعنده) والتاء المربوطة (كما في طالبة) إذ يحدث اللبس فيهما عندما لا توضع نقطتان للتاء المربوطة. وكذلك التفريق بين التاء المربوطة التي يوقف عليها بالهاء (مثل فاطمة) والتاء المفتوحة التي يوقف عليها بالتاء (كما في قالت وأوقات).

وقد يُعد من المشكلات وصل الحروف بعضها ببعض، لأنه يجعل من الكلمة وحدة إملائية لا وحدة معجمية أو صرفية بحتة، ويصعب تحليل كلمات مثل (سألتك) إلى عناصرها الأولى من صرفية ومعجمية.

ومن المشكلات أن الواو تمثل حرفين هما واو المد وواو اللين، وأن الياء تمثل حرفين كذلك هما ياء المد وياء اللين، مع أن اللين ذو طبيعة صامتة والمد ذو طبيعة صائتة. (٧)

ومن المشكلات زيادة حروف لا تنطق، كزيادة ألف بعد واو الجماعة التي يُسند الفعل الماضي والمضارع المجزوم أو المنصوب إليها (دَعَوْا، لم يدْعُوا، لن تدعوا). وربما كانت هذه الزيادة ذات فائدة في التمييز بين واو الجماعة التي أُسند إليها الفعل والواو الأصلية التي ينتهي بها الفعل الناقص غير المسند، مثل يدعو ويرجو. إلا أنه صار من الأخطاء الشائعة إدخال هذه الألف بعد واو الفعل غير المسند (يدعو)، وبعد جمع المذكر السالم المرفوع المضاف (مهندسوا المشروع).

وعكس ذلك نقص بعض الحروف كنقص الألف في أسماء الإشارة (مثل هذا وهذه). وقد أصبحت هذه الكلمات وأشباهها بحكم الشيوع والإلف بها شيئاً معتاداً لا يثير في نفس العربي أية مشكلة، خلا بعض الصعوبات في مراحل التعليم الأولى. وربما كانت هذه الكلمات وأضرابها مشكلة لدى المتعلمين من الأجانب. وكاختفاء لام التعريف -نطقاً- قبل الحروف الشمسية وهو من قبيل الإدغام<sup>□□□</sup>.

ومن الأخطاء الشائعة وضع ياء بعد ضمير المخاطبة (قرأتي، لكي)، والخطأ في كتابة "ابن" بين علمين، والطريقة التي تكتب بها بعض الألفاظ مثل مئة بالألف، و"داود" و"طاووس" بواو واحدة، والتي والذي والذين بلامين فتلتبس بلفظي المثني للذين واللتين.

والخطأ في كتابة الاسم المنقوص، ومسائل الفصل والوصل في أمثال: لئن، ولئلا وطالما وإنما وثلاثمئة.

---

□□□ يُنظر في مشكلات الرسم والأخطاء الشائعة الكتب الآتية: الإملاء العربي لأحمد قبش، والإملاء والترقيم في الكتابة العربية لعبد العليم إبراهيم، وقواعد الإملاء لعبد السلام هارون

ويمكن التفريق بين الواو المدية والواو اللينة، والياء المدية، والياء اللينة، بوضع ضمة قبل واو المد، وكسرة قبل ياء المد، كما فعل القدماء.

إن المشكلات السابقة ليست بالمستعصية، ويمكن للبرامج الحاسوبية أن تقوم بمعالجة كثير منها، وقد أحرز نجاح لا يستهان به في هذا المجال. وبقي بعضها عقبة أمام برامج التحليل اللغوي.

كما يجدر التنبيه هنا إلى أن علامات الترقيم تساعد كالضبط تماماً على الفهم السليم للنصوص، إذا ما أحسن استخدامها عند حدود المعاني الجزئية وحدود المعاني الكلية. وهي علامات مهمة في تحليل النصوص حاسوبياً.

## ٥. الخلاصة

لا تواجه جهود لسانيي الحاسوب العرب تحديات كبرى تعود إلى عدم كفاءة نظام كتابة العربية في تمثيل صوت الكلمة العربية أو صرفها. ومع ذلك قصرت جهودهم عن تطوير نظم التحليل اللغوي التي تضاهي النظم التي طورت للغة الإنجليزية. ولما تظهر حتى الآن تطبيقات الذكاء الاصطناعي العربية في الفهم الآلي التي يمكن توظيفها في تطبيقات الترجمة الآلية والبحث عن النصوص والتعرف على الحروف، وغيرها من التطبيقات. إن هذه النتيجة تدعو إلى دراسة أسباب هذا القصور ومعالجة أسبابه.

## ٦. المراجع

1. Abderahman. W. (1991) A concise dictionary of scientific roots: Graeco – Latin – English – Arabic. A new approach to the study of some aspects of Neologisation. Librairie du Liban.

٢. غانم قدوري الحمد، "علم الكتابة العربية"، دار عمار، عمان، ٢٠٠٤، ص ١٠٥.

٣. محمد سالم العوفي: تطور كتابة المصحف الشريف وطباعته، مجمع الملك فهد لطباعة المصحف الشريف، المدينة المنورة . ١٤٢١هـ - ٢٠٠٠م
٤. محمد أحمد أبو عبيد، مجلة دراسات في اللغة العربية وآدابها، فصلية محكمة، العدد السادس عشر، شتاء / ٢٠١٤م.
٥. تمام حسان، "اللغة العربية معناها ومبناها"، عام الكتب، القاهرة، ٢٠٠٤، ص ٤٠.
٦. ابن جني، سر صناعة الإعراب، بتحقيق حسن هندأوي، دار القلم، دمشق، ١٩٨٥ ج ١ ص ١٧.
٧. كمال بشر، علم الأصوات، دار غريب، القاهرة، ٢٠٠٠م، ص ١٦٤.





## أبحاث جامعة ليدز في مجال لسانيات المدونات العربية

د. إيريك أتويل (Eric Atwell) (\*)

E.S.Atwell@leeds.ac.uk

أ. عبد الله بن يحيى الفيضي (\*\*)

ayjfaifi@imamu.edu.sa

- 
- (\*) أستاذ مشارك في قسم الحاسب الآلي في جامعة ليدز ببريطانيا. متخصص في معالجة اللغة الطبيعية، ومهتم بمجال التنقيب في النصوص وتحليل البيانات، وله اهتمام كذلك باللغة العربية، وقد أجرى العديد من الأبحاث حول هذه الموضوعات، كما أشرف على مجموعة من طلاب الدكتوراه الذين أسهموا في بحث هذه الموضوعات.
- (\*\*) معهد تعليم اللغة العربية، جامعة الإمام محمد بن سعود الإسلامية، الرياض.
- متخصص في معالجة اللغة الطبيعية، شارك في عدد من المشاريع العلمية في مجال معالجة اللغة العربية حاسوبياً، له عدة أبحاث منشورة حول مدونات المتعلمين، إضافة إلى مشاركته في تحكيم عدد من الأبحاث العلمية. أنشأ المدونة اللغوية لتعليمي اللغة العربية (WWW.arabiclearnercorpus.com)، إضافة إلى بعض التطبيقات الحاسوبية في مجال المدونات اللغوية، مهتم بعلم اللغة الحاسوبي، وتعليم اللغة بمساعدة الحاسب.



## الملخص

طور الباحثون في مجال لسانيات المدونة ومعالجة اللغة الطبيعية في بريطانيا الكثير من المدونات اللغوية والأدوات الحاسوبية للبحث في اللغة الإنجليزية. وفي جامعة ليدز<sup>□□□</sup>، أردنا أن تمتد هذه الأبحاث لتشمل اللغة العربية أيضاً؛ ولأن هذا يتطلب بناء مدونات لغوية عربية فقد أنشأنا عدة مدونات منها: مدونة تعليم العربية بواسطة الحاسب الآلي، والمدونة اللغوية للعربية المعاصرة، ومدونة الإنترنت العربية، ومدونة اللغة العربية حول العالم، والمدونة اللغوية لتدريس معلومات عن الإسلام، والبنك الشجري للخطاب العربي، والمدونة اللغوية لمتعلمي اللغة العربية، والمدونة العربية لنصوص القرآن الكريم، ومدونة الإحالة الثنائية لضمائر القرآن الكريم، ومدونة الترابط الدلالي بين آيات القرآن الكريم، ومدونة القرآن الكريم الموسومة بموضوع الآيات مع الترجمة الإنجليزية، ومدونة جامعة الملك سعود للغة العربية الفصحى. كذلك طورنا مجموعة من الأدوات لتحليل النصوص العربية، منها: الكشف السياقي aConCorde للغة العربية، والمحلل الصريفي SALMA لوسم أقسام الكلام في المدونات العربية، وجدول وسم الأخطاء في مدونات المتعلمين العربية، والوسم الصوتي والمقطعي للغة العربية، وتدريب برامج التخاطب الآلي على مدونة عربية، والوسم الدلالي والتمثيل المعرفي للقرآن الكريم. وقد استخدمت هذه المدونات اللغوية والأدوات الحاسوبية لإجراء العديد من الأبحاث في مجال

---

□□□ جامعة ليدز هي أقدم وأكبر جامعة في مقاطعة Yorkshire في إنجلترا. افتتحت الجامعة عام ١٩٠٤، ويعمل فيها حالياً ما يربو على سبعة آلاف أكاديمي وموظف، كما يدرس فيها أكثر من اثنين وثلاثين ألف طالب من ١٤٦ دولة. تتميز الجامعة كذلك بأبحاثها الواسعة في عدد من الموضوعات الحيوية كالهندسة، والطب، واللغة، وغيرها. ويعد مركز "اللغة في جامعة ليدز" واحداً من أكبر المراكز البريطانية التي تقدم أبحاثاً ومنحاً في الجوانب المتعلقة بدراسة اللغة، وينتمي لهذا المركز أكثر من مئة باحث في العلوم الإنسانية، والآداب، والتربية، والعلوم الاجتماعية والنفسية، وعلوم الأحياء، والحاسب الآلي، والهندسة.

لسانيات المدونات العربية، ومن ذلك على سبيل المثال تعليم اللغة العربية، والمقارنة بين اللغتين الإنجليزية والعربية. ويعد موقع المدونة العربية لنصوص القرآن الكريم أحد أبرز الأمثلة على هذه الأبحاث، حيث يُستخدم على نطاق واسع من قبل الباحثين في اللغويات العربية، وكذلك علماء الشريعة الإسلامية، وعامة الناس في البلدان الإسلامية وغيرها من البلدان حول العالم. وقد قادتنا هذه الأبحاث إلى اقتراح موضوع "فهم النصوص الدينية" باعتباره أحد التحديات الجديدة والهامة في مجال البحث في لسانيات المدونات اللغوية.

## ١ - المقدمة

يمكن تعريف المدونة اللغوية (corpus) بأنها مجموعة من النصوص اللغوية -أو الأحاديث الشفهية- الطبيعية، التي تجمع لأغراض محددة، وتحفظ بطريقة قابلة للقراءة والبحث حاسوبياً (Ju-; Leech, 1992; McEnery et al., 2006; rafsky and Martin, ٢٠٠٨). وتتعدد التسميات العربية للفظة corpus، لكن تسميتها "مدونة لغوية" قد تكون الأكثر شيوعاً حسبما يرى محمود إسماعيل صالح (العصيمي، قيد النشر)، وكذلك وردت في مجموعة من معاجم ترجمة المصطلحات اللغوية والحاسوبية (انظر مثلاً البعلبكي، ١٩٩٠: ١٢٨، والمبارك، ١٩٩٥: ٦٧، والزهيرى، ٢٠٠٦: ٣٦٥، والفهرى، ٢٠٠٩: ٦٤).

لم تعد خافية على الباحثين اللغويين تلك الأهمية التي بات يحظى بها الجانب الحاسوبي في الدراسات اللغوية، فمثلاً يرى Kilgariff (٢٠٠٧) أن أدوات معالجة اللغة الطبيعية وكذلك اللغويات الحاسوبية قد دخلت إلى مجالات البحث اللغوي كالمتممر في باحة المدرسة، مما جعل كل ما هو غير حاسوبي إما خاضعاً لها، أو متواطئاً معها، أو مُهمَّشاً. وقد أدرك الباحثون في جامعة ليدز هذه الأهمية، فطوّروا أدوات حاسوبية ومصادر لغوية بداية باللغة الإنجليزية، ثم توسعوا في هذه الأدوات والمصادر لتشمل عدداً من اللغات ومنها اللغة العربية. وفيما يتعلق بلسانيات

المدونات العربية فثمة عدة وحدات بحثية في الجامعة تشترك جميعاً في العمل على هذا المجال، ومنها قسم الحاسب الآلي، وقسم اللغات والثقافات والمجتمعات، وقسم الدراسات العربية والإسلامية والشرق أوسطية، ومركز دراسات الترجمة، وقسم اللغويات والصوتيات، ومعهد الذكاء الاصطناعي والأنظمة الحيوية. وتتبنى هذه الأقسام في جامعة ليدز عدداً من المشاريع البحثية في مجالات رئيسة مثل: بناء المدونات العربية، وتطوير أدوات تحليل النصوص العربية، إضافة إلى إجراء الدراسات العلمية في مجال لسانيات المدونات العربية.

نسرد في الفصل الثاني من هذا البحث ثبثاً بأهم المدونات اللغوية العربية التي أنشئت في جامعة ليدز، ثم نخصص الفصل الثالث لاستعراض مجموعة من الأدوات الحاسوبية المصممة لتحليل النصوص العربية، وفي الفصل الرابع نتحدث عن أبرز الموضوعات البحثية التي استفادت من هذه المدونات اللغوية والأدوات الحاسوبية، ثم نختم بأهم مجالات البحث اللغوي التي تسعى لدراساتها مستقبلاً.

## ٢ - المدونات اللغوية العربية

أنشئت في جامعة ليدز مجموعة من المدونات اللغوية العربية لتكون مصدراً مساعداً لإجراء الدراسات على اللغة العربية، ونسرد فيما يلي هذه المدونات مع إيراد أمثلة للأبحاث التي استفادت من كل واحدة منها.

**مدونة تعليم العربية بواسطة الحاسب الآلي (ABC: Arabic By Computer)**  
بدأت مشاركة جامعة ليدز في أبحاث لسانيات المدونات العربية من خلال مشروع مدونة النصوص العربية الفصحى لتعليم اللغة العربية بمساعدة الحاسب Arabic By Computer. وكان الهدف من هذا المشروع إنشاء مصدر لغوي لمتعلمي اللغة العربية، يحتوي على قاعدة بيانات للنصوص العربية ومعجم للمفردات (Brockett et al ١٩٨٩). وعند إنشاء المدونة، كان عرض النصوص العربية وتحريرها يتطلب أجهزة وبرامج خاصة للحاسب الآلي من شركة أبل مآكنتوش،

وفي ثمانينات القرن الماضي لم تكن إدارة الحاسب الآلي في جامعة ليدز توفر حاسبات لتعليم اللغة، حيث كان استخدام هذه الأجهزة مقصوراً على أبحاث العلوم والهندسة. ومن هنا رأينا أهمية إيجاد وسائل تمكن المختصين من الوصول إلى المدونات العربية بسهولة ودون أية قيود لكونها مصدراً لغوياً مهماً في البحث والتعليم.

**المدونة اللغوية للعربية المعاصرة (CCA: the Corpus of Contemporary Arabic)**  
كانت أبحاث لسانيات المدونات في جامعة ليدز قد طوّرت في بداياتها بعض المدونات وأدوات التحليل الخاصة باللغة الإنجليزية، ولكي يتم التوسع في هذه الأدوات لتشمل اللغة العربية فقد قامت بإنشاء أول مدونة لغوية عربية وإتاحة تنزيلها مجاناً، وتشمل مليون كلمة من اللغة العربية المعاصرة (Al-Sulaiti and Atwell ٢٠٠٥، ٢٠٠٦). صممت هذه المدونة لتكون قابلة للمقارنة مع مدونة Lan-caster-Oslo-Bergen Corpus (LOB) للإنجليزية البريطانية المعاصرة التي تشمل على مليون كلمة، وكذلك مدونة Brown للإنجليزية الأمريكية المعاصرة، وفيها مليون كلمة أيضاً (Leech et al 1983a). وعوضاً عن استنساخ نفس الأنواع الأدبية للنصوص الموجودة في هاتين المدونتين، فقد استُطلعت آراء مجموعة من المستخدمين المحتملين لهذه المدونة في مجالي تعليم اللغة وتحليل النصوص العربية، للتعرف على مدى تفضيلهم لأنواع أدبية محددة، ومن ثم ضُمّنت هذه الأنواع في المدونة. وقد استُخدمت المدونة اللغوية للعربية المعاصرة من قبل عدد من الباحثين في مجال لسانيات المدونات العربية وذلك لعدة أغراض بحثية، منها على سبيل المثال: تعليم الهجاء والمفردات في اللغة العربية (Erradi et al ٢٠١٢)، وتصنيف المعاجم العربية (Attia et al ٢٠١١)، وعمل معجم لما يكتب متصلاً بالعربية (Elarian and Idris ٢٠١٠)، وترجمة المجازات ذات الصبغة الثقافية في المقالات العلمية (Merakchi and Rogers ٢٠١٣)، والاختلافات المعجمية في قسمي الشؤون الدولية والرياضة في الصحف العربية (Abdul Razak ٢٠١١)،

ومنها أيضاً الدراسات القائمة على المدونات في علم اللغة الاجتماعي (Friginal and Hardy ٢٠١٤).

### مدونة الإنترنت العربية (Arabic Internet Corpus)

كانت المدونة الوطنية البريطانية (BNC: British National Corpus) في تسعينات القرن الماضي تمثل معياراً مُعْتَمَداً في مجال لسانيات المدونات الإنجليزية، ولم يكن هناك تمويل وخبرات كافية لبناء مدونات عامة وكبيرة لأي من اللغات الأخرى لتحاكي المدونة الوطنية البريطانية مئة مليون كلمة). وللتغلب على ذلك فقد طُوِّرت منهجية لاستخدام محتويات شبكة الإنترنت باعتبارها مدونة لغوية (Baroni and Bernardini ٢٠٠٤)، مما طلب تحديد قائمة من الكلمات تمثل اللغة الهدف وتُستخدَم من قبل برامج خاصة تقوم بجمع مواد المدونة، وذلك بإرسال مجموعة من هذه الكلمات للبحث عنها في محركات البحث المعروفة على شبكة الإنترنت، مثل Google وYahoo وBing وغيرها، ومن ثم تنزيل صفحات النتائج، وتصفية النصوص، ثم تجميعها في مدونة واحدة. وفي جامعة ليدز أُتُبِعَتْ هذه المنهجية (باعتبار الإنترنت مدونة لغوية كبرى) لجمع مدونات من الإنترنت للغة العربية، والصينية، والفرنسية، والألمانية، والإيطالية، والإسبانية، والبولندية، والروسية (Sharoff ٢٠٠٦). وهي مدونات متاحة للجميع من خلال واجهة لكشاف السياقات والمتصاحبات على الموقع التالي: <http://corpus.leeds.ac.uk/inter-net.html>. وتشمل مدونة الإنترنت العربية ١٧٦ مليون كلمة، وقد أضيفت الأصول المعجمية إلى كلماتها في وقت لاحق باستخدام برنامج التحليل الصريح SALMA (Sawalha and Atwell ٢٠١٣a).

### مدونة اللغة العربية حول العالم (World Wide Arabic Corpus)

جمع الباحثون في جامعة ليدز مدونة لغوية عربية من أنحاء العالم (Atwell et al ٢٠٠٧) مماثلة لمدونة اللغة الإنجليزية حول العالم World Wide English



Corpus. وتشمل مدونات فرعية تتألف كل واحدة من مئتي ألف كلمة من كل دولة، وذلك لدراسة اختلاف اللهجات من بلد إلى بلد. وقد استُخدمت هذه المدونة لدراسة التنوع بين لغات الأقطار العربية (أو ما يعرف باللهجات) في استخدام أدوات الوصل (Hassan et al ٢٠١٠، ٢٠١٣)، وكذلك التباين في اللغة العربية وإنجليزية العرب في العالم العربي (Atwell et al ٢٠٠٩).

**المدونة اللغوية لتدريس معلومات عن الإسلام** (Corpus for Teaching about Islam) استُخدمت شبكة الإنترنت - باعتبارها مدونة لغوية - لجمع مدونة تخصصية تضم النصوص المستخدمة لتدريس معلومات حول الإسلام للمستوى الجامعي (Atwell et al ٢٠١١)، وذلك من أجل تأليف موسوعة جامعية لاستخدامها في تدريس الطلاب عن الإسلام والمسلمين، وتشمل اللغة، واللسانيات، ودراسات المناطق.

#### **البنك الشجري للخطاب العربي** (Arabic Discourse Treebank)

يحتاج تحليل الخطاب في اللغة العربية إلى نوع مختلف من وسم الكلمات (وهو ما يعرف بالتحشية annotation)، وقد أنشأت الباحثان Al-Saif and Markert (٢٠١٠) البنك الشجري للخطاب العربي، وهي مدونة تشمل ٥٣٧ نصاً إخبارياً قد وُسمت جميع أدوات الوصل فيها، إضافة إلى العلاقة التي تربط كل أداة مع طرفيها. وقد تطلب هذا تطوير أداة حاسوبية لوسم الخطاب في النصوص العربية، وكذلك إنشاء موقع على شبكة الإنترنت لنشر المدونة.

#### **المدونة اللغوية لمتعلمي اللغة العربية** (ALC: Arabic Learner Corpus)

أنشئت المدونة اللغوية لمتعلمي اللغة العربية لتكون مصدراً لغوياً لأبحاث تعليم اللغة العربية، وكذلك للبحث في مجال معالجة اللغة الطبيعية (Alfaifi and Atwell 2013a, Alfaifi et al 2014). يمكن تنزيل نصوص المدونة من شبكة الإنترنت مجاناً، وهي تتألف من مجموعة من المواد المكتوبة والمنطوقة التي حررها متعلمو اللغة العربية في المملكة العربية السعودية خلال العامين ٢٠١٢ و٢٠١٣، وتضم ١٥٨٥

نصاً (٢٨٢، ٧٣٢ كلمة)، شارك في تحريرها ٩٤٢ طالباً من ٦٧ جنسية، و٦٦ لغة أم مختلفة، ويبلغ متوسط طول النص الواحد ١٧٨ كلمة. وتضم نصوص المدونة نوعين أدبيين: الأول سرديّ يحكي فيه الطالب رحلته في إحدى الإجازات، والثاني يناقش فيه الطالب اهتماماته الدراسية ومستقبله العلمي. ويستطيع الباحثون من خلال المعلومات الوصفية للطلاب والنصوص (metadata) التعرف على خصائص النص اللغوي وكذلك مؤلف النص، مما يضيف عمقاً أكثر لتحليل بيانات المدونة، كما تفيد أيضاً في مجالي التعرف الآلي على النص المكتوب بخط اليد، والتعرف على الكلام المنطوق. كما أن أصول الأوراق - المكتوبة يدوياً من قبل الطلاب - متاحة للتنزيل بعد أن أُدخِلت عن طريق الماسح الضوئي، ثم حفظت في ملفات بصيغة PDF. وكذلك الحال بالنسبة للتسجيلات الصوتية لأولئك الطلاب الذين أذنوا بنشرها على الإنترنت، فهي متاحة للتنزيل في صيغة صوتية رقمية MP3 (ما مجموعه ثلاث ساعات ونصف تقريباً).

سُمِّيت ملفات المدونة بطريقة تسمح بالتعرف على الخصائص الأساسية للنص وكذلك المؤلف، نحو: S038\_T2\_M\_Pre\_NNAS\_W\_C، وهي بالترتيب من اليسار: رقم الطالب، رقم النص، جنس الطالب، المرحلة العامة، ناطق بالعربية باعتبارها لغته الأم أو ناطق بغيرها، نوع النص (مكتوب أم منطوق)، مكان تحرير أو تسجيل النص (في الصف أو في المنزل). يمكن تنزيل نصوص المدونة من موقعها الرئيس (<http://www.arabiclearnercorpus.com>)، كما يمكن البحث فيها عن طريق الإنترنت من خلال الموقع التالي (<http://www.alcsearch.com>) أو موقع SketchEngine (Kilgariff et al ٢٠١٤). استُخدمت هذه المدونة في تقنية تهدف للتعرف على النصوص التي حررها الناطقون باللغة العربية وتمييزها عن تلك التي حررها الناطقون بغير العربية (Malmasi and Dras ٢٠١٤)، وكذلك تصحيح الأخطاء ألياً (Mohit et al ٢٠١٤).

## المدونة العربية للغة القرآن الكريم (Quranic Arabic Corpus)

بعد إصدار المدونة اللغوية للعربية المعاصرة (Corpus of Contemporary Arabic) ظهرت عدة أنواع من المدونات العربية المتاحة للاستخدام، لكن نصوص القرآن الكريم، وكذلك نصوص اللغة العربية الفصحى، لم تجد نفس العناية من قبل المهتمين ببناء المدونات العربية. ولذا تعد المدونة العربية للغة القرآن الكريم Quranic Arabic Corpus (Dukes et al 2013) من أشهر المشاريع في هذا الجانب، فقد بُنيَ هذا المصدر اللغوي من خلال شراكة انطلقت من جامعة ليدز. وتحتوي هذه المدونة عدة طبقات من الوسم، مثل أقسام الكلام بعد تجزئة الكلمات بناءً على الوحدات الصرفية (Dukes and Habash, 2010)، والتحليل النحوي القائم على التوابع (Dukes and Buckwalter, 2010, Dukes et al 2010)، والمعاني الإنجليزية لمفردات القرآن العربية، وكذلك الترجمة الإنجليزية للآيات، وتضم المدونة أيضاً تسجيلات صوتية لتلاوة الآيات، إضافة إلى تصنيف الموضوعات في القرآن الكريم. وكان الدافع لهذا المشروع إنتاج مصدر يساعد على فهم القرآن الكريم، وإجراء المزيد من الأبحاث على نصوصه. ويمكن القول بأنه يختلف عن المشاريع الأخرى بتوفيره مادة لغوية أكثر عمقاً، تقوم على تحليل قواعد اللغة العربية حسب المنهج العربي المعروف بالإعراب. وقد بات بالإمكان من خلال تبني هذا النهج تحفيز الباحثين اللغويين والشرعيين للعمل على موضوع التحشية بأسلوب تعاوني عن طريق الإنترنت، حيث يُستخدم الوسم الآلي المبني على قواعد محددة في هذه الطريقة الجديدة لتحشية المدونة لغوياً، مع إجراء تدقيق يدوي أولي، ثم القيام بالتصحيح عن طريق الإنترنت من قبل الباحثين المتعاونين. وقد استفاد الوسم الصريح لمدونة العربية لنصوص القرآن الكريم من الجهد الذي قدمه مئة متطوع تقريباً، وذلك على شكل اقتراحات لتدقيق الوسم اللغوي للمدونة. كما كان هناك دور إشرافي لعدد قليل من المختصين في عملية الوسم، بحيث يقبلون اقتراحات التصحيح المقدمة عن طريق المتعاونين أو يرفضونها. كما استفادت

المدونة كذلك من الكم الكبير للتراث النحوي العربي الذي قُدِّم من خلال تعليقات الباحثين على نصوص المدونة.

التحدي الأبرز الذي واجه عملية تحشية نصوص القرآن عن طريق الإنترنت تمثل في الحاجة إلى بناء برنامج خاص، عبارة عن منصة عمل تساعد على الوسم بأسلوب تعاوني، ومن هنا بدأت فكرة بناء منصة التحليل اللغوي متعدد الوسائط (Dukes and At-) (LAMP: Linguistic Analysis Multimodal Platform) well (٢٠١٢). استخدمت المدونة العربية لنصوص القرآن الكريم كمصدر معياري لعدد من الأبحاث التي أجريت على اللغة العربية الفصحى، ومن هذه الأبحاث استخراج جذور الكلمات العربية (Yusof et al ٢٠١٠)، والتحليل النحوي للغة العربية (Mohammed and Omar 2011, Rabiee 2011)، واستخراج البصمة الأسلوبية في اللغة العربية (Alqurneh et al forthcoming)، وتحليل الترابط في دراسات الترجمة العربية (Tabrizi and Mahmud ٢٠١٣)، والتلخيص الآلي (El-Haj et al forthcoming)، وتحليل الصيغ الشفهية (Bannister ٢٠١٤). كما كان للمدونة أثر اجتماعي كبير تمثل في مليون زيارة لموقع المدونة خلال عام واحد، ومن ضمن هؤلاء الزوار ناطقون بلغات غير العربية قادتهم للاستفادة من المدونة رغبتهم في الوصول إلى فهم أعمق لنص أصيل من اللغة العربية الفصحى، وذلك بالاستفادة من التحشية اللغوية المستخدمة في المدونة.

**مدونة الإحالة الثنائية لضمائر القرآن الكريم** (QurAna: Quran Pro-) (noun Anaphoric Co-Reference Corpus)

تعد مدونة الإحالة الثنائية لضمائر القرآن الكريم (Sharaf and Atwell 2012a) مدونة غنية من حيث التحشية، فقد وسمت جميع الضمائر الشخصية فيها بمعلومات عن العائد الذي يشير إليه الضمير، سواء أكان العائد سابقاً للضمير أم كان لاحقاً له وهو الغالب في المدونة. كما شمل الوسم كذلك العائد نفسه، أي

الكلمة أو الجملة التي تحوي المعنى، أو الشخص، أو الكيان، أو الفكرة التي يشير إليها الضمير، ولهذا سميت بمدونة الإحالة الثنائية لشمول الضمير وعائده في التحشية. في هذه المدونة وسم الباحثون ما يزيد عن ٢٤.٥٠٠ ضمير بمعلومات حول العائد، وكذلك أكثر من ألف عائد - تتكون من أسماء وعبارات - بمعلومات عن الأشخاص، أو الكيانات، أو المفاهيم، وذلك لجميع الأسماء أو الجمل التي تشير إليها هذه الضمائر. لم يكن من السهل تحديد العائد لجميع الضمائر الشخصية في القرآن الكريم، لكن كتب التفسير وكذلك كلام العلماء حول هذه الضمائر كانا بمثابة الدليل للقائمين على تحشية المدونة. وفي بعض الحالات الخاصة حيث يكون للضمير أكثر من عائد، فقد اتبعنا ما ورد في تفسير ابن كثير الذي يعد مؤلفاً مرجعياً في تفسير القرآن الكريم، وقد اعتمد عليه بدرجة كبيرة في تحديد العائد، مع الأخذ كذلك بدرجة الاتفاق بين العاملين على تحشية المدونة حول عائد الضمير. وتعد هذه المدونة أول مدونة للغة العربية الفصحى يمكن تنزيلها مجاناً، مع كونها موسومة بهذا النوع من المعلومات عن الضمير وعائده.

### مدونة الترابط الدلالي بين آيات القرآن الكريم (QurSim: Quran) (Verse Similarity Corpus)

مدونة الترابط الدلالي بين آيات القرآن الكريم (Sharaf and Atwell 2012b) تمثل إضافة جديدة في مجال التحشية اللغوية على النص القرآني، ففي هذه المدونة - المتاحة مجاناً للباحثين اللغويين المهتمين ببحث موضوع التشابه اللغوي والعلاقات الدلالية بين آيات القرآن الكريم - وسمت الآيات التي بينها علاقة من ناحية الدلالة، وقد اعتمد الباحثون على تفسير ابن كثير في استخراج العلاقات بين آيات القرآن الكريم، حيث يشير ابن كثير إلى الآيات ذات الصلة في معرض تفسيره لكل آية، وقد مكن ذلك من استخراج العلاقات المرجعية بين هذه الآيات والتي زاد عددها عن ٧.٦٠٠ علاقة مشتركة. نشرت "مدونة الترابط الدلالي بين آيات القرآن الكريم" على الموقع التالي على شبكة الإنترنت (TextMiningTheQuran).

com) حيث يستطيع المستخدم رؤية شبكة العلاقات المباشرة وغير المباشرة لأي آية من آيات القرآن الكريم. وقد أظهر تحليل المدونة أن ٣٣٪ فقط من الآيات المرتبطة ببعضها تشترك كلماتها في نفس الجذر، مما يدل على أن الارتباط بينها يتجاوز التطابق المعجمي للألفاظ إلى المعنى الدلالي والمجال المعرفي. يمكن استخدام هذه المدونة لاستخراج الموضوعات الواردة في القرآن الكريم وتصويرها من خلال العلاقات الدلالية بين الآيات (Panju 2014)، وبما أن ابن كثير قد تحدث في تفسيره عن علاقات النص القرآني وهونص عربي فصيح، فهذه العلاقات مصدر جيد يمكن الاستفادة منه في أبحاث ترجمة معاني القرآن الكريم، ومن ذلك على سبيل المثال أن الآيات التي بينها ارتباط في اللغة العربية ينبغي أن تبقى كذلك عند ترجمة معاني الآيات إلى أي لغة أخرى، وبالتالي فقد تكون هذه المدونة مصدراً للبحث في ترجمة معاني القرآن الكريم للغات الأخرى، وخصوصاً في موضوع الآيات المرتبطة دلاليًا والعلاقة بينها.

### مدونة القرآن الكريم الموسومة بموضوع الآيات مع الترجمة الإنجليزية (Qurany: Quran Corpus Annotated with English Transla- tions and Verse Topics)

مدونة القرآن الكريم الموسومة بموضوع الآيات مع الترجمة الإنجليزية Qurany (Abbas 2009, Abbas and Atwell 2013) عبارة عن أداة ثنائية اللغة (إنجليزي - عربي) ذات قدرة على تحسين الدقة والاستجابة (recall and precision) عند البحث في موضوعات القرآن الكريم، وقد أمكن تحقيق هذا من خلال الجمع بين عدة أنواع من التحشية، حيث وسمت كل آية من آيات القرآن الكريم بمعلومات عن الموضوع الدلالي الذي تحتويه، والذي استُخرج من "مصحف التجويد"، وهو مرجع موثوق يحوي فهرساً لما يقارب ألفاً ومئة من المفاهيم والموضوعات مع الآيات المرتبطة بها. وقد حوّل الباحثان فهرس هذا المصحف إلى تصنيف يعرضها بصورة شجري يمكن استخدامه من خلال موقع المدونة

على شبكة الإنترنت، حيث يستطيع المستخدم التنقل بين شجرة التصنيف لإيجاد الفكرة المطلوبة، ومن ثم الوصول إلى قائمة بالآيات المرتبطة بها. كما وسمت كل آية بثماني ترجمات إنجليزية مشهورة لمعاني القرآن الكريم، فبات بالإمكان إيجاد الآيات القرآنية من خلال البحث في هذه الترجمات الإنجليزية باستخدام أي كلمة أو كلمات مفتاحية، كما يمكن للمستخدم مشاهدة مجموعة من مرادفات الكلمات المفتاحية المستخدمة في البحث، والاستفادة منها في توسيع نطاقه وبالتالي تحسين جودة النتائج. بإمكان المستخدم أيضاً الوصول إلى بيانات هذه المدونة وتنزيل محتواها على شكل صفحات مكتوبة بلغة النصوص المتشعبة HTML، وتشتمل كل صفحة من هذه الصفحات على آية واحدة مع ترجماتها الإنجليزية، وكذلك قائمة مصحف التجويد للموضوعات المرتبطة بها. تتوافق لغة HTML هذه مع معايير محرك البحث Google، حيث يمكنك قصر البحث في هذا المحرك على موقع إنترنت محدد باستخدام العبارة (site) مع كلمات البحث، فعلى سبيل المثال تقوم عبارة البحث التالية:

"prayer site:http://www.comp.leeds.ac.uk/nora/html" بإيجاد جميع الآيات التي ورد ذكر الصلاة في ترجمات معانيها، أو ورد ذلك في قائمة موضوعات مصحف التجويد.

### مدونة جامعة الملك سعود للغة العربية الفصحى (KSUCCA King) (Saud University Corpus of Classical Arabic)

يمكن دراسة الأنماط المعجمية في القرآن الكريم باستخدام أي كشاف للسياقات متوافق مع اللغة العربية مثل aConCorde، لكن الباحثين اللغويين وصناع المعاجم يحتاجون إلى مدونات كبيرة لدراسة المتصاحبات اللغوية والنماذج السياقية لمفردات اللغة. وفي اللغة الإنجليزية على سبيل المثال، أنشئت المدونة الوطنية البريطانية لتكون أساساً للبحث المعجمي، وهي تحتوي على مئة مليون

كلمة، أما في اللغة العربية الفصحى فالنص القرآني يحتوي على خمسين ألف كلمة تقريباً (بناء على طريقة احتساب حدود الكلمة)، ولأن البحث في الدلالة التوزيعية لمفردات المعجم يحتاج إلى عدد كبير نسبياً من الأمثلة لكل كلمة أو متصاحب لغوي للتمكن من دراسته، ولأن كثيراً من الكلمات والعبارات في القرآن الكريم لم ترد إلا في مواضع قليلة؛ فقد تعاونت جامعة ليدز مع جامعة الملك سعود لإنشاء مدونة تحوي خمسين مليون كلمة من اللغة العربية الفصحى القرية من فترة نزول القرآن الكريم، وهذه المدونة هي مدونة جامعة الملك سعود للغة العربية الفصحى (Alra-biah et al 2013, 2014a,b). وقد سمحت لنا هذه المدونة بإيجاد كثير من الأمثلة والسياقات لكلمات وردت في القرآن الكريم، وهي مدونة يمكن تنزيل محتوياتها من موقعها الخاص على شبكة الإنترنت (<http://ksucorpus.ksu.edu.sa>)، كما يمكن البحث فيها باستخدام موقع SketchEngine (Kilgarrieff et al 2014). هذه المدونة تعد مفتاحاً للدراسات اللغوية التاريخية القائمة على المدونات (Alrabiah et al 2014b). وكذلك دراسات الدلالة التوزيعية لمفردات القرآن الكريم.

### ٣. أدوات تحليل النصوص العربية

من المهم في ظل تنامي استخدام المدونات اللغوية وجود أدوات حاسوبية تستطيع الاستفادة من هذه المصادر اللغوية، وبسبب الخصائص الفريدة للغة العربية فإنها تحتاج إلى أدوات تستطيع التعرف عليها ومعالجتها بطريقة صحيحة، وقد صمم الباحثون في جامعة ليدز مجموعة من الأدوات المخصصة لمعالجة اللغة العربية نستعرضها في هذا القسم.

**الكشاف السياقي aConCorde للغة العربية** (aConCorde Concordance for Arabic)

من الملاحظ أن كشافات السياقات المتوفرة قبل عدة سنوات لم تكن مصممة



للتعامل مع الخصائص المميزة للنص العربي، ومن ذلك مثلاً اختلاف رسم الحروف العربية عن تلك اللاتينية، والتباين في معايير تمثيلها حاسوبياً، واختلاف رسم بعض الحروف بناءً على موضعها في الكلمة، ومن تلك الخصائص أيضاً عدم رسم الصوائت القصيرة (الحركات) في كثير من الأحيان، والتباين في طريقة استخدام علامات الترقيم، وكذلك كتابة النص العربي من اليمين إلى اليسار، والذي أبرز الحاجة إلى تبديل نافذتي "الكلمات السابقة" و"الكلمات اللاحقة" في برامج كشافات السياقات لتلائم اتجاه النص العربي. ولهذا فقد بُنيَ كشاف السياقات aConCorde في قسم الحاسب الآلي بجامعة ليدز، وهو أداة مجانية ومفتوحة المصدر، صممت في المقام الأول لمساعدة الباحثين في لسانيات المدونات العربية (Roberts et al 2005, 2006). وقد أشادت دراسة علمية أجريت حول كشافات السياقات في ذلك الوقت (Wiechmann and Fuhs 2006) ببرنامج aConcorde الذي يوفر دعماً كاملاً للنصوص العربية، ويتمثل هذا الدعم في عدة جوانب منها واجهة المستخدم التي يمكن تحويلها بين اللغتين العربية والإنجليزية، وتبنيه لترميز الحروف المعياري Unicode، إضافة إلى دعمه لمجموعة من الخطوط العربية، وكذلك اتجاه النص العربي من اليمين إلى اليسار. وبعد عقد من الزمن أُجريَ تقييم آخر للأدوات المستخدمة في تحليل المدونات العربية (Alfaifi and Atwell 2014a) حيث وجدت الدراسة أن معظم الأدوات ما زالت تعاني من مشاكل في دعم اللغة العربية. وقد استُخدم aConcorde في عدة دراسات في مجال لسانيات المدونات العربية، ومن ذلك الكشف عن الأنماط المعجمية الشائعة في النصوص العربية (El-Haj et al forthcoming)، والكلمات والعبارات المفتاحية (Ali 2012)، والتعرف على أساليب الجرائم من خلال مدونة إخبارية لتقارير الجرائم (Alruily 2012).

## المحلل الصرفي SALMA لوسم أقسام الكلام في المدونات العربية

(SALMA: Standard Arabic Language Morphological Analysis)

أسهم الباحثون في جامعة ليدز في مشروع وسم مدونة LOB للإنجليزية البريطانية المعاصرة (Atwell 1982, Leech et al 1983b)، وقد منحهم ذلك خبرة استفادوا منها في وسم المدونات العربية (Atwell 2008, Atwell et al)، حيث جرى العمل على تطوير أداة لتحليل الصرفي، وكذلك وسم أقسام الكلام في المدونات العربية، إضافة إلى بعض الوظائف الأخرى. وكانت البداية بوسم عينة قصيرة من القرآن الكريم إضافة إلى نص إخباري، لتكون هاتان العينتان بمثابة معيار يمكن الاعتماد عليه لإجراء مقارنة بين المحللات الصرفية الحالية (Sawalha and Atwell 2008). كما جرت مقارنة بين منهجين لتحليل اللغة العربية صرفياً، يقوم الأول على أساس القواعد اللغوية، بينما يعتمد الثاني على المدونات اللغوية (Sawalha and Atwell 2009). قادت هذه الدراسات إلى تطوير أداة جديدة أطلق عليها SALMA، وهي أداة لتحليل الوحدات الصرفية الصغرى في اللغة العربية -Arabic Corpus Part-of-Speech Tagging and Morphological Analysis، مع وسمها باستخدام جدول الوسم الصرفي لأقسام الكلام المضمن في هذه الأداة.

والاسم SALMA يشير إلى: محلل ليدز الصرفي لصوالة وأتويل Sawalha وAtwell Leeds Morphological Analysis (Sawalha and Atwell 2013a) أو المحلل الصرفي للغة العربية الفصحى -Standard Arabic Language Morphological Analysis (Sawalha et al 2013). ويشتمل على مجموعة من الأدوات لتحليل المدونات العربية، ومنها جدول معياري لوسم أقسام الكلام، يأخذ بعين الاعتبار السمات الصرفية الأصلية للغة العربية (Sawalha and Atwell 2013b)، وقد تضمن هذا العمل تحليلاً لأبحاث اللغويين العرب القدماء واستخدامها كأساس للجدول الحالي الذي يعطي تصنيفاً شاملاً ومفصلاً لبنية الكلمة في اللغة العربية.

ومع وجود عدد من الجداول الأخرى لوسم أقسام الكلام، والتي أُعدت للقيام بمهام محددة في اللغة العربية، إلا إنها في الغالب إما مستفادة من نماذج أعدت للغة الإنجليزية، أو إنها تغطي جزءاً محدوداً فقط من الدراسات الصرفية العربية، ولهذا فقد أنشئ مقياس لمقارنة وتقييم جداول أقسام الكلام كجزء من المشروع الحالي. وقد أضيف إلى أداة التحليل الصرفي وجدول أقسام الكلام معجم عربي كبير مستمد من المصادر المعجمية مفتوحة المصدر على شبكة الإنترنت، وكذلك المعاجم التراثية للغة العربية (Sawalha and Atwell 2010b)، كما أضيفت كذلك ميزة لعرض للتحليل الصرفي العربي بطريقة مرئية، حيث تظهر الوحدات الصرفية مميزة بألوان مختلفة عن بعضها ليسهل تمييزها بصرياً (Sawalha and Atwell 2012). وللتحقق من جودة عمل هذه الأدوات على المدونات الكبيرة، فقد تمت تجربتها على مدونة الإنترنت العربية التي يصل حجمها إلى ١٧٦ مليون كلمة (Sawalha and Atwell 2013a)، كما استُخدمت في مجموعة من الأبحاث في مجال لسانيات المدونات العربية، ومنها إنشاء قوائم بالمفردات اللغوية لتعليم اللغة العربية (Kilgariff et al ٢٠١٤)، وتعليم الهجاء والمفردات العربية (Er-radi et al 2012)، وكذلك التحليل النحوي العربي (Rabiee 2011)، وأيضاً تحليل وسائل التواصل الاجتماعي العربية (El-Beltagy and Ali 2013).

### **جدول وسم الأخطاء في مدونات المتعلمين العربية (Arabic Learner Corpus Error Tag-Set)**

أنشئ جدول تصنيف الأخطاء لاستخدامه في وسم الأخطاء اللغوية في مدونات المتعلمين العربية (Alfaifi et al 2013، Alfaifi and Atwell 2012، 2014b)، وقد استفاد هذا التصنيف من مجموعة من تصنيفات الأخطاء المصممة لمدونات المتعلمين، ليكون مناسباً للأخطاء اللغوية التي يقع فيها متعلمو اللغة العربية. استخدم هذا التصنيف لتطوير أداة هي عبارة عن محرر لوسم أخطاء الطلاب في المدونة اللغوية لمتعلمي اللغة العربية (Alfaifi and Atwell 2015)، والذي يمكن الاستفادة منه في وسم أي مدونة من مدونات المتعلمين.

## الوسم الصوتي والمقطعي للغة العربية (Arabic Phonetic and Prosodic Tagging)

نوع آخر من أنواع الوسم يتمثل في كتابة النصوص صوتيًا - ومنها نصوص القرآن الكريم - وذلك لعدة أهداف وتطبيقات كالتمكن من قراءتها بصوت مسموع، والتعرف الآلي على الكلام المنطوق. يستخدم علماء الأصوات الألفبائية الصوتية الدولية (IPA International Phonetic Alphabet) لكتابة النصوص المنطوقة في مختلف لغات العالم، ولأن رسم اللغة العربية لا يعبر عن طريقة اللفظ بدقة كبيرة، إضافة إلى أنه لا يوجد ربط مباشر بين كل واحد من الحروف المكتوبة في اللغة العربية مع رموز الألفبائية الصوتية الدولية API، فقد بُني نظام دقيق للربط بينهما مع كتابة النصوص العربية صوتيًا وبطريقة آلية، وقد بني هذا النظام على تحليل لتلاوة القرآن الكريم وأحكام التجويد كالوقف والغنة ونحوهما، إضافة إلى الاستفادة من مجالي اللغويات العربية وعلم الأصوات الحديث (Brierley et al 2012a, b). ويتضمن النص القرآني على وجه الخصوص رموزًا مقطعية تشير إلى عدة أنواع من الوقف أو حدود الجملة التي ينبغي مراعاتها عند قراءة الآيات بصوت مسموع، وهي جزء من أحكام التجويد التي تضبط عملية النطق والوقف أثناء تلاوة القرآن الكريم. وقد استخدم نظام الربط بالألفبائية الصوتية الدولية، وكذلك رموز المقاطع، في التحشية الصوتية والمقطعية للمدونة العربية للقرآن الكريم الموسومة بحدود الوقف في النص القرآني (Brierley et al 2012a, b). يمكن أن تساعد هذه المدونة غير الناطقين باللغة العربية على تعلم التلاوة الصحيحة لآيات القرآن، كما يمكن أن تُستخدم لتدريب أدوات الوسم المقطعي لنصوص اللغة العربية بما في ذلك اللغة العربية المعاصرة (Sawalha et al 2012a, b).

## تدريب برامج التخاطب الآلي على مدونة عربية (-Arabic Corpus Trained Chatbots)

من الاستخدامات المبتكرة للمدونات استخدامها في تدريب برامج التخاطب الآلي عن طريق الإنترنت، والتي تقوم على أنظمة تعلم الآلة Machine Learning (Abu Shawar and Atwell 2005a)، ومن ثم تستخدم هذه البرامج بعد التدريب كأداة لاستكشاف المدونات نفسها (Abu Shawar and Atwell 2005b). وكمثال على ذلك، فقد جرى تدريب برنامج للتخاطب الآلي على مدونة قرآنية، وهو ما أنتج لنا نظام تخاطب عربي يعطي إجابات من القرآن الكريم (Abu Shawar and Atwell ٢٠٠٤)، كما دُرِّبَتْ نسخة أخرى منه على مدونة للأسئلة الأكثر شيوعاً عن الحوسبة العربية، ليقوم بالإجابة على أي سؤال حول هذا الموضوع (Abu Shawar and Atwell 2009).

## الوسم الدلالي والتمثيل المعرفي للغة القرآن الكريم (-Semantic Tagging and Knowledge Representation for Quranic Arabic)

تم إضافة الكثير من مستويات الوسم اللغوي للنص القرآني مثل: أقسام الكلام، والصرف، والإحالات المرجعية، والكتابة الصوتية، والحدود المقطعية للجمل، والبنية النحوية، كما أضيفت عدة أنواع من التحشية التي تمثل معلومات معرفية في القرآن، ومنها تصنيف الكيانات الاسمية المعروفة بالضمائر الشخصية، وكذلك موضوعات الآيات، وعلاقاتها الدلالية، وأيضاً ترجمة معاني القرآن الكريم إلى اللغة الإنجليزية، ومن هذه الترجمات ما هو على مستوى الكلمة ومنها ما هو على مستوى الآية. ويهدف الباحثون في جامعة ليدز إلى توحيد جميع هذه التحشيات اللغوية والمعرفية (Abbas et al 2013, Abbas and Atwell 2013) لإنشاء تصور للتمثيل المعرفي في القرآن الكريم، وذلك للتمكن من إجراء وسم دلالي له (Sharaf and Atwell 2009, Alrehaili and Atwell 2013, 2014).

#### ٤. البحث اللغوي باستخدام المدونات العربية وأدوات تحليل نصوصها

استُخدمت المدونات اللغوية والأدوات الحاسوبية التي طورها الباحثون في جامعة ليدز لإجراء العديد من الأبحاث في مجال لسانيات المدونات العربية، ومن ذلك تعليم اللغة العربية، والمقارنة بين اللغتين الإنجليزية والعربية، وهو ما سنتحدث عنه في هذا القسم.

#### تعليم اللغة العربية باستخدام المدونات

من خلال المشروع الرائد ABC (Arabic By Computer) بنى الباحثون قاعدة بيانات للنصوص العربية ومعجم للمفردات لاستخدامهما في تعليم وتعلم اللغة العربية (Brockett et al 1989). وفي الآونة الأخيرة استطعنا في جامعة ليدز تجربة كشف للسياقات وبرامج للتخاطب الآلي، مستخدمين مدونات على شبكة الإنترنت، وذلك بهدف تعليم اللغة العربية (Al-Sulaiti et al 2005, 2007)، ومن خلال الأبحاث القائمة على المدونات أيضاً أنشئت قوائم للمفردات اللغوية التي تخدم متعلمي اللغة العربية (Kilgariff et al 2013)، وقد توصلنا مع مجموعة من معلمي ومتعلمي اللغة العربية في قسم اللغة العربية بجامعة ليدز الذين يستخدمون المصادر اللغوية للمدونات العربية في تعلم وتعليم هذه اللغة. كما أن المجتمع المحلي للمسلمين أقام مدرسة لتعليم الأطفال اللغة العربية في أيام السبت من كل أسبوع، لتمكينهم من قراءة وفهم القرآن الكريم، وقد استمتع الطلاب بنظام التخاطب المبكر على شبكة الإنترنت، والذي جرى تدريبه على مدونة قرآنية ليعطي إجابات بنفس اللغة العربية الفصحى المستخدمة في القرآن الكريم (Abu Shawar and Atwell 2004).

#### مقارنة اللغتين العربية والإنجليزية القائم على المدونات

كان البحث في جامعة ليدز قد بدأ بمجال لسانيات المدونات الإنجليزية، (انظر مثلاً Lee et al 1983a.b)، ثم تبعه اهتمام بإجراء مقارنات بين اللغتين

العربية والإنجليزية تقوم على مدونات لهاتين اللغتين، وتشمل هذه المقارنات تأثير اللغة العربية على إنجليزية العرب، والتنوع في استخدام اللغة الإنجليزية في العالم العربي (Atwell et al 2009)، كما تشمل كذلك جداول الوسم الصرفي وأقسام الكلام في اللغتين العربية والإنجليزية (Atwell 2008, Sawalha and Atwell 2013c)، وأيضاً التمثيل البصري للنبر والوقف في نصوص المدونات العربية والإنجليزية من خلال تمييز مواضعها في النص بعلامات متباينة الألوان (Brier-ley et al 2012c).

## الخاتمة

تستخدم أجهزة الحاسب الآلي والمدونات اللغوية على نطاق واسع في أبحاث اللسانيات، وقد طوّر الباحثون في جامعة ليدز عدة مدونات لغوية وبرامج حاسوبية على شبكة الإنترنت للاستفادة منها في مجال لسانيات المدونات العربية، وهي أدوات مفتوحة المصدر ومتاحة على شبكة الإنترنت وليست تجارية، آملي أن يسهم ذلك في استخدامها على نطاق واسع. ويعتبر وسم وتمثيل الجوانب الدلالية والمعرفية في النصوص العربية - وخصوصاً في القرآن الكريم والنصوص الدينية الأخرى - تحدياً للأبحاث القادمة، كما يمثل فهم القرآن حاسوبياً التحدي الأكبر في مجال لسانيات المدونات العربية.





## المراجع

١. البعلبكي، رمزي منير (١٩٩٠) معجم المصطلحات اللغوية (إنجليزي عربي). دار العلم للملايين، بيروت.
٢. الزهيري، نبيل (٢٠٠٦) معجم المصطلحات اللغويات في المعلوماتية. مكتبة لبنان، بيروت.
٣. صالح، محمود إسماعيل (قيد النشر) المدونات اللغوية وكيفية الإفادة منها، تحرير صالح العصيمي، المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها. مركز الملك عبد الله بن عبدالعزيز الدولي لخدمة اللغة العربية، الرياض.
٤. الفهري، عبد القادر الفاسي (٢٠٠٩) معجم المصطلحات اللسانية (إنجليزي - فرنسي - عربي). دار الكتاب الجديد، بيروت.
٥. المبارك، مبارك (١٩٩٥) معجم المصطلحات اللغوية (فرنسي - إنجليزي - عربي). دار الفكر، بيروت.
6. Abbas. N and E Atwell. 2013. 'Annotating the Arabic Quran with a classical semantic ontology.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.
7. Abbas. N. L Aldhubayi. H Al-Khalifa H. Z Alqassem. E Atwell. K Dukes. M Sawalha. and M Sharaf. 2013. 'Unifying linguistic annotations and ontologies for the Arabic Quran.' Proceedings of WACL2 Second Workshop on Arabic Corpus Linguistics.

8. Abbas. N. 2009. 'Quran Search for a Concept Tool and Website'. MRes Thesis. School of Computing. University of Leeds.
9. Abdul Razak. Z. 2011. 'Modern media Arabic: a study of word frequency in world affairs and sports sections in Arabic newspapers.' PhD Thesis. University of Birmingham.
10. Abu Shawar. B and E Atwell. 2004. 'An Arabic chatbot giving answers from the Quran.' Proc TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles.
11. Abu Shawar. B and E Atwell. 2005a. 'Using corpora in machine-learning chatbot systems.' International Journal of Corpus Linguistics. vol. 10. pp. 489-516.
12. Abu Shawar. B and E Atwell. 2005b. 'A chatbot system as a tool to animate a corpus.' ICAME Journal: International Computer Archive of Modern and Medieval English Journal. vol. 29. pp.5-24.
13. Abu Shawar. B and E Atwell. 2009. 'Arabic Question-Answering via Instance Based Learning from an FAQ Corpus.' Proceedings of CL2009 Corpus Linguistics.
14. Al-Saif. A. and K Markert. 2010. 'The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic.' Proceedings of LREC'2010: Language Resources and Evaluation Conference.

15. Al-Sulaiti. L and E Atwell. 2006. 'The design of a corpus of contemporary Arabic.' International Journal of Corpus Linguistics. vol. 11. pp. 135-171.
16. Al-Sulaiti. L. A Roberts. and E Atwell. 2005. 'The use of corpora and concordance in the teaching of contemporary Arabic.' Proceedings of EuroCALL'2005.
17. Al-Sulaiti. L. A Roberts. B Abu Shawar. and E Atwell. 2007. 'The Use of Corpus. Concordancer and Chatbot in the Teaching of Contemporary Arabic.' Proceedings of CL'2007 Corpus Linguistics
18. Al-Sulaiti. L. and E Atwell. 2005. 'Extending the Corpus of Contemporary Arabic.' Proceedings of CL'2005 Corpus Linguistics.
19. Alfaifi. A and E Atwell. 2012. 'Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors'. Proceedings of ICCA'2012 International Computing Conference in Arabic.
20. Alfaifi. A and E Atwell. 2013a. 'Arabic Learner Corpus v1: A New Resource for Arabic Language Research.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.
21. Alfaifi. A and E Atwell. 2013b. 'Arabic Learner Corpus: Texts Transcription and Files Format.' Proceedings of CORPORA'2013 International Conference on Corpus Linguistics.

22. Alfaifi. A and E Atwell. 2014a. 'Tools for Searching and Analysing Arabic Corpora: an Evaluation Study.' Proceedings BAAL-CUP'2014 British Association for Applied Linguistics and Cambridge University Press Applied Linguistics Workshop.
23. Alfaifi. A and E Atwell. 2014b. 'An evaluation of the Arabic error tagset v2.' Proceedings of ACL'2014 American Association for Corpus Linguistics.
24. Alfaifi. Abdullah and Atwell. Eric. 2015. Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. The 8th Saudi Students Conference. 31 January – 1 February 2015. London. UK.
25. Alfaifi. A. E Atwell. and G Abuhakema. 2013. 'Error Annotation of the Arabic Learner Corpus: A New Error Tagset. Language Processing and Knowledge in the Web. vol. 8105. pp.14-22. Springer.
26. Alfaifi. A. E Atwell. and I Hedaya. 2014. 'Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners.' Proceedings of LCSAW'2014 Learner Corpus Studies in Asia and the World.
27. Ali. I. 2012. 'Application of a Mining Algorithm to Finding Frequent Patterns in a Text Corpus: A Case Study of Arabic.' International Journal of Software Engineering and Its Applications. vol.6(3). pp.127-134.

28. Alqurneh. A. A Mustapha. M Murad. and N Sharef. Forthcoming. 'Stylometric model for detecting oath expressions: A case study for Quranic texts.' Literary and Linguistic Computing journal.
29. Alrabiah. M. A Al-Salman. and E Atwell. 2013. 'The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.
30. Alrabiah. M. A Al-Salman. E Atwell. and N Alhelewh. 2014a. 'KSUCCA: A Key To Exploring Arabic Historical Linguistics.' International Journal of Computational Linguistics. vol. 5. pp.27-36.
31. Alrabiah. M. N Alhelewh. A Al-Salman. and E Atwell. 2014b. 'An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus.' International Journal of Computational Linguistics. vol. 5. pp.1-13.
32. Alrehaili. S and E Atwell. 2013. 'Linguistics features to confirm the chronological order of the Quran.' Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.
33. Alrehaili. S and E Atwell. 2014. 'Computational ontologies for semantic tagging of the Quran.' Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts.

34. Alruily, M. 2012. 'Using Text Mining to Identify Crime Patterns from Arabic Crime News Report Corpus.' PhD Thesis. De Montford University.
35. Attia, M. P Pecina, L Tounsi, A Toral, and J Van Genabith. 2011. 'Lexical Profiling for Arabic.' Proceedings of eLex'2011 Electronic Lexicography in the 21st Century.
36. Atwell, E. C Brierley, K Dukes, M Sawalha, and A Sharaf. 2011. 'An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet.' Proceedings of NITS'2011 3rd National Information Technology Symposium, Riyadh.
37. Atwell, E. J Arshad, C Lai, L Nim, N Rezapour Asheghi, J Wang, and J Washtell. 2007. 'Which English dominates the World Wide Web, British or American?' Proceedings of CL'2007 Corpus Linguistics.
38. Atwell, E. K Dukes, A Sharaf, N Habash, B Louw, B Abu Shawar, A McEnery, W Zaghouani, and M El-Haj. 2010. 'Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence.' Proceedings of GCCR'2010 Grand Challenges in Computing Research.
39. Atwell, E. L Al-Sulaiti, and S Sharoff. 2009. 'Arabic and Arab English in the Arab World.' Proceedings of CL2009 Corpus Linguistics.

40. Atwell. E. L Al-Sulaiti. S Al-Osaimi. and B Abu Shawar. 2004. 'A review of Arabic corpus analysis tools'. Proceedings of TALN'2004: Traitement Automatique des Langues Naturelles.
41. Atwell. E. N Abbas. B Abu Shawar. L Al-Sulaiti. A Roberts. and M Sawalha. 2008. 'Mapping Middle Eastern and North African Diasporas.' Proceedings of BRISMES'2008 British Society for Middle Eastern Studies.
42. Atwell. E. (ed.) 1993. 'Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledge Based Systems Initiative.' 146pp. Leeds University Press.
43. Atwell. E. 1982. LOB Corpus Tagging Project: Manual Postedit Handbook. Department of Linguistics and Modern English Language. University of Lancaster.
44. Atwell. E. 1993. 'The HEFC's Knowledge Based Systems Initiative.' AISBQ: Artificial Intelligence and Simulation of Behaviour Quarterly. vol. 83. pp.29-34.
45. Atwell. E. 2008. 'Development of tag sets for part-of-speech tagging.' Ludeling A; Kyto M (ed.) Corpus Linguistics: An International Handbook. Volume 1. pp.501-526. Mouton de Gruyter.
46. Atwell. E. 2011. 'Exploiting New Technology and Innovation for Detecting Terrorist Activities.' Counter Terror Expo. London.



47. Bannister, A. 2014. 'An Oral-Formulaic Study of the Quran.' Lexington.
48. Baroni, M and S Bernardini. 2004. 'BootCaT: Bootstrapping corpora and terms from the web.' Proceedings of LREC'2004 Language Resources and Evaluation Conference.
49. Brierley, C, E Atwell, C Rowland, and J Anderson. 2013. 'Semantic Pathways: a Novel Visualization of Varieties of English.' ICAME Journal of the International Computer Archive of Modern English, vol. 37, pp.5-36.
50. Brierley, C, M Sawalha, and E Atwell. 2012a. 'Boundary Annotated Quran Corpus for Arabic Phrase Break Prediction.' Proceedings of IVACS'2012 Inter-Varietal Applied Corpus Studies.
51. Brierley, C, M Sawalha, and E Atwell. 2012b. 'Open-source boundary-annotated corpus for Arabic speech and language processing.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.
52. Brierley, C, M Sawalha, and E Atwell. 2012c. 'Visualisation of Prosody in English and Arabic Speech Corpora.' Proceedings of AVML'2012 Advances in Visual Methods for Linguistics.
53. Brierley, C, M Sawalha, B Heselwood, and E Atwell. forthcoming. 'A verified Arabic-IPA mapping for Arabic tran-

- scription technology. informed by Quranic recitation. traditional Arabic linguistics. and modern phonetics.' Journal of Semitic Studies.
54. Brockett A. E Atwell. O Taylor. and M Page. 1989. 'An Arabic text database and glossary system for students.' Proceedings of the Seminar on Bilingual Computing in Arabic and English.
  55. Danso. S. E Atwell. O Johnson. A ten Asbroek. S Soromekun. K Edmond. C Hurt. L Hurt. C Zandoh. C Tawiah. J Fenty. S Etego. S Agyei. and B Kirkwood. 2013. 'A semantically annotated verbal autopsy corpus for automatic analysis of cause of death.' ICAME Journal of the International Computer Archive of Modern and Medieval English. vol. 37. pp.37-69.
  56. Dukes. K and E Atwell. 2012. 'LAMP: a multimodal web platform for collaborative linguistic analysis.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.
  57. Dukes. K and N Habash. 2010. 'Morphological Annotation of Quranic Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.
  58. Dukes. K and T Buckwalter. 2010. 'A Dependency Treebank of the Quran using Traditional Arabic Grammar.' Proceedings of INFOS'2010 7th Informatics and Systems.

59. Dukes. K. E Atwell. and A Sharaf. 2010. 'Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.
60. Dukes. K. E Atwell. and N Habash. 2013. 'Supervised collaboration for syntactic annotation of Quranic Arabic.' Language Resources and Evaluation Journal. vol. 47. pp.33-62.
61. El-Beltagy. S. and A Ali. 2013. 'Open issues in the sentiment analysis of Arabic social media: A case study.' Proceedings of IIT'2013 Innovations in Information Technology.
62. El-Haj. M. U Kruschwitz. C Fox. Forthcoming. Creating language resources for under-resourced languages: methodologies. and experiments with Arabic. Language Resources and Evaluation journal.
63. Elarian. Yousef S. and Fayeze M. Idris. 2010. 'A Lexicon of Connected Components for Arabic Optical Text Recognition' First International Workshop on Frontiers in Arabic Handwriting Recognition. 22 August 2010. Istanbul.
64. Erradi. A. S Nahia. H Almereghi. and L Al-kailani. 2012. ArabicTutor: a Multimedia m-Learning Platform for Learning Arabic Spelling and Vocabulary. Proceedings of ICMCS'2012 International Conference on Multimedia Computing and Systems.

65. Friginal. E and J Hardy. 2014. 'Corpus-based Sociolinguistics: A Guide for Students.' Routledge.
66. Garside. R and N Smith. 1997. 'A hybrid grammatical tagger: CLAWS4.' in Garside. R. G Leech and A McEnery (eds.) 'Corpus Annotation: Linguistic Information from Computer Text Corpora.' Longman. London. pp. 102-121.
67. Greene. B and G Rubin. 1971. 'Automatic grammatical tagging of English.' Technical report. Department of Linguistics. Brown University.
68. Hassan. H. N Daud. and E Atwell. 2010. 'Connectives in the World Wide Arabic corpus.' Proceedings of IVACS 2010 Inter-Varietal Applied Corpus Studies.
69. Hassan. H. N Daud. and E Atwell. 2013. 'Connectives in the World Wide Web Arabic corpus.' World Applied Sciences Journal (Special Issue of Studies in Language Teaching and Learning), vol. 21, pp.67-72.
70. Jurafsky. D. and Martin. J. H. 2009. 'Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics'. New Jersey: Prentice Hall.
71. Karlsson. F. A Voutilainen. J Heikkila. and A Anttila (eds.). 1995. 'Constraint Grammar: A Language-Independent System for Parsing Running Text.' Mouton de Gruyter. Berlin and New York.

72. Kilgarriff. A. 2007. 'Re: [Corpora-List] history of corpus linguistics.' Corpora-List Archive. 6 January 2007.
73. Kilgarriff. A. V Baisa. J Bušta. M Jakubíček. V Kovář. J Michelfeit. P Rychlý. and V Suchomel. 2014. 'The Sketch Engine: ten years on.' *Lexicography journal* vol.1(1). pp.1-30.
74. Kilgarriff. A. F Charalabopoulou. M Gavriliidou. J Jonassen. S Khalil. S Johansson. R Lew. S Sharoff. R Vadlapudi. and E Volodina. 2013 'Corpus-based vocabulary lists for language learners for nine languages.' *Proceedings of LREC'2013 Language Resources and Evaluation Conference*.
75. Leech. G. 1992 'Corpora and theories of linguistic performance'. in Svartvik. J. 'Directions in Corpus Linguistics'. pp 105-22. Mouton de Gruyter. Berlin.
76. Leech. G. R Garside. and E Atwell. 1983a. 'Recent developments in the use of computer corpora in English language research.' *Transactions of the Philological Society*. 1983. pp.23-40.
77. Leech. G. R Garside. and E Atwell. 1983b. 'The Automatic Grammatical Tagging of the LOB Corpus.' *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*. vol. 7. pp.13-33.
78. Malmasi. S. and M Dras. 2014. Arabic Native Language Identification. In the proceedings of the EMNLP 2014

- Workshop on Arabic Natural Language. 25 October 2014. Doha. Qatar.
79. McEnry, T. R Xiao and Y Tono. 2006. «Corpus-Based Language Studies: An advanced resource book». Routledge. London & New York.
  80. Merakchi, K and M Rogers. 2013 “The translation of culturally bound metaphors in the genre of popular science articles: A corpus-based case study from Scientific American translated into Arabic.’ Intercultural Pragmatics journal. vol.10(2). pp.341-372.
  81. Mohammed, M and N Omar. 2011. ‘Rule based shallow parser for Arabic language.’ Journal of Computer Science. vol.7(10). pp.1505-1514.
  82. Mohit, B. A Rozovskaya, N Habash, W Zaghouani and O Obeid. 2014. The First QALB Shared Task on Automatic Text Correction for Arabic. In the proceedings of the EMNLP 2014 Workshop on Arabic Natural Language. 25 October 2014. Doha. Qatar.
  83. Panju, M. 2014. ‘Statistical Extraction and Visualization of Topics in the Quran Corpus.’ MMath Thesis. University of Waterloo.
  84. Rabiee, H. 2011. Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic. Proceedings of RANLP’2011 Recent Advances in Natural Language Processing.

85. Roberts. A. L Al-Sulaiti. and E Atwell. 2005. 'aConCorde: towards a proper concordance of Arabic.' Proceedings of CL'2005 Corpus Linguistics.
86. Roberts. A. L Al-Sulaiti. and E Atwell. 2006 'aConCorde: Towards an open-source, extendable concordancer for Arabic.' Corpora journal. vol. 1. pp. 39-57.
87. Sawalha. M and E Atwell. 2008. 'Comparative evaluation of Arabic language morphological analysers and stemmers.' Proceedings of COLING'2008 Computational Linguistics.
88. Sawalha. M and E Atwell. 2009. 'Linguistically informed and corpus informed morphological analysis of Arabic.' Proceedings of CL'2009 Corpus Linguistics.
89. Sawalha. M and E Atwell. 2010a. 'Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.
90. Sawalha. M and E Atwell. 2010b. 'Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic.' Proceedings of LREC'2010 Language Resources and Evaluation Conference.
91. Sawalha. M and E Atwell. 2011. 'Morphological analysis of classical and modern standard Arabic.' Proceedings OF ICCA'2011 International Computing Conference in Arabic.

92. Sawalha. M and E Atwell. 2012. 'Visualization of Arabic Morphology.' Proceedings of AVML'2012 Advances in Visual Methods for Linguistics.
93. Sawalha. M and E Atwell. 2013a. 'Accelerating the processing of large corpora: using grid computing for lemmatizing the 176 million words Arabic Internet Corpus.' Proceedings of WACL'2 2nd Workshop of Arabic Corpus Linguistics.
94. Sawalha. M and E Atwell. 2013b. 'A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging.' Word Structure journal. vol. 6. pp.43-99.
95. Sawalha. M and E Atwell. 2013c. 'Comparing morphological tag-sets for Arabic and English.' Proceedings of CL'2013 Corpus Linguistics.
96. Sawalha. M. C Brierley. and E Atwell. 2012a. 'Predicting phrase breaks in classical and modern standard Arabic text.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.
97. Sawalha. M. C Brierley. and E Atwell. 2012b. 'Prosody prediction for Arabic via the open-source boundary-annotated Qur'an corpus.' Journal of Speech Sciences. vol. 2. pp.175-191.
98. Sawalha. M. C Brierley. and E Atwell. 2014. 'Automatically generated. phonemic Arabic-IPA pronunciation tiers for the



boundary annotated Qur'an dataset for machine learning,' Proceedings of LRE-Rel'2: 2nd Workshop on Language Resource and Evaluation for Religious Texts.

99. Sawalha. M. E Atwell. and M Abushariah. 2013. 'SALMA: Standard Arabic Language Morphological Analysis.' Proceedings ICCSPA'2013 International Conference on Communications. Signal Processing. and their Applications. pp.1-6.
100. Sharaf. A and E Atwell. 2009. 'A Corpus-based Computational Model for Knowledge Representation of the Quran'. Proceedings of CL'2009 Corpus Linguistics.
101. Sharaf. A and E Atwell. 2012a. 'QurAna: Corpus of the Quran annotated with Pronominal Anaphora.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.
102. Sharaf. A and E Atwell. 2012b. 'QurSim: A corpus for evaluation of relatedness in short texts.' Proceedings of LREC'2012 Language Resources and Evaluation Conference.
103. Sharoff. S. 2006. 'Open-source corpora: using the net to fish for linguistic data.' International Journal of Corpus Linguistics 11 (4). pp. 435-62.
104. Tabrizi. A. and R Mahmud. 2013. 'Issues of coherence analysis on English translations of Quran.' Proceedings of ICC-

- SPA'2013 International Conference on Communications.  
Signal Processing. and their Applications.
105. Wiechmann. D and S Fuhs. 2006. 'Concordance Software.'  
Corpus Linguistics and Linguistics Theory journal. vol.2.  
pp109-130
106. Yusof. R. R Zainuddin. M Baba. and Z Yusoff. 2010.  
'Quranic words stemming.' Arabian Journal for Science and  
Engineering. vol.35(2). pp.37-49.



## قواعد البيانات الإلكترونية للمخطوطات التراثية العربية والإسلامية؛ الحاضر والمستقبل

د. سامح عويضة (\*)

s.awaida@qu.edu.sa

---

(\*) أستاذ مساعد في قسم هندسة الحاسوب في جامعة القصيم، المملكة العربية السعودية. حصل على درجة الدكتوراة في هندسة وعلوم الحاسب الآلي من جامعة الملك فهد للبترول والمعادن، والبيكالوريوس والماجستير في الهندسة الكهربائية من جامعة هارتفورد، الولايات المتحدة الأمريكية. في السابق عمل محاضراً في جامعة الملك فهد وجامعة الأميرة سمية. تشمل اهتماماته البحثية التعرف على الأنماط ومعالجة الصور والأنظمة المدمجة. بالإضافة إلى اثنتين من براءات الاختراع العالمية، نشر أكثر من خمسة عشر بحثاً علمياً في مجلات ومؤتمرات دولية.



## ملخص

إنَّ المشتغل بالتراث العربي والإسلامي يلاحظ نقصاً واسعاً في جهود رقمنة المخطوطات، وندرة في الاستفادة من البرمجيات والخوارزميات الحاسوبية ذات العلاقة بمعالجة الوثائق التراثية؛ على النقيض من مدى استفادة العلوم الإنسانية عند الغرب لهذه التقنية. يذكر الباحث في هذا المقال واقع رقمنة الوثائق التراثية من قواعد البيانات الإلكترونية لفهارس المخطوطات، وقواعد المخطوطات الرقمية، مع ذكر أبرز المزايا والعيوب. كما يقترح المواصفات المناسبة لبناء قاعدة بيانات متنوعة الصور للمخطوطات العربية والإسلامية المسوَّحة ضوئياً، مع تحديد المتطلبات التي تمكن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية وعلوم الحاسب الآلي معاً. تمَّ تنظيم قاعدة البيانات الإلكترونية المقترحة للمخطوطات التراثية لتقوم وفق مهام محدَّدة مسبقاً تساعد في تطوير البرمجيات الخاصَّة بباحثي الحاسب الآلي، وتمكينهم من خدمة التراث العربي والإسلامي؛ والتي منها - على سبيل المثال - التعرف على الكتابة آلياً، والبحث عن صور الكلمات دون تعرف، وتحليل صور الوثائق، والتعرف على الناسخ والتحقق منه، وغيرها من مستجدَّات البحث العلمي.



## ١. مقدمة

إن العناية بتراث علمائنا السابقين، بحفظ أصول كتبهم الخطية من الواجبات الكفائية، ويشمل ذلك إيقاف طلبة العلم النبهاء عليها، واستخدام الوسائل العصرية التي توسع نطاق ذلك، وتذلل عقباته، كالبرامج الحديثة التي تعين على الوقوف على المعلومات بأسرع وقت، وعلى وجه فيه دقة [١].

ويقدر عدد المخطوطات العربية والإسلامية اليوم بثلاثة ملايين مخطوطة. تتطرق العربية منها إلى موضوعات شتى، إذ تتسع دائرة عناوينها لتشمل كافة مجالات العلوم الإنسانية تقريباً، بينما تحتوي المخطوطات الإسلامية على نصوص بعدة لغات - إضافة إلى اللغة العربية - كالفارسية والتركية العثمانية والأوزبكية وعدة لغات أخرى. هذا، وتوجد مجموعة ضخمة من المخطوطات والوثائق الإسلامية والتاريخية في المكتبات والمتاحف العالمية يجري حالياً مسحها وتخزينها رقمياً - لأغراض الحفظ وتيسير الحصول عليها من قبل الباحثين والمهتمين في جميع أنحاء العالم - من خلال المكتبات الرقمية المتوفرة على الشبكة العنكبوتية الإلكترونية (الإنترنت). لما كانت المخطوطات الأصلية بين يدي الباحثين ولا سيما مجاميع المخطوطات، كانوا على مقدرة من معرفة عمر المخطوط بنوع الورق والحبر، وتحديد اسم الناسخ الذي يكون مسجلاً في آخر المجموع غالباً، فلما توزعت رسائل المجموع بالتصوير الضوئي فقدنا أسماء الناسخ في عدد من الرسائل ما عدا الأخيرة من المجموع، مما يدعو إلى الحاجة إلى استخدام التقنية في التعرف على ناسخي هذه الرسائل.

تشكل تقنيات المسح الضوئي والرقمنة للمخطوطات وسيلة وصول لكنوز التراث المخفية في مكتباتنا. وقد اهتم حافظي زهير في مقاله " دور تكنولوجيا المعلومات في حفظ المخطوطات العربية " [٢] بذكر بعض التجارب الرائدة في استخدام



النظم الآلية في معالجة المخطوطات العربية وإتاحتها، مع ذكر متطلبات رقمنة المخطوطات وخصوصياتها ومواصفاتها. ومع هذه الجهود المباركة، فإن مجرد عرض صور من المخطوطة لا يعد كافياً، لأن الصور غير قادرة على نقل العديد من الجوانب الهامة للمخطوطة؛ بما في ذلك وصف المخطوط، ووصف ورقه واكتماله، ومحتوى الصور، والكتابة والخط، وذكر تاريخ ومكان كتابة المخطوط بفهرس قابل للبحث فيه، وغيرها من المعلومات الهامة. لذا، فينبغي ألا يكون الهدف الرئيسي من رقمنة الوثائق التراثية مقتصرًا على توفير صور المخطوطات والوثائق فحسب، ولكن على توفير وصف لمحتوى المخطوطات يتميز بالدقة والشمولية، وكذلك على توفير نصوص المخطوطات لكي يتنسى للباحثين في المستقبل البحث في صور المخطوطات بسهولة ويسر، والعمل على توفير الترجمة (الآلية) لهذه النصوص لتذليل العلوم الإسلامية والعربية لسكان العالم أجمع.

نظرًا للتطورات الأخيرة في علوم الحاسب الآلي، وخاصة علم التعرف على الأنماط (Pattern Recognition)، فقد أصبح من الممكن لهذه العلوم خدمة باحثي التراث العربي والإسلامي بصورة كبيرة. نذكر تاليًا أهم هذه العلوم وطرق إفادتها لباحثي العلوم الإنسانية:

#### أ. التعرف على النصوص المكتوبة آلياً (Automatic Text Recognition) :

تقوم هذه البرامج بتحويل صور المخطوطات المسوَّحة ضوئياً إلى نص إلكتروني ممَّا يتيح البحث فيه، والتعديل عليه، وطباعته ونشره إلكترونياً ضمن عدد من العمليات المفيدة. ولا شك أن برمجيات التعرف على النص المكتوب آلياً من أكثر الأمور التي تفيد الباحثين في العلوم الإنسانية، إذ تعد بديلاً عن عملية النسخ والتفريغ اليدوي للمخطوط التي قد تستغرق الأيام أو الأشهر والسنين. وهي ميزة تفوق الوصف في تفسير علم البحث في المخطوطات، ولك أن تتصور كم في هذا من تذليل للعلوم إن توفّر البحث في قاعدة بيانات تضم عدداً ضخماً من الصور

آلياً، وخاصةً إن استطاعت هذه العملية الآلية التفوق على نتائج التفريغ اليدوي في الدقة. ومما تعتمد عليه نتائج هذه البرمجيات في نجاحها دقة نتائج برمجيات أخرى، مثل تحليل صور الوثائق ومعالجة اللغة الطبيعية.

#### ب. البحث عن صور الكلمات ( Word Spotting ) :

كثيراً ما يحتاج الباحثون في العلوم الإنسانية إلى البحث عن كلمة أو جملة محددة في المخطوط دون معرفة نص المخطوط بأكمله، أي دون أن تضطرهم لقراءة المخطوط بأكمله أو التعرف على نصه آلياً للبحث عما يحتاجون. ولتطبيق هذه العملية نستخدم صوراً للكلمة أو الجملة المراد البحث عنها ونستخدم برمجيات علم التعرف على الأنماط للبحث عنها وتحديد مكانها في المخطوط.

#### ج. تحليل صور الوثائق ( Document Image Analysis ) :

يهدف علم تحليل صور الوثائق إلى تحليل صور المخطوطات التراثية ووصف هذه الصور للمساعدة في علوم أخرى. فعلى سبيل المثال، تستطيع برمجيات تحليل صور الوثائق التعرف على مواقع الفقرات الكتابية في الصفحة وتحديد إحداثيات محيط هذه الفقرات، كما تقوم بتقسيم الفقرات إلى أسطر منفردة لتجهيزها للتعرف الآلي على الكتابة، بالإضافة إلى تحديد الرسومات (مثل الأختام أو الأشكال) في صور الصفحات مع وصفها وصفاً مفيداً، والكثير من الأمور الأخرى التي تساعد برمجيات التعرف على الأنماط.

#### د. معالجة اللغة الطبيعية ( Natural Language Processing ) :

يستخدم علم معالجة اللغة الطبيعية قواعد اللغة العربية ونتائج التحليل الإحصائي للنصوص الإسلامية في تمكين أجهزة الحاسب الآلي من فهم نصوص الوثائق التراثية واستكشاف العلاقات والأنماط المفيدة، وهو من العلوم المساعدة المهمة في علم التعرف على الأنماط. فعلى سبيل المثال، تُستخدم برمجيات معالجة اللغة الطبيعية في تحسين دقة نتائج التعرف على الكتابة آلياً.

## د. التعرف على الناسخ والتحقق منه (Writer Identification and)

(Verification) :

تعنى عملية التعرف على الناسخ أو الخاط بتحديد آلياً من مجموعة من النّسّاخ من خلال تحليل عينات من صور خطوطهم ضمن قاعدة بيانات لخطوط النّسّاخ. كما تعني عملية التحقق من الناسخ التأكيد الآلي لصحة نسبة المخطوط لناسخ معيّن من خلال مقارنتها مع عينات سابقة لصور خطّه. ممّا قد يعطي قيمة علمية للمخطوط كأن يكون الناسخ ممن عُرِف بفن معين فيسهل على الباحث التعامل مع المخطوط في فهم النص وتوجيه عبارته من خلال ما عُرِف من منهج الناسخ في منسوخاته أو منهج المؤلف في مؤلفاته السابقة، وتزيد أهمية هذا البند في الاهتمام لاسم الناسخ أو المؤلف فيما عُرِف عليه من مخطوطات ناقصة الأول والآخر؛ فيتم التعرف عليه من خلال وجود قاعدة ضخمة من صور المخطوطات - التي عرف اسم ناسخها أو مؤلفها - تعرض على إثر بعضها فيظهر اسم الناسخ أو المؤلف، وقد يهتم هذا العلم - بدلاً من تحديد اسم الناسخ - بتحديد عصر أو قرن المخطوط، أو نوع الخط المستخدم في الكتابة، ممّا يساعد في تثمين قيمة المخطوطات الأثرية وكشف المزيف منها.

إنّ أحدث التقنيات حالياً لا يمكنها التعامل مع المخطوطات والوثائق التاريخية في البحث عن النصوص، أو معرفة المحتوى، فضلاً عن ترجمته بصورة ناجحة. لذا فلا بدّ من بناء قاعدة بيانات رقمية ضخمة للباحثين تحتوي على الوصف المناسب لكل مخطوط بالإضافة إلى نصوص المخطوط، مع تجهيز الأدوات المناسبة للباحثين في مجال علوم الحاسب الآلي والتي تساعدهم في بناء البرمجيات الخاصة بالبحث والتحليل والترجمة والتعرف على الأنماط آلياً.

نقوم بتقسيم ما تبقى من هذا المقال على النحو التالي؛ نناقش في الفصل الثاني واقع فهارس المخطوطات الرقمية مع ذكر أبرز المؤلفات والمواقع المعنية

بفهارس المخطوطات العربية والإسلامية، كما نخصص الفصل الثالث لذكر جهود الباحثين والمؤسسات في قواعد بيانات المخطوطات الرقمية، مع إيـراد أبرز هذه القواعد وميزاتها وعيوبها، ونذكر في الفصل الرابع الجهود المبذولة في محاولة جمع قواعد بيانات خاصة بخطوط العلماء السابقين وناسخي المخطوطات، ثم نقدم في الفصل الخامس مقترحاً لمواصفات قاعدة بيانات مستقبلية للمخطوطات الإسلامية والعربية وعرضها بالتقنية الرقمية، مع الخيارات والميزات التي تمكن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية وعلوم الحاسب الآلي معاً، وأخيراً، يتم عرض النتائج والتوصيات في الخاتمة في الفصل السادس.

## ٢ - فهارس المخطوطات الرقمية

يُعدُّ فن فهرسة المخطوطات علماً قائماً بذاته، إذ يعنى بضبط المخطوطات وتوثيقها، ثم تسجيلها في قوائم، أو بطاقات، أو كتب لحفظها. فيقوم المفهرس بوضع كل ما يتعلق بالمخطوط من معلومات في هذه الفهارس، كعنوان المخطوط، واسم المؤلف، واسم الناسخ، وتاريخ النسخ، وعدد الأوراق، وعدد السطور في كل صفحة، وحجم الورق، ونوع الورق، ونوع الحبر، وحال المخطوط، كما أنه يضع اقتباساً من بداية المخطوط ونهايته، وغير ذلك من أوصاف يراها المفهرس، وهي تختلف من مكتبة إلى أخرى. وقد ذكر صلاح الدين المنجد في كتابه "قواعد فهرسة المخطوطات العربية" [٣] تاريخ فهرسة المخطوطات العربية والإسلامية، بالإضافة إلى بعض القواعد والأساليب المستخدمة في فهرسة المخطوطات.

ومع عدم توفر مصادر رسمية لعدد فهارس المخطوطات في العالم، إلا أنه يمكننا تقدير عددها بالآلاف، بل قد يصل إلى عشرات الآلاف؛ بدليل أن مؤسسة الفرقان جمعت في كتابها "المخطوطات الإسلامية في العالم" ما يزيد على خمسة عشر ألف فهرس للمخطوطات الإسلامية [٤]. كذلك الفهارس التي جمعتها مؤسسة آل البيت للفكر الإسلامي في فهرسها الشامل وصلت إلى ألف وستمائة

فهرسٍ للمخطوطات [٥] مع أنها لم تستوعب جميع العلوم الإسلامية. هذا بالإضافة إلى العديد من المكتبات الخاصة والشخصية التي لم يوقف على فهارس مخطوطاتها.

### **جهود الباحثين المنشورة في جمع فهارس المخطوطات العربية والإسلامية**

تزخر المكتبات بمخطوطات التراث العربي والإسلامي في مختلف بقاع العالم، لذا فإنك لا تكاد تزور دولة من هذه الدول إلا وتجد فيها فهرسًا أو أكثر لمخطوطات مكتباتها، إمّا باللغة العربية وإمّا بلغة الدولة الأمّ. وقد قام الباحثون على مدى التاريخ بعدة محاولات لجمع هذه الفهارس؛ وذلك للتعريف بها وتسهيل عمل الباحثين والمهتمين بهذا العلم، نذكر على سبيل المثال كتاب "تاريخ الأدب العربي" [٦]، لمؤلفه المستشرق الألماني كارل بروكلمان، الذي صدر الجزء الأساسي منه في مجلدين عامي ١٨٩٨ و ١٩٠٢ ثم تلتها المجلدات الثلاثة الكبيرة أعوام: ١٩٣٧ و ١٩٣٨ و ١٩٤٢ م. وقد يتصور الباحث للوهلة الأولى أنّ هذه الموسوعة كتاب في تاريخ الأدب، بينما هي في الحقيقة سجل للمصنفات العربية المخطوط منها والمطبوع. ومن هذه الجهود أيضًا كتاب "تاريخ آداب اللغة العربية" لجورجي زيدان [٧]، و "تاريخ التراث العربي" لفؤاد سزكين [٨] الذي طبع باللغة الألمانية، ثم تولت جامعة الإمام محمد بن سعود ترجمته إلى العربية، وصدر منه عشرة أجزاء. كما قام السيد رزق الطويل في كتابه "مقدمة في أصول البحث العلمي وتحقيق التراث" [٩] بتكريس قسم منه لذكر فهارس المخطوطات الإسلامية. وبالجملّة، فإنّ المؤلفات في هذا الباب كثيرة، وقد تصل إلى خمسين مؤلفًا وزيادة، نذكر في العناوين التالية أبرزها.

### **فهارس "معهد المخطوطات العربية"**

يُعدّ معهد المخطوطات العربية واحدًا من أقدم المراكز المعنية بالمخطوطات في الوطن العربي. تأسس المعهد سنة ١٩٤٦ وهو ملحق بالمنظمة العربية للتربية

والثقافة والعلوم مع تمتعه بشخصية معنوية مستقلة. يُعنى المعهد بالتراث العربي المخطوط بمختلف أصعده؛ جمعاً، وإتاحة، وصيانة، وترميمًا، وفهرسةً، وتعريفًا، ودراسةً، وتوظيفًا. للمعهد عدد من الإصدارات المهمة بفهارس المخطوطات العربية، نذكر منها "فهارس المخطوطات العربية في العالم"، و"دليل مكاتب المخطوطات في الوطن العربي"، وعدد من فهارس المخطوطات في دول العالم مثل "فهرس مخطوطات جامعة الإسكندرية" وغيرها، بالإضافة إلى مؤلفات فيصل الحفيان - مدير معهد المخطوطات العربية -.

ولا شك أنّ من أهمّ مؤلفات المعهد "فهرس المخطوطات المصورة" [١٠]، وهو فهرس للمخطوطات المصورة في معهد المخطوطات العربية ومقسّم حسب الفن من تاريخ، وعلوم، وعارف عامّة، وسيرة نبوية، وأدب، وفقه، وحديث، وتفسير. وقدّ تمت طباعة هذا الفهرس على عدّة مراحل زمنية تجاوزت خمسة عقود من الزمن، إذ تمّ طباعة المجلد الأول في عام ١٩٥٤ م، وثمّ طبع آخر مجلد عام ٢٠١١ م. وممّا يعيب مطبوعات المركز الانقطاع الطويل في أعمالهم، وعدم الاستمرارية في الجهود المبذولة. كما أنّ المعهد من أوائل الذين تبنّوا سياسة رقمنة المخطوطات والفهارس منذ عشرات السنين، ولكننا لا نجد لهم برنامجًا متميزًا من إصداراتهم في هذا الباب حتى يومنا هذا.

وللفائدة، فقد نشر المعهد موقعًا باسم "خزانة المخطوطات" [١١]، يذكرون فيه القدرة على البحث الرقمي عن المخطوطات التابعة للمعهد، وهو قيد التعديل والتجريب منذ فترة، ونتائج غير مرضية. كما أنّ موقع المعهد على الإنترنت مليء بالروابط التي لا تعمل والمعلومات الناقصة.

### "المخطوطات الإسلامية في العالم" مؤسسة الفرقان

قامت مؤسسة الفرقان للتراث الإسلامي [١٢] بعمل مسح وإحصاء شامل لمجموعات المخطوطات الإسلامية في أنحاء العالم شملت فهارس مئة وستة بلدان،

ثم نشرت نتائج هذه الدراسة في كتاب "المخطوطات الإسلامية في العالم" بنسخته الإنجليزية [١٣] والعربية [٤]. وتذكر الموسوعة دراسات حول عدد المخطوطات، واللغات التي كتبت بها، كاللغة العربية والفارسية والتركية والأوردوية والسواحيلية؛ وأماكن تواجد تلك المخطوطات، كما تناولت الدراسة المجموعات التي فهرست، وعناوين الفهارس التي تم نشرها، والتي لم تصدر بعد في تلك المكتبات، إضافة إلى ذكر بعض عناوين نواذر المخطوطات. وقد ذكرت مؤسسة الفرقان أنها قامت بجمع هذا الكتاب مما يزيد على خمسة عشر ألف فهرس للمخطوطات الإسلامية. مع الإشارة إلى أن كتاب "المخطوطات الإسلامية في العالم" هو أقرب لموسوعة تقوم بالتعريف بمكتبات العالم التي تحتفظ بفهارس للمخطوطات الإسلامية مع التعريف بهذه الفهارس، من كونها مجموعاً لفهارس المخطوطات كما هو حال "فهارس آل البيت" التي نناقشها تالياً - مثلاً -.

### فهارس آل البيت

ومن المشاريع المؤسسية المهمة بحصر المخطوطات العربية الإسلامية استناداً على الفهارس المطبوعة والمخطوطة للمكتبات العامة والخاصة، مشروع "الفهرس الشامل للتراث العربي الإسلامي المخطوط" الذي سبقت الإشارة إليه، والمعروف باسم "فهارس آل البيت" [٥] إذ قامت بجمعه مؤسسة آل البيت للفكر الإسلامي، فخرج في خمسة وعشرين مجلد. وقد اعتمدت المؤسسة في تحقيق هذا العمل على ما جمعته من الفهارس والمصادر والدراسات المختصة التي جاوزت الألف وستمئة فهرس للمخطوطات في سبع عشرة لغة. وقد قامت المؤسسة بطباعة المشروع على مراحل عدة حسب الموضوع، فطبعت فهرس مخطوطات التفسير وعلومه في مجلدين، وفهرس المصاحف المخطوطة ومخطوطات رسم المصحف في مجلد، وفهرس مخطوطات الحديث النبوي الشريف وعلومه ورجاله في ثلاثة مجلدات، وفهرس المخطوطات العربية في مكتبة تشستريتي في إيرلندا في ثلاثة مجلدات، وفهرس مخطوطات التجويد في مجلد، وفهرس مخطوطات القراءات في مجلد،

وفهرس مخطوطات السيرة والمدائح النبوية في مجلدين، وفهارس مخطوطات الفقه وأصوله في اثنتي عشرة مجلد. ومع ضخامة عمل مؤسسة آل البيت في فهارسها، إلاَّ إنَّه لا يزال ناقصاً فلم يشمل جميع العلوم بعد، بل إنَّه متوقف حالياً كما علمتُ من المؤسسة نفسها، وآخر هذه المجلدات من الفهرس طبع قبل عشر سنين في عام ٢٠٠٤ م. ومن أبرز المآخذ على هذا الفهرس أنَّه لم يعتمد على مشاهدة المخطوطات التي تمت فهرستها؛ بل اقتصر على النقل من الفهارس الأخرى؛ ممَّا أدى إلى استمرار بعض الأخطاء كنسبة بعض المخطوطات إلى غير أصحابها - مثلاً - اعتماداً على ما ذكر في الفهارس الأخرى.

وعلى الرغم من عدم فهرسة بعض المكتبات لمخطوطاتها أو عدم نشر هذه الفهارس، فإنَّ الكثير من مكتبات العالم التي تحوي مخطوطات عربية وإسلامية صنعت فهرساً ورقياً مطبوعاً لهذه المخطوطات، لكنَّ ظاهرة رقمنة فهارس المخطوطات لا تزال محدودة جداً؛ فالمكتبات ما زالت تعتمد على طباعة الفهارس ورقياً، مما يجبر الباحثين على البحث اليدوي في هذه الفهارس، فيستهلك منهم الكثير من الوقت والجهد، عدا عن ندرة بعض هذه الفهارس المطبوعة لقدمها أو نفاد المطبوع منها. ولذا قام عدد من المؤسسات بمحاولة جمع هذه الفهارس في مطبوع موحد يسهِّل على المهتمين عملية البحث، وسنحاول في هذا القسم التعريف بأشهر فهارس المخطوطات الرقمية. وبالرغم من نشر هذه المجاميع لفهارس المخطوطات، فإنَّ خروج غالب هذه المحاولات بصورة ورقية لا يزال يشكل عائقاً في عملية البحث العلمي.

### الجمع اليدوي لفهارس المخطوطات

من جهود رقمنة فهارس المخطوطات، قيام بعض المؤسسات والمواقع المهمة بالتراث العربي والإسلامي بجمع فهارس المخطوطات المطبوعة ومسحها ضوئياً ووضعها على صورة ملفات (PDF أو Word) على الشبكة الإلكترونية. ومع أنَّ



غالب هذه الملفات هي صور لصفحات هذه الفهارس، مما يعني عدم القدرة على البحث إلكترونياً فيها، إلا إنَّ بعضها قد تمّ تفرّيفه إلى نصّ إلكتروني، إمّا آلياً أو يدوياً أو آلياً قد تمّت مراجعته يدوياً. إنَّ هذه الخطوات - وإن كانت أولية في رقمنة الفهارس - إلا أنَّها مهمّة، خاصة في الفهارس المطبوعة القديمة وضرورية لرقمنة فهارس المخطوطات كلياً، ولن تكون تامة حتى يتم تفرّيفها كلها، وتدقيقها، ووضعها في قواعد بيانات منظمة. نذكر على سبيل المثال فهارس المخطوطات في الموقع الإلكتروني لـ "مركز ودود للفهارس وكتب التحقيق" [٢٤]، وفهارس المخطوطات المجموعة في برنامج "المكتبة الشاملة" [١٤]، وموقع "جامع المخطوطات الإسلامية" [١٥]، وبعض المنتديات الإلكترونية المهتمة بالعلوم الشرعية [١٦].

### قواعد فهارس المخطوطات الرقمية

وبعد أن وقفنا على أهم الجهود في جمع فهارس المخطوطات الإسلامية المطبوعة وتصويرها ضوئياً، نذكر هنا قواعد فهارس المخطوطات الرقمية، وهي خطوة أساسية في رقمنة المخطوطات، إذ تسمح هذه القواعد بالبحث في حقول بطاقة كل مخطوط على حدة، مثل اسم المؤلف وعنوان المخطوط وسنة النشر وغيرها من المعلومات، إلا إنَّ هذه الفهارس تقتصر عادة إلى خدمة تصفح صور المخطوطات أو تحميلها.

### مكتبة الفرقان الرقمية

من هذه القواعد نجد مكتبة الفرقان الرقمية [١٧] إذ يسمح الموقع الإلكتروني بالبحث في المخطوطات. إلا إنَّ واجهة البحث حالياً تعمل باللغة الإنجليزية (يذكر الموقع أن الواجهة العربية تحت التطوير) ولا شك أنَّ الموقع بحاجة إلى إضافة الكثير من المخطوطات ليصبح مرجعاً معتبراً في البحث الرقمي في فهارس المخطوطات، إذ يذكر الموقع أنه يحوي فهارس ٢٧٠٥٥ مخطوطاً، وهذا العدد في ازدياد. يحوي الموقع الكثير من الخيارات التي تمكن الباحثين من البحث حسب المؤلف، والموضوع،

والمكتبة، ولغة المخطوط، وغيرها من الخيارات المفيدة. كما تظهر نتائج البحث معلومات المخطوط بطريقة منظمة، مع فرز كل جزء من بطاقة المخطوط (العنوان، اسم المؤلف، اسم الناسخ، ...) بحقل منفرد، مع إمكانية ربط بعض هذه الحقول إلكترونياً، كما يظهر في رسم توضيحي ١.

Manuscript Details		Read Catalogue Introduction
Country:	Algeria	
City:	Bejaia	
Library:	Private library of Shaikh Lmuhub Ulahbib	
Catalogue:	Catalogue of Islamic Manuscripts in the private library of Sheikh Lmuhub Ulahbib, Bejaia - Algeria	
Prepared by Djamel-Eddine Mechehed, Edited by Ayman Fuad Sayyid		

MANUSCRIPT DETAILS	
SP 18407	Manuscript Id
[ كتاب * في الفقه ]	Title
محمد الموهوب بن البشير بن لحيب، 1237هـ/1821م: تاريخ وفاة المؤلف الميلادي 1899.	Author
الفقه.	Subject
1298هـ/1880م: تاريخ التأليف الميلادي 1881	Date of Authorship
27	Number of folios
25	Lines
17cm x 21cm	Size
11cm x 17cm	Written Surface
Maghribi	Script Type
أسود، الأحمر	Ink colour
Arabic	Language
<p>La bibliothèque savante de cheikh al Muhub ; lettrés locaux et culture écrite au kabyli au milieu du XIX<sup>ème</sup> siècle. A paraître ( publisud, Paris) les manuscrits de botanique et médecine en kabyli au XIX<sup>ème</sup> siècle, ANNALI, instituto universitario orientale, fascicolo 1-4. Napoli 1999.</p> <p>ص 78-98: ANNALI, les manuscrits de botanique et médecine en kabyli au XIX<sup>ème</sup> siècle, Napoli 1999.</p> <p>ج 59, ص 69-92.</p>	
Condition	نسيئة، وبها نثر من أولها وآخرها.
Notes & Comments	مكان التأليف : قرية تالة ورزار أيت ورتيلان - منطقة القبائل - الجزائر. التعاون بالأحمر.
Library	الشيخ الموهوب أولحيب الخاصة/بجاية
Classmark	F0010

رسم توضيحي ١ : صورة توضيحية لنظام البحث في مكتبة الفرقان الرقمية.

## موقع يوسف زيدان الإلكتروني

يقوم موقع يوسف زيدان الإلكتروني [١٨] بتوفير قاعدة بيانات لفهارس المخطوطات ، ويتيح للمتصفح البحث فيها، ويحوي هذا الموقع مجموعة قليلة من فهارس المخطوطات العربية والإسلامية، وهي: فهرس مكتبة رفاعة الطهطاوي (قراية ١٨٠٠ مخطوطة) ، وفهرس المخطوطات العلمية بمكتبة بلدية الإسكندرية (٤٧٠ مخطوطة) ، وفهرس مخطوطات شبين الكوم (٣٥٠ مخطوطة) ، وفهرس مخطوطات الفقه بمسجد أبي العباس المرسى (٨٠٠ مخطوطة) ، وفهرس مخطوطات المعهد الديني بسُموحة (٢٠٤ مخطوطة) ، وفهرس مخطوطات دار الكتب بطنطا (٤٤٧ مخطوطة). ويظهر رسم توضيحي ٢ نظام البحث في فهارس المخطوطات في موقع يوسف زيدان.

### فهارس المكتبات الخطية

#### بحث متقدم

##### إشارة:

يشتمل هذا الباب من الموقع ، على فهارس وصفية كاملة لمكتبات خطية مهمة . ويمكن البحث في محتوى هذا الباب ، بكلمة من عنوان المخطوطة ، أو كلمة من اسم مؤلفها ، أو كلمة من بدايتها أو موضوع المخطوطة أو رقم حفظها.

جميع الفهارس	حدد الفهرس
<input type="text"/>	العنوان
<input type="text"/>	المؤلف
<input type="text"/>	الموضوع
<input type="text"/>	البداية
<input type="text"/>	رقم الحفظ
<input type="button" value="بحث"/>	

رسم توضيحي ٢ : صورة توضيحية لنظام البحث في موقع يوسف زيدان.

## "خزانة التراث" لمركز الملك فيصل

كما قام مركز الملك فيصل للبحوث والدراسات الإسلامية بالرياض في العام ١٩٩٢ م بإنشاء قاعدة بيانات باسم "خزانة التراث" [١٩] تشتمل على فهرس المخطوطات الإسلامية في جميع المكتبات والخزانات ومراكز المخطوطات في العالم، حتى بلغ عدد عناوين هذه المخطوطات أكثر من مائة وعشرين ألف. تتضمن خزانة التراث فهرس مخطوطات عربية وإسلامية من مختلف أنحاء العالم، بالإضافة إلى فهرس لمخطوطات مكتوبة بلغات أجنبية ترجمها المركز إلى العربية. كما يوفر البرنامج مداخل بحث متعددة يستطيع الباحث من خلالها الوصول إلى المعلومة التي يريدها بكل يسر وسهولة، مثل البحث في الفن، وفي عنوان المخطوط، والعنوان الفرعي، وفي اسم المؤلف، واسم الشهرة، وتاريخ الوفاة، وقرن الوفاة، واسم المكتبة، واسم الدولة، واسم المدينة، ورقم الحفظ. ويظهر رسم توضيحي ٣ عدّة صور لشاشات برنامج "خزانة التراث". ومن أبرز عيوب برنامج "خزانة التراث" - مع تميزه في مجاله - أنّه لا يعمل إلا على نظام تشغيل الويندوز فقط، وأنّه لم يتم تحديثه منذ سنين، مما يحدّ من استخدام البرنامج على أكثر الأجهزة الحديثة.



رسم توضيحي ٣: عدّة صور توضيحية لشاشات برنامج "خزانة التراث".

### "خزانة المآجد للتراث" لمركز جمعة الماجد

ومن البرامج المميزة في فهرس المخطوطات برنامج "خزانة المآجد للتراث" [٢٠] المصمم من قبل شركة الدار العربية لتقنية المعلومات للتعريف بمخطوطات مركز جمعة الماجد للثقافة والتراث والبحث فيها. يحوي البرنامج أشمل قاعدة بيانات إلكترونية لمخطوطات مركز جمعة الماجد. تذكر بطاقة المخطوط عنوان المخطوطة، والعنوان البديل، والمؤلف، والناسخ، وبدايتها ونهايتها، والملاحظات المتعلقة بالمخطوطة، وعدد الأوراق، واللغة، ويظهر رسم توضيحي ٤ مثلاً لإحدى بطاقات المخطوطات في البرنامج. ويقدم البرنامج خدمة البحث عن أي مخطوطة وذلك بالإشارة إلى ما يراد البحث فيه من العنوان أو العنوان البديل أو بداية المخطوط أو نهايته أو عن طريق مؤلفه أو ناسخه أو جميع ما سبق، مما يتيح للباحث

سرعة الوصول للمخطوطة المطلوبة. إلا إنَّ هذا البرنامج كسابقه -برنامج "خزانة التراث" لمركز الملك فيصل - يعتمد على نظام تشغيل معيَّن بدلاً من جعله على الشبكة العنكبوتية العالمية ممَّا يحدُّ فعلياً من فائدته؛ لغياب التحديثات له.

243171	رقم المادة
تاريخ النسخة وتاريخ النسخ	التاريخ الرئيسي
البقي: نعم، إن وصلنا من عصر الكائن العنكبوتي الشافعي سراج الدين وأبو	الترجمة
800	رقم الأوعية
العربية	اللغة / لغة المادة

رسم توضيحي ٤ : صورة توضيحية لبطاقة مخطوط في برنامج "خزانة المآجد للتراث".

### فهارس مخطوطات مكتبة المسجد النبوي

ومن الجهود الحديثة في رقمنة فهارس المخطوطات جهود الرئاسة العامة لشؤون المسجد الحرام والمسجد النبوي، إذ قاموا برقمنة فهارس مخطوطات مكتبة المسجد النبوي وتوفيرها للباحثين على شبكة الإنترنت [٢١]. تقوم المؤسسة دورياً بإضافة بطاقات للمخطوطات، وفي زيارة حديثة للموقع تم حصر فهارس ٦٢٢٠ مخطوطاً. يسمح الموقع بالبحث في الفن، وفي عنوان المخطوط، وفي أول المخطوط وآخره، أو في جميع بيانات المخطوط السابقة. ويظهر رسم توضيحي ٥ مثلاً للبحث في موقع مكتبة المسجد النبوي. يفتقد الموقع إلى تنوع مجالات البحث، كما تفتقد بطاقات المخطوطات إلى الشمول، وهناك نقص في كثير منها. إلا إنَّ الاستمرار في تحديث هذا العمل وتحسينه قد يجعله من المراجع الرئيسة في فهارس المخطوطات الرقمية في المستقبل.



### ٣. قواعد المخطوطات الرقمية

اختلف الباحثون في تقدير عدد فهارس المخطوطات العربية والإسلامية، كما اختلفوا في تقدير عدد المخطوطات نفسها. ويرجع هذا الاختلاف إلى عدة أسباب، أهمها عدم وجود مؤسسة رسمية تعنى بإحصاء دقيق ودوري لهذه المخطوطات. كذلك من الأسباب عدم الاتفاق في تحديد ما يراد إحصاؤه، فهل يراد حصر المخطوطات الإسلامية بلغاتها المختلفة أو إدخال المخطوطات العربية بعلمها كافة؟ وأيضاً الاختلاف في تحديد القرون المراد إحصاء مخطوطاتها، ووجود عدد من المخطوطات ذات الملكية الخاصة أو المخطوطات المفقودة أو المسروقة التي لا يمكن إحصاؤها إلى غيرها من الأسباب التي تجعل التفاوت -حتى في الأرقام التقديرية للمخطوطات- كبيراً بين المصادر المختلفة. فمنهم من يقدّرها بمليون مخطوطة، أو ثلاثة ملايين، ومنهم من يوصل عددها إلى خمسة ملايين مخطوطة، مع شبه اجماعهم على أن عدد المخطوطات العربية والإسلامية بالملايين. ويذكر الباحثون أن أكثر هذه المخطوطات في تركيا (٢٠٠,٠٠٠ إلى ٣٠٠,٠٠٠)، كما يقدّرون أن إيران والهند واليمن يملكون مئات الآلاف من المخطوطات كذلك [٢٣]. وفي الدول العربية، فقد انتشر مؤخراً في الجامعات والمؤسسات التعليمية الاهتمام بجمع المخطوطات الأصلية وفهرستها، فعلى سبيل المثال تقدّر عدد المخطوطات في جامعة الملك سعود في المملكة العربية السعودية بأكثر من ٣٧ ألف مخطوطة أصلية ومصوّرة. أمّا بشأن المخطوطات المصوّرة، فقد بدأت تظهر جهود أفراد يملكون عشرات أو مئات الألوف من المخطوطات.

### الجمع اليدوي للمخطوطات

يقوم بعض الباحثين والمهتمين بالمخطوطات بالمسح الضوئي لهذه المخطوطات وجمع الصور لها في مواقع على الإنترنت. نذكر على سبيل المثال منها موقع "جامع المخطوطات الإسلامية" [١٥]، وموقع "مركز ودود للفهارس وكتب التحقيق" [٢٤]، وقسم المخطوطات في موقع "ملتقى أهل الحديث" [٢٥]، وقسم



المخطوطات في موقع "مكتبة الألوكة" [٢٦]. إلا أن هذه المخطوطات غير مفهرسة، ولا تتيح للباحث البحث في حقول معينة فيها، كما أن الكثير من هذه المواقع لا يملك وصفًا مفصلاً للمخطوطة، وليست موضوعة في قاعدة بيانات إلكترونية؛ مما يقلل من فائدة هذه المخطوطات في البحث العلمي.

### جهود الباحثين في علوم الحاسب الآلي في رقمنة المخطوطات

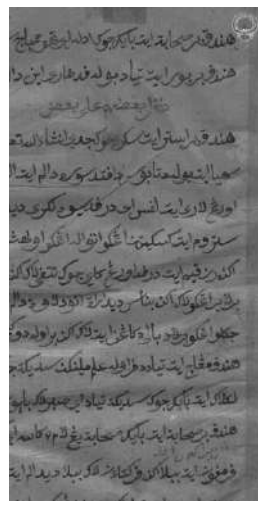
من اهتمامات الباحثين في علوم الحاسب الآلي تطوير البرمجيات المختلفة المعنية بخدمة المخطوطات العربية والإسلامية. إذ من الممكن استخدام قاعدة البيانات في مجال التعرف الآلي على نصوص المخطوطات، وفي تحليل ومعالجة صور صفحات المخطوطات آلياً، والتعرف الآلي على مواقع الكتابة والرسومات والصور والجداول وغيرها وتقسيم صور المخطوطات آلياً إلى مناطق أو أجزاء وتوصيف هذه الأجزاء، كما يقومون بتطوير أساليب البحث الآلي النصي في صور المخطوطات التي تم مسحها ضوئياً، والتعرف على كلمات صور المخطوطات آلياً، بالإضافة إلى تطوير أساليب التعرف الآلي على خط الناسخ، والتأكد منه في المخطوطات العربية والإسلامية. إلا أن تطبيق هذه العلوم بحاجة إلى قاعدة بيانات للمخطوطات بمواصفات خاصة ستم مناقشتها في الفصل الخامس. لذا قام بعض الباحثين في علوم الحاسب الآلي بمحاولات فردية لبناء هذه القواعد، نذكر في الفقرات القادمة أمثلة عليها، ولكن من الملاحظ أن هذه الجهود عادة ما تخدم جانباً ما في علوم الحاسب الآلي مع عجز في خدمة جوانب أخرى، كما أن الكثير منها يحوي عدداً قليلاً جداً من المخطوطات، بل إن الكثير منها لا يتجاوز جهده المخطوط الواحد فقط.

من الأمثلة على هذه الجهود والمحاولات ما قام به قسم علوم الحاسب الآلي وتقنية المعلومات في جامعة مالاي [٣٠] في ماليزيا إذ جمعوا ١٦٩ مخطوطة من المخطوطات التاريخية والدينية في اللغة الماليزية ولغات الدول المحيطة بها.

بالإضافة إلى توفير قاعدة البيانات لهذه المخطوطات على الإنترنت، إذ قام الباحثون بتوفير البيانات الوصفية المناسبة لكل مخطوطة باستخدام مبادرة ترميز النصوص (TEI) [٢٧] التي ستم مناقشتها بتفصيل في الفصل الخامس.

أمّا حسين نفشي وزملاؤه في مختبر سينكروميديا التابع لجامعة كيوييد في كندا [٢٨] فقد جمعوا ١٥ صورة من صور المخطوطات التاريخية والكتب التراثية الفارسية كانت في خزانة السجلات التاريخية والوثائق والمخطوطات القديمة لميرزا محمد كاظميني في يزد، إيران. ويظهر في الصور المجموعة آثار الرطوبة والطمس والتلاشي ونزيف الحبر نظراً لتقدم الزمان. ومع قلة عدد الصور في قاعدة البيانات فإنّ الباحثين يأملون أن تكون هي النواة الأولى في سلسلة من قواعد البيانات التي توفر صور المخطوطات؛ للمساهمة في علم تحليل الصور والوثائق الإلكترونية والتعرف على نص الكتابة والكاتب.

كما قام الباحثون رضا مقدّم وزملاؤه في مختبر سينكروميديا وجامعة ماكغيل في كندا بتقديم قاعدة بياناتهم المسماة "ابن سينا" [٢٩]، وأدخلوا فيها مخطوط "كشف التمويهات في شرح التنبيهات" من تأليف سيف الدين أبي الحسن علي بن أبي علي بن محمد بن سالم الأمدي (ت ٦٣١هـ) وهي تحوي ردّاً وتعليقاً على شرح الرازي لكتاب الإشارات لابن سينا. تتكون قاعدة البيانات هذه من ٥١ صفحة، قام الباحثون باستخراج ٢٠٧٢٢ جزءاً من كلمة عربية منها. تمّ التعرف على نص المخطوط آلياً والتحقق من نتائج التعرف يدوياً. يظهر رسم توضيحي ٦ أمثلة على جهود الباحثين في علوم الحاسب الآلي المذكورة سابقاً في رقمنة المخطوطات.



(ج) رضا مقدّم وزملاؤه  
٢٩

(ب) حسين نقشي  
وزملاؤه ٢٨

(أ) جامعة مالايا  
٣٠

رسم توضيحي ٦ : صور لبعض المخطوطات من قواعد بيانات رقمية من قبل باحثي في علوم الحاسب الآلي.

### جهود المؤسسات والجامعات في رقمنة المخطوطات

تتابعت في الآونة الأخيرة عدد من المؤسسات والجامعات على بناء قواعد بيانات رقمية ضخمة للمخطوطات العربية والإسلامية تميّزت باحتوائها على مئات أو آلاف المخطوطات، مع توصيف لهذه المخطوطات، وإمكانية البحث في الحقول المختلفة، وتصفّح صور المخطوطات. ورغم أهمية هذه الخصائص في قواعد البيانات الرقمية، إلّا أنّ الملاحظ أنّ هذه الجهود موجهة في الأصل لباحثي العلوم الشرعية مع عدم أخذ باحثي علوم الحاسب الآلي بعين الاعتبار. فمما يعيب هذه القواعد عدم اكتمال وصف كثير من المخطوطات أو عدم تدقيق ذلك الوصف؛ فقد لا تحتوي على اسم الناسخ مثلاً، أو على حقل خاص بنص الصفحة، أو توصيف لأجزاء الصفحات المختلفة. كما إنّ كثيراً من هذه المواقع لا تسمح

بتنزيل المخطوط كاملاً؛ فكيف بتنزيل مجموعة ضخمة من المخطوطات ( المئات أو الآلاف ) التي يحتاجها باحث علوم الحاسب الآلي في برمجياته؟ كما إنَّ كثيراً من صور المخطوطات معروضة بدقة منخفضة، مقارنة بالمعايير المتبعة في مسح المخطوطات والوثائق التاريخية، وعلى الكثير منها علامات مائية كبيرة الحجم تؤثر على صورة المخطوط، ممَّا يؤكد الحاجة إلى إعادة تصميم قواعد البيانات هذه لخدمة الباحثين في العلوم المختلفة على أفضل وجه ممكن.

وهنا تظهر مدى الحاجة في ربط تواصل بين المبرمج والباحث في العلوم الإنسانية والشرعية، وهو عمل لا يتم إلا بوجود تعاون بين مؤسسات تعنى بعلوم البحث التراثي بالتعاون مع مبرمجين أكفاء يؤمنون بسمو الهدف وثمرته الحقيقية، وهذه المؤسسات كثيرة في بلادنا العربية ولكنها تقتصر إلى تنظيم يجمع ألبتها ويجمع جهودها.

### نظام المخطوطات لجامعة الملك سعود

نذكر على سبيل المثال في هذا الباب موقع " المخطوطات " التابع لجامعة الملك سعود في المملكة العربية السعودية [٣١]. وهو نظام قامت بإنشائه جامعة الملك سعود، وهو متاح للاستخدام من قبل جميع الباحثين في أنحاء العالم. يحوي النظام أكثر من إحدى عشرة ألف مخطوطة، ويستطيع المستخدم تصفح المخطوطات، وقراءتها ومعرفة بعض المعلومات التفصيلية عنها من مؤلف وعدد صفحات وغير ذلك. ومن مميزات هذا النظام أنَّ خدمته تعمل على شبكة الإنترنت ممَّا يسمح للمستخدمين بالدخول من أي مكان، كما إنَّ واجهة المستخدم تدعم اللغتين العربية والإنجليزية.

يتميز نظام المخطوطات بعرض معلومات المخطوط، بالإضافة إلى صور المخطوط كما هو مبين في رسم توضيحي ٧، فتظهر المعلومات بعض التفاصيل عن المؤلف والكتاب بالإضافة إلى معلومات الناسخ حين توفرها. كما يسمح الموقع

للباحثين والمهتمين بتفريغ نص المخطوط وإرساله إلى نظام المخطوطات؛ لمراجعته حتى يتمكن توفير خدمة البحث في نصّ المخطوط في المستقبل. لذا يمتلك نظام المخطوطات في جامعة الملك سعود عدداً من المقومات اللازمة للباحثين في علوم الحاسب الآلي لتطوير بحوثهم في خدمة المخطوطات العربية الإسلامية، لكنّه بحاجة إلى عدد من التعديلات والتصحيحات لجعله ملائماً لهذه الأبحاث.



جامعة  
الملك سعود  
King Saud University

مخطوطات

MAKHTOTA

[الصفحة الرئيسية](#) | [استعراض المخطوطات](#) | [ابحث عن مخطوطة](#) | [احصائيات](#) | [خريطة الموقع](#) | [دليل الإستخدام](#) | [رأيك يهمنا](#)

مخطوطات > الاختيارات > الصفحة رقم 6

**عنوان المخطوطة:** الاختيارات

**رقم المصنف:** 1 2173 ل

**المؤلف:** ابن اللطام - علي بن محمد

**الرقم العام:** 329

**التاريخ المقترن باسم المؤلف:** 803 هـ

**المراجع:** معجم المؤلفين 7 : 206

**الوصف:** نسخة جيدة ، نسخ معتاد ، بها نقص

**الوصف المادي:** 121 ق ، 22 س ؛ 15 × 21 سم

**الموضوع:** المذهب الخليلي 2- فقه المذاهب الإسلامية

**الإحالات:** 1 - المؤلف ب - تاريخ النسخ

**اسم الناشر:**

**تاريخ النسخ:** 1232 هـ



أول السابق 1 6 130 التالي آخر

احصل على الصفحة بصيغة بي دي إف

هذه المخطوطات هي خدمة مقدمة من جامعة الملك سعود تهدف لخدمة الباحثين وتوفير آلية للحصول على المخطوطات النادرة بطريقة سهلة وميسرة نرجو استخدام هذه الخدمة بطريقة معتدلة ومسؤولة.

ساهم في خدمة الباحثين بتحويل الصورة إلى نص مكتوب قابل للبحث

## رسم توضيحي ٧: صورة توضيحية لنظام المخطوطات في جامعة الملك سعود.

وبالرغم من توفر هذه المقومات في نظام المخطوطات التابع لجامعة الملك سعود، إلا أنّ الوضع الحالي للنظام لا يسمح للباحثين باستخدام المخطوطات العربية والإسلامية. إذ يسمح النظام حالياً فقط بتصدير كل صفحة من المخطوط على انفراد، ولا يسمح بتصدير المخطوط بأكمله فضلاً عن تصدير مجموعة من

المخطوطات دفعة واحدة. كما إنَّ الصور المعروضة ممسوحة بدقة منخفضة (٩٦ نقطة لكل إنش) مقارنة بالمعايير المتبعة في مسح المخطوطات والوثائق التاريخية (٤٠٠ نقطة لكل إنش) ممَّا يؤثر على قابلية استخدام هذه الصور في الأبحاث العلمية، ناهيك أنَّ القائمين عليها قاموا بإضافة علامة مائية كبيرة الحجم تؤثر على الكتابة في المخطوط كما هو مبين في رسم توضيحي ٨. وممَّا يلاحظ أيضًا أنَّ المعلومات المقدمة عن المخطوط قليلة، وفي أحيان كثيرة تنقصها بعض البيانات خاصَّة اسم الناسخ.



رسم توضيحي ٨: مثال على صور إحدى المخطوطات من قاعدة بيانات جامعة الملك سعود.

### مجموعة المخطوطات الإسلامية لجامعة ميتشغان

ومن جهود الجامعات في توفير قواعد بيانات إلكترونية للمخطوطات، نذكر "مجموعة المخطوطات الإسلامية" في جامعة ميتشغان الأمريكية [٣٢]. تملك الجامعة ١٧٩٥ مخطوطاً إسلامياً قامت بعرض ٩١٢ مخطوطاً على الشبكة

الإلكترونية - إلى تاريخنا الحالي - وتوفيرها مجاناً للباحثين والمهتمين، كما يسمح النظام بتحميل صفحة من المخطوط أو المخطوط كاملاً بالإضافة إلى عرض فهرس لكل مخطوط (رسم توضيحي ٩) مع إمكانية التبرع بإضافة الفهارس أو نص المخطوط. تتميز مجموعة المخطوطات الإسلامية بوضوح صور المخطوطات وجودتها، كما إنَّ العلامة المائية موضوعة خارج إطار صفحة المخطوط؛ فلا تؤثر على الأبحاث المعنية بالمخطوط، ومما يعيها عدم شمولية بطاقات المخطوط كما إنَّ الملفات المحملة تكون بصيغة (PDF) بدلاً من صيغ الصور المناسبة للأبحاث العلمية (JPG, TIF, PNG). ومن الجدير بالذكر أنَّ موقع مجموعة المخطوطات الإسلامية في جامعة ميتشغان على الشبكة الإلكترونية يحوي العديد من المعلومات القيمة، والقوائم المفيدة في اللغة الإنجليزية للمهتمين بأبحاث المخطوطات الإسلامية والعربية، مثل البحث عن المخطوطات في الفهارس، وقائمة بأسماء فهارس المخطوطات، وفهارس المخطوطات الإلكترونية وقواعد البيانات الإلكترونية للمخطوطات الإسلامية، إضافة إلى العديد من الدراسات المهمة بالخط العربي، والكثير من المواضيع الأخرى التي لها تعلق بهذا الفن.



رسم توضيحي ٩: مجموعة المخطوطات الإسلامية في جامعة ميتشغان الأمريكية.

يلاحظ من الأمثلة السابقة مدى اهتمام المؤسسات والمكتبات والجامعات مؤخراً في تملك المخطوطات العربية والإسلامية ورقمنتها (تحويلها إلى صور حاسوبية). وللمزيد من هذه الأمثلة، نحيل القارئ إلى موقع مكتبة جامعة ميتشغان الأمريكية [٣٣] فقد ذكر الموقع قائمة من مواقع الإنترنت المحتوية على قواعد بيانات للمخطوطات الإسلامية الرقمية، مع ترتيبها على حسب الدولة، والجدير بالذكر أن غالب هذه الجهود مقدّم من قبل جامعات، وأن غالبها في دول غير عربية وهي المخطوطات العربية والإسلامية الموجودة في أوروبا، ممّا يدل على حجم المشكلة ومدى تقصير الجامعات والمؤسسات العربية تجاه هذا التراث العظيم.

### برنامج مكتبة الملك عبد العزيز العامة بالرياض

أصدرت مكتبة الملك عبد العزيز العامة بالرياض [٣٤] برنامجاً رقمياً لعرض المخطوطات التابعة للمكتبة كإصدار أولي تجريبي منذ ما يقارب خمس سنين، على أن يتمّ تحسين البرنامج وإصدار نسخة نهائية للجمهور للاستفادة منه، ولكن لم يتم إصدار النسخة النهائية إلى يومنا هذا للأسف. يحوي البرنامج ما يقارب ٢٠٠٠ مخطوطة مختلفة، ويسمح بالبحث في اسم المخطوط، أو اسم المؤلف، أو رقم الحفظ، كما يظهر من رسم توضيحي ١٠. يتميز البرنامج بإمكانية رؤية صور المخطوط مع بعض الخصائص التي تفيد القارئ، مثل تغيير نسبة السطوع، ونسبة التباين، وتكبير الصورة. يفتقد البرنامج إلى توصيف دقيق لبطاقات المخطوطات، فهو يكتفي فقط بتحديد ثلاثة حقول (اسم المخطوط، اسم المؤلف، ورقم الحفظ)، ويعرض باقي المعلومات كملف من غير تنسيق. كما إنّ مرور الأعوام من غير إصدار البرنامج للجمهور أو إعلان تحديثات عليه تقلّل من أهميته في مجال رقمنة المخطوطات العربية والإسلامية.





(ب) شاشة عرض المخطوط



(أ) شاشة البحث مع الشاشة الرئيسية للبرنامج

رسم توضيحي ١٠ : برنامج مكتبة الملك عبد العزيز العامة بالرياض.

### مشاريع مستقبلية في رقمنة المخطوطات

ثمة عدد من المؤسسات التي تذكر مشاريع مستقبلية في رقمنة المخطوطات العربية والإسلامية، نذكر على سبيل المثال منها مؤسسة الفرقان التي أعلنت شروعاتها في بناء أكبر قاعدة بيانات للتراث الإسلامي على الشبكة العنكبوتية [١٧]، كما قامت جامعة ماكغيل في كندا بتأسيس مشروع "العلوم العقلانية في الإسلام" (Rational Sciences in Islam) [٣٥] والذي يهدف إلى دراسة التراث الفلسفي والعلمي والديني للمسلمين بطريقة شاملة. ويتوقع القائمون على هذا المشروع أن يرفد الباحثين والمهتمين بعشرات الآلاف من المخطوطات الإسلامية بصورة ميسرة وعلمية، ولقد عمدوا إلى تقسيمه لعناصر ثلاثة، منها اثنان خاصان برقمنة المخطوطات الإسلامية:

- التقاليد العلمية في المجتمعات الإسلامية (STIS): يهدف إلى دراسة تأثير العلوم الإسلامية على العلوم الحديثة في أوروبا، وتأثير الكتابات الغربية ما قبل القرن الثاني عشر الهجري على الباحثين في الدول الإسلامية.

- مبادرة لبناء قاعدة بيانات في مخطوطات الفلسفة الإسلامية ما بعد الكلاسيكية (PIPDI): يقوم الباحثون بإنشاء البنية التحتية العلمية المناسبة لإجراء دراسات منهجية في النصوص الفلسفية الإسلامية التي يرجع تاريخها إلى ما بين القرن الخامس إلى القرن الثالث عشر الهجري. وتشتمل قاعدة البيانات على مخطوطات في المنطق، وفي فلسفة اللغة، ونظرية المعرفة، والأخلاق، والميتافيزيقيا، والفلسفة الطبيعية وعلم الكونيات، وفلسفة العقل.

- مبادرة المخطوطات العلمية الإسلامية (ISMI): يهدف هذا المشروع إلى توفير مجموعة واسعة من المعلومات حول العلوم الدقيقة في العالم الإسلامي قبل العصر الحديث، من خلال تقديم هذه العلوم والمخطوطات إلى الباحثين والخبراء بلا مقابل عن طريق الإنترنت، وإلى تثقيف الجمهور في جميع أنحاء العالم في العلوم الإسلامية. وينتظر أن تحتوي قاعدة البيانات على أعمال ما يقارب ١٧٠٠ مؤلفاً وعالمًا في مختلف أرجاء العالم الإسلامي من إسبانيا الإسلامية إلى الهند وحدود الصين، منذ بداية القرن الأول الهجري حتى القرن الثاني عشر الهجري. وتتراوح موضوعات هذه الأعمال بين علم الفلك، والرياضيات، والفيزياء، والجغرافيا، والميكانيكا، وعدد من التخصصات ذات الصلة.

وبعد ذكر جهود ومحاولات الباحثين والمؤسسات في رقمنة المخطوطات مع ذكر أمثلة على أبرزها، ومزايا وعيوب هذه المحاولات. ويندرج في هذا الباب كذلك الجهود المبذولة في محاولة جمع قواعد بيانات خاصة بخطوط العلماء السابقين وناسخي المخطوطات. وهذا ما سيتم تفصيله في الفصل القادم.

### جمع نماذج من خطوط العلماء

اهتم العلماء السابقون بتحصيل نماذج من خطوط العلماء والنسخ والمشاهير

وجمعها في كتب، وقد دعت الحاجة في زماننا الحالي إلى بناء قاعدة بيانات رقمية تحوي عددًا كبيرًا من الخطوط الموجودة في الوثائق التاريخية العربية والإسلامية بمعايير محدّدة، مع توصيف كل هذه الخطوط؛ لاستخدامها من قبل علماء الحاسب الآلي والاستفادة منها. إنَّ بناء قاعدة بيانات تعد عملية أساسية للأبحاث المتعلقة بتحديد هوية صاحب الخط والتعرف على الكتابة في اللغة العربية، إذ من الممكن استخدام قواعد بيانات الخطوط في الوثائق التاريخية العربية والإسلامية المكتوبة بخط اليد في المجالات البحثية التالية:

١. تحديد هوية كاتب المخطوطات العربية والإسلامية (Writer Identification).
٢. التأكد من هوية الكاتب (Writer Verification).
٣. التعرف على الخطوط المزيّفة وتمييزها، والمساعدة في كشف المخطوطات المزورة.
٤. قاعدة بيانات مرجعية للباحثين؛ لاختبار أنظمتهم وأساليبهم المطورة.

ورغم اعتناء العلماء بخطوط العلماء والنسّاخ منذ مئات السنين، وجمع هذه الخطوط وطباعتها في كتب مفردة، إلّا أنّه لا توجد للآن قاعدة بيانات متاحة للباحثين تعنى بخطوط النسّاخ والكتّاب في المخطوطات التراثية. نذكر في هذا الفصل أمثلة على كتب مطبوعة اعتنت بجمع خطوط العلماء والنسّاخ مع ميزات هذه الكتب، ثم نذكر بعض الجهود المبدئية في بناء قاعدة بيانات رقمية مختصة بخطوط كتّاب المخطوطات العربية والإسلامية.

### **كتب مطبوعة اعتنت بجمع خطوط العلماء والنسّاخ**

من جهود علمائنا السابقين في هذا المجال ما جمعه الحافظ المؤرخ محمد بن علي ابن طولون الدمشقي "نوادير الإجازات والسماعات" [٣٦] من خطوط علماء مكة ومصر والشام ما بين عام ٩٢٠-٩٢٦ هـ. حُقِّق الكتاب عام ١٤١٩ هـ من قبل محمد مطيع الحافظ. ويحوي الكتاب ست عشرة إجازة، بالإضافة إلى سماعات

لأحاديث وتراجم وأسانيد لأمهات الكتب الإسلامية. ويدل هذا الجمع على اهتمام المتقدمين بخطوط علمائهم وتوثيق إجازاتهم وأسانيدهم توثيقاً صحيحاً.

ومن أمثلة جهود علمائنا المعاصرين قيام خير الدين الزركلي بتأليف كتابه "الأعلام" [٣٧] إذ حوى ثمانية مجلدات من تراجم المشاهير والأعلام من العلماء في مختلف الفنون، كالأنبياء، والملوك والسلاطين، والشعراء، وغيرهم. كما قام بتوثيق وإضافة صور لعينات من خطوط مئات من الأعلام قديماً وحديثاً.

كما قام ستيفن ليدر وزملاؤه بجمع وانتخاب السماعات الدمشقية الموجودة في دار الكتب الظاهرية في مخطوطات عام ٥٥٠ إلى ٧٥٠ هـ. قام المؤلفون في كتابهم "معجم السماعات الدمشقية" [٣٨] بتحليل ١٣٥٠ سماعاً بالإضافة إلى توصيف كل سماع بما فيه الشيخ المسمّع، وقارئ السماع، وكاتبه، وناقله، والمستمع، ورقم المخطوط، ورقم السماع، ومكان السماع، وزمان السماع، وغير ذلك.

ومثلهم جمع عبد الله وجاسم الكندري نماذج وأمثلة من خطوط العلماء من القرن الخامس إلى العاشر الهجري، إذ حوى كتابهما "خطوط العلماء من القرن الخامس إلى العاشر الهجري، نماذج وأمثلة" [٣٩] ثلاثمائة نموذج لخطوط العلماء من السماعات المختلفة، مع توصيف وتفرغ نص كل سماع. ويظهر أن كتاب الكندريين اعتمد بكثرة على الاقتباس من كتاب "معجم السماعات الدمشقية" لليدر وزملائه لكثرة النقول. ويظهر رسم توضيحي ١١ نماذج مقتبسة من الكتب المذكورة آنفاً مع التمثيل ببعض خطوط العلماء.



(ب) كتاب الزركلي ٣٧



(د) كتاب عبد الله وجاسم الكندري ٣٩



(i) كتاب الحافظ ابن طولون الدمشقي ٣٦



(ج) كتاب ستيفن ليدر وزملاؤه ٣٨

رسم توضيحي ١١ : أمثلة من كتب مطبوعة اهتمت بجمع خطوط العلماء.

### قواعد بيانات رقمية معنية بخطوط العلماء

اعتنى الباحثون في علوم الحاسب الآلي بالتعرف على الناسخ في المخطوطات الغربية، وكثرت الأبحاث في هذا المجال - نذكر على سبيل المثال منها [٤١ - ٤٣]، ولكن اعتنائهم بالمخطوطات العربية والإسلامية يكاد يكون منعدماً. فعدا عن بحث خاص بصاحب المقال [٤٤]، وبحث آخر عُرِضَ حديثاً [٤٥]، لا تجد بحثاً مختصاً - فضلاً عن قاعدة بيانات رقمية مختصة - في هذا المجال.

قام الباحث في بحثه "التعرف على الكاتب مستقلاً عن النص في المخطوطات العربية التراثية وتأثير زيادة عدد الكتاب" - Text Independent Writer Identification of Ancient Arabic Manuscripts and the Effect of Writers Increase [٤٤] - ببناء قاعدة بيانات للوثائق التاريخية العربية والإسلامية المحتوية على عينات مختلفة من تأليف العلماء العرب والمسلمين. تحوي قاعدة البيانات صوراً لمئتي مخطوطة مختلفة، تم اختيار خمس صور من كل مخطوطة؛ ليصبح عدد صور قاعدة البيانات ١٠,٠٠٠ صورة لمئتي خط ناسخ. واستخدمت قاعدة البيانات المطورة لفحص هوية الكاتب والتحقق منها. وعند استخدام نظام تمييز النمط وبرمجيات التعرف على الكاتب المطورة تم تحقيق نسبة نجاح تقديرها: ٩٣,٩٥٪ في التعرف على الكاتب. وتعد هذه النتائج المشجعة دليلاً على قدرة البرمجيات والحاسب الآلي على تمييز الخطوط في المخطوطات والوثائق التراثية. إلا أن هذه النتائج بحاجة إلى عدة تجارب معتمدة على قاعدة بيانات تحوي عشرات الآلاف من المخطوطات؛ للتوثق من النتائج، وهو الذي ندعو المؤسسات البحثية التي تُعنى بالتراث وتحرس على إفادة الباحثين إلى تبنيه وإقامة خطة لتفعيله على أرض الواقع لما لنتائجه - بإذن الله - من حفظ لتراثنا في صورة علمية وعملية، ولما فيه من تقديم التيسير على باحثي التراث في إنجاز أعمالهم في تحقيق تراث الأمة، ولما فيه من بناء نظام قادر على تقديم خدمة مهمة للباحثين في تمييز خطوط العلماء في المستقبل.

وبعدما قمنا في الفصول السابقة بذكر واقع وحاضر قواعد البيانات الإلكترونية للمخطوطات التراثية، وميزات الجهود الحالية، وأبرز العيوب والعوائق الحالية، نذكر في الفصل القادم مقترحاً لمواصفات قاعدة بيانات مستقبلية للمخطوطات الإسلامية والعربية، وعرضها بالتقنية الرقمية، مع الخيارات والميزات التي تمكن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية، وعلوم الحاسب الآلي معاً.

### **بناء قاعدة بيانات للمخطوطات الإسلامية وتوصيفها**

في هذا الفصل نقترح بناء قاعدة بيانات رقمية للمخطوطات الإسلامية والعربية بمواصفات ومعايير عالمية معتمدة؛ تسمح بتطويرها واستخدامها في بحوث الحاسب الآلي، بالإضافة إلى الاستفادة منها من قبل الباحثين والمهتمين بالتراث العربي والإسلامي. تبدأ عملية بناء قاعدة البيانات بمسح صور المخطوطات آلياً وتخزينها على الشبكة العنكبوتية، فيتم مسح هذه الصور باتباع المعايير العالمية في مسح المخطوطات والوثائق التاريخية ( بدقة ٤٠٠ نقطة لكل إنش). وبعد القيام بمسح صور المخطوطات آلياً، نقوم بتوصيف رقمي للمخطوط وصوره للاستفادة منه على أفضل وجه. ولتحقيق هذه الغاية، لا بد من تعريف القارئ بمبادرة ترميز وتبادل النصوص الرقمية (Text Encoding Initiative - TEI) [٢٧].

### **مبادرة ترميز وتبادل النصوص الرقمية**

قبل إنشاء مبادرة ترميز النصوص بوصفها نتاج عمل مجلس تعاوني مشترك، لم يجد علماء العلوم الإنسانية معايير موحدة لترميز النصوص الإلكترونية تخدم أهدافهم الأكاديمية. وفي عام ١٩٨٧ م، قام مجموعة من العلماء يمثلون العلوم الإنسانية، واللسانيات، والحوسبة باعتماد مجموعة من المبادئ التوجيهية المعروفة باسم "مبادئ باوكيبسي"، ثم تطويرها، وتم إعلان مبادرة ترميز النصوص الرقمية [٤٦].

تقوم مبادرة ترميز النصوص على تحديد مجموعة من العلامات التي يمكن إدراجها في التمثيل الإلكتروني للنصوص، وذلك لتعريف بنية النص ووصفه بطريقة دقيقة، إذ تعتمد برامج الكمبيوتر على وجود مثل هذه العلامات الواضحة لمعالجة النصوص بطريقة صحيحة، فبدونها يبدو النص الرقمي سلسلة من الأحرف غير المتميزة، ودون معنى لأجهزة الحاسب.

يعتبر ترميز "وصف المخطوط" (Manuscript Description) [٤٧] من الوحدات المتخصصة المندرجة ضمن نظام "ترميز النص الإلكتروني وتبادله" التي يمكن استخدامها لتقديم معلومات وصفية مفصلة عن الوثائق التراثية والمخطوطات المكتوبة بخط اليد. وعلى الرغم من أن عنصر "وصف المخطوط" صمّم أصالةً لترميز المخطوطات في التقاليد الأوروبية في العصور الوسطى، فإنّ برمجياته عامّة وشاملة؛ إذ يمكن أن تمتد إلى جميع الوثائق التاريخية بغض النظر عن لغتها. هذا الترميز يسمح للمستخدم بالبحث عن معلومات المخطوط من عنوان، ومواد، وتاريخ ومكان المنشأ. وللتعريف بعنصر "وصف المخطوط"، نقوم بالتعريف ببعض حقوله:

- msDesc: حقل يتضمن وصفاً أو أكثر للمخطوط.
- msIdentifier: معرف للمخطوط، يحتوي على المعلومات المطلوبة لتحديد المخطوط.
- head: يحتوي على أي نوع من العناوين، مثل عنوان القسم، أو عنوان القائمة، عنوان المخطوط.
- msContents: يتضمن محتويات المخطوط، كما يصف المحتوى الفكري للمخطوط أو جزء من المخطوط.
- physDesc: يحتوي حقل الوصف المادي على الوصف المادي الكامل للمخطوط أو جزء من المخطوط، ينقسم اختياريًا باستخدام عناصر أكثر تخصصًا.



- history: مجموعة من العناصر تصف التاريخ الكامل لمخطوط أو جزء من مخطوط.
- msPart: حقل يحتوي على معلومات حول جزء من المخطوط.
- dimensions: مواصفات أبعاد المخطوطة، أو صفحة معينة من المخطوط، أو جزءاً منها.
- locus: يحدد موضع وإحداثيات موقع داخل جزء من المخطوط.
- material: وصف للمواد التي يتكون منها المخطوط.
- watermark: وصف لأي علامة مائية أو رمز مماثل في المخطوط.
- origDate: يحتوي على أي شكل من أشكال التاريخ، وتستخدم لتحديد تاريخ المنشأ للمخطوطة.
- origPlace: يحتوي على أي طريقة لوصف المكان، وتستخدم لتحديد مكان الأصل للمخطوطة.

### الرموز المطلوبة لتوصيف المخطوطات العربية والإسلامية

نذكر في الجداول التالية بعض الحقول المطلوبة عند توصيف المخطوطات العربية والإسلامية. مع الإشارة إلى أن بعض هذه الصفات قد تحتوي على أكثر من مدخل، فعلى سبيل المثال، قد يحتوي المخطوط على عدة عناوين له، وقد يكون للمؤلف أكثر من لقب. كما نشير إلى أن بعض هذه الحقول قد لا تتوفر في بعض المخطوطات، أو لا يتم تعبئتها عند توصيفها. إلا أن النظام يسمح بإضافتها أو تعديلها في المستقبل مما يعطي مرونة لقاعدة البيانات المقترحة.

#### أ. معلومات خاصة بالمخطوط :

عنوان المخطوطة	الموضوع	المستودع	معلومات الكتاب المطبوع
عرض الحاشية	الملخص	حالة المخطوط	صنف الورق
اللغة	عدد اللوحات	أبعاد اللوحات	مقاصد المخطوط
ذكر أول المخطوط	ذكر عبارة الختام	اسم المكتبة	اسم الدولة
اسم المدينة	رقم الحفظ	خط النسخة	رقم الفلم
تاريخ النسخ	تاريخ التأليف	تواريخ أخرى	مصدر معلومات المخطوط
السماعات	الإجازات	التملكات	الوقفيات

#### ب. معلومات خاصة بالمؤلف / الناسخ / صاحب الحاشية / الخ :

الاسم	الكنية	اللقب	اسم الشهرة
-تاريخ الميلاد	تاريخ الوفاة	مصادر	نوع الخط

#### ج. معلومات خاصة بصفحة أو منطقة أو جزء من المخطوط :

عدد الأسطر	امتداد السطر	ارتفاع السطر	عدد الكلمات لكل سطر
سمك القلم	عرض الحاشية	إحداثيات محيط المنطقة	نص الكلام في كل سطر
السماعات	الإجازات	التملكات	الوقفيات
الخط المستخدم	الموضوع	رموز أو رسومات	إحداثيات كل سطر

#### مقترح المهام الخاصة بأبحاث الحاسب الآلي

وأخيراً، نقوم هنا باقتراح وتوصيف عدد من المهام القادرة على استثمار برمجيات الحاسب الآلي في مساعدة الباحثين في مجال العلوم الإنسانية والتراث العربي والإسلامي. ولتحقيق هذه المهام، فلا بدّ من سماح قاعدة البيانات المقترحة للمخطوطات الإسلامية للباحثين بتحميل مجموعة من صور المخطوطات ومعلومات هذه الصور بأعداد تصل إلى عشرات الآلاف بسهولة وسرعة، مما يتطلب أخذ

هذه المتطلبات بعين الاعتبار عند تصميم قاعدة البيانات الرقمية، وهي الميزة التي تفتقدها جميع قواعد البيانات الرقمية للمخطوطات الإسلامية والعربية - للأسف - حالياً.

كما تتضمن هذه المهام تقسيم المدخلات إلى مجموعات منفصلة للتدريب، والتحقق، والاختبار، وغيرها للمساعدة في تطوير البرمجيات بطريقة علمية دقيقة. مع الإشارة إلى أن عدد الصور أو النصوص المختارة في الفقرات القادمة هي أعداد مقترحة، وهي أعداد تقديرية للتمكن من تحقيق نتائج مرضية على مستوى عالمي في مجال التعرف الآلي، وتعتبر الزيادة على هذه الأعداد المقترحة أمراً مفضلاً. وتالياً نحدد الخطوط العريضة لخمس من المهام المقترحة الخاصة بأبحاث الحاسب الآلي:

#### أ. التعرف على الكتابة آلياً (Automatic Text Recognition)

تعدّ مهمة التعرف الآلي على نصوص المخطوطات من أهم النتائج المتوقعة لعلم الحاسب الآلي في خدمة التراث العربي والإسلامي، وهي في الوقت نفسه من أكثرها صعوبة بلا شك. فإن تحقيق نتائج مرضية في هذا المجال يعدّ شائعاً في صور الوثائق المكتوبة بخط اليد في اللغة العربية عموماً، وفي المخطوطات التراثية خصوصاً، إذ تحتاج هذه المهمة إلى الكثير من الوقت والجهد لتحضير صور المخطوطات - بالإضافة إلى حاجتها لتوفر نصوص هذه الصور - قبل استطاعة الباحث في الحاسب الآلي البدء في تطوير برمجياته. كما يحتاج إلى تحديد إحداثيات كل سطر في صور صفحات المخطوط (بداية السطر ونهايته وارتفاعه)، وتحديد نص كل صورة سطر على انفراد. والحقيقة أن الجهد الكبير لتجهيز مثل هذه المتطلبات لا بد أن يعتمد على تعاون الباحثين في أنحاء العالم؛ لكي يكتب له النجاح، ولذا تعتمد كثير من الجهود الغربية في مثل هذه الأمور على العمل التطوعي للباحثين عن طريق الإنترنت، ممّا يستلزم بناء الموقع بطريقة تسمح للمشاركة في هذه الجهود، بالإضافة إلى وجود نظام محكم للتأكد من نتائج هذه الجهود.

بعد تجهيز البيانات اللازمة لمهمة التعرف على الكتابة آلياً، يتم تقسيم الصور إلى أربع مجموعات منفصلة، نذكرها هنا مع ذكر قيم تقريبية لعدد المخطوطات في كل مجموعة (بافتراض وجود ٤٠ ألف صورة للمخطوطات على الأقل في قاعدة البيانات المقترحة): مجموعة التدريب (٥٠٪ من صور المخطوطات)، مجموعة التحقق (١٠٪)، مجموعة الاختبار (٢٠٪)، والمجموعة الإضافية (٢٠٪). مع ملاحظة أن يتم اختيار هذه الصور من أكبر عدد ممكن من المخطوطات، مع اختلاف العصور الزمنية، والفنون، ومكان المخطوط، وأسلوب المخطوط، وغيره، حيث يسمح التنوع بتطوير برمجيات ذات مرونة قابلة للتعرف على أكبر عدد من المخطوطات. يستخدم الباحثون عادةً مجموعة التدريب لتطوير برمجياتهم وتعديل معاييرها؛ لتتوافق مع صور المخطوطات المقدمة لها، ويستخدمون مجموعة التحقق للتأكد من سلامة نتائجهم على مخطوطات لم تقدم لبرمجياتهم من قبل، ومن ثمّ يتم اختبار أنظمتهم على مجموعة الاختبار ومقارنة النتائج. وتحافظ الجهة المطوّرة لقاعدة البيانات على المجموعة الإضافية من غير نشرها، لاستخدامها في مقارنة نتائج البرمجيات المطوّرة في مسابقات عالمية مستقبلية، إذ تساهم مثل هذه المسابقات على تشجيع البحث العلمي في هذه المجالات وتقدمه.

#### ب. البحث عن صور الكلمات (Word Spotting)

يمكن استخدام صور ونصوص مهمة "التعرف على الكتابة آلياً" السابقة في تقييم وتطوير مهمة البحث عن صور الكلمات، ولكن مع وضع مجموعة من الكلمات الاختبارية (١٠٠ - ٥٠٠ كلمة) بالاعتبار للبحث عنها في هذه الصور. ويفضل أن تكون هذه الكلمات المختارة من الكلمات المشتهرة عند باحثي التراث عند البحث في المخطوطات، ومن الممكن استخدام مهمة "معالجة اللغة الطبيعية" لمعرفة أشهر هذه الكلمات. وبعد تحديد هذه الكلمات يأتي دور تحديد إحداثيات هذه الكلمات في صور المخطوطات المختارة لكي يتم تطوير البرمجيات المختصة بالبحث عن صور الكلمات.

### ج. تحليل صور الوثائق (Document Image Analysis)

يتم في مهمة تحليل صور الوثائق اختيار أكبر تنوع ممكن من أشكال صور المخطوطات، إذ يتمكن الباحثون من الاستفادة من هذه الصور في تطوير العديد من التقنيات والتطبيقات في مجال تحليل صور الوثائق، بما في ذلك على سبيل المثال لا الحصر: تجزئة الصفحات والفقرات والأسطر، والتعرف على الرموز (مثل التملكات والأختام)، والاسترجاع المعتمد على الشكل، وتمييز النص عن الرسومات، والتحليل المنطقي للوثائق، وتحديد تشابه الصفحات، ومعالجة المعلومات غير النصية. وتحتاج كل من هذه التطبيقات أولاً إلى تحديد البيانات والمعلومات المختصة بهذا التطبيق. فعلى سبيل المثال، يستلزم تطبيق التعرف على الرموز تحديد الإحداثيات المئات أو الآلاف من هذه الرموز الموجودة في صور المخطوطات، مع إعطاء الوصف المناسب لكل من هذه الرموز.

### د. معالجة اللغة الطبيعية (Natural Language Processing)

تعد هذه المهمة من أسهل المهام في جانب التحضير المبدئي؛ إذ لا تستلزم صوراً للمخطوطات، وإنما تحتاج إلى نصوص هذه المخطوطات، والتي تتوفر بكثرة في المكتبات الإلكترونية للتراث العربي والإسلامي. ويفضل التنوع في اختيار آلاف النصوص لهذه المهمة من حيث تاريخها، والفن المختار، وما إلى ذلك.

### هـ. التعرف على الكاتب والتحقق منه (Writer Identification and Verification)

نحتاج في مهمة التعرف على الكاتب والتحقق منه إلى إضافة ترميز اسم الكاتب (ورقمه التسلسلي) لكل صورة في المخطوط، مع الاهتمام بالإكثار من المخطوطات المتنوعة للكاتب نفسه. وهذا يتم عن طريق البحث عن النسخ الأكثر من المخطوطات في قواعد بيانات المخطوطات الرقمية. إذ تساعد المخطوطات المتنوعة للكاتب نفسه في التأكد من قدرة النظام الآلي المطور على التعرف على الكاتب لا على المخطوط أو سنة النسخ. ويفضل أن يتم اختيار ٢٠ صورة من داخل

المخطوط (١٠،٠٠٠ مخطوط) مع استثناء صور البداية والنهاية؛ لاحتوائها عادةً على كتابات لغير الناسخ مثل التملكات والوقفات التي تؤثر على نتائج البحث، ثمّ تقسّم الصور المختارة إلى المجموعات الأربع؛ مجموعة التدريب (٧ صور لكل مخطوط)، مجموعة التحقق (٣ صور لكل مخطوط)، مجموعة الاختبار (٥ صور لكل مخطوط)، والمجموعة الإضافية (٥ صور لكل مخطوط).



## خاتمة

قام الباحث في هذا المقال بتفصيل واقع رقمنة الوثائق التراثية، مع اقتراح الخطوات اللازمة في المستقبل للاستفادة المثلى من البرامج الحاسوبية الحديثة. إذ تمّت دراسة فهارس المخطوطات الرقمية مع ذكر أبرز المؤلفات والمواقع المعنية بفهارس المخطوطات العربية والإسلامية، وذكر أبرز المزايا والعيوب لهذه المؤلفات والمواقع، وخاصةً

مكتبة الفرقان الرقمية وفهارس مخطوطات مكتبة المسجد النبوي المتميزتين في هذا المجال. وكما فصل المقال جهود الباحثين والمؤسسات في بناء قواعد بيانات المخطوطات الرقمية، مع إيراد أبرز هذه القواعد وميزاتها وعيوبها، وخاصةً نظام المخطوطات التابع لجامعة الملك سعود الذي نحسبه قادراً على أن يصبح مرجعاً معتبراً في البحث الرقمي في المخطوطات التراثية شريطة زيادة مخطوطاته كثيراً، وشريطة المراجعة الدقيقة لبطاقات مخطوطاته، وإضافة السمات التقنية الحديثة اللازمة المبينة في الفصل الخامس لها أيضاً. يُلاحظ في جمع الجهود المبذولة في بناء قواعد بيانات خاصة بخطوط العلماء؛ مدى قلة هذه الجهود، والحاجة الماسة لتأسيس قاعدة بيانات رقمية لخدمة بحوث التعرف الآلي على الكاتب في المخطوطات التراثية.

ونظراً لعدم قدرة أي من قواعد بيانات الرقمية للمخطوطات التراثية الموجودة حالياً على استيفاء حاجة الباحثين؛ فقد قام الباحث بعرض مقترح لمواصفات قاعدة بيانات رقمية مستقبلية للمخطوطات الإسلامية والعربية، مع تحديد الخيارات والميزات التي تمكّن قاعدة البيانات المقترحة من خدمة باحثي العلوم الإسلامية وعلوم الحاسب الآلي معاً. تحتوي قاعدة البيانات المقترحة للمخطوطات



التراثية على مهام محدّدة مسبقاً تساعد في تطوير البرمجيات الخاصّة بباحثي الحاسب الآلي: كالتعرف على الكتابة آلياً، والبحث عن صور الكلمات، وتحليل صور الوثائق، ومعالجة اللغة الطبيعية، والتعرف على الكاتب والتحقق منه. ونتمنى أن تتواصل الجهود في الاتجاه المقترح لاستكمال بناء قاعدة بيانات رقمية للمخطوطات التراثية بالموصفات المطلوبة، حتّى يتسنى للباحثين في علوم الحاسب الآلي تطوير البيئة الحاسوبية والبرمجيات اللازمة لخدمة التراث العربي والإسلامي، والقيام بالواجب المرجو تجاه الأمّة.

## شكر وتقدير

يتقدم معدّ البحث بالشكر لمركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية لدعمهم لهذا المقال خاصّة وللغة العربية والدراسات اللغوية الحاسوبية عامّة. ويذكر كذلك بوافر الامتنان كلّ من ساهم في إثراء هذا البحث من أفكار أو مراجعة أو معلومات قيمة، ونخصّ بالشكر منهم الباحثين في العلوم الشرعية: الشيخ مشهور بن حسن آل سلمان، وسامي الأسعد، والباحث في العلوم اللغوية عماد السوايعر، والباحث في علوم الحاسب الآلي يوسف العريان. شكرنا أيضاً موجّه إلى جامعة القصيم لتوفيرها المرافق الحاسوبية. نسأل الله عزّ وجل أن يجعل عملهم خالصاً لوجهه.



## المصادر والمراجع

١. الشيخ مشهور بن حسن آل سلمان. الحاسوب وكتب التراث الخطية. موقع فضيلة الشيخ مشهور بن حسن آل سلمان. [متصل] ٢٠٠٩. [تاريخ الاقتباس: ١٩ ١١، ٢٠١٤]. <http://www.mashhoor.net>
٢. دور تكنولوجيا المعلومات في حفظ المخطوطات العربية. حافظي زهير. ١٤، القاهرة : البوابة العربية للمكتبات والمعلومات، ٢٠٠٧، Cybrarians Journal.
٣. د. صلاح الدين المنجد. قواعد فهرسة المخطوطات العربية. الثاني. بيروت، لبنان : دار الكتاب الجديد، ١٩٧٦.
٤. مجموعة مساهمين، تحرير جيفري روبر، ترجمة عبد الستار الحلوجي. المخطوطات الإسلامية في العالم - النسخة العربية. لندن : مؤسسة الفرقان للتراث الإسلامي، ١٩٩٧-٢٠٠٢. صفحة ٣٤٩٦. ١-٩٧٨-١٠٢-٨٧٣٩٩٢-٤٥-٦.
٥. مؤسسة آل البيت. الفهرس الشامل للتراث العربي الإسلامي المخطوط (فهارس آل البيت). عمان : المجمع الملكي لبحوث الحضارة الإسلامية، ٢٠٠٤.
٦. كارل بروكلمان. تاريخ الأدب العربي. [المترجمون] عبد الحليم النجار و رمضان عبد التواب. الطبعة الخامسة. القاهرة : دار المعارف، ١٩٧٧.
٧. جرجي زيدان. تاريخ آداب اللغة العربية. [المحرر] ضيف شوقي. القاهرة : دار الهلال، ١٩٥٧.
٨. فؤاد سزكين. تاريخ التراث العربي. [المحرر] عرفة مصطفى و سعيد

عبد الرحيم. [المترجمون] محمود حجازي. الرياض : جامعة الإمام محمد بن سعود الإسلامية، ١٩٩١.

٩. السيد رزق الطويل. مقدمة في أصول البحث العلمي وتحقيق التراث. الطبعة الثانية. القاهرة : المكتبة الأزهرية للتراث، ٢٠٠٤.

١٠. معهد المخطوطات العربية. فهارس المخطوطات. الإصدارات. [متصل]  
<http://www.manuscriptsinstitute.org/> [٢٠١٤، ١٢ ٦].  
Esdatat\_fahares.aspx

١١. معهد المخطوطات العربية. خزانة المخطوطات. قواعد البيانات. [متصل]  
<http://www.manuscriptsinstitute.org/> [٢٠١٤، ١٢ ٦].  
database\_kazana.aspx

١٢. معالي الشيخ أحمد زكي يمانى. مؤسسة الفرقان للتراث الإسلامى. [متصل]  
<http://www.al-furqan.com> [٢٠١٤، ١٠ ٠٩].

13. Various Authors. Edited by Geoffrey Roper. World Survey of Islamic Manuscripts. 4th Edition. London : Al-Furqan Islamic Heritage Foundation ، 1992-1994 ، صفحة 2522. 978-1-873992-04-3.

١٤. برنامج المكتبة الشاملة. المكتبة الشاملة. [متصل] برعاية المكتب التعاوني للدعوة والإرشاد وتوعية الجاليات بحى الروضة. [تاريخ الاقتباس: ١٠ ٠٩،  
<http://shamela.ws> [٢٠١٤].

١٥. مشرف الشهري. الصفحة الرئيسية. موقع جامع المخطوطات الإسلامية. [متصل] [تاريخ الاقتباس: ١١ ٨، ٢٠١٤].  
<http://wqf.me>

١٦. فهارس المخطوطات من جميع انحاء العالم. ملتي أهل التفسير. [متصل]  
[تاريخ الاقتباس: ١٠ ٠٩، ٢٠١٤]. <http://vb.tafsir.net/tafsir28947/#>.  
VDbETRb-29Z

17. Al-Furqan Digital Library Portal. Al-Furqan Islamic Heritage Foundation[.2014, 10 09. [تاريخ الاقتباس: 10 09].  
[متصل] [http://www.al-furqan.com/world\\_library/lang/en](http://www.al-furqan.com/world_library/lang/en).

١٨. موقع يوسف زيدان. [متصل] [تاريخ الاقتباس: ١١ ٥، ٢٠١٤]. <http://www.ziedan.com>.

١٩. مركز الملك فيصل للبحوث والدراسات الإسلامية. برنامج خزانة التراث.  
[قرص مدمج] الرياض : مركز الملك فيصل، ١٩٩٢.

٢٠. مركز جمعة الماجد للثقافة والتراث. برنامج خزانة الماجد للتراث. [قرص  
مدمج] الرياض : شركة الدار العربية لتقنية المعلومات.

٢١. الرئاسة العامة لشؤون المسجد الحرام والمسجد النبوي. خدمة البحث بفهارس  
المخطوطات. خدمات المكتبة والباحثين. [متصل] [تاريخ الاقتباس: ١١ ٥،  
٢٠١٤]. [http://eservices.wmn.gov.sa/serv\\_1/L\\_S004.aspx](http://eservices.wmn.gov.sa/serv_1/L_S004.aspx).

22. Evyn Kropf. Manuscript catalogues online. University of Michigan Library[.2014, 6 28. [تاريخ الاقتباس: 6 28].  
[متصل] <http://guides.lib.umich.edu/islamicmsstudies/onlinecatalogues>.

٢٣. الشنطي، عصام. توزيع المخطوطات العربية والإسلامية في العالم. دار  
التأصيل. [متصل] [تاريخ الاقتباس: ١١ ٩، ٢٠١٤]. <http://taaseel.com/?PUrl=Bahs&ID=27>.

٢٤. موقع شيخة المَرِّي رحمها الله تعالى. الصفحة الرئيسية. مركز ودود للمخطوطات. [متصل] [تاريخ الاقتباس: ١١٩ ، ٢٠١٤]. <http://wadod.com>
٢٥. قسم المخطوطات. ملتقى أهل الحديث. [متصل] [تاريخ الاقتباس: ١١٩ ، ٢٠١٤]. <http://www.ahlalhdeeth.com/vb/forumdisplay.php?f=20>
٢٦. سعد الحميد و خالد الجريسي. قسم المخطوطات. مكتبة الألوكة. [متصل] [تاريخ الاقتباس: ١١٩ ، ٢٠١٤]. <http://www.alukah.net/library/8010>
27. Lou Burnard و Syd Bauman. TEI P5: Guidelines for electronic text encoding and interchange. International Consortium : TEI Consortium.2007 .
28. H. Ziaei Nafchi .R. Farrahi Moghaddam و M. Cheriet. Historical Document Binarization Based on Phase Information of Images [المحرر] Jong-Il Park و Junmo Kim. Computer Vision - ACCV 2012 Workshops. Berlin. Heidelberg : Springer Berlin Heidelberg.1-12 الصفحات ، 2013 .
29. IBN SINA: A Database for Research on Processing and Understanding of Arabic Manuscripts Images. R. Farrahi Moghaddam ، وآخرون. New York. USA : ACM Press.2010 . الصفحات 11-18.
30. Faculty of Computer Science & Information Technology - University of Malaya. Home Page. Digital Library of Malay Manuscripts [تاريخ الاقتباس: 25 7، 2014]. <http://mymanuskrip.fsktm.um.edu.my/Greenstone/images/html/mymanuskrip.htm>.

٣١. جامعة الملك سعود. المخطوطات. عمادة التعاملات الإلكترونية و الاتصالات.  
[متصل] [تاريخ الاقتباس: ٩٢، ٢٠١٤]. <http://makhtota.ksu.edu.sa>
32. E Kropf. Islamic Manuscripts Collection. University of Michigan Library. [متصل] [تاريخ الاقتباس: <http://guides.lib.umich.edu/content.php?pid=465626&sid=3812450>, 28 6, 2014].
33. Evyn Kropf. Online Collections of Digitized Islamic Manuscripts. University of Michigan Library. [متصل] [تاريخ الاقتباس: <http://guides.lib.umich.edu/islamic-msstudies/onlinecollections>, 15 8, 2014].
٣٤. مجلس إدارة مكتبة الملك عبد العزيز العامة. الصفحة الرئيسية. مكتبة الملك عبد العزيز العامة بالرياض. [متصل] [تاريخ الاقتباس: ١٢٦، ٢٠١٤].  
<http://www.kapl.org.sa>
35. Rational Sciences in Islam. McGill University. [متصل] [تاريخ الاقتباس: <http://islamsci.mcgill.ca/RASI>, 8 11, 2014].
٣٦. محمد بن علي ابن طولون الدمشقي. نوادر الإجازات والسماعات. [المحرر] د. محمد مطيع الحافظ. الطبعة الأولى. دمشق: دار الفكر، ١٩٩٨. صفحة ٩٦.
٣٧. خير الدين الزركلي. الأعلام، قاموس تراجم لأشهر الرجال والنساء من العرب والمستعربين والمستشرقين. الطبعة الخامسة عشر. بيروت: دار العلم للملايين، ٢٠٠٢. صفحة ٣٤٩. المجلد ٨.
٣٨. ستيفن ليدر، ياسين السواس و مأمون الصاغرجي. معجم السماعات الدشقية. الطبعة الأولى. دمشق: المعهد الفرنسي للدراسات العربية بدمشق، ٢٠٠٠. صفحة ٥٤٠.



٣٩. عبد الله الكندري و جاسم الكندري. خطوط العلماء من القرن الخامس إلى العاشر الهجري، نماذج وأمثلة. الطبعة الأولى. المنامة : دار البشائر الإسلامية، ٢٠١٤. صفحة ٨٣٩.

40. Using Codebooks of Fragmented Connected-Component Contours in Forensic and Historic Writer Identification. Lambert Schomaker ,Katrin Franke و Marius Bulacu. 6. U.S.A. : ScienceDirect .2007 ,Pattern Recognition Letters.28 ، المجلد ،
41. Automatic Handwriting Identification on Medieval Documents. M. Bulacu و L. Schomaker. Modena. Italy : IEEE. 14 .2007th International Conference on Image Analysis and Processing. ICIAP2007.279-284 .الصفحات .
42. An Expert Vision System for Analysis of Hebrew Characters and Authentication of Manuscripts. Laurence Likforman-Sulem ,Henri Maître و Colette Sirat. 2.1991 . Pattern Recognition.121-137 ، المجلد 24 ، الصفحات .
43. Hermite and Gabor Transforms for Noise Reduction and Handwriting Classification in Ancient Manuscripts. V. Eglin ,S. Bres و C. Rivero. 2-4 .New York : Springer101- . 122International Journal of Document Analysis and Recognition (IJDAR.9 المجلد ) (
44. Text Independent Writer Identification of Ancient Arabic Manuscripts and the Effect of Writers Increase. Sameh Awaida. Sousse. Tunisia : IEEE2013 . ICAI'2013

- The International Conference on Artificial Intelligence.
45. Writer Identification for Historical Arabic Documents.  
D. Fecker وآخرين. Stockholm. Sweden : IEEE ,August.  
2014. Proc. of 22nd International Conference of Pattern  
Recognition (ICPR..)
46. Descriptive metadata: Emerging standards. J.R. Ahron-  
heim. 5 ,U.S.A. : Elsevier ,1998 ,Journal of Academic Li-  
brarianship.395 ،المجلد 24 ،صفحة
47. Text Encoding Initiative. Manuscript Description. P5:  
Guidelines for Electronic Text Encoding and Interchange.  
[متصل]http://www.tei-  
c.org/release/doc/tei-p5-doc/en/html/MS.html.  
[تاريخ الاقتباس: 12 11 ،2014. 16 9 ،2014.]



## استقصاء تقنيات معالجة اللغات الطبيعية وتطبيقاتها في اللغة العربية

د. أمجد أبو جبارة (\*)

amjbara@umich.edu

---

(\*) شركة مايكروسوفت

حصل أمجد أبو جبارة على درجتي الماجستير والدكتوراة في تخصص معالجة اللغات واسترجاع المعلومات من قسم هندسة وعلوم الحاسوب بجامعة ميشيغان الأمريكية في العامين ٢٠٠٩ و ٢٠١٣، كما حصل على درجة البكالوريوس في هندسة الحاسوب من الجامعة الإسلامية في فلسطين في العام ٢٠٠٥. يعمل حالياً كباحث تطبيقي في مجال تطوير محركات البحث في شركة مايكروسوفت الأمريكية. تشمل اهتماماته البحثية حالياً تقنيات استرجاع المعلومات ومعالجة اللغات، و تطبيقات تعلم الآلة.



## ملخص

يهدف هذا البحث إلى التعريف بمعالجة اللغات الطبيعية (Natural Language Processing) باعتبارها أحد أهم مجالات علوم الحاسوب المتفرعة عن مجال الذكاء الاصطناعي (Artificial Intelligence) وباعتبارها تمثل نقطة التقاء بينه وبين مبحث اللغويات (Linguistics).

يسلط البحث الضوء على الجهود البحثية التي تستهدف اللغة العربية بالتحديد، مع تبيان ما للغة العربية من خصوصيات يلزم الباحثين الانتباه لها ومراعاتها عند نقل التقنيات المستخدمة في اللغات الأخرى إلى اللغة العربية.

يستعرض هذا البحث أهم الموضوعات الفرعية المندرجة تحت مبحث معالجة اللغات الطبيعية كتقطيع الكلام إلى وحدات، وتصنيف أقسامه، وإعراب الجمل، وفهم المعاني، وتوليد الكلام، إلخ. كما يستعرض الوظائف الأساسية التي ينبغي للباحثين في هذا المجال الإلمام بها ومعرفة التقنيات والطرق المختلفة لتنفيذها. كما يستعرض البحث -أيضاً- أهم وأشهر التطبيقات التي تعتمد بشكل أساسي على تقنيات معالجة اللغات.

يعتمد هذا البحث على عرض خلاصات ما ورد في أهم الكتب والبحوث المرجعية في المجال، وعقد المقارنات والمفاضلات بين الطرق المختلفة التي اقترحها الباحثون وعرض نتائج تقييم العديد من هذه الطرق.

يتضمن البحث كذلك استعراضاً مقتضباً لأهم الأدوات والمكتبات البرمجية والمجموعات النصية والبيانية المستخدمة بشكل متكرر من قبل الباحثين. هذا مع ترك إشارة إلى كل مرجع وكل أداة حتى يستطيع القارئ المهتم اللجوء إلى المصادر للحصول على المزيد من المعلومات التفصيلية في كل موضوع.

## ١. مقدمة إلى علم معالجة اللغات الطبيعية

هو أحد علوم الحاسوب ويعتبر أحد فروع العلم الناتجة عن التقاء مبحثي الذكاء الاصطناعي واللغويات. فعلم الذكاء الاصطناعي يهتم بتطوير خوارزميات وتقنيات تجعل الحاسوب قادراً على تحليل وفهم البيانات التي يستقبلها، وقادراً على توظيف هذا الفهم لاتخاذ قرارات ذكية وراشدة في ظروف وسياقات مختلفة - قد لا يكون تعرض لها من قبل - دون أن يكون قد تمت تغذية بأوامر برمجية محددة تخبره ماذا يفعل في كل ظرف [١].

أما علم اللغويات فهو العلم الذي يهتم بفهم مكونات اللغة، وأقسام الكلام، والعلاقات القواعدية والمعنوية بين تلك الأقسام، وأنماط التعبير، وطرق استخدام الألفاظ المختلفة في السياقات المختلفة، إلخ.

وبالتالي فإن علم معالجة اللغات الطبيعية هو علم تطبيقي يُعني باستخدام تقنيات الذكاء الاصطناعي - وعلى رأسها تقنية تعلم الآلة - من أجل حوسبة المفاهيم النظرية الموجودة في علوم اللغويات الخاصة باللغات الطبيعية التي يستخدمها البشر في التخاطب بحيث يصبح الحاسوب قادراً على تحليل وتمثيل وتوليد النصوص المدخلة أو المكتوبة أو المنطوقة بتلك اللغات.

وتدخل تقنية معالجة اللغات في كثير من التطبيقات الهامة التي أصبحت تشغل حيزاً كبيراً من الحياة اليومية لمستخدمي الكمبيوتر والإنترنت، ومن هذه التطبيقات: محركات البحث، والترجمة الآلية، والتلخيص الآلي، والإجابة الآلية عن الأسئلة، وأنظمة الإعلانات الذكية عبر الإنترنت، وتحليل المشاعر وقياس الرأي العام، وتصنيف موضوعات الأخبار، وتقنية البريد الإلكتروني من الرسائل الضارة، إلى آخر قائمة تطول من التطبيقات الهامة.

## نبذة تاريخية

علم اللغويات -والذي يستمد مجال معالجة اللغات أسسه النظرية منه- علم قديم جداً يصل عمره إلى مئات السنوات أو ربما أكثر من ذلك. أما علم معالجة اللغات الطبيعية فظهر بطبيعة الحال بعد وقت قصير من ظهور وانتشار الحواسيب إذ تعود أصوله إلى خمسينيات القرن الماضي عندما وضع العالم الشهير آلن تيورنج Alan Turing معياراً يمكن من خلاله الحكم على ذكاء الحواسيب [٢-٣]. يتمثل هذا المعيار في اختبار يتنافس فيه الحاسوب مع إنسان على إجابة أسئلة يلقاها عليهما حكم باستخدام لغة طبيعية ودون أن يعرف الحكم من المتسابقين هو الذي يجيب. يعتبر تيورنج في معياره هذا أن الحاسوب ذكي بقدر كافٍ إذا كان الحكم عاجزاً عن التمييز بين إجابات الإنسان وإجابات الحاسوب.

هذا المعيار الذي وضعه تيورنج يستدعي أن يكون الحاسوب قادراً على فهم اللغة الطبيعية التي يتلقى فيها الأسئلة وأن يكون قادراً على توليد الإجابات بلغة طبيعية كذلك. وقد شكل هذا التحدي أحد أوائل محفزات البحث في مجال معالجة اللغات الطبيعية.

وكان من أوائل التطبيقات التي انصب عليها اهتمام الباحثين في فترة الخمسينات والستينات من القرن الماضي فكرة الترجمة الآلية من لغة إلى لغة أخرى وكانت أولى التجارب التي نجحت نجاحاً مبدئياً في هذا المجال بحث صدر عن جامعة جورج تاون نجح فيه الباحثون في ترجمة قرابة ٧٠ جملة من اللغة الروسية إلى اللغة الإنجليزية [٤-٥]. ومن التطبيقات الأولى التي لاقَت اهتماماً أيضاً - أنظمة "إجابة الأسئلة" Question Answering [٦-٧]، ثم أنظمة تحليل وفهم الحوار [٨-٩]، وأدوات تقطيع الكلام وتحديد أصنافه وإعراب الجمل [١٠-١٣]، ثم تطبيقات التلخيص الآلي [١٤]، وأنظمة استرجاع البيانات [١٥]، ثم تطبيقات تحليل المشاعر التي تزامن ظهورها مع انتشار شبكة الإنترنت [١٦-١٧].



وقد كان التركيز في البدايات ينصب بشكل أساسي على تحديد مجموعة من القوانين التي يصوغها الباحثون بشكل يدوي (Rule-based Approach) ويلقونها للحاسوب في صورة برمجية حتى يتمكن من خلال تنفيذها من فهم أو توليد الكلام، وكان وضع هذه القوانين يتطلب من الباحثين فهماً عميقاً للغة وقواعدها ومعاني كلماتها وكيف يمكن أن تتغير هذه المعاني في السياقات المختلفة.

ولكن مع ظهور تقنيات تعلم الآلة Machine Learning في أواسط الثمانينات ومع تطور سرعات وذواكر الحواسيب بدأ الاعتماد على وضع قوانين مصاغة يدوياً يتراجع بشكل تدريجي لصالح الطرق الإحصائية التي تترك للحاسوب مهمة استنباط أو تعلم القوانين بشكل آلي من خلال الاطلاع على كميات ضخمة من النصوص والبحث فيها عن خصائص وأنماط متكررة ذات مدلول إحصائي (كمجيء الصفة بعد الموصوف بشكل متكرر في اللغة العربية مثلاً). وهذا يعني أنه لم يعد من المهم أن يكون لدى مطوري أنظمة معالجة اللغات فهم عميق وكامل لكافة قواعد اللغة ومعانيها، وإنما يلزمهم امتلاك القدرة على تحويل النصوص إلى صيغ إحصائية تتمثل فيها أهم خصائص وأنماط اللغة المتكررة.

مؤخراً بدأت تظهر بحوث جديدة تتبنى طرق هجينة تعتمد على كل من القوانين المصاغة يدوياً والأنماط الإحصائية التي يتم تعلمها آلياً باستخدام تقنيات تعلم الآلة للحصول على نتائج أفضل.

### معالجة اللغة العربية

ظهر الاهتمام بتقنيات معالجة اللغة العربية متأخراً بقدر ملحوظ عن اللغات الأخرى، وكانت البدايات الأولى في التسعينيات من القرن الماضي، وتناولت الأبحاث وقتها بعض جوانب اللغة كالتحليل الصرفي Morphological Analysis [١٨-٢٠] وتجذير الكلمات stemming [٢١] والتمثيل الحاسوبي لقواعد النحو [٢٢].

وتزايد الاهتمام بمعالجة اللغة العربية بشكل مطرد مع دخول الألفية الثانية، لاسيما مع تزايد الاهتمام بدراسة المنطقة العربية ولغتها، وتوفرت أموال وافرة لدعم البحوث الخاصة بمعالجة اللغة العربية من جهات حكومية وجامعات ومؤسسات بحثية عالمية، كما نُظمت العديد من ورش العمل الخاصة بدراسة تقنيات معالجة اللغة العربية على هوامش المؤتمرات الكبرى، وتناولت الأبحاث في أكثرها نقل التقنيات المستخدمة لمعالجة اللغات الأخرى وتطبيقها على اللغة العربية بعد تكييفها وتحويلها لتناسب مع خصوصيات اللغة العربية، ومن المهام الرئيسية التي تناولتها الأبحاث تقطيع الكلام، وتحديد أصناف الكلمات، والإعراب، وتحديد الروابط المعنوية بين أجزاء الكلام، ونسبة الضمائر إلى ما تعود عليه. كما تناولت الأبحاث العديد من التطبيقات الهامة كالترجمة الآلية بين اللغة العربية واللغات الأخرى [٢٣-٢٥]، والتلخيص الآلي [٢٦-٢٧]، وتحليل المشاعر واكتشاف الآراء [٢٨-٢٩]. هذا بالإضافة إلى تناول جوانب بحثية تخص اللغة العربية كالتشكيل الآلي للكلمات [٣٠-٣٢]، وكتمييز ومعالجة اللهجات المختلفة للغة العربية [٣٣-٣٦].

### ويميز الباحثون بين ثلاثة أشكال للغة العربية [٣٧]:

أولاً: العربية الفصحى التراثية Classical Arabic وهي اللغة المستخدمة في النصوص الدينية والنصوص التاريخية والكتب القديمة، وتعتبر اللغة العربية القديمة من أقدم أشكال اللغة التي حافظت على وجودها وانتشارها واستعمالها من قبل أعداد كبيرة من الناس بشكل يومي، ويعود الفضل في هذا إلى كونها لغة القرآن الكريم ولغة نصوص السنة النبوية التي يتداولها ويستخدمها المسلمون - الذين يمثلون أغلبية العرب- بشكل يومي خلال ممارستهم للشعائر الدينية.

ثانياً: العربية الفصحى الحديثة Modern Standard Arabic وهي اللغة الرسمية في كل الدول العربية، وتعتبر لغة ثانية في بعض الدول غير العربية، وهي اللغة المستخدمة في الكتابة الرسمية والتعليم والصحف ووسائل الإعلام. وتمتلك

الفصحى الحديثة نفس خصائص الفصحى التراثية من حيث قواعد اللغة وبناء الجمل وتصريف الكلمات وأصوات الحروف، ولكنها تختلف في كثير من الكلمات والتعابير، حيث إن ثمة الكثير من الكلمات والتعابير الموجودة في العربية التراثية تراجع استخدامها في اللغة الحديثة، ودخلت على اللغة كلمات وتعابير جديدة مقتبسة من لغات أخرى، أو في إطار الاستجابة الطبيعية لمتطلبات التطور المعرفي والحضاري.

ثالثاً: اللهجات العربية العامية Dialectal Arabic وتمثل اللغات المحلية المستخدمة في التخاطب اليومي بين الأفراد في المعاملات الحياتية المختلفة، وهي تختلف عن بنية وقواعد العربية الفصحى كما تختلف اللهجات العامية عن بعضها من نطاق جغرافي لآخر ومن بلد لآخر. وقد اشتهر بين الباحثين في اللغة العربية تمييزاً لـ ٦ لهجات رئيسية في العالم العربي: المصرية، والخليجية، والشامية، والمغربية، والعراقية، واليمينية. وتوجد عدة تقسيمات أخرى، ولكن يعتبر هذا التقسيم أكثرها شيوعاً. وكل واحدة من اللهجات تندرج تحتها لهجات فرعية تختلف من منطقة إلى أخرى، كما تختلف بحسب المستوى الاجتماعي ومكان المعيشة: المدينة أم القرية أم البادية.

وعليه فإن اللغة العربية تتمثل فيها ظاهرة يعرفها علماء اللسانيات ويسمونها "ازدواج اللسان" Diglossia وهي تعني أن أهل منطقة معينة يتحدثون لغتين بينهما اختلاف واضح من حيث القواعد وبنية الجمل والتراكيب وتصريف الكلمات دون أن يشعروا أنهم يتحدثون لغتين مختلفتين وإنما يتصرفون باعتبارهم يتحدثون لغة واحدة [٣٨]، مما يوجد تحديات بحثية لا سيما في ما يتعلق بمعالجة النصوص المنشورة في وسائل التواصل الاجتماعي التي يكثر فيها الخلط بين العربية الفصحى واللهجة العامية [٣٤].

## ٢. الوظائف الرئيسية في معالجة اللغات الطبيعية

الفصول التالية من هذا البحث ستتناول المهام والوظائف الرئيسية التي يتكرر الاحتياج إليها في أنظمة معالجة اللغات، مع استعراض مقتضب للطرق المختلفة التي توصل إليها الباحثون للقيام بهذه المهام مع تسليط الضوء على الأبحاث التي استهدفت اللغة العربية بشكل خاص في كل من هذه المهام. كما يتضمن هذا الاستعراض تبياناً لبعض الوظائف والمهام الإضافية التي تلزم عند معالجة اللغة العربية واللغات الشبيهة بها على وجه الخصوص.

### تسوية الكتابة Orthographic Normalization

وتعتبر هذه العملية من أول العمليات التي يتم إجراؤها على النصوص المكتوبة قبل البدء بمعالجتها بهدف التأكد من تنقية النص من الشوائب الكتابية كالرموز الزائدة وعلامات الترقيم غير الهامة لعملية المعالجة مثلاً، وكذلك من أجل التأكد من توحيد الأنماط المختلفة لكتابة الشيء الواحد. وقد تختلف إجراءات التسوية بحسب ما يحتاجه التطبيق. ولكن بشكل عام تتضمن إجراءات التسوية إزالة بعض الرموز الغريبة التي لا فائدة من وجودها في النص، وتوحيد طريقة كتابة علامات الترقيم، وتحويل الحروف الكبيرة Capital Letters في اللغة الانجليزية إلى حروف صغيرة Small Letters إذا كان التطبيق لا يحتاج إلى التفريق بينهما.

وقد بينت بعض البحوث المتعلقة بمعالجة اللغة العربية أن إجراء عمليات التسوية على النصوص العربية له تأثير ملحوظ على جودة وكفاءة عمليات المعالجة اللاحقة للنص [٣٩].

ومن أمثلة عمليات التسوية في اللغة العربية إزالة التطويل من الحروف التي يتم إضافة تطويل لها (عن طريق زر  $z + \text{shift}$  في الكيبورد مثل: "سيارة" حيث يظهر فيها طول زائد لحرف الياء) [٤٠]. ومن الأمثلة أيضاً توحيد طريقة كتابة الحروف التي يكثر وقوع الخلط فيها في اللغة العربية كالهزمة (الخلط بين

الوصل والقطع، والخلط بين الهمزة على ألف حال كونها مفتوحة أو مضمومة - حيث تكون الهمزة أعلى الألف - وحال كونها مكسورة - حيث تكون الهمزة أسفل الألف -)، كذلك الخلط بين الألف المقصورة والياء، وبين الهاء والتاء المربوطة. فيتم توحيد كتابة كل هذه الحروف بتحويل كل حالة يحصل فيها خلط إلى صورة موحدة في كل النص. ومن أمثلة عمليات التسوية كذلك إزالة التشكيل ( خاصة في حال كونه موجوداً لبعض الحروف أو الكلمات فقط دون أخرى).

### النسخ الحريفي Transliteration

ويقصد بالنسخ الحريفي كتابة الكلمة باستخدام حروف لغة أخرى غير اللغة الأصلية للكلمة عن طريق تحويل كل حرف إلى حرف أو أكثر يقابله في اللغة الأخرى وغالباً ما يكون لفظ هذا المقابل قريب من لفظ الحرف الأصلي في اللغة الأصلية. من أمثلة ذلك تحويل الحرف "ب" في اللغة العربية إلى "b" أو تحويل الحرف "ش" إلى "sh".

ويقصد بالنسخ الحريفي في سياق معالجة اللغات كتابة نص لغة ما باستخدام رموز الـ ASCII والتي تقتصر على الحروف الانجليزية الصغيرة والكبيرة والأرقام وعلامات الترقيم وعدد من الرموز الخاصة الأخرى، يتم اللجوء إلى النسخ الحريفي في تقنيات معالجة اللغات لأسباب عديدة منها وجود أنظمة حاسوب ولغات برمجة لا تدعم سوى ترميز الـ ASCII، ولا تدعم الترميز العالمي Unicode الذي يدعم حروف كل اللغات، وحتى في الحالات التي يتوفر فيها الترميز العالمي فما زال العديد من الباحثين الذين يتعاملون مع لغة غير لغتهم يلجأون إلى استعمال النسخ الحريفي لتسهيل التعامل مع اللغة الغربية عنهم.

واللغة العربية واحدة من اللغات التي يستعمل فيها النسخ الحريفي بشكل متكرر، وأكثر قواعد النسخ الحريفي المستخدمة من قبل الباحثين الأجانب هي طريقة Buckwalter Transliteration [٤٢]. أو أحد الطرق المبنية عليها (مثال

Habash-Soudi-Buckwalter [٤٢]. ووفق طريقة تحويل Buckwalter يكون لكل حرف عربي ما يقابله (قد يكون حرف أو أكثر أو حرف ورمز) في ترميز الـ ASCII، فمثلاً كلمة "سما" تنسخ حرفياً إلى "samaA"، وكلمة "كان" تنسخ إلى "kaAna"، وهكذا. ويكون التحويل لكل كلمة تحويلاً فريداً بحيث يمكن عكس عملية التحويل لاستعادة الكلمة الأصلية كما هي.

ونحيل المطورين المهتمين بتطوير وبرمجة أدوات معالجة اللغة العربية إلى البرنامج المشار إليه في [٤٣] والذي يقوم بتحويل النص العربي المرمز باستخدام الـ Unicode إلى ترميز ASCII وفق طريقة Buckwalter وبالعكس. بالإضافة إلى هذا فإن الواجهة البرمجية (API) المرفقة بـ "المجموعة النصية العربية القرآنية" تتضمن برنامجاً يقوم بنفس المهمة باستخدام طريقة مشتقة من طريقة Buckwalter بعد إدخال تغييرات طفيفة عليها [٤٤].

### تقطيع الكلام Tokenization

وتسمى هذه العملية أيضاً بالتحليل اللفظي Lexical Analysis، ويقصد به تقطيع النص إلى وحدات Tokens تتكون كل وحدة منها من أحرف أو أرقام أو رموز متصلة بالكلمات أو الأعداد أو علامات الترقيم، مع تحديد موضع بداية ونهاية كل وحدة. وتختلف درجات التقطيع التي يتم إجراؤها على النص بحسب ما يتطلبه التطبيق، ومنها "التقطيع البسيط" وغالباً ما تستخدم فيه التعبيرات النمطية Regular Expressions لتقطيع النص إلى الوحدات التي تفصل بينها المسافة Whitespace أو نهاية السطر، كذلك يتم فصل علامات الترقيم عن الكلمات (كالفاصلة المتصلة بالكلمة)، بحيث تصبح الكلمة وحدة وعلامة الترقيم وحدة أخرى منفصلة عنها، وكذلك فصل الأرقام عن الكلمات إذا كانت متصلة بها، وكذلك فصل بعض أدوات الربط التي قد تتصل بالكلمات أحياناً كما في اللغة العربية مثلاً حين يتصل حرف العطف "و" بالكلمة التي تليه كما في كثير من

النصوص العربية. وقد تشمل عمليات التقطيع البسيط إزالة بعض الرموز الغريبة أو استبدالها كلها برمز واحد بهدف تقليل درجة التشويش noise في النص.

وقد تتطلب بعض التطبيقات عملية "تقطيع متقدم" تذهب إلى أبعد من مجرد تقطيع النص إلى كلمات تفصلها مسافة Whitespace [٤٥-٤٦]، وفيها يتم تقطيع الكلمة نفسها إلى وحدات أصغر إذا كانت الكلمة ناتجة عن تركيب عدة مكونات كدخول "ال" التعريف على الكلمة أو الضمائر المتصلة كما في اللغة العربية. وفي هذه الحالة تتطلب عملية التحليل إجراء تحليل صريفي Morphological Analysis للتعرف على المكونات التي تتركب منها الكلمات حتى يكون من الممكن تقطيعها على هذا النحو.

الشكل ١ يوضح مثالاً لجملة باللغة العربية ونتيجة تقطيعها تقطيعاً بسيطاً وتقطيعاً متقدماً، وتوجد بين هذين المستويين من التقطيع مستويات عدة يتم اللجوء إلى ما يناسب التطبيق من بينها [٣٧].

الجملة	سيلقي المدير كلمة، وستتناول كلمته شرحاً للخطة الجديدة.
تقطيع بسيط	سيلقي   المدير   كلمة   ،   و   ستتناول   كلمته   شرحاً   للخطة   الجديدة   .
تقطيع متقدم	س   يلقي   ال   مدير   كلم   ة   ،   و   س   تتناول   كلم   ة   ه   شرح   ل   ال   خط   ة   ال   جديد   ة   .

شكل ١: مثال يوضح كيفية تطبيق عمليات التقطيع البسيط والمتقدم على نص عربي.

جدير بالذكر أن عملية تقطيع النص - حتى في أبسط أشكالها - تعتبر عملية صعبة ومعقدة للغاية في اللغات التي لا تستخدم فواصل محددة بين الكلمات ومن أمثلة هذه اللغات اللغة الصينية واليونانية التراثية واللغة التايلاندية وغيرها.

ومن الأدوات التي يمكن أن يستعملها الباحثون والمطورون لتقطيع النص العربي أداة TOKAN [٤٧]، وهي تحتاج إلى إجراء تحليل صريفي للنص حتى

تتمكن من تقطيعه، وهو ما توفره أداة MADA وهي متاحة مع TOKAN كباقة واحدة للتحميل من [٤٨]، وتحتوي الباقة -بالإضافة إلى المحلل الصريفي والمقطع- على أدوات أخرى، منها أداة لتسوية النص Orthographic Normalization، وأداة لتحويل النصوص العربية إلى ترميز ASCII وفق طريقة Buckwalter كما شرحنا في الفصل السابق، وغير ذلك.

ومن الأدوات المتاحة أيضًا AMIRA-TOK وهي إحدى الأدوات التي تتضمنها باقة AMIRA [٤٩] التي تضم أدوات لتنفيذ العديد من المهام الأساسية في معالجة اللغة العربية، وهي متاحة للتحميل من [٥٠]، وتصل دقة التقطيع في AMIRA إلى  $F1 = 99.2\%$  مع العلم بأن التقطيع الذي يجريه البرنامج لا يتعامل مع كل حالات التصريف وبالتحديد ما كان داخليا كـ Inflectional Morphology.

### التحليل الصريفي Morphological Analysis

يقصد بالصرف Morphology تحويل الأصل أو الجذر من الكلام إلى أبنية وأشكال مختلفة تحمل دلالات معنوية مرتبطة بمعنى الجذر، ومن أشكال الصرف: الجمع والتثنية والتأنيث واسم الفاعل واسم المفعول والصيغ الزمنية للأفعال: مضارع وماضي وأمر، وصيغ المبالغة، إلخ [٥١]. ومثال على ذلك في اللغة العربية الجذر "ذهب" الذي يمكن أن يأتي في التصريفات التالية: "يذهب" و "يذهبون" و "يذهبان" و "يذهبان" و "ذاهب" و "ذاهبا" و "ذاهبا" و "ذاهبا" ... إلخ.

وتهدف عملية التحليل الصريفي للكلمات إلى دراسة بنية الكلمة بغرض التعرف على القسم الصريفي لها، كتحديد هل هي جمع أم مفرد، صيغة تذكير أم تأنيث، صيغة ماضٍ أم مضارع أم أمر للأفعال ... إلخ، كما تهدف إلى تحديد جذر الكلمة وتحديد الزوائد التي أدخلت على الجذر.



### ويميز الدارسون لعلم التحليل الصرفي بين نوعين من التحليل [٣٧]:

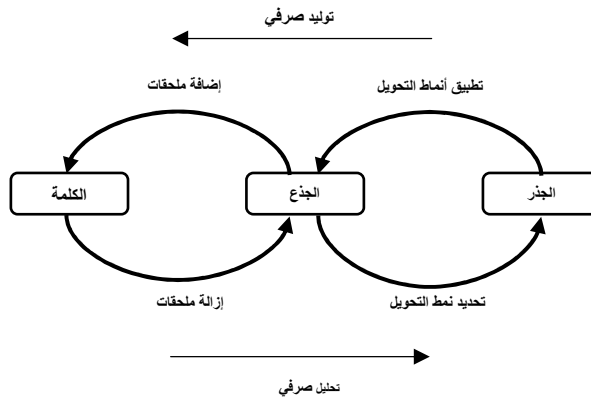
أولاً: تحليل معتمد على النوع أو الشكل Form-based Morphology : ويقصد به تحديد مجموع "المورفييمات" Morhpemes المكونة للكلمة وأنواعها. و"المورفييم" هو أقل لفظة في اللغة تقيد معنىً متضمناً فيها (قد تكون جزءاً من كلمة - كضمير متصل مثلاً). والمورفييمات التي تتكون منها الكلمات تشمل الجذع Stem مثل "كتب" في كلمة "تكتبين"، والاضافات السابقة prefixes مثل "التاء" التي تدخل على الجذع للدلالة على كون الفعل مضارع للأنثى الغائبة أو المذكر المخاطب، والإضافات اللاحقة suffixes مثل "ين" فتصبح الكلمة "تكتبين" للدلالة على توجيه خطاب للأنثى في صيغة الفعل المضارع. والإضافات الداخلية infixes كدخول الألف على الفعل فتحوله إلى صيغة اسم الفاعل كما في "كتب" - "كاتب"، ويعتبر من المورفييمات أيضاً الزوائد التي تحمل خصائص كلمة منفصلة ولكنها تأتي متصلة مع الكلمة في الكتابة مثل "أل" التعريف والضمائر المتصلة مثل "هم" في "كتابهم".

ثانياً: تحليل معتمد على الوظيفة Function-based Morphology: ويقصد بها تحديد الدور الإعرابي والدلالة المعنوية التي تنتج عن دخول كل مورفييم على الكلمة، وتنحصر وظائف المورفييمات المختلفة في ثلاثة أنواع رئيسية:

أولاً: تصريف اشتقاقي Derivational Morphology: ويقصد به أن دخول المورفييمات على الكلمة يصنع منها كلمة جديدة، فمثلاً "جامع" أصلها "جمع" ودخول مورفييم زائد عليها جعلها تعنى مكان الصلاة في المعنى الدارج للكلمة، و"جامعة" كذلك أصلها "جمع" ودخل عليها مورفييمان فأصبحت تعني مكان للتدريس في المعنى الدارج للكلمة.

ثانياً: تصريف إعرابي Inflectional Morphology: ويقصد به دخول المورفييمات على الكلمة بغرض تحديد الزمن، أو الجنس، أو حالة الخطاب، أو العدد كما أوضحنا في الأمثلة السابقة.

ثالثاً: تصريف إلحاقى Cliticization Morphology: ويقصد به أن يلحق بالكلمة مورفيم يحمل الخصائص المعنوية والإعرابية للكلمات المنفصلة. وضربنا آنفاً مثال دخول أل التعريف على أول الكلمة أو أحد الضمائر المتصلة على آخرها. الشكل ٢ يوضح نظام الاشتقاق والتحويل بين الكلمة وجذعها وجذرها ودور عمليات التصريف في هذا الاشتقاق [٥٢].



## شكل ٢: نظام الاشتقاق الصرفي والعلاقة بين الكلمة وجذعها وجذرها

ولعلم الصرف في اللغات المختلفة قواعد وقوانين تضبط عمليات الصرف وتحدد أشكال وأنواع الصرف التي تصلح لكل سياق وطريقة تنفيذ هذا الصرف. وبقدر ما تكون هذه القواعد قليلة ومحددة تكون حوسبة التحليل الصرفي أسهل. وتعتبر اللغة الانجليزية مثلاً من اللغات التي لها نظام تصريف سهل. في حين تعتبر اللغة العربية من اللغات الغنية من الناحية الصرفية بمعنى أن هناك الكثير من الأشكال والأبنية الصرفية التي يمكن تكوينها من الجذر الواحد مقارنة بمعظم اللغات الأخرى، وهناك الكثير من القواعد والقوانين التي ينبغي مراعاتها عند تصريف الكلمات. ولذلك يعتبر التحليل الصرفي من أبرز وأهم المهام التي يُعنى

بها الباحثون في مجال معالجة اللغة العربية [٣٧] ولذلك تجد أن بحوثاً كثيرة قد تناولت هذه الجزئية (أمثلة: [٢٠، ٥٣-٥٤]) واقترحت طرق متعددة لإجراء التحليل الصرفي للنصوص العربية. ويمكن للقارئ المهتم أن يرجع إلى أحد الاسقصاءات المنشورة في هذا الموضوع والتي حصرت الطرق المختلفة وقارنت بينها (مثال: [٥٢]). حيث اقترح المؤلفون تصنيف الطرق المختلفة لحوسبة عمليات التحليل الصرفي إلى أربعة أقسام:

أولاً: البحث في جدول شامل Table Lookup Approach: وفيها يتم تخزين كل الكلمات العربية الممكنة في جدول كبير يحتوي أيضاً تفكيك كل كلمة إلى مكوناتها من المرفيمات. هذه الطريقة تحتاج إلى جهد بشري كبير لبناء هذا الجدول.

ثانياً: الطريقة اللغوية Linguistic Approach: وفيها يقوم خبراء في اللغة وقواعد الصرف فيها بصياغة مجموعة من القوانين التي تحكم عمليات الصرف لأكثر عدد ممكن من الكلمات. ويتم تغذية الحاسوب بهذه القوانين حتى يستخدمها لتحليل الكلمات [٥٥].

ثالثاً: طريقة التوافيق Combinatorial Approach: وفي هذه الطريقة يتم تجريب كل المجموعات التي من الممكن تشكيلها من حروف الكلمة، ويتم بعد ذلك مقارنة كل مجموعة من تلك المجموعات بقائمة شاملة بالجذور العربية roots، وقوائم للملحقات suffixes [٥٦-٥٧]. وهذه الطريقة مكلفة من حيث احتياجها لطاقة حوسبة أعلى حتى تتمكن من اختبار كل المجموعات الممكنة في وقت معقول.

رابعاً: طريقة الأنماط Pattern-based Approach: وتعتمد هذه الطريقة على تحديد عدد من الأنماط المتكررة في الكلمات العربية من حيث التركيب الصرفي وتطبيق هذه الأنماط على الكلمات من أجل تحليلها [٥٨].

ونحيل القارئ المهتم إلى المصادر المشار إليها في هذا الفصل للتعرف أكثر على الطرق المستخدمة في التحليل الصرفي للكلمات العربية.

ومن أدوات التحليل الصريفي المتاحة للباحثين والمطورين بافتا MADA و AMIRA اللتان تمت الإشارة إليهما في الفصل السابق، ومنها أيضاً BAMA Stan-) SAMA و [41,59] (Morphological Analyzer Buckwalter Arabic (Morphological Analyzer dard Arabic [٦٠] وهو مبني على BAMA إلا إنه لا يتوفر مجاناً حالياً.

ومن الأدوات أيضاً ElixirFM ويتوفر له عرض لتجريبه من خلال واجهة يمكن تشغيلها عبر المتصفح [٦١]، كما أنه متاح للتحميل المجاني [٦٢].

### التجذيع Stemming

هي عملية يتم إجراؤها على الكلمة وتهدف إلى تحويل الكلمة إلى صورة أبسط بحيث تكون أقرب ما يمكن إلى جذع أو جذر الكلمة وذلك من خلال إزالة الزوائد الصرفية التي تدخل على أصل الكلمة. فمثلاً جذع كلمة computer في اللغة الإنجليزية هو comput، وكذلك جذع كلمة computing هو comput، ومثلهم أيضاً كلمات computed و computers و computation كل هذه الكلمات لها جذع واحد هو comput. ونلاحظ هنا أن comput في حد ذاتها ليست كلمة صحيحة في اللغة الانجليزية وليست هي الجذر الذي اشتقت منه الكلمات الأخرى ولكنها قريبة جداً من الجذر.

وتقوم عملية التجذيع هذه على افتراض أن الكلمات المشتقة من نفس الأصل تحمل نفس المعنى أو معاني متقاربة جداً في الغالبية العظمى من الحالات، وعليه فإن الهدف الأساسي من عملية التجذيع هو تحويل كل كلمة في النص إلى أقرب ما يمكن من أصل هذه الكلمة بحيث لا يعود هناك اختلاف بين computer و-comput computing و-ers، وعند مقارنة النصوص لاحتساب مدى تشابه نص مع نص آخر يتم التعامل مع كل هذه الكلمات على أنها كلمة واحدة.

ويتم اللجوء إلى عملية التجذيع في كثير من تطبيقات معالجة اللغات، ومن أمثلة ذلك عمليات الفهرسة indexing التي تقوم بها أنظمة استرجاع المعلومات Information Retrieval كمحركات البحث وما شابهها، فأنت عندما تبحث عن كلمة Computers عبر أحد محركات البحث، فلن يكون من الخطأ أن يعرض لك محرك البحث نتائج فيها كلمة Computer أو Computing [٦٣-٦٤]. ومثل ذلك أيضاً التطبيقات التي تحتاج إلى احتساب مقدار التشابه بين النصوص، كتطبيقات تصنيف النصوص Text Classification، وتطبيقات تجميع النصوص المتشابهة Text Clustering، وكذلك تطبيقات كشف النسخ في المنشورات Plagiarism Detection.

وتعتمد أغلب خوارزميات تجذيع الكلمات على مجموعة من القواعد والقوانين المعدة بشكل يدوي، وتختلف هذه القواعد من لغة إلى لغة، ولكن الرابط المشترك بينها كلها أنها تحاول أن تزيل الزوائد الصرفية التي تدخل على الكلمة. ويعتبر هذه النوع من الخوارزميات الأشهر والأكثر استخداماً بسبب سهولة تطبيقه وسرعة تشغيله على النصوص وعدم حاجته إلى قوائم كلمات أو قواميس. وتعتبر خوارزمية Porter التي اقترحها Matrin Porter (عام ١٩٨٠) أشهر هذه الخوارزميات للغة الانجليزية وما زالت مستخدمة على نطاق واسع حتى اليوم [٦٥].

وهناك مجموعة أخرى -ولكنها أقل شيوعاً- من خوارزميات التجذيع تعتمد على رصد خصائص إحصائية للكلمات، وذلك من خلال إيجاد الكلمات المتشابهة التي تظهر معاً بشكل متكرر في النصوص المختلفة ثم اختيار الكلمة الأقصر من بين كل هذه الكلمات لتكون هي الجذع الذي ترد إليه كل الكلمات الأخرى [٦٦-٦٧]. ولكن هذه الطريقة تحتاج إلى توفر كميات كبيرة جداً من النصوص تكفي لاحتساب علاقات إحصائية ذات معنى بين الكلمات. والميزة الأساسية لهذا النوع من الخوارزميات هو قدرته على تجذيع التصريفات الغير قياسية مثل foot و feet

وهو ما تفشل الخوارزميات المعتمدة على القواعد المصاغة يدويًا في التعامل معه. الميزة الأخرى لهذا النوع من الخوارزميات هو أن الخوارزمية نفسها يمكن تطبيقها على أكثر من لغة. في حين أن القواعد المصاغة يدويًا تختلف بحسب اللغة وتحتاج إلى خبير لغوي حتى يتمكن من تحديد تلك القواعد.

اللغة العربية تعتبر من اللغات التي يصعب تطبيق التجذيع فيها لعدة أسباب أهمها الثراء الصرفي لها وكون الزوائد الصرفية تدخل على الكلمة في أولها ووسطها وآخرها، مما يجعل صياغة قواعد للتجذير عملية صعبة ومعقدة جدًا، هذا بالإضافة إلى الجدل بين الباحثين حول درجة التجذيع اللازمة في اللغة العربية بحيث تحقق ما تحتاجه تطبيقات استرجاع المعلومات أو تصنيف النصوص بين من يرى أن الاقتراب من صورة الجذع يكفي وبين من يرى ضرورة إيجاد جذر الكلمة الثلاثي أو الرباعي [٦٨].

وقد اقترحت عدة طرق لتجذيع الكلمات العربية منها ما يعتمد على قوائم معدة يدويًا تحتوي كل كلمة وجذعها (أو جذرها) [٦٨-٦٩]، ومنها ما يعتمد على صياغة قواعد شبيهة بما تقوم به خوارزمية porter في اللغة الانجليزية (مثال [٥٨])، ومنها كذلك ما يعتمد على الطرق الإحصائية [٧٠-٧١].

ومن أدوات التجذيع المتاحة للاستعمال من قبل الباحثين والمطورين أداة Khoja Stemmer [٧٢] التي تستخدم مجموعة من القوانين المصاغة يدويًا لحذف الزوائد، ومجموعة من الأنماط والقوائم لاستخراج الجذع (أو الجذر).

#### تأصيل الكلمة Lemmatization

وهي عملية شبيهة جدًا من حيث الهدف بعملية التجذيع Stemming ولكنها تختلف عنها من عدة نواحي أهمها أن الأصل الناتج عن العملية هو كلمة صحيحة في اللغة وتكون في الغالب هي أبسط شكل للكلمة، ففي المثال الوارد في الفصل

السابق، فإن أصل كلمة computers و compute (قارنه بالجذع comput). الفارق الآخر هو أنها تقوم بتحويل الكلمة إلى كلمة أخرى إذا استدعى الأمر ذلك بشكل لا تستطيع أغلب خوارزميات التجذيع القيام به. مثال ذلك تحويل is و are و am إلى be وكذلك better إلى good وهكذا.

وتعتمد أغلب خوارزميات تأصيل الكلمات على استخدام قوائم بالكلمات المختلفة وأصولها مع دمج ذلك مع بعض القوانين البسيطة لجعل حجم تلك القوائم أصغر.

### تصنيف أقسام الكلام Part of speech tagging

وهي واحدة من العمليات الأساسية اللازمة للكثير من تطبيقات معالجة اللغات، ويتم فيها تصنيف كل كلمة في النص بحسب حالتها الصرفية وبحسب سياقها الإعرابي، كتصنيف الكلمة كفعلٍ ماضٍ أو اسم بصيغة الجمع أو أداة عطف ... إلخ.

والصعوبة الأساسية في هذه العملية تكمن في ضرورة أخذ السياق الذي ترد فيه الكلمة في عين الاعتبار وذلك بسبب الغموض الذي يكتنف بعض الكلمات إذا ما عوملت منفصلةً عن سياقها. مثال لذلك في اللغة العربية كلمة "ذهب" ففي بعض السياقات هي اسم معدن ثمين، وفي سياقات أخرى هي فعلٌ ماضٍ للمفرد الغائب.

يعرض شكل ٣ مثالين أحدهما باللغة الإنجليزية والآخر باللغة العربية يوضحان نتيجة تطبيق عملية تصنيف أقسام الكلام على النص. الرموز الظاهرة في المثالين مثل VBZ و NN و \$PRP وغيرها تسمى Tags، وكل منها يرمز إلى أحد أقسام الكلام Part-of-speech، فمثلاً VBZ تشير إلى الفعل المضارع للمفرد في اللغة الإنجليزية.

My son also likes eating meat

./ My/PRP\$ son/NN also/RB likes/VBZ eating/VBG meat/NN

تعجبني الأفكار الإبداعية ويسعدني أن أساهم في إغاجها

تعجبني/ VBP الأفكار/ DTNN الإبداعية/ DTNN و/ CC يسعدني/ VBP ان/ IN

اساهم/ VBP في/ IN انجاحها/ PUNC ./ NNP

### شكل ٣ : نتيجة تطبيق عملية تصنيف أقسام الكلام على جملة إنجليزية وأخرى عربية

ويتم تعريف الأقسام Parts-of-speech التي يتم تصنيف الكلمات إليها من قبل علماء اللغويات، الذين يعرفون مجموعة من الأقسام تسمى TagSet. ولكل لغة من اللغات أصناف مختلفة من الكلام تختلف بحسب الثراء الصرفي والقواعدي لتلك اللغة. كما أن اللغة الواحدة قد يلجأ العلماء فيها إلى تعريف أكثر من Tag-Set تختلف في درجة تفصيلها. فمثلاً أبسط TagSet في اللغة العربية ممكن أن تحتوي ثلاثة أقسام فقط: اسم وفعل وحرف. في حين أنه من الناحية النظرية ممكن تعريف عدد قد يزيد على ٣٣٠,٠٠٠ قسم!

ومن أشهر التقسيمات المستعملة في اللغة العربية Buckwalter TagSet بصورتها المخصصة للكلمات المقطعة Tokenized (وهي الأكثر شيوعاً وتضم قرابة ٥٠٠ قسم) وغير المقطعة Untokenized وتصل أقسامها إلى عشرات الآلاف. وفي حالة تقسيم الكلام Tokenization فإن الكلمة الواحدة تنقسم إلى مكوناتها الصرفية وكل من تلك المكونات يكون تابعاً لقسم مختلف.

هذا بالإضافة إلى العديد من التقسيمات الأخرى التي حاول بعضها اختصار Buckwalter TagSet بهدف تقليل عدد التقسيمات وتسهيل عملية التصنيف [٧٦-٧٣].



وتستخدم معظم خوارزميات تصنيف أقسام الكلام حالياً تقنيات تعلم الآلة، وذلك من خلال إعداد كمية كبيرة من النصوص والاستعانة بخبراء في اللغة لتحديد القسم الذي تتبع له كل كلمة في النص، ثم يتم استخدام هذا النص المصنف لتعليم خوارزميات تعلم الآلة كيفية إجراء التصنيف لنصوص جديدة غير تلك التي تعلمت منها، ومن أشهر تقنيات تعلم الآلة المستعملة هنا: Hidden Markov Models [٧٧-٧٨] و Conditional Random Fields [٧٩].

وغالباً ما تسمى المجموعات النصوية المستعملة في تعليم خوارزميات تعلم الآلة على تصنيف أقسام الكلام والإعراب وما شابههما بـ Treebanks، ومن أشهر هذه المجموعات في الإنجليزية الـ Penn Treebank<sup>□□□</sup> ومن أشهرها في اللغة العربية Penn Arabic Treebank<sup>□□□</sup>.

ومن أدوات تصنيف أقسام الكلام المتاحة للمطورين أدانا AMIRA و MADA وقد سبقت الإشارة إليهما في فصول سابقة. ومنها أيضاً Stanford Log-linear Part-Of-Speech Tagger<sup>□□□</sup> والذي تصل دقته إلى ٩٦,٥٪ على النصوص العربية [٨٠] ويعتبر من أفضل الأدوات المتاحة وأسهلها في الاستخدام، كما أنه يدعم لغات أخرى.

### التشكيل الآلي Automatic Diacritization

وهي من الوظائف المهمة في اللغة العربية، وذلك لأن نطق أحرف اللغة العربية يختلف باختلاف تشكيل الحرف، وتختلف تبعاً لذلك المعاني والبنية الإعرابية للجملة والعديد من الخصائص اللغوية للنص. ومع ذلك فإن أغلب النصوص العربية المنتشرة في الكتب وعلى الإنترنت تخلو من التشكيل إما كلياً أو جزئياً، أما النصوص المشكلة تشكيلاً كلياً فهي نادرة جداً.

- 
- (1) <http://www.cis.upenn.edu/~treebank/>
  - (2) <https://catalog.ldc.upenn.edu/LDC2003T06>
  - (3) <http://nlp.stanford.edu/software/tagger.shtml>

وهذه الظاهرة في اللغة العربية تتسبب في حالة التباس كبيرة تجعل من الضروري الاعتماد بشكل أكبر على السياق لفك الالتباسات. ومن أشهر التطبيقات التي يعتبر توفر التشكيل فيها أساسياً وضرورياً: توليد الأصوات العربية Arabic Speech Synthesis، حيث يكون من الضروري تحديد تشكيل كل حرف في الكلمة حتى تتمكن آلة توليد الصوت من إخراج الصوت المناسب للحرف. وكحل لهذه المشكلة اقترحت بعض البحوث بناء أنظمة لإضافة التشكيل الآلي، وأغلب هذه الطرق تعتمد على تقنيات تعلم الآلة أو تقنيات هجينة تجمع بين تعلم الآلة وبين القواعد المصاغة يدوياً، وتستخدم أغلب هذه الطرق خصائص لفظية Lexical، وخصائص صرفية Morphological، وخصائص نحوية Syntactic لتدريب خوارزميات تعلم الآلة على إجراء عمليات التشكيل الآلي. ومن أمثلة الأبحاث التي تناولت الموضوع ونحيل القارئ المهتم إلى مطالعتها: [٨١-٨٣].

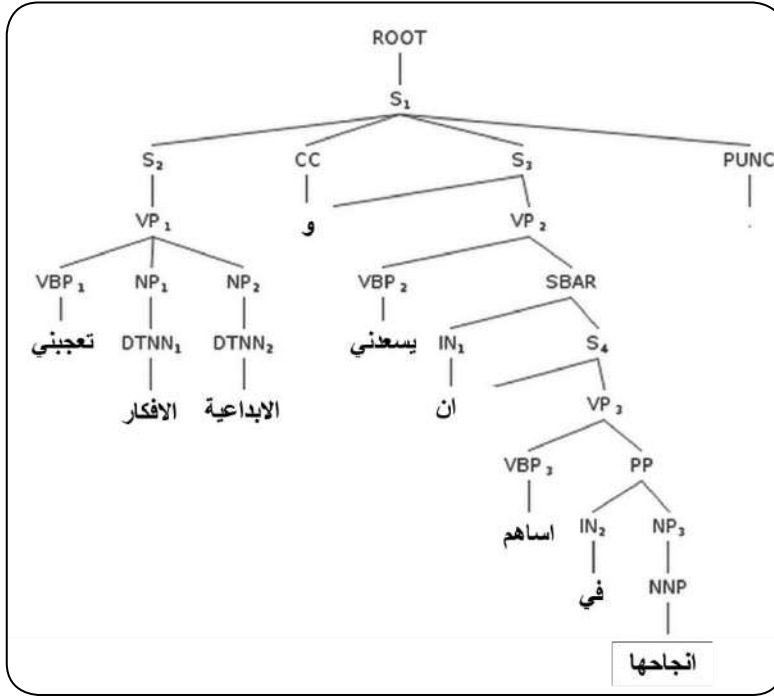
وتصل دقة أنظمة التشكيل الآلي إلى ٨٨,٢٪ وذلك في تحديد تشكيل نهايات الكلمات كما في الطريقة المنشورة في [٨٤] التي تستخدم تقنيات التعلم العميق Deep Learning لتنفيذ المهمة.

كما تتضمن باقة MADA التي أشرنا إليها سابقاً أداة للتشكيل الآلي تصل دقتها في كشف حركات نهايات الكلمات إلى ٨٥,١٪ ويمكن للباحثين والمطورين استعمالها.

### تحليل البناء النحوي Syntactic Parsing

وتهدف هذه العملية إلى كشف بنية الجملة، وتحديد العلاقات النحوية بين الكلمات والعبارات في الجملة، كتحديد الأقسام البسيطة التي تتكون منها الجمل المركبة وأدوات الربط بين تلك الأقسام، مثال ذلك الجمل التي تحتوي على شرط وجواب شرط وأداة شرط تربط بينهما. وكذلك تحديد الكلمات التي تكون معاً عبارة اسمية Noun Phrase أو عبارة فعلية Verb Phrase. وغالباً ما يتم تمثيل

هذه العلاقات في صورة شجرة جذرها في الأعلى ويمثل كل تفرع من تفرعات الشجرة قسم من الجملة، كما تمثل كل ورقة من أوراق الشجرة النهائية كلمة من كلمات الجملة أو علامات الترقيم فيها (والقسم Part of speech الذي تتبع له). النقاط الداخلية في الشجرة تمثل بنى نحوية في الجملة فمثلاً النقطة  $S$  تمثل جملة، والنقطة  $VP$  تمثل جملة فعلية، والنقطة  $SBAR$  تمثل جملة تابعة كجملة جواب الشرط أو الجملة المعطوفة. الشكل ٤ يوضح التمثيل الشجري لنتيجة تحليل البناء النحوي لجملة عربية.



شكل ٤ : التمثيل الشجري للبناء النحوي لجملة عربية

ويلاحظ أن عملية الإعراب النحوي تتضمن تصنيف أقسام الكلام لكل كلمة وتزيد عليها بالذهاب إلى مستوى أعلى من التحليل يشمل تصنيف أقسام أشباه الجمل والتعبيرات الاسمية أو الفعلية أو أدوات الربط، وغيرها مع تمثيل هذا كله في شكل شجري يعبر عن البنية النحوية للجملة. وتسمى هذه العملية أحياناً بالإعراب العميق Deep Parsing وذلك بغرض التفرقة بينها وبين الإعراب السطحي -Shal low parsing الذي يكتفي بتحديد أقسام الكلمات وتحديد العبارات الاسمية والعبارات الفعلية دون تحديد البنية الكاملة للجملة.

وتعتمد تقنيات الإعراب النحوي على تعريف قواعد النحو الحاكمة للغة وتمثيلها بشكل يفهمه الحاسوب ثم تطبيق هذه القواعد على الجمل لتحديد الإعراب الصحيح للجملة. وتنقسم خوارزميات الإعراب النحوي من حيث كيفية استخدام قواعد النحو إلى خوارزميات تستعمل طريقة الصعود -Bottom up وأخرى تستعمل طريقة النزول Top-down. الأولى تبدأ من الكلمات (أوراق الشجرة) وتبدأ بتطبيق القواعد النحوية التي تحتوي تلك الكلمات صعوداً حتى الوصول إلى جذر الشجرة (الذي يمثل جملة مكتملة). أم الطريقة الثانية فتبدأ من القواعد النحوية التي تتضمن جذر شجرة (أي تركيبة جملة كاملة) نزولاً حتى الوصول إلى الأوراق (الكلمات).

وتحتاج خوارزميات الإعراب أن يتم تزويدها بقواعد النحو الخاصة باللغة حتى تتمكن من تطبيقها، ويتم تمثيل قواعد النحو حاسوبياً على شكل مجموعة من القوانين التي تُعرّف البنى المسموح وجودها في الجملة. الشكل يحتوي مثالاً إيضاحي لبعض قواعد النحو وطريقة تمثيلها وكيفية استخدامها لإعراب جملة عربية [٨٥].

وتحتاج عملية الإعراب إلى معرفة المعنى المقصود كذلك وليس قواعد النحو فقط، فالجملة الواحدة - إذا ما تجاهلنا معناها المقصود - فإنها قد تحتل أكثر من تركيب نحوي.

ولذلك فإن العديد من خوارزميات الإعراب شائعة الاستخدام حالياً تلجأ إلى الطريقة الإحصائية Statistical Approach والتي تهدف إلى إعطاء قيمة احتمالية لكل قاعدة من قواعد التركيب النحوي بحسب مقدار تكرار استعمالها في سياق الجملة التي يراد إعرابها، كما أنها تعطي درجة احتمالية لكل شجرة من الشجرات النحوية العديدة التي يمكن بناؤها للجملة الواحدة، وعادة ما تستخدم خوارزمية بحث مثل Viterbi algorithm للبحث عن أفضل شجرة في فضاء الشجرات المحتملة (دون الحاجة لاحتساب احتمالية كل شجرة بالكامل). ويتم تعلم القيم الاحتمالية للقواعد من خلال استخدام مجموعات نصية مصنفة ومعرية Treebanks كتلك المستخدمة في عملية تصنيف أقسام الكلام التي أشرنا إليها في الفصل السابق.

القواعد وشرحها باللغة العربية		الشجرة الإعرابية لـ "أشرب شايًا لذيذاً"
S => VP	الجملة يمكن أن تكون جملة فعلية	<pre> graph TD     root --&gt; S     S --&gt; VP     VP --&gt; V     VP --&gt; NP     V --&gt; اشرب     NP --&gt; N     NP --&gt; ADJ     N --&gt; شاي     ADJ --&gt; ساخن </pre>
VP => V NP	الجملة الفعلية تتكون من فعل وعبارة اسمية	
NP => N ADJ	العبارة الاسمية يمكن أن تتكون من اسم وصفة	
NP => N ADJ	العبارة الاسمية يمكن أن تتكون من اسم وصفة	
اشرب <= V	«أشرب» فعل	
شاي <= N	«شاي» اسم	
لذيذ <= ADJ	«لذيذ» صفة	

شكل ٥ : مثال يوضح طريقة تمثيل قواعد اللغة وتطبيقها لإعراب جملة

ومن أدوات الإعراب المتاحة للباحثين والمطورين<sup>[٨٦]</sup> The Stanford Parser، وهي تدعم اللغة العربية إلى جانب الإنجليزية والصينية والإسبانية، ويمكن استخدام نفس الأداة لتصنيف أقسام الكلام كذلك Part-of-speech tagging، ومن الأدوات أيضاً<sup>[٨٦]</sup> Bikel's Parser والتي تدعم اللغة العربية والإنجليزية.

### تحليل العلاقات الاعتمادية (الإعراب) Dependency Parsing

وهو نوع آخر من التحليل البنائي للجمل ولكنه يهدف إلى تحديد العلاقات بين الكلمات وتبيان أي الكلمات أو التعابير تعتمد معنوياً أو نحوياً على كلمات أو تعابير أخرى مع توضيح نوع الارتباط الذي يجمع بينهما، ومن أمثلة ذلك ارتباط الاسم بالفعل في جملة عندما يكون الاسم فاعلاً كما في "شرب الطفل الحليب"، الفعل "شرب" مرتبط بعلاقة اعتمادية مع الاسم "الطفل" حيث أن الطفل هو الفاعل، وبالمثل كلمة "الحليب" مرتبطة بالفعل "شرب" لكونها المفعول به. وقد تكون العلاقات التي تربط مكونات الجملة علاقات معنوية Semantic أو علاقات نحوية Syntactic أو علاقة تصريفية Morphological.

ويوضح الشكل ٦ مثال لجملة عربية (آية قرآنية من سورة الملك) وإعرابها والعلاقات الاعتمادية بين مكوناتها<sup>[٨٦]</sup>.

---

(1) <http://nlp.stanford.edu/software/lex-parser.shtml>

(2) <http://web.mit.edu/6.863/tools/dbparser/>

المثال مأخوذ من موقع المجموعة النصية القرآنية

<http://corpus.quran.com>



في أغلب التطبيقات، ولكن قد تحتاج بعض المجالات إلى تحديد أصناف أخرى من أسماء الأعلام فمثلاً في التطبيقات الطبية تجد أصناف من مثل أسماء الجينات وأسماء البروتينات وأسماء الأمراض وأسماء الأدوية.

شكل ٧ يعرض مثالاً لجملة عربية مأخوذة من خبر صحفي، والنتيجة المتوقعة بعد تطبيق أحد الخوارزميات العامة لتمييز أسماء الكيانات على تلك الجملة.

النص	عقدت جامعة الدول العربية اجتماعاً برئاسة أحمد العربي لمناقشة أوضاع القدس
بعد تحديد أسماء الأعلام	عقدت [جامعة الدول العربية] اجتماعاً برئاسة [أحمد العربي] لمناقشة أوضاع [القدس] مكان

شكل ٧: مثال يوضح نتيجة تطبيق خوارزمية تحديد أسماء الأعلام على نص عربي.

وتحتاج العديد من التطبيقات لهذه العملية بشكل أساسي حتى لا تخطئ بين أسماء الكيانات التي ينبغي معاملتها كتعبير واحد كل الوقت، وبين الكلمات المتفرقة، ومن هذه التطبيقات أنظمة اشتقاق المعلومات -Information Extrac- tion، وأنظمة الإجابة الآلية للأسئلة Question Answering Systems ومحركات البحث Search Engines والترجمة الآلية Machine Translation.

وكغيرها من وظائف معالجة اللغات، فإن التقنيات المستخدمة لتمييز أسماء الكيانات تنقسم بشكل أساسي إلى خوارزميات تستخدم قواعد وقوانين مصاغة يدوياً تعرف الحاسوب على الأنماط النصية التي تظهر فيها أسماء الكيانات، وخوارزميات أخرى إحصائية Statistical -وهي المستخدمة الآن- وتعتمد على تقنيات تعلم الآلة. ومن أشهر التقنيات المستخدمة حالياً -Conditional Ran- dom Fields إلا أنها تحتاج إلى توفر كميات كبيرة من النصوص التي يتم تحديد أسماء الكيانات فيها يدوياً لاستخدامها لتدريب الخوارزمية. ومن الطرق التي



أظهرت جدوى وهي أيضاً تحتاج إلى كمية كبيرة من النصوص -ولكن كمية الجهد اليدوي اللازم فيها أقل- تلك الطرق التراكمية التي تبدأ بمعرفة كمية قليلة من أسماء الكيانات، وتتعلم الأنماط التي تظهر فيها بشكل متكرر، ثم تستخدم هذه الأنماط للتعرف على المزيد من أسماء الكيانات، ثم منها تتعرف على مزيد من الأنماط وهكذا. وقد لخص البحث الاستقصائي المشار إليه في [٩٢] أبرز الطرق المستخدمة لتحديد أسماء الكيانات وقارن بين تلك الطرق، وعلى الرغم من كون البحث منشوراً في ٢٠٠٧ إلا أن تقنيات تحديد أسماء الكيانات كانت قد وصلت في ذلك الوقت إلى درجة عالية جداً من الدقة، ولم تشهد أي تطور كبير خلال الفترة التالية.

أما في اللغة العربية فقد ظهر الاهتمام بهذا الموضوع متأخراً بعض الشيء، وقد رصد الاستقصاء المشار إليه في [٩٣] أبرز ما يخص اللغة العربية في هذا الموضوع من تحديات، ولخص أهم الطرق المستخدمة فيه وعرض تقييماً لكل منها.

وتعتبر عملية تمييز أسماء الكيانات في اللغة العربية أصعب نسبياً من اللغات الأخرى، ومن أسباب ذلك مثلاً أن اللغة العربية لا يوجد في طريقة كتابتها أي علامة تميز أسماء الكيانات على خلاف اللغة الإنجليزية مثلاً والتي تكتب فيها أسماء الكيانات مبدوءةً بأحرف كبيرة Capital Letters، وهي خاصية من أهم الخصائص التي تعطيها خوارزميات التعلم الآلي وزناً كبيراً بين الخصائص اللغوية الأخرى. ومنها كذلك الثراء الصريفي للغة العربية والذي يسمح بدخول الزوائد على أسماء الكيانات وليس على الكلمات العادية فقط، مثل دخول الباء على "مكة" فتصبح "بمكة" أو دخول الناء على "القدس" وإزالة أل التعريف منها فتصبح "قدسنا"، وهكذا. ومن التحديات أيضاً ندرة الموارد ذات الجودة العالية التي تحتوي قوائم بالأسماء كما يوجد في اللغات الأخرى، ومنها كذلك غياب التشكيل عن معظم النصوص العربية وينتج عنه غموض في كثير من الكلمات يجعل

عملية الاعتماد على السياق لحل الغموض عملية مهمة وحساسة، وتحديات أخرى مشروحة بالتفصيل في [٩٣].

ومن الطرق التي اقترحها الباحثون لتمييز أسماء الكيانات في اللغة العربية ما طرح في [٩٤-٩٥] وهي تعتمد على مجموعة من القواعد المصاغة يدوياً، وما هو مطروح في [٩٦-٩٨] وهي طرق تعتمد على تقنيات تعلم الآلة.

### وظائف أخرى

بالإضافة إلى الوظائف التي استعرضناها في هذا الفصل هناك وظائف أخرى يتكرر الاحتياج إليها في معالجة اللغات، ونذكرها هنا أسماءها مع ترك إشارات لبعض المراجع المفيدة فيها. ومنها تحديد الإشارات المتعددة للشيء الواحد Co-reference resolution (ارجع للاستقصاء [٩٩])، وتحليل الخطاب Discourse Analysis (ارجع لكتاب [١٠٠])، وتمييز وتقسيم موضوعات النص Topic Recognition & Segmentation [١٠١-١٠٢]، وحل التباس معاني الكلمات Word Sense Disambiguation (ارجع للاستقصاء [١٠٣])، وغيرها. وهذه الوظائف المشار إليها هنا لم تتل حظاً من الاهتمام البحثي فيما يتعلق باللغة العربية وهي بلا شك مجالات هامة ولازمة للغة العربية وتحتاج إلى من يبادر ويتصدر للبحث فيها.

### ٣. أشهر تطبيقات معالجة اللغات

وأخيراً، نستعرض هنا أشهر تطبيقات معالجة اللغات التي شهدت نشاطاً بحثياً في اللغة العربية. وتستخدم هذه التطبيقات في أغلب خطوات تنفيذها الوظائف التي استعرضناها في الباب السابق.

#### الترجمة الآلية Machine Translation

كما ذكرنا في بداية هذا البحث، فقد كانت الترجمة الآلية من أوائل تطبيقات معالجة اللغات التي لاقت اهتماماً كبيراً في الستينات وأوائل السبعينات من القرن

الماضي، ثم خفت الاهتمام بها بسبب النتائج المخيبة للآمال في ذلك الوقت، ثم عادت الأبحاث في هذا الموضوع لتنتعش في التسعينات من القرن الماضي وحتى يومنا هذا من قرننا الحالي. ويعتبر مجال الترجمة الآلية من المجالات البحثية الساخنة جداً في الوقت الحالي، وقد حققت نجاحات وقفزات كبيرة وملموسة، ولم تعد تلك التقنيات مقصورة على البحث العلمي بل أصبحت من المنتجات المهمة في الصناعة، وتوجد في الأسواق خدمات ترجمة آلية تقوم على تطويرها والترويج لها شركات كبيرة مثل Bing Translator من مايكروسوفت و Google Translate وغيرها.

ويعود الفضل في تحقيق القفزة الكبيرة في مجال الترجمة الآلية إلى بروز وانتشار الطرق الإحصائية Statistical Methods ، فبعد أن كانت عملية الترجمة تتطلب فهماً عميقاً لقواعد ومعاني وتراكيب اللغة المصدر واللغة الهدف وصياغة قواعد وقوانين معقدة لترجمة النص، فإن الطرق الإحصائية أتاحت ترك استنباط هذه القوانين والقواعد لخوارزميات التعلم الآلي، التي تستنبط هذه القوانين من خلال معالجة عدد كبير من المجموعات النصية المتوازية Parallel Corpora بعد إجراء عملية ربط لكل جملة في المستند المكتوب باللغة المصدر إلى الجملة المقابلة لها في اللغة الهدف، وفعل ذلك الربط أيضاً على مستوى الكلمات. وقد سهل انتشار الانترنت وتقنيات النشر الالكتروني كثيراً من إمكانية توفير مثل هذه المجموعات النصية، ومن أشهر المصادر لهذه المجموعات وثائق الأمم المتحدة التي تتم ترجمتها أولاً بأول إلى كل لغات العالم الأساسية، وكذلك ترجمات الكتب، وكذلك المواقع الالكترونية التي تحتوي صفحات متعددة اللغات، وغيرها.

وكحال غيرها من تطبيقات معالجة اللغات فإن الفترة الأخيرة شهدت ميلاً للجمع بين الطرق المعتمدة على القواعد المصاغة يدوياً، والطرق المعتمدة على تعلم الآلة، بحيث تصاغ بعض القواعد الضابطة التي تقيد وتحكم استخدام نتائج التعلم الآلي بهدف منع الوقوع في أخطاء يسهل ضبطها بالقوانين اللغوية.

وقد شهدت الترجمة الآلية من وإلى اللغة العربية اهتماماً متزايداً مؤخراً ومن أمثلة ذلك: ترجمة عربي-إنجليزي [١٠٤-١٠٦] وعربي-فرنسي [١٠٧]، وعربي-صيني [١٠٨].

### التلخيص الآلي Automatic Summarization

تهدف عمليات التلخيص الآلي إلى اختصار الكلام، وتقليل كمية النص إلى مقدار محدد يختاره المستخدم أو يفرضه التطبيق مع الحفاظ على أهم المعاني والمفاهيم الواردة في النص. ويفرق الباحثون بين نوعين من التلخيص الآلي: أولاً: الاختصار الاقتباسي Extractive Summarization وفيه تكون جمل الملخص مقتطعة كما هي من نص المصدر، أي إن عملية التلخيص في هذه الحالة ما هي إلا عملية ترتيب للجمل حسب أهميتها ومقدار ما تحتويه من معلومات هامة وتغطي جوانب الموضوع. ثانياً: الاختصار الخلاصي Abstractive Summarization وفيه قد تكون جمل الملخص مولدة آلياً أو ناتجة عن إحداث تغييرات وعمليات حذف أو إضافة أو تعديل على الجمل القادمة من النص المصدر، وهي أصعب بمراحل من التلخيص الاقتطاعي لأنها تتضمن عملية توليد الجمل، وتحتاج إلى توفير سبل ضمان رصانتها وتماسكها وسلامتها النحوية وصحة معانيها وعدم تناقضها مع المعلومات الواردة في المصدر.

والتلخيص الآلي -لاسيما التلخيص الاقتطاعي- واحد من الموضوعات التي شهدت قدراً من الاهتمام في اللغة العربية، وجاءت أغلب الأبحاث في الموضوع لتطبق الطرق المستخدمة في اللغات الأخرى مع إضافة بعض عمليات المعالجة التي تخص اللغة العربية، ومن أمثلة الأبحاث التي تناولت التلخيص العربي [١٠٩-١١٢].

### تصنيف النصوص Text Classification

تقوم أنظمة تصنيف النصوص بتحليل النص وتصنيفه بحسب موضوعه أو محتواه إلى أصناف محددة تختلف من تطبيق لآخر، ومن أمثلة ذلك تصنيف الأخبار

بحسب موضوعها إلى أخبار سياسية، أو رياضية، أو علمية، إلخ. وتعتمد أغلب هذه الأنظمة اليوم على تقنيات تعلم الآلة المخصصة لعمليات التصنيف-Classification، وتستخدم في ذلك مجموعة من الخصائص من أشهرها مجموعات الكلمات n-grams المفردة والمزدوجة والثلاثية (و غالباً يتم ذلك بعد تجذيع الكلمات-Stemming)، وخصائص أخرى تختلف بحسب التطبيق.

و تصنيف النصوص العربية هو أيضاً أحد الموضوعات التي لاقت اهتماماً من قبل عدد من الباحثين [١١٣-١١٥].

### تحليل المشاعر واستكشاف الآراء Sentiment Analysis

وهي عملية يتم فيها تحليل المحتوى النصي بهدف الكشف عما يحمله من مشاعر سلبية أو إيجابية أو محايدة، وتستخدم هذه العمليات بشكل كبير في تحليل المراجعات التي يكتبها الناس على الإنترنت تعليقاً على ما يشترونه من منتجات، أو تلك التي يعلق بها الناس على ما يشاهدونه من أفلام مصورة أو مقاطع صوتية، وكذلك في قياس الرأي العام تجاه القضايا العامة والقضايا السياسية، وتفيد هذه العمليات صناع القرار في الشركات أو الدوائر الرسمية لاتخاذ ما يلزم من قرارات وسياسات. وقد ازداد الاهتمام بهذا الموضوع بشكل كبير مع انتشار الإنترنت، ومع ثورة شبكات التواصل الاجتماعي تحديداً، حيث أصبحت هذه الشبكات مليئة بالنص المحمل بالمشاعر والآراء التي يمكن لصناع القرار الوصول إليها وتحليلها والخروج باستنتاجات من خلالها.

وتعتمد هذه الأنظمة على تحديد الكلمات والتعبيرات السلبية والإيجابية في اللغة فمثلاً كلمة "سعيد" هي كلمة إيجابية وفي المقابل كلمة "حزين" هي كلمة سلبية. وقد تعتمد قطبية الكلمات على الموضوع فمثلاً عندما يوصف سيناريو فيلم ما بأنه "غير متوقع" فهذا قد يكون مؤشر إيجابي يدل على أنه مشوق، في المقابل عندما يوصف مقود التحكم في السيارة بأن "غير متوقع" فهذا تعبير سلبي يدل على أنه خطير ويصعب التحكم فيه بدقة عند القيادة.

وقد ظهر اهتمام كبير مؤخراً بتطبيقات تحليل المشاعر واستكشاف الآراء في اللغة العربية وظهرت العديد من البحوث التي اقترحت آليات لبناء قوائم بالتعابير الإيجابية والسلبية المستخدمة في اللغة العربية [١١٦-١١٧]، وأخرى تناولت تحليل نصوص شبكات التواصل الاجتماعي كمنتديات الحوار ومواقع التدوين القصير لقياس الرأي العام [١١٨-١٢٠].



## خاتمة

عرّفنا في هذا البحث الاستقصائي بتقنيات معالجة اللغات الطبيعية، واستعرضنا أهم الوظائف الأساسية التي يلزم إجراؤها على النصوص في أغلب تطبيقات معالجة اللغات، وأوضحنا ما للغة العربية من خصوصيات في كل من تلك الوظائف مع ترك إشارات مرجعية للبحوث التي تناولت كل وظيفة حتى يعود إليها القارئ المهتم بمعرفة التفاصيل. كما أشرنا إلى العديد من الوظائف الأخرى التي لم يتطرق إليها البحث في اللغة العربية فيما نعلم وتحتاج إلى مبادرة الباحثين لطرق بابها. ثم عرضنا نماذج لتطبيقات مهمة لتقنيات معالجة اللغات لاسيما تلك التي اهتم بها الباحثون في اللغة العربية، وتركنا كذلك إشارات مرجعية لتلك الأبحاث التي يمكن أن يرجع إليها القارئ المهتم بالتفاصيل.





## المراجع

1. S. J. a. N. P. Russell. Artificial Intelligence: A Modern Approach. Pearson Education. 2003.
2. A. M. Turing. "Computing machinery and intelligence." Mind. pp. 433-460. 1950.
3. A. C. I. a. A. V. Saygin. "Turing Test: 50 Years Later." Minds and Machines . vol. 10. no. 4. 2000.
4. 701"Translator." IBM. 8 January 1954. [Online]. Available: [http://www-03.ibm.com/ibm/history/exhibits/701/701\\_\\_translator.html](http://www-03.ibm.com/ibm/history/exhibits/701/701__translator.html). [Accessed 3 December 2014.]
5. J. Hutchins. "The first public demonstration of machine translation: the Georgetown-IBM system. 7th January 1954." in AMTA conference. 2004.
6. W. Lehnert. "WHAT MAKES SAM RUN? SCRIPT BASED TECHNIQUES FOR QUESTION ANSWERING." in Theoretical Issues in Natural Language Processing; Supplement. 1975.
7. K. R. McKeown. "Paraphrasing Using Given and New Information in a Question-Answer System." in 17th Annual Meeting of the Association for Computational Linguistics. 1979.

8. L. Karttunen. "DISCOURSE REFERENTS." in INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS COLING 1969: Preprint No. 60. 1969.
9. J. H. Clippinger. "SPEAKING WITH MANYTONGUES: SOME PROBLEMS IN MODELING SPEAKERS OF ACTUAL DISCOURSE." in Theoretical Issues in Natural Language Processing. 1975.
10. S. K. a. R. Simmons. "A computational approach to grammatical coding of English Words." JACM. vol. 10. p. 334-337. 1963.
11. B. B. G. a. G. M. Rubin. "Automated grammatical tagging of English." Department of Linguistics. Brown University. 1971.
12. K. W. Church. "A stochastic parts program and noun phrase parser for unrestricted text." in The second conference on Applied natural language processing. 1988.
13. S. J. DeRose. "Grammatical category disambiguation by statistical optimization." Computational Linguistics. vol. 14. no. 1. pp. 31-39. 1988.
14. J. Tait. "Automatic summarization of english texts." PhD thesis. University of Cambridge. Cambridge. UK. 1983.
15. G. a. M. M. J. Salton. Introduction to modern information retrieval. McGill . 1983.

16. P. Turney. "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." in The 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia. Pennsylvania. 2002.
17. L. L. a. S. V. Bo Pang. "Thumbs up? Sentiment classification using machine learning techniques." in EMNLP. 2002.
18. R. Al-Shalabi and M. Evens. "A Computational Morphology System for Arabic." in Workshop on Computational Approaches to Semitic Languages. 1998.
19. M. Smets. "Paradigmatic Treatment of Arabic Morphology." in Workshop on Computational Approaches to Semitic Languages. 1998.
20. K. R. Beesley. "Arabic Finite-State Morphological Analysis and Generation." in COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. 1996.
21. K. R. Beesley. "Consonant Spreading in Arabic Stems." in 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Volume 1. 1998.
22. H. El-Shishiny. "A Formal Description of Arabic Syntax in Definite Clause Grammar." in COLNG 1990 Volume

3. Papers presented to the 13th International Conference on Computational Linguistics. 1990.
23. R. a. M. E. a. D. J. a. S. D. a. M. S. a. S. R. a. M. J. a. Z. F. O. a. C.-B. C. Zbib. "Machine Translation of Arabic Dialects." in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012.
24. W. Salloum and N. Habash. "Elissa: A Dialectal to Standard Arabic Machine Translation System." in Proceedings of COLING 2012: Demonstration Papers.
25. A. Ittycheriah and S. Roukos. "A Maximum Entropy Word Aligner for Arabic-English Machine Translation." in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005.
26. A. Azmi. "Ikhtasir — A user selected compression ratio Arabic text summarization system." in International Conference on Natural Language Processing and Knowledge Engineering. 2009. NLP-KE 2009.. 2009.
27. J. D. Schlesinger. D. P. O'Leary and J. M. Conroy. "Arabic/English Multi-document Summarization with CLASSY - The Past and the Future." Computational Linguistics and Intelligent Text Processing. pp. 568-581. 2008.

28. G. Badaro, R. Baly, H. Hajj, N. Habash and W. El-Hajj. "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining." in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014.
29. A. Abu-Jbara, B. King, M. Diab and D. Radev. "Identifying Opinion Subgroups in Arabic Online Discussions." in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013.
30. D. Vergyri and K. Kirchhoff. "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition." in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004.
31. D. Vergyri and K. Kirchhoff. "Arabic Diacritization Using Weighted Finite-State Transducers." in Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004.
32. K. Shaalan, H. M. Abo Bakr and I. Ziedan. "A Hybrid Approach for Building Arabic Diacritizer." in Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, 2009.
33. N. Habash, O. Rambow and G. Kiraz. "Morphological Analysis and Generation for Arabic Dialects." in Pro-

- ceedings of the ACL Workshop on Computational Approaches to Semitic Languages. 2005.
34. F. O. Zaidan and C. Callison-Burch. "Arabic Dialect Identification." *Computational Linguistics*. vol. 40. no. 1. 2014.
  35. E. Mohamed. B. Mohit and K. Oflazer. "Transforming Standard Arabic to Colloquial Arabic." in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012.
  36. N. Habash. M. Diab and O. Rambow. "Conventional Orthography for Dialectal Arabic." in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. 2012.
  37. N. Habash. *Introduction To Arabic Natural Language Processing*. 2010.
  38. A. Farghaly and K. Shaalan. "Arabic Natural Language Processing: Challenges and Solutions." *ACM Transactions on Asian Language Information Processing (TALIP)*. vol. 8. no. 4. 2009.
  39. A. a. H. N. El Kholy. "Orthographic and morphological processing for English--Arabic statistical machine translation." *Machine Translation*. vol. 26. pp. 25-45. 2012.

40. Y. a. I. F. Elarian. "A Lexicon of Connected Components for Arabic Optical Text Recognition." in First International Workshop on Frontiers in Arabic Handwriting Recognition. Istanbul. Turkey. 2011.
41. T. Buckwalter. "Buckwalter Arabic Morphological Analyzer Version 1.0." Linguistic Data Consortium (LDC). 2002.
42. N. a. S. A. a. B. T. Habash. "On Arabic Transliteration." in Arabic computational morphology. Springer. 2007. pp. 15-22.
43. A. Roberts. "Buckwalter2Unicode." 15 September 2004. [Online]. Available: <http://www.redhat.com/archives/fedora-extras-commits/2007-June/msg03617.html>. [Accessed 4 12 2014.]
44. T. Q. A. Corpus. "Java API - Buckwalter Transliteration." [Online]. Available: <http://corpus.quran.com/java/buckwalter.jsp>. [Accessed 5 12 2014.]
45. M. A. Attia. "Arabic Tokenization System." in The 5th Workshop on Important Unresolved Matters. Prague. Czech Republic. 2007.
46. N. a. S. F. Habash. "Arabic preprocessing schemes for statistical machine translation." in NAACL-HLT. New York City. NY. USA. 2006.



47. N. a. R. O. a. R. R. Habash. "Mada+ token: A toolkit for arabic tokenization. diacritization. morphological disambiguation. pos tagging. stemming and lemmatization." in The 2nd International Conference on Arabic Language Resources and Tools (MEDAR). Cairo. Egypt. 2009.
48. N. a. R. O. a. R. R. Habash. "Morphological Tagging for Arabic." [Online]. Available: [http://www1.cs.columbia.edu/~rambow/software-downloads/MADA\\_\\_Distribution.html](http://www1.cs.columbia.edu/~rambow/software-downloads/MADA__Distribution.html). [Accessed 5 12 2014.].
49. M. Diab. "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization. POS tagging. and base phrase chunking." in 2nd International Conference on Arabic Language Resources and Tools. 2009.
50. M. Diab. "AMIRA 2.1." 25 10 2011. [Online]. Available: [http://www.ibridgenetwork.org/columbia/ir\\_\\_ms-12s-2-1](http://www.ibridgenetwork.org/columbia/ir__ms-12s-2-1). [Accessed 5 12 2014.].
51. S. R. Anderson. A-morphous morphology. Cambridge University Press. 1992.
52. I. A. a. A.-K. I. A. Al-Sughaiyer. "Arabic morphological analysis techniques: A comprehensive survey." Journal of the American Society for Information Science and Technology. vol. 55. no. 3. pp. 189-213. 2004.

53. K. a. H. N. Dukes. "Morphological Annotation of Quranic Arabic." in European Languages Resources Association (ELRA). Valletta. Malta. 2010.
54. N. a. E. R. a. H. A. Habash. "A Morphological Analyzer for Egyptian Arabic." in Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology. Montreal. Canada. 2012.
55. M. El-Affendi. "An algebraic algorithm for Arabic morphological analysis." The Arabian Journal for Science and Engineering, vol. 16, no. 4B, p. 605– 611, 1991.
56. S. a. A.-A. F. Al-Fedaghi. "A new algorithm to generate root-pattern forms." in In Proceedings of the 11th National Computer Conference , Dhahran. Saudi Arabia. 1989.
57. R. Al-Shalabi. "Design and implementation of an Arabic morphological system to support natural language processing." Doctoral dissertation. Illinois Institute of Technology. Chicago. IL. 1996.
58. I. a. A.-S. I. Al-Kharashi. "Rule merging in a rule based Arabic stemmer." in In Proceedings of the 19th International Conference on Computational Linguistics (COLING-02). Taipei. Taiwan. 2002.

59. T. Buckwalter. "Buckwalter Arabic Morphological Analyzer Version 2.0." Linguistic Data Consortium. 2004. [Online]. Available: <https://catalog.ldc.upenn.edu/LD-C2004L02>. [Accessed 5 12 2014.]
60. D. G. B. B. S. K. A. B. S. K. Mohamed Maamouri. "LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1." Linguistic Data Consortium. 2010. [Online]. Available: <https://catalog.ldc.upenn.edu/LD-C2010L01>. [Accessed 5 12 2014.]
61. O. a. B. V. a. B. T. Smrž. "ElixirFM." [Online]. Available: <http://quest.ms.mff.cuni.cz/cgi-bin/elixir/index.fcgi?mode=home>. [Accessed 5 12 2014.]
62. O. a. B. V. a. B. T. Smrž. "ElixirFM: Functional Arabic Morphology," 3 7 2013. [Online]. Available: <http://sourceforge.net/projects/elixir-fm/>. [Accessed 6 12 2014.]
63. D. A. Hull. "Stemming algorithms: A case study for detailed evaluation." JASIS. vol. 47, no. 1, pp. 70-84. 1996.
64. W. a. P. R. Kraaij. "Viewing stemming as recall enhancement." in ACM SIGIR96. 1996.
65. P. Willett. "The Porter stemming algorithm: then and now." Electronic library and information systems. vol. 40, no. 3, pp. 219-223. 2006.

66. F. C. L. M. F. a. W. P. Ekmekcioglu. "Stemming and n-gram matching for term conflation in Turkish texts." Information Research News. vol. 7. no. 1. pp. 2-6. 1996.
- .76 "Corpus-based stemming using cooccurrence of word variants." ACM Transactions on Information Systems. vol. 16. no. 1. pp. 61-81. 1998.
68. I. A. a. E. M. W. Al-Kharashi. "Comparing words, stems, and roots as index terms in an Arabic information retrieval system." Journal of the American Society for Information Science. vol. 45. no. 8. pp. 548-560. 1994.
69. T. Buckwalter. "Qamus: Arabic lexicography." [Online]. Available: <http://www.qamus.org/>
70. A. N. a. A.-F. De Roeck. "A morphologically sensitive clustering algorithm for identifying Arabic roots." in ACL 2000. Hong Kong. 2000.
71. L. S. a. B. L. a. C. M. E. Larkey. "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis." in 25th annual international ACM SIGIR conference on Research and development in information retrieval. 2002.
72. S. Khoja. "Arabic stemmer." [Online]. Available: <http://zeus.cs.pacificu.edu/shereen/research.htm#stemming>. [Accessed 6 12 2014.]

73. M. a. H. K. a. J. D. Diab. "Automatic tagging of Arabic text: From raw text to base phrase chunks." in HLT-NAACL 2004: Short Papers. 2004.
74. N. a. R. R. M. Habash. "Catib: The columbia arabic treebank." in The ACL-IJCNLP 2009 Conference Short Papers. 2009.
75. S. a. G. R. a. K. G. Khoja. "An Arabic Tagset for the Morphosyntactic Tagging of Arabic." Lancaster University. 2001.
76. O. a. Z. P. Smrz. "Sherds from an arabic treebanking mosaic." Prague Bulletin of Mathematical Linguistics. 2002.
77. K. W. Church. "A stochastic parts program and noun phrase parser for unrestricted text." in The second conference on Applied natural language processing. Stroudsburg, PA. 1988.
78. S. J. DeRose. "Grammatical category disambiguation by statistical optimization." Computational Linguistics. vol. 14. no. 1. p. 31-39. 1988.
79. J. a. M. A. a. P. F. C. Lafferty. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." in 18th International Conference on Machine Learning. 2001.
80. K. a. K. D. a. M. C. D. a. S. Y. Toutanova. "Feature-rich part-of-speech tagging with a cyclic dependency net-

- work." in the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 2003.
81. I. Zitouni. J. S. Sorensen and R. Sarikaya. "Maximum Entropy Based Restoration of Arabic Diacritics." in Workshop on Computational Approaches to Semitic Languages. Sydney-Australia. 2006.
  82. K. a. E. G. Adali. "Vowel and Diacritic Restoration for Social Media Texts." in Workshop on Language Analysis for Social Media (LASM) , Gothenburg. Sweden. 2014.
  83. A. a. R. M. a. M. R. H. a. R. A. Al Sallab. "Automatic Arabic diacritics restoration based on deep nets." in EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). Doha. Qatar. 2014.
  84. M. A. a. A. S. A. A. a. R. H. M. a. R. A. Rashwan. "Automatic Arabic diacritics restoration based on deep nets." in ANLP. 2014.
  85. S. a. M. H. a. A. M. S. Alqrainy. "Context-Free Grammar Analysis for Arabic Sentences." International Journal of Computer Applications (0975 - 8887), vol. 53, no. 3, pp. 7 - 11. 2012.
  86. D. M. Bikel. "Intricacies of Collins' parsing model." Computational Linguistics. vol. 30. p. 479-511. 2004.

87. Y. a. H. N. a. R. O. Marton. "Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features." *Computational Linguistics*. 2013.
88. S. Tratz. "A Cross-Task Flexible Transition Model for Arabic Tokenization. Affix Detection. Affix Labeling. POS Tagging. and Dependency Parsing." in *The Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*. 2013.
89. E. Mohamed. "The Effect of Automatic Tokenization. Vocalization. Stemming. and POS Tagging on Arabic Dependency Parsing." in *The Fifteenth Conference on Computational Natural Language Learning*. Portland. Oregon. USA. 2011.
90. a. A. E. a. M. S. A.-B. Dukes. "Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank." in *the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta. Malta. 2010.
91. A. F. a. S. N. A. a. X. E. P. Martins. "Concise integer linear programming formulations for dependency parsing." in *The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009.

92. D. a. S. S. Nadeau. "A survey of named entity recognition and classification." *Lingvisticae Investigationes*. vol. 30. no. 1. pp. 3 - 26. 2007.
93. K. Shaalan. "A Survey of Arabic Named Entity Recognition and Classification." *Computational Linguistics*. vol. 40. no. 2. 2013.
94. K. a. H. R. Shaalan. "Person name entity recognition for Arabic." in *Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Semitic 2007. Stroudsburg, PA. 2007.
95. K. a. H. R. Shaalan. "Arabic named entity recognition from diverse text types." *Advances in Natural Language Processing*. vol. 5221. 2008.
96. Y. M. D. a. P. R. Benajiba. "Arabic named entity recognition: An SVM-based approach." in *Arab International Conference on Information Technology*. Hammamet. 2008.
97. Y. M. D. a. P. R. Benajiba. "Arabic named entity recognition using optimized feature sets." in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*. Stroudsburg, PA. 2008.
98. Y. a. P. R. 2. Benajiba. "ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag Information." in



- Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence, Mumbai, 2007.
99. P. Elango. "Coreference resolution: A survey." University of Wisconsin. Madison, WI. 2005.
  100. D. a. T. D. a. H. H. E. Schiffrin. The handbook of discourse analysis. John Wiley & Sons. 2008.
  101. J. a. R. C. Arguello. "Topic segmentation of dialogue." in HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech. 2006.
  102. J. C. Reynar. "Topic segmentation: Algorithms and applications." University of Pennsylvania. 1998.
  103. R. Navigli. "Word sense disambiguation: A survey." ACM Computing Surveys (CSUR), vol. 41, no. 2. 2009.
  104. I. a. Z. R. a. G. J. Badr. "Segmentation for English-to-Arabic statistical machine translation." in 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. 2008.
  105. I. a. Z. R. a. G. J. Badr. "Syntactic phrase reordering for English-to-Arabic statistical machine translation." in 12th Conference of the European Chapter of the Association for Computational Linguistics. 2009.

106. H. a. L. A. Al-Haj. "The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation." Machine translation. vol. 26. 2012.
107. S. a. E. I. A. a. N. H. Hasan. "Creating a large-scale Arabic to French statistical machine translation system." in Fifth Int. Conf. on Language Resources and Evaluation (LREC). 2006.
108. N. a. H. J. Habash. "Improving Arabic-Chinese statistical machine translation using English as pivot language." in Proceedings of the Fourth Workshop on Statistical Machine Translation. 2009.
109. H. a. N. O. a. B. P. Oufaida. "Multilingual Summarization Experiments on English, Arabic and French (R)sum) Automatique Multilingue Expérimentations sur l'Anglais, l'Arabe et le Français [in French." in Proceedings of TALN 2014 (Volume 2: Short Papers). Marseille. France. 2014.
110. L. a. F. C. a. E.-H. M. a. G. G. Li. "ulti-document multilingual summarization corpus preparation. Part 1: Arabic, English, Greek, Chinese, Romanian." in MultiLing 2013 Workshop on Multilingual Multi-document Summarization. Sofia. Bulgaria. 2013.
111. F. S. a. L. G. Douzidia. "Lakhas, an Arabic summarization system." in DUC2004. 2004.

112. J. D. a. O. D. P. a. C. J. M. Schlesinger. "Arabic/English multi-document summarization with CLASSY—the past and the future." in *Computational Linguistics and Intelligent Text Processing*. Springer. 2008. pp. 568--581.
113. L. Khreisat. "Arabic text classification using N-gram frequency statistics a comparative study." in *Conference on Data Mining| DMIN-06*. 2006.
114. A. El-Halees. "Arabic text classification using maximum entropy." *The Islamic University Journal*. vol. 15. no. 1. pp. 157--167. 2007.
115. S. A. A. A.-T. ., A. K. M. S. a. A.-R. Al-Harbi. "Automatic Arabic text classification." in *The 9th International Conference on the Statistical Analysis of Textual Data*. Lyon. France. 2008.
116. G. a. B. R. a. H. H. a. H. N. a. E.-H. W. Badaro. "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining." in *The EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Doha. Qatar. 2014.
117. M. a. D. M. Abdul-Mageed. "SANA: A Large Scale Multi-Genre. Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis." in *The Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. Reykjavik. Iceland. 2014.

118. E. a. R. V. Refaee. "Evaluating Distant Supervision for Subjectivity and Sentiment Analysis on Arabic Twitter Feeds." in EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). Doha, Qatar. 2014.
119. A. a. K. B. a. D. M. a. R. D. Abu-Jbara. "Identifying Opinion Subgroups in Arabic Online Discussions." in the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Sofia, Bulgaria. 2013.
120. A. a. D. K. Mourad. "Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs." in the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Atlanta, Georgia. 2013.
- .121 "ARNE: A tool for named entity recognition from Arabic text." in Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4). San Diego, CA. 2012.
122. A. Andreyewsky. "SUBCLASSIFICATION OF PARTS OF SPEECH IN RUSSIAN: VERBS." in COLING. 1965.



## استخدام القواعد الاستنباطية في تحسين أداء التشكيل الآلي

أ. عمر السيد شعبان<sup>(\*)</sup>

omar.s.shaaban@gmail.com

---

(\*) باحث في مجال الحوسبة العربية، درس الهندسة البرمجية في جامعة الملك فهد للبترول والمعادن وحصل منها على درجة البكالوريوس عام ٢٠٠٩. حصل على الماجستير من نفس الجامعة عام ٢٠١٣ في تخصص علوم الحاسب الآلي، وكانت رسالة الماجستير عن التشكيل الآلي في اللغة العربية. من اهتماماته الأخرى التعرف الآلي على الكلام، والتصحيح الآلي للنصوص.



## ملخص (١)

يُعتبر التشكيل الآلي في اللغة العربية أحد أهم المباحث في مجال الحوسبة العربية، حيث يُمثل مرتكزاً رئيساً لمباحث أخرى في نفس المجال كالتعرف الآلي على الكلام العربي، والنطق الآلي، وفي مرحلة تحليل كلام اللغة المصدر للترجمة الآلية. كما يُمكن الاستفادة منه من قبل دارسي اللغة العربية حيث يُعتبر غموض معنى الكلمة غير المُشكَّلة تحدياً كبيراً لديهم وخاصة المبتدئين منهم.

قام الباحثون باتباع طرق عديدة في التشكيل الآلي أغلبها وأنجحها يعتمد على الطرق الإحصائية والتعلم الآلي. كما قام بعضهم بمحاولة اتباع الطرق النحوية التي تعتمد على قواعد اللغة العربية وإن كان نجاحهم في ذلك مقصوراً على الجمل القصيرة البسيطة. وربما كان السبب الرئيس في ذلك هو الغنى اللغوي الذي تتميز به اللغة العربية عن بقية اللغات.

في هذا البحث نُقدم طريقة تعتمد على استنباط قواعد سياقية وحرفية عن طريق استخراجها من مكانز اللغة العربية المُشكَّلة يدوياً. وقد قمنا باستخدام تلك القواعد في مرحلتين: مرحلة التنظيف وهي المرحلة التي يتم فيها استبعاد التشكيل حسب درجة نُدرته، ومرحلة التطبيق وفيها يتم تطبيق هذه القواعد على الأحرف غير المُشكَّلة في النص.

وتظهر نتائج التجارب التي قمنا بها تحسناً كبيراً في نسبة التشكيل تصل إلى ١١٪ في المتوسط، وكذلك تحسناً في نسبة الخطأ الكلمي بمتوسط ٧٪، وفي نسبة الخطأ الحرفي بمتوسط ١٠٪.

□□□ يعتمد هذا البحث على رسالة بعنوان "استرجاع التشكيل آلياً للنصوص العربية" قدمت لجامعة الملك فهد للبترول والمعادن ومنح بموجبها الباحث درجة الماجستير تحت إشراف د. حسني المحتسب.





## مقدمة

تعد معالجة اللغة الطبيعية (Natural Language Processing NLP) أحد المواضيع المهمة في علوم الحاسوب واللسانيات الحاسوبية، وقد تطورت كثيراً على مدار العقود الثلاثة الماضية. وتكمن أهمية معالجة اللغة العربية في انتشار الإنترنت والأجهزة المحمولة التي تتطلب المزيد من التفاعل الطبيعي مع المستخدمين. فبعض تطبيقات معالجة اللغة الطبيعية مخصصة لأهداف لغوية مستقلة كالمدقق والمصحح الإملائي والنحوي وبعضها تستخدم كأدوات مساعدة لتحقيق أغراض أكثر تعقيداً كالاعتراف الآلي على الكلام (Automatic Speech Recognition ASR) حيث يستطيع الشخص إطلاق أوامر صوتية أو إملاء نص للحاسوب، وكذلك مثل تحويل النص إلى كلام (Text-to-Speech TTS) حيث يقوم الحاسوب بنطق نص ما أو إعلام المستخدم بأمر ما.

## التشكيل في الكتابة العربية

تتكون الأبجدية العربية من ٢٨ حرفاً، منها ٢٥ حرف صامت، و٣ أحرف علة (صائتة) (الألف، والواو، والياء)، ويمكن للواو والياء أن يكونا ساكنين (شبه صائتة) كذلك [١].

يمكن لكل حرف صامت أن يُشكل بإحدى ١٤ طريقة كما يبينها الجدول ١، ويمكن تصنيف هذه الطرق في خمس مجموعات: الأولى تمثل الحركات البسيطة وهي الفتحة والضمّة والكسرة، والثانية تمثل حركتي التقطيع وهما السكون والشدة، والثالثة تمثل التنوين (تنوين فتح، تنوين ضم، تنوين كسر)، أما المجموعتين الرابعة والخامسة فتمثلان الحركات المركبة من الشدة مع حركة من الحركات البسيطة.

يمكن أن نقسم عملية تشكيل الكلمة في اللغة العربية إلى قسمين: تشكيل صريفي، وتشكيل إعرابي. ونعني بالتشكيل الصريفي التشكيل الداخلي للكلمة الذي لا يتغير بتغير موقع الكلمة في الجملة. أما التشكيل الإعرابي فيكون عادة في الحرف الأخير من الكلمة ويمكن أن يتغير بتغير موقع الكلمة في الجملة. ويمكن اعتبار القسم الأول العامل الأساس في تحديد معنى الجملة ككل. ويمكن عند التشكيل الآلي، استخدام الطرق الإحصائية لاستنتاج القسم الأول، أما القسم الثاني فلا يمكن استنتاجه إلا باستخدام القواعد النحوية.

#### جدول ١ تقسيم التشكيل العربي

الرقم	القسم	الحركة	مثال
١	الحركات البسيطة	الفتحة (ـَ)	بَ
٢		الضمة (ـُ)	بُ
٣		الكسرة (ـِ)	بِ
٤	حركات التقطيع	السكون (ـْ)	بْ
٥		الشدة (ـً)	بً
٦	التنوين	تنوين الفتح (ـً)	بً
٧		تنوين الضم (ـٌ)	بٌ
٨		تنوين الكسر (ـٍ)	بٍ
٩	الشدة + الحركات البسيطة	الشدة + الفتحة (ـً)	بً
١٠		الشدة + الضمة (ـٌ)	بٌ
١١		الشدة + الكسرة (ـٍ)	بٍ
١٢	الشدة + التنوين	الشدة + (ـً)	بً
١٣		الشدة + تنوين الضم (ـٌ)	بٌ
١٤		الشدة + تنوين الكسر (ـٍ)	بٍ

## تعريف المسألة

نهدف في هذا البحث إلى تطبيق التشكيل الآلي على النصوص العربية، وبمعنى أدق؛ يمكن وصف مسألة التشكيل الآلي بأنها عملية استرجاع التشكيل الناقص من النصوص غير المشكلة (أو المشكلة جزئياً). يبين مثالا على نص غير مشكل (أ)، والتشكيل المفترض بعد تطبيق التشكيل الآلي (ب).

خَيْرُ النَّاسِ أَنْفَعُهُمُ لِلنَّاسِ	خير الناس أنفعهم للناس
(ب) المخرج	(أ) المدخل

### رسم توضيحي ١ : مثال على مخرجات عملية التشكيل الآلي

التشكيل الآلي هو أحد قضايا معالجة اللغة الطبيعية التي يمكن اعتبارها مستقلة لها تطبيقاتها الخاصة، كما يمكن اعتبارها أداة مساعدة لتطبيقات أكثر تعقيداً. وتظهر هذه المسألة في اللغات السامية بشكل خاص كالعربية والعبرية، وكذلك في بعض اللغات الأخرى مثل اليونانية والكورية [٢]. وباعتبار اللغة العربية يُحذف التشكيل عادة تاركاً مهمة تحديد معنى الجملة لفطنة القارئ، وهذه ليست مشكلة للقارئ الخبير الذي يستطيع بخبرته تحديد المعنى المقصود بناء على السياق، ولكن تبرز مشكلة تقدير التشكيل للقارئ المبتدئ - وكذلك للحاسوب - حيث لا يمكن تحديد معنى الجملة بسهولة. ومن هذا المنطلق تظهر الحاجة إلى التشكيل الآلي.

### تطبيقات التشكيل الآلي

يمكن استخدام التشكيل الآلي بصورة مستقلة أو بصورة مساعدة لمسائل أخرى، فعند تطبيقه يمكن أن يقلل من غموض النص ويساعد في تحديد المعنى المقصود، وهذا مفيد بحد ذاته. كذلك فإن تشكيل النص هو أمر أساسي في مسألتنا التعرف الآلي على الكلام وتحويل النص إلى كلام والترجمة الآلية. إذ تُستخدم

الطرق الإحصائية في التعرف على الكلام على الأغلب والتي تعتمد على وجود مكنز مُشكَّل، ولأجل الحصول على هذا المكنز فإن جهود الباحثين تستنزف في بنائه بصورة يدوية، وهذا أمر شاق للغاية. أما عند تحويل النص إلى كلام فيتطلب أن يكون النص المراد نطقه مشكلاً تشكيلاً كاملاً. ولذا، فإن بناء نظام للتشكيل الآلي إضافة ضرورية لتطبيقات وبحوث الصوتيات العربية.

### أساليب التشكيل الآلي

لقد ازدادت الأبحاث المنشورة المتعلقة بالتشكيل الآلي، سواء للعربية أو غيرها، في العقد الأخير، وفي هذا الفصل نعرض آخر الأبحاث في هذا الموضوع. إذ نعرض الأبحاث المتعلقة باللغة العربية في القسم ١، ٢، بينما نعرض بعض الأبحاث المتعلقة باللغات الأخرى في القسم ٢، ٢. أما القسم ٢، ٣ فيعرض مقارنة بين الطرق المستخدمة في الأبحاث التي اطلعنا عليها.

### التشكيل الآلي في اللغة العربية

قام رشوان وآخرون [١] بعرض نظام ستوكاستيكي للتشكيل الآلي ذي الوضعين المزدوجين، حيث يقوم الوضع الأول بالبحث في قاموس يحتوي على كلمات كاملة ومشكَّلة، باستخدام البحث الشعري  $\times A$ ، وتقدير الاحتمال التسلسلي طويل المدى لأجل الحصول على أقرب تشكيل ممكن. وفي حال لم تكن الكلمة (أو سلسلة الكلمات) موجودة في القاموس (Out-Of-Vocabulary OOV)؛ يقوم الوضع الثاني بتحليل كل كلمة إلى جميع احتمالاتها الصرفية ثم يعيد البحث في القاموس لكل احتمال. يقول المؤلفون بأن النظام حقق نسبة خطأ ١، ٣٪ على مستوى الكلمات، و٥، ١٢٪ على مستوى الأحرف.

واتبع الزيتوني وساريكايا [٣] طريقة الإنتروبيا القصوى (Maximum Entropy) لاستنتاج التشكيل. وفي هذه الطريقة يمكن استخدام معلومات متعددة بصورة تكاملية كالخصائص الكلمية، والقطعية، ووسوم أجزاء الكلام. وعلى هذا؛

قام المؤلفان بتعريف مسألة التشكيل الآلي على أنها مسألة تصنيف، واستخدما مصنف MaxEnt، كما استخدما في تجاربهما المكنز LDC Arabic Treebank، وبناء عليه تم التوصل إلى درجة خطأ كالتالي: ١٧, ٢٪ على مستوى الكلمة باعتبار التشكيل الإعرابي و ٧, ٢٪ بتجاهله، درجة خطأ ١, ٥٪ على مستوى الأحرف باعتبار التشكيل الإعرابي و ٢, ٢٪ بتجاهله.

واستخدم الشافعي وآخرون [٤] نماذج ماركوف المخفية (Hidden Markov Models HMM) لحل مسألة التشكيل الآلي. تطلب هذا الأسلوب مكنزاً كبيراً من النصوص المشكلة لأجل استخراج الخصائص التي ستستخدم في النمذجة، وقد قام بعض المؤلفين باستخدام نصوص القرآن الكريم للنمذجة والاختبار. كما استعمل المؤلفون سلسلة الكلمات غير المشكلة كخصيصة، واعتبروا أن الحالة المخفية هي السلسلة ذاتها مشكّلة، وبناء على ذلك؛ توصلوا إلى نسبة خطأ ١, ٤٪، انخفضت بعد ذلك إلى ٢, ٥٪ باستخدام مرحلة ما قبل المعالجة وسلاسل ثلاثية مختارة.

وقد وصف عطية [٥] نظام ArabDiac، وهو مشكّل تجاري يُستخدم في النسخ النطقي للكلام، وقد استعمل أسلوباً يدمج بين القواعد والإحصاءات لأجل استخراج التشكيل الأقرب لنص ما. ويعمل هذا النظام في أربع مراحل: تُعاد الكلمات المختصرة والأرقام في المرحلة الأولى إلى ألفاظها، وفي المرحلة الثانية يقوم المحلل المفرداتي بتحديد التشكيل الأنسب للكلمة وكذلك الأجزاء الصرفية لها أو تحديد ما إذا كانت الكلمة معرّبة أم لا. وفي المرحلة الثالثة يقوم الواسم بتحديد وسوم أجزاء الكلام لكل كلمة غير معرّبة ومن ثم يقوم المحلل الإعرابي بتحديد التشكيل الإعرابي، أما في حالة كون الكلمة معرّبة فتستخدم الإحصاءات والقواعد الصوتية لاستنتاج أفضل تشكيل لها. وأخيراً تُحوّل النصوص المشكلة إلى النسخ النطقي في المرحلة الرابعة باستخدام قواعد خاصة آخذة في الاعتبار تأثير

نطق الكلمة على الكلمة التالية لها. حصل المؤلف على دقة ٩٧٪ (لو تضع هنا نسبة الخطأ أيضاً لتسهيل المقارنة) دون اعتبار أواخر الكلمات، و٨٨٪ (لو تضع هنا نسبة الخطأ أيضاً لتسهيل المقارنة) باعتبارها.

### التشكيل الآلي في اللغات الأخرى

بالرغم من أن أكثر الأبحاث المتعلقة بالتشكيل الآلي خاصة باللغة العربية؛ إلا إن عددًا من الباحثين قد درسوا هذه المسألة على لغات أخرى كالسنديّة والأردية، والتي تُشبه العربية في الكتابة، مما يجعلها مرتبطة بموضوع البحث بشكل كبير.

فقد قام جافيد وآخرون [٦] بدراسة مسألة التشكيل الآلي للغة السنديّة، وهي لغة يُتحدث بها في باكستان وأجزاء من الهند، فقاموا ببناء نظام يستخدم الـ WordNet والتي تخزن العلاقات الدلالية بين الكلمات، واستخدموا فيه ثلاثة مكانز: الأول سَمَوهُ CRITICAL ويستخدم للكلمات الغامضة والهامة، والثاني وقد سَمَوهُ HOMONYMY يستخدم للكلمات الهامة المتشابهة في الكتابة المختلفة في المعنى، أما الثالث WNL فيستخدم للكلمات القياسية. وبحسب تجاربهم؛ حصل المؤلفون على نسبة خطأ ٧١,٠٪ على مستوى الكلمات، و٣٩,٣٪ على مستوى الأحرف.

وركز هاريتل وآخرون [٧] على اللغات السامية وبشكل خاص اللغة السيريانية. إذ استخدموا طريقة نماذج ماركوف المشروطة (Conditional Markov Mod- els CMM) في التشكيل الآلي والتي تتطلب مكنزاً مشكلاً، واعتمدت الكلمات (أو الأجزاء) المشكلة السابقة على الكلمة في بناء تلك النماذج. وقد حقق المؤلفون نسبة خطأ ١٥٪ على مستوى الكلمة للعربية، و١٠,٥٪ للسيريانية.

كما قام علي [٨] بدراسة المسألة للغة الأردية، فبين كيف قام ببناء نظام للتشكيل الآلي قد اعتمد على مُعجم ومكنز مشكل وموسوم يدوياً. ويمكن تلخيص

طريقة عمله كما يلي: أولاً؛ تُزال علامات التشكيل من النص قبل معالجته، ومن ثم يقوم الواسم المعتمد على HMM بوسم الكلمات، ثم يتم البحث عن الكلمة ووسمها في المعجم لكي يتم الحصول على الكلمة مشكّلة، وإذا لم يُعثر على الكلمة يتم تحليل الكلمة باستخدام القواعد وإلا فتُستخدم الإحصاءات لتشكيلها. وقد وصلت دقة هذا النظام إلى ٩٥٪ (لو تضع هنا نسبة الخطأ أيضاً لتسهل المقارنة) بحسب ما ذكر المؤلف.

### قواعد التشكيل الاستنباطية

يتميز التشكيل في اللغة العربية بدرجة كبيرة من التعقيد، إلا إن المتأمل يمكنه أن يلاحظ الكثير من القواعد والأنماط المتكررة التي تُتبع في التشكيل إلى حد كبير، ومن ذلك - على سبيل المثال - أن حرف التاء المربوطة، الذي يكون دائماً في نهاية الكلمة، لا بُدَّ أن تسبقه حركة الفتح كما في "حَمَزَة" و"رَغْبَة"، وقس على هذا كثير من القواعد التي يمكن استنباطها عن طريق الاستقراء.

### الخصائص

للتعرف على قواعد التشكيل الاستنباطية لا بد من تحديد الخصائص التي تتكون منها هذه القواعد، فقاعدة التاء المربوطة السابقة لا تتطلب سوى معرفة أن الحرف التالي هو تاء مربوطة. ولكن يمكن للقاعدة أن تحتوي على خصائص أكثر من ذلك بكثير. لهذا قمنا بتحديد ١٦ خاصية متعلقة بالحرف الحالي وما قبله وما بعده، والكلمة الحالية وما قبلها وما بعدها، وخصائص أخرى كطول الكلمة وهي كالتالي:

١. الحرف الحالي: ويمكن أن يحتوي كل الأحرف العربية بجميع أشكالها وهي ٣٦ شكلاً، وهي الأحرف الثمان والعشرون مضافاً إليها الهمزات بأشكالها الست، والألف المقصورة، والتاء المربوطة.



٢. موقع الحرف الحالي: ويحسب الموقع من بداية الكلمة ويأخذ قيماً من ١ إلى طول الكلمة، دون أخذ حركات التشكيل والكشيدة (حرف التطويل) في الاعتبار.
٣. الحرفان السابقان للحرف الحالي، والحرفان التاليان له: حيث يمكنهم - جميعاً - أن يكون أحد الأشكال المذكورة في النقطة ١، بالإضافة إلى القيمة الفارغة N إذا كان الحرف غير متوفر.
٤. حركات التشكيل للحرفين السابقين والتاليين (إن توفرت): وتأخذ أحد ١٤ قيمة ممكنة أو القيمة الفارغة N إن لم تتوفر.
٥. الكلمة الحالية: ويمكن أن تأخذ أحد ١٠٠٠ قيمة هي الكلمات الأكثر استخداماً، وإن لم تكن ضمن تلك الكلمات فتأخذ القيمة الفارغة N.
٦. الكلمتان السابقتان والكلمتان التاليتان: وهي كالكلمة الحالية في عدد القيم الممكنة.
٧. طول الكلمة الحالية (دون أخذ الحركات ولا الكشائد في الاعتبار).

### استخراج القواعد

بعد تحديد الخصائص التي ستؤخذ في الاعتبار عند استقرار القواعد آلياً نقوم باستخراجها عن طريق برنامج خاص كتبناه لقراءة نصوص المكنز المشكل سابقاً وتسجيل القواعد الممكنة دون استبعاد أي منها، ثم تخزينها في ملفات نصية تحتوي تلك القواعد وعدد مرات تكرارها، ونسبة نجاحها. يوضح جدول ٢ بعضاً من هذه القواعد.

### التطبيق

قمنا باستخدام هذه القواعد الاستنباطية لأجل تحسين نتائج التشكيل الآلي وذلك بتطبيقها على مرحلتين:

١. مرحلة التنظيف: وهي مرحلة استبعاد الحركات التي يستحيل أو يندر حصولها. ويمكن تحديد القيمة الأدنى التي يمكن عن طريقها اعتبار الحركة نادرة أم من نسبة النجاح أو من عدد مرات التكرار. وتكمن أهمية هذه المرحلة في أن كثيراً من أدوات التشكيل الآلي تقوم بالتشكيل بطريقة إحصائية لا تراعي احتمالية وجود حركة ما على حرف ما. وقد لوحظ وجود الكثير من الحالات التي لا يمكن حصولها.

٢. مرحلة التشكيل: وهي مرحلة تطبيق القواعد على الأحرف التي ليس لها تشكيل، سواء كان ذلك بسبب استبعادها في المرحلة الأولى أو بسبب نقص نسبة التشكيل في النص. ويمكن أيضاً تحديد القيمة الأدنى لنسبة التشكيل وعدد مرات التكرار التي يجب توافرها قبل تطبيق قاعدة ما.

جدول ٢ أمثلة على قواعد التشكيل الاستنباطية باعتبار ٤ خصائص هي الحرف الحالي وموقعه، والحرفان السابق واللاحق (مرتبة بنسبة النجاح تنازلياً).

الحرف الحالي	موقع الحرف الحالي	الحرف السابق	الحرف التالي	حركة الحرف الحالي	نسبة النجاح	عدد مرات التكرار
ل	٢	ا	م	ـَ	٩٩,٧٣٥	٦٦٥٠٤٣
و	١	N	ا	ـَ	٩٩,٨٨٦	٦١٠٦٣٦
ق	١	N	ا	ـَ	٩٩,٨٥	٤٢٩٣٦٣
ل	٢	ع	ى	ـَ	٩٩,٧٧٩	٤٠٨٣٨١
ك	١	N	ا	ـَ	٩٩,٨٨	٣٠٧٩٣١
ه	٣	ن	N	ـُ	٩٩,٣٥٦	٢٧٩٢٦١
و	١	N	أ	ـَ	٩٩,٨٥٤	٢٢٩٣٧٤
ل	٢	ا	ح	ـُ	٩٩,٦٠٢	٢٢٧٩٩٢
و	١	N	إ	ـَ	٩٩,٩٢١	٢٢١٤٦١

الحرف الحالي	موقع الحرف الحالي	الحرف السابق	الحرف التالي	حركة الحرف الحالي	نسبة النجاح	عدد مرات التكرار
أ	١	N	ي	ـَ	٩٩,٩١٧	٢١٥١٦٧
و	٢	ق	ل	ـُ	٩٩,٨١٥	٢١١٣٩٨
ل	٢	ا	ع	ـُ	٩٩,٦٩٢	٢٠٥٧٤٥
و	١	N	ه	ـَ	٩٩,٤٩	٢٠٢٩٨٢
ه	٢	ل	N	ـُ	٩٩,٩٤٢	١٩٤٨٨٠
ل	٢	إ	ى	ـَ	٩٩,٨٢٥	١٩٢٧١٨
م	٢	ل	N	ـُ	٩٩,٢٥٣	١٩١١٠٢
أ	٢	ل	ن	ـَ	٩٩,٨٧٧	١٨٧٤٢٣
و	١	N	م	ـَ	٩٩,٨	١٨٣٠٤٠
ف	١	N	إ	ـَ	٩٩,٩٥٧	١٧٧٥٤٩

ويجدر ملاحظة أننا في مرحلة التنظيف قد استبعدنا الخصائص المتعلقة بحركات التشكيل، وذلك بسبب أن النص المراد تنظيفه قد لا يكون مُشكلاً كلياً (على عكس المكنز الذي تم استخراج القواعد منه)، مما قد يؤدي إلى استبعاد الكثير من الحركات الممكنة بطريقة خاطئة.

### طريقة التقييم

لتقييم أداء البرنامج قمنا ببناء نص اختباري تم تشكيله يدوياً، ثم استخدمنا ٦ أدوات للتشكيل الآلي لتشكيل النص غير المشكّل آلياً، وهذه الأدوات الست هي كالتالي:

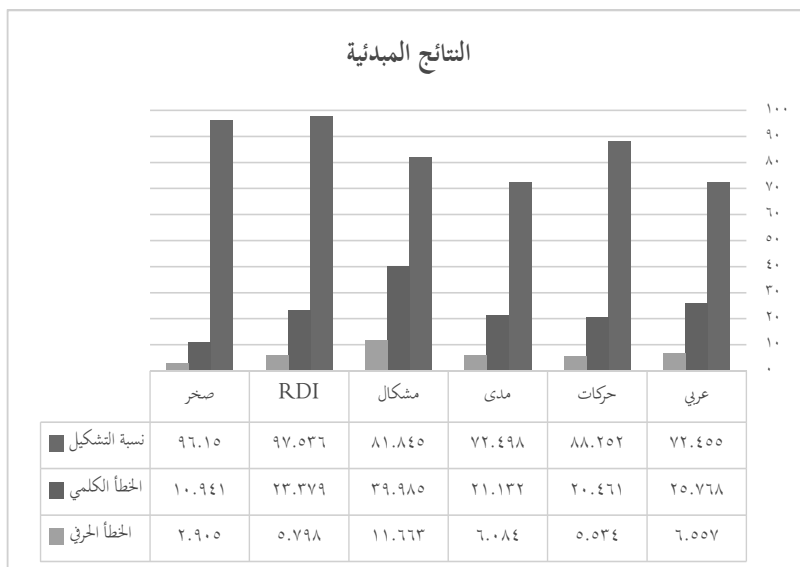
١. عربي <http://www.arabinlp.com>
٢. حركات <http://harakat.ae>
٣. مدى <https://flintbox.com/public/project/8348>

٤. مشكال <http://tahadz.com/mishkal>

٥. RDI <http://www.rdi-eg.com/technologies/diac.aspx>

٦. صخر <http://arabdiac.sakhr.com.eg>

ونظراً لمحدودية توافر بعض هذه الأدوات فقد قمنا بالتواصل مع أصحابها، الذين وافقوا مشكورين على تشكيل النص الاختباري باستخدام أدواتهم، وإرسال النص المُشكَّل إلينا. ويوضح الشكل التالي أداء كل أداة على النص الاختباري. ويمكن ملاحظة أن مُشكِّل صخر حقق الأداء الأفضل مقارنة بأقرانه حيث بلغت نسبة التشكيل فيه ٩٦٪ ونسبة الخطأ الكلمي ١١٪ ونسبة الخطأ الحرفي ٣٪.



رسم توضيحي ٢ : نتائج تشكيل النص الاختباري باستخدام الأدوات الست

### النص الاختباري

يتكون النص الاختباري من حوالي ٥٠٠ جملة (أو مقطع)، تم اختيارها عشوائياً من المكنز، والذي روعي فيه التنوع ليشمل نصوصاً إخبارية ودينية وثقافية واقتصادية ورياضية. وقد تم استبعاد هذا النص من المكنز عند استخراج القواعد الاستنباطية. ويوضح الجدول التالي بعض الإحصائيات عن هذا النص.

#### جدول ٣ إحصائيات عن النص الاختباري

٧٧,٧٣٢	عدد الأحرف العربية
١٦,٢٤٢	عدد الكلمات العربية
٧,٦٤٠	عدد الكلمات الفريدة العربية
٤٩٥	عدد الجمل
٩٩,٣٪	نسبة التشكيل

### المقاييس المستخدمة

درج الباحثون في مجال التشكيل الآلي على استخدام مقياسين رئيسيين لتحديد دقة التشكيل وهما نسبة الخطأ الكلامي ("WER" Word Error Rate) ونسبة الخطأ الحرفي (Character Error Rate / Diacritic Error Rate)، ويتم قياسهما عن طريق المعادتين التاليتين:

ويوضح المثال في الجدول التالي كيفية حساب المقياسين في سياق التشكيل الآلي:

جدول ٤ مثال لتوضيح طريقة حساب نسبة الخطأ الكلمي ونسبة الخطأ الحرفي

النص الأصلي	لكل مجتهد نصيب
النص المُشكَّل آلياً	لُكِّلْ مَجْتَهَدٌ نَصِيبٌ
التشكيل الصحيح للنص	لُكِّلْ مَجْتَهَدٌ نَصِيبٌ
نسبة الخطأ الكلمي	$66.7\% \equiv 0.667 = 2 / 3$
نسبة الخطأ الحرفي	$17\% \equiv 0.17 = 2 / 12$

ورغم أن هذين المقياسين هما الأشهر والأكثر استخداماً إلا إنهما لا يكفيان لمعرفة درجة أداء البرنامج الذي يقوم بالتشكيل، فمن الممكن مثلاً أن يقوم البرنامج بتشكيل نص بنسبة ٥٠٪ دون أن تزيد نسبة الخطأ في التشكيل عن ١٪. لذا رأينا أنه من الضروري استخدام مقياس جديد ( بالتكامل مع المقياسين السابقين ) هو نسبة التشكيل ( Diacritization Level (DL) )، والتي يمكن حسابها بالمعادلة التالية:

حيث يدخل في المُشكَّلة الأحرف الأحرف التي لا تُشكَّل عادة إما بسبب وضوحها كالكسرة في همزة إنسان، أو التي لا تظهر نُطقاً كألف واو الجماعة وحرف الواو في جَاوُوا.

وتجدر الإشارة إلى أن هذا المقياس يَصُعبُ حسابه آلياً بدقة، حيث إن الأحرف المُشكَّلة ضمنياً لا يمكن التعرف عليها بسهولة، لذلك فقد قمنا باستخدام قواعد مُبسطة لتحديد تلك الأحرف، وقد حققت نتائج مقبولة إلى حد كبير.

## النتائج والتحليل

لاختبار أثر القواعد الاستنباطية في تحسين نتائج التشكيل الآلي باستخدام الأدوات الست سابقة الذكر؛ قمنا بإجراء تجربتين على النصوص المشكّلة آلياً: الأولى باستخدام حد أدنى للندرة ١٪، والثانية باستخدام حد أدنى ٠٪. ويمثل هذا الحد الأدنى الدرجة التي يتم عندها تحديد ما إذا كان تشكيل حرف معين نادراً أم لا؛ حيث إنه كان نادراً فإنه يتم استبعاده. ويظهر الرسمان التوضيحيان ٣ و٤ نتائج كلتا التجربتين باستخدام المقاييس الثلاثة: درجة التشكيل، ونسبة الخطأ الكلمي، ونسبة الخطأ الحرفي.

ويمكننا أن نلاحظ أن نسبة التشكيل تتحسن بشكل كبير (في التجربتين) على أغلب الحالات خاصة أن أكثر هذه الأدوات لم تقم بتشكيل النص بشكل كامل، وذلك باستثناء مُشكّلي صخر و RDI الآليين اللذين انخفضت فيهما نسبة التشكيل بشكل طفيف.

أما بالنسبة لنسبة الخطأ الكلمي فنرى انخفاضاً يتراوح بين ١-٢٪ لأغلب الحالات، وقد يصل إلى ٦٪ في حالة أداة مشكال، وزيادة ١٪ تقريباً في حالة أداة صخر. أما في التجربة الثانية، حيث قمنا بخفض الحد الأدنى للندرة، فنلاحظ انخفاضاً في حدود ١٪ لجميع الأدوات الست.

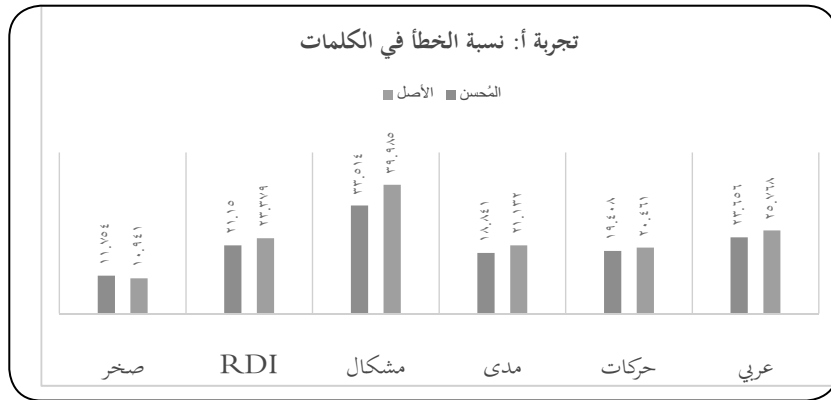
وأخيراً، نرى نسبة الخطأ الحرفي تنخفض أيضاً بحوالي ٥، ٠٪ في كلتا التجربتين، لأغلب الحالات باستثناء حالة مشكل صخر.

ويلاحظ من هذه النتائج أن انخفاض نسبة الندرة يزيد من نسبة التشكيل بشكل كبير، حيث وصلت إلى حوالي ٩٥٪ في جميع الأدوات الست، وهذا يمثل تغيراً كبيراً يصل إلى أكثر من ١٠٪ لجميع الأدوات باستثناء صخر و RDI. ويعود ذلك

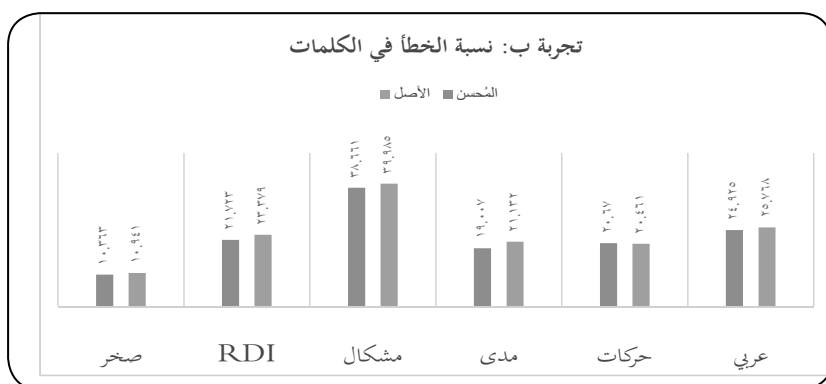
لكون خفض نسبة الندرة يُبقي على درجة أكبر من التشكيل الأصلي، وذلك قبل أن يتم تطبيق القواعد الاستنباطية على النص.

ويُلاحظ كذلك أن انخفاض نسبة الندرة يُقلّل من التحسن في نسبة الخطأ، سواء الكلمي أو الحرفي، حيث يؤدي ذلك إلى الإبقاء على أخطاء التشكيل الموجودة مُسبقاً دون إزالتها.





رسم توضيحي ٣: نتائج استخدام القواعد الاستنباطية في التجربة أ (حيث تم استخدام ١% كحد أدنى في مرحلة التنظيف)



رسم توضيحي ٤: نتائج استخدام القواعد الاستنباطية في التجربة ب (حيث تم استخدام ٠% كحد أدنى في مرحلة التنظيف)

وَيُمْكِنُ أَنْ نَقَارِنَ دَرَجَةَ التَّحْسُنِ لِكُلِّ مَقْيَاسٍ مِنْ جَدُول ٥ وَجَدُول ٦. حَيْثُ فَمِنَّا فِي حَالَةِ الْمَقْيَاسِ الْأَوَّلِ (نِسْبَةُ التَّشْكِيلِ)؛ بِحِسَابِ دَرَجَةِ التَّحْسُنِ بِطَرَحِ الْفَرْقِ بَيْنَ الْقِيَمَتَيْنِ ثُمَّ قَسَمْتَهُ عَلَى الْقِيَمَةِ الْأَصْلِيَّةِ. أَمَّا فِي حَالَةِ نِسْبَةِ الْخَطَأِ (سَوَاءٌ كَانَتْ فِي الْكَلِمَاتِ أَوْ الْأَحْرَفِ) فَقَدْ قَمْنَا بِضَرْبِ النَّاتِجِ بِإِشَارَةِ سَالِبَةٍ، لِأَنَّ التَّحْسُنَ هُنَا هُوَ دَرَجَةُ النِّقْصِ وَلَيْسَ دَرَجَةُ الزِّيَادَةِ.

**جدول ٥ درجة التحسن في التجربة أ**

اسم الأداة	نسبة التشكيل	نسبة الخطأ (كلمات)	نسبة الخطأ (أحرف)
عربي	٢٧,٨٪	٨,٢٪	٧,٢٪
حركات	٥,٩٪	٥,١٪	٨,٤٪
مدى	٢٧,٨٪	١٠,٨٪	١٩,٧٪
مشكال	١١,٤٪	١٦,٢٪	٢٣,٠٪
RDI	٣,٥-	٩,٥٪	٨,٧٪
صخر	٣,٢-	٧,٤-	٥,٠-
المتوسط	١١,٠٣٪	٧,٠٧٪	١٠,٣٣٪

**جدول ٦ درجة التحسن في التجربة ب**

اسم الأداة	نسبة التشكيل	نسبة الخطأ (كلمات)	نسبة الخطأ (أحرف)
عربي	٣٢,٢٪	٣,٣٪	٢,٤٪
حركات	١٠,٦٪	١,٠-	٠,٧٪
مدى	٣٢,٣٪	١٠,١٪	١٩,٠٪
مشكال	١٥,٤٪	٣,٣٪	٧,٤٪
RDI	٠,٢-	٧,١٪	٧,٧٪
صخر	٠,٥-	٥,٣٪	٧,٦٪
المتوسط	١٤,٩٧٪	٤,٦٨٪	٧,٤٧٪

## الخلاصة

في هذا العمل البحثي قمنا بتقديم طريقة القواعد الاستنباطية، وهي قواعد تشكيلية تم استخراجها من أحد المكانز، وتحدد مدى تأثير عدد من الخصائص (كنوع الحرف وموقعه في الكلمة) على التشكيل. قمنا باستخدام تلك القواعد في مرحلتين: مرحلة التنظيف وهي المرحلة التي يتم فيها استبعاد التشكيل حسب درجة ندرته، ومرحلة التطبيق وفيها يتم تطبيق هذه القواعد على الأحرف غير المشكّلة في النص.

وتظهر نتائج التجارب التي قمنا بها تحسناً كبيراً في نسبة التشكيل تصل إلى ١١٪ في المتوسط، وكذلك تحسناً في نسبة الخطأ الكلمي بمتوسط ٧٪، وفي نسبة الخطأ الحرفي بمتوسط ١٠٪.



## المراجع

1. M. Rashwan, M. Al-Badrashiny, M. Attia, S. Abdou and A. Rafea. "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features." Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 1, pp. 166-175, jan. 2011.
2. Wikipedia. "Diacritic --- Wikipedia. The Free Encyclopedia." 16 January 2012. [Online]. Available: <http://en.wikipedia.org/wiki/Diacritic>. [Accessed 1 February 2012.]
3. I. Zitouni and R. Sarikaya. "Arabic Diacritic Restoration Approach Based on Maximum Entropy Models." Computer Speech & Language, vol. 23, no. 3, pp. 257-276, 2009.
4. M. Elshafei, H. Al-Muhtaseb and M. Al-Ghamdi. "Statistical Methods for Automatic Diacritization of Arabic Text." 2006.
5. M. Attia. "Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications." PhD. Dissertation, RDI, Cairo, Egypt, 2005.
6. J. Mahar and G. Memon. "Lexicon Based Diacritic Resto-

- rations Using WordNet For Sindhi." International Journal Of Academic Research. vol. 3. no. 2. pp. 37-43. March 2011.
7. R. A. Haertel. P. McClanahan and E. K. Ringger. "Automatic Diacritization for Low-Resource Languages Using a Hybrid Word and Consonant CMM." Stroudsburg. PA. USA. 2010.
  8. A. Ali. "Automatic Urdu Diacritization." M.Sc. Thesis. Center for Research in Urdu Language Processing. Lahore. Pakistan. 2009.
  9. O. Shaaban. Automatic Diacritics Restoration for Arabic. Dhahran: King Fahd University of Petroleum and Minerals. 2013.

## الترجمة الآلية من العربية وإليها

أ.د. محمد زكي خضر<sup>(\*)</sup>

khedher@ju.edu.jo

---

(\*) الجامعة الأردنية

حصل على البكالوريوس في الهندسة من جامعة بغداد عام ١٩٦٥ وعلى الدكتوراه من بريطانيا عام ١٩٧٢ وعمل في جامعة الموصل بالعراق وحصل على مرتبة الأستاذية عام ١٩٨١، وهو يعمل في الجامعة الأردنية منذ عام ١٩٩٢. نشر أكثر من ١٠٠ بحث وألف ٢٠ كتاباً. عمل في مجال المعالجة الآلية للغة العربية ونال جوائز عالمية وهو عضو في جمعيات علمية عالمية ويرأس تحرير المجلتين الدوليتين للتطبيقات الإسلامية في علم الحاسوب وتقنياته باللغتين العربية والإنجليزية.





## ملخص

حصل تقدم هائل للترجمة الآلية خلال السنوات الماضية خاصة بين اللغات الأوروبية. وفي الوقت نفسه حصل تقدم كبير في الترجمة الآلية للغات الشرقية كالصينية واليابانية والكورية. كما توفرت أنظمة متعددة مساعدة لأعمال المترجمين المختصين وأنظمة التصحيح اللغوي وغيرها، مما يساعد في عملية الترجمة اليدوية. أما الترجمة من اللغة العربية وإليها فهناك محاولات للترجمة الآلية، لكن معظم هذه الأنظمة تنتج لغة ركيكة وغير دقيقة.

تعتمد أنظمة الترجمة الآلية على أسس مختلفة، فبعضها يستند إلى المعاجم مع مساعدة نحوية لتحليل الجملة ثم إعادة تركيب ترجمتها وبعضها يعتمد الترجمة بالاستناد إلى الذخيرة اللغوية المحتوية على نصوص ترجمت من قبل مترجمين من البشر باستعمال عمليات إحصائية وغير ذلك من الوسائل حيث تمثل المعاجم ركيزة رئيسية في الترجمة، إضافة إلى البرمجيات اللغوية المتخصصة.

هذا البحث هو استكمال لبحث سابق (١) ويهدف إلى تقديم فكرة عن تاريخ الترجمة الآلية بين اللغات العالمية وتطورها وما وصل إليه التقدم في هذا المجال كما يناقش الوضع بالنسبة للترجمة الآلية من اللغة العربية وإليها والمشاكل التي تعاني منها ومن ثم يبين الخطوات التي ينبغي اتباعها للتقدم في هذا المجال، ثم يعرج إلى مؤشرات عن مستقبل الترجمة الآلية من اللغة العربية وإليها وما يحتاج ذلك من جهود.



## مقدمة

إن اختلاف اللغات التي يتكلم بها البشر آية من آيات الله تعالى في الأرض قال تعالى: "وَمِنْ آيَاتِهِ خَلْقُ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافُ السِّنِّكُمْ وَأَلْوَانِكُمْ إِنَّ فِي ذَلِكَ لَآيَاتٍ لِّلْعَالَمِينَ" (الروم/ ٢٢)، ولذلك احتاج الناس الترجمة منذ الزمن الغابر الذي تعددت فيه لغات البشر لغرض التواصل بينهم، فعن خارجة بن زيد بن ثابت قال: قال زيد بن ثابت: "أمرني رسول الله -صلى الله عليه وسلم- فتعلمت له كتاب يهود، وقال: إني والله ما آمن يهود على كتابي، فتعلمته، فلم يمر بي إلا نصف شهر حتى حدقته، فكنت أكتب له إذا كتب، وأقرأ له إذا كتب إليه" -رواه أبو داود-.

الترجمة هي نقل معاني نص من لغة إلى لغة أخرى مع مراعاة للدقة والأسلوب. ويحتاج ذلك فهم النص الأصلي والتعبير عن المحتوى والأسلوب بلغة أخرى. فالمرجم يجب أن يتقن اللغتين المترجم منها والمترجم إليها وهو يتبع إحدى طريقتين: إما الترجمة الحرفية والالتزام بمعاني مفردات النص الأصلي ونقلها إلى اللغة الثانية أو فهم المعنى العام ثم التعبير عنه باللغة الثانية بأسلوب المترجم نفسه (١).

يتعرف المترجم على الرموز المكتوبة في الترجمة الكتابية والأصوات المنطوقة في الترجمة الشفوية، لكن اللغة العربية المكتوبة بحروف غير مشكولة تستوجب معرفة من القارئ لكي يفهم المقصود بدون تشكيل. وهو بذلك يستعمل خبرات ذاتية من قواعد اللغة لكي يستعملها بشكل ذاتي وسريع. أما المترجم للكلام المنطوق فهو يستطيع تمييز الجملة التي يترجمها إن كانت استفهامية أم خبرية أم تعجبية دون وجود علامات استفهام أو تعجب (٢).

بعد ذلك يرجع المترجم ذهنياً أو واقعياً إلى الوحدات المعجمية وهي الكلمات والتعبيرات الاصطلاحية ويفهم معانيها في سياقاتها اللغوية والاجتماعية المختلفة،

فالكلمة الواحدة قد تعني أشياء كثيرة يقيدها السياق الذي ترد فيه. فمثلاً "عين" قد تعني عين الإنسان أو الحيوان وقد تعني عين الماء وقد تعني في الصرف عين الفعل (٢). وهذه الترادفات تمثل مشكلة كبيرة بالنسبة للترجمة عامة والترجمة الآلية خاصة.

ولحل هذه المشكلة يقوم المترجم بتحليل المفهوم للوصول إلى كنهه وما يتفرع عنه من ظلال في المعنى، وخاصة المجازية منها، وذلك حتى تتحدد العلاقة بين هذا المفهوم الأساس وما يتفرع منه وبين المفاهيم التي يمكن أن تتصل به بشكل أو بآخر، أي العلاقة بينه وبين مجموعات المفاهيم الأخرى.

فالمفهوم الذي يرتبط بجسم واحد بعينه، أو يدل عليه، هو مفهوم إفرادي، مثل "عطارد". أما إذا ارتبط المفهوم بعدة أجسام بينها نوع من التجانس الذي يضعها في مجموعة واحدة، فإن المفهوم يكون عاماً (بمعناها الضيق هنا)، مثل "كوكب". فهذا المفهوم يدل على عدد من الأجسام التي تدور حول الشمس. ولتحديد المقصود يمكن أن يكون التعريف مركّزاً أو موسعاً. وعلى ذلك فإن دقة التعريف تحدد المفهوم. والتعريف قد يكون شاملاً في بعض الأحيان، وفي غالب الأحيان لا يخلو من النقص أو أن يكون تعريفاً بالضد أو ما يعرف بالتعريف السلبي أو أن يكون فضفاضاً أو ضيقاً أو يأخذ بمبدأ الإحلال والتعويض إلى غير ذلك من الصفات التي يتصف بها التعريف. وهذا التعريف قد يكون واضحاً في ذهن المترجم وقد يكون مشوشاً فيتصرف في ترجمته لهذا المفهوم وقد يعبر عنه بمفهوم يفهمه هو ويكون قريباً أو بعيداً من المفهوم الأصلي.

ولقد سبق الإمام أبو حامد الغزالي اللسانيين المحدثين في تحديد العلاقة بين المفهوم واللفظ حيث يقول: "الشيء له في الوجود أربع مراتب: الأولى حقيقته في نفسه والثانية ثبوت مثال حقيقته في الذهن وهو الذي يعبر عنه بالعلم والثالثة تأليف صوت بحروف تدل عليه وهو العبارة الدالة على المثال الذي في النفس،

والرابعة تأليف رقوم تدرك بحاسة البصر دالة على اللفظ وهو الكتابة، فالكتابة تتبع للفظ إذ تدل عليه، واللفظ تبع للعلم إذ يدل عليه، والعلم تبع للمعلوم إذ يطابقه ويوافقه، وهذه الأربعة متطابقة متوازية، إلا أن للأولين وجودان حقيقيان لا يختلفان بالأعصار والأمم، والآخرين وهما اللفظ والكتابة يختلفان بالأعصار والأمم لأنهما موضوعان بالاختيار "أ.ه (٣)

كان للتطور في تمثيل اللغويات الذي قاده نعوم شومسكي تأثيراً على نظريات اللسانيات الحديثة. فقد أعاد شومسكي تعريف الأهداف من النظريات اللسانية لكي تأخذ بعين الاعتبار خلفية المتكلمين بالنسبة للغتهم بدل البحث في المكانز اللغوية للوصول إلى المطابقة المطلوبة. فالفرد قد يسمع جملة بلغته الذاتية فيفهم معناها بسهولة رغم أنه لم يسمعها من قبل (١٤).

إن قواعد أية لغة محدودة لكن الجمل التي يمكن أن تتركب باستخدام تلك القواعد لا نهاية لها. أما المتكلمون باللغة فيتعلمون بالاستنتاج، فمثلاً يمكن للمتكلم أن يشعر بالفرق بين معنيين مختلفين في جملة غامضة مثل "حضر ابن عمي الصغير" فيمكن أن يقصد أن الذي حضر هو الابن الصغير للعم أو هو ابن أصغر الأعمام. إن القواعد اللغوية يجب أن تحاكي قابلية المتكلم بلغته الأم بحيث تستطيع أن تميز الغموض في مثل هذه الجملة وأن تعطي نموذجين للتركيبتين المحتملتين.

وقد أجبرت الحواسيب اللسانيين أن يكونوا أكثر دقة في وصفهم لقواعد اللغة التي سبق وأن تصوروا أنهم يعرفونها عن اللغة. ولكن لا يزال الشوط بعيداً في دقة القواعد التي وصلوا إليها لتمثيل اللغة بشكل صادق.

إن التقدم في الوصف الدقيق والمعمق للغات الطبيعية وتوفر التقنية الحاسوبية لمعالجة اللغات الطبيعية اجتماعاً في تكوين بيئة مكنت من الوصف المعمق للغات مع أسس قواعدية لتلك اللغات.

تساعد بعض العمليات في التعرف على المفهوم مثل التعرف على الوحدات النحوية وعلى وظائفها مثل المسند والمسند إليه والتكملة والمضاف والمضاف إليه، والجار والمجرور والصفة والموصوف وغير ذلك. وعلى المترجم أن يفهم معنى كل واحد من هذه العناصر ووظيفته اللغوية والنحوية في النص. فالصيغة الخبرية مثلاً قد تأتي للإخبار كما تستعمل للطلب غير المباشر كما في عبارة: "أرى أن لديك قلمًا زائدًا"، بل وللدعاء كذلك مثل: "رحم الله فلانًا" (٤). ومن الأمور المهمة في الترجمة معرفة حقل النص ومجال تخصصه. مثلاً إذا كان النص في علم الأحياء فإنك تجد أن بعض الكلمات تختلف معانيها في هذا النص عن حقل تجارة المواد الغذائية أو في حقل من حقول العلوم الزراعية. وهكذا.

المحور الثاني في الترجمة هو التعبير عن محتوى النص باللغة الهدف أي اللغة المترجم إليها. ويحتاج هذا معرفة بإنتاج المقابلات الصوتية أو الكتابية والنحوية والمعجمية والبلاغية. وعلى المترجم أن يقوم باختيار العبارات المناسبة في اللغة التي يترجم إليها التي توافق المفهوم الذي توصل إليه في الخطوات السابقة. "وعند ذلك تبرز براعة المترجم في اللغتين حيث يجب أن يتقن ما يسمى بالتقابل المعجمي أو الاصطلاحي، فالتعبير العربي (رجع بخفي حنين) ليس له مقابل حريفي بأية لغة، ولو ترجم ترجمة حرفية لأصبحت الترجمة مضحكة. فهو يجب أن يفهم مثل هذه العبارات ويجب أن يعرف العبارات الدارجة التي تقابلها في اللغة الثانية أو على الأقل يكون هو تعبيراً من كلمات تلك اللغة بما يعبر عن المفهوم بدقة كافية مستعملاً القواعد النحوية والصرفية في اللغة التي يترجم إليها" أ.هـ (١).

أما التعامل مع الجمل الطويلة فهو من أهم مشاكل الترجمة حيث يحتاج ذلك إلى معالجات ذهنية معقدة. فحينما تحوي الجملة الطويلة على عدة مفاهيم فإن المفهوم الواحد ربما يتجزأ ليتكامل بين بدء الجملة ونهايتها. وقد يتضمن المفهوم الواحد مفاهيم ثانوية. وقد يكون المفهوم المهم مندساً بين عدد من المفاهيم الأقل

أهمية منه. وهكذا فإن الجملة الطويلة ربما تستعصي على الفهم والتحليل وتحتاج براعة عند التركيب في اللغة المترجم إليها.

يمكننا أن نذكر أن المترجمين لا يقومون كلهم بالخطوات نفسها، بل إن كل مترجم له أسلوبه الذي يتبعه في الترجمة. ماذا نفعل تمامًا حينما نحاول فهم نص ما من النصوص؟ إن الطريقة التي يتعامل بها الإنسان ذهنيًا مع النصوص ما زالت غير معروفة على وجه الدقة وتخضع لنظريات غير مؤكدة حتى الآن. وليس هناك أساليب متفق عليها حتى الآن يمكن أن ننقلها بشكل من الأشكال إلى الحاسوب في الترجمة الآلية. لذلك فإن فهم النص يعتبر مشكلة كبرى في الترجمة، فهناك في كل اللغات لبس دلالي نتيجة تعدد معاني بعض الكلمات وتحديد الجهة التي تعود لها بعض الضمائر (١). وهناك غموض آخر بسبب بناء أو تركيب الجملة يسمى اللبس النحوي، حيث تعني الجملة أكثر من معنى أو يمكن تفسيرها بأكثر من طريقة، كما أن هناك لبس على المستوى الصريفي خاصة في الأفعال المتعدية لمفعولين والصيغ التي تعني اسم الفاعل واسم المفعول في الوقت نفسه وصيغ المبني للمجهول والربط بين عدة وحدات لغوية مثل تغير معنى الفعل حسب حرف الجر الذي يأتي بعده مثل "رغب في" و"رغب عن"، وهناك مصادر أخرى عديدة للبس (٥).

إن حرف الجر قد يأخذ معاني مختلفة بين اللغات المختلفة، فمثلا اللام في جملة: "أعطيت تفاحة لأحمد" تعني اللام "إلى to" بينما في جملة "أكلت تفاحة لأحمد" تعني أن التفاحة هي "ملك أحمد of".

### الحاسوب في خدمة الترجمة

الترجمة الآلية اليوم لا ترقى إلى المستوى الذي يضاهي المترجمين من البشر. لكن الحاسوب اليوم بإمكانه أن يقدم عونًا لا بأس به للمترجم. ومن هذا العون ما يأتي: (١)



١. الترجمة الآلية التي تحتاج إلى تحرير لاحق، أي الحاجة إلى مراجعة بشرية بعد الترجمة الآلية.
٢. تحرير أولي ثم القيام بالترجمة الآلية بعد التحرير السابق، وذلك بتبسيط الجمل الطويلة وتحديد معاني الكلمات ذات المعاني المتعددة.
٣. التفاوض مع الآلة بين المترجم والحاسوب، حيث هناك برامج حاسوبية ذوات إمكانيات حوارية بأن تعطي الترجمة جملة بعد جملة، ويتوقع من المترجم أن يوافق أو يعدل عليها.
٤. قيام المترجم بالترجمة وتساعد الآلة، كأن تعطي الآلة للمترجم الكلمات ذات العلاقة من المعجم مع مرادفاتها، وهو يختار من بينها.
٥. الخدمات الحاسوبية الأخرى وسنورد بعض ما يتوفر من برمجيات تخدم الترجمة في ما بعد.

### نشأة الترجمة الآلية وتطورها

كان أول من استخدم الحاسوب في الترجمة فعلياً هو وارن ويفر عام ١٩٤٧، فقد كتب في عام ١٩٤٩ مذكرة لزملائه بأربعة مقترحات لتطوير أنظمة الترجمة الآلية لكي تتقدم أكثر من الترجمة كلمة كلمة وذلك بفحص المفهوم المباشر للكلمات الغامضة. كما أشار إلى التناظر بين تركيب المخ البشري والآلات المنطقية ثم انتهى بأن الترجمة الآلية ممكنة التحقيق. وفي تلك الفترة كانت هناك حاجة سريعة للترجمة غير الدقيقة نظراً لحاجة الولايات المتحدة للاطلاع على اتصالات الاتحاد السوفياتي، وفي الوقت نفسه ظهرت الحواسيب. ولذلك كانت أنظمة الترجمة من الروسية إلى الإنكليزية أولى المحاولات، فقد أجريت بنجاح في جورج تاون أول ترجمة من اللغة الروسية إلى الإنكليزية عام ١٩٥٤، وفي عام ١٩٥٥ أجريت في الاتحاد السوفياتي أول تجربة في الترجمة الآلية من الإنكليزية إلى

الروسية باستعمال قاموس يحوي ٢٣٠٠ كلمة. وأشارت التجربة إلى ضرورة تمثيل التراكيب اللغوية بمستوى بناء الجملة وبالمستوى المعجمي، كما اتضح أن مشكلة الغموض هي مشكلة فعلية رغم أنها لم تقدر حق قدرها حينئذ، حيث كانت الفترة من ١٩٥٦ إلى ١٩٦٦ فترة التوقعات الكبيرة من الترجمة الآلية التي لم يتحقق ما كان يؤمل من الحصول على دقة بنسبة ٩٥٪ دون معالجة مسبقة للنصوص. لقد أوضحت الأبحاث أن تركيب اللغات أكثر تعقيداً مما كان متصوراً، وأن المترجمين من البشر يستعملون الكثير من المعرفة في حقل المادة المترجمة والمعرفة بالحياة العادية أكثر مما كان متصوراً. وأدى ذلك إلى تباطؤ في أبحاث الترجمة الآلية وربما إهمال الموضوع حتى عام ١٩٧٥ (١).

عادت الأبحاث على أشدها في أوروبا وكندا واليابان بين عامي ١٩٧٥ و ١٩٨٥ باستعمال الأنظمة الخبيرة وأبحاث معالجة اللغات الطبيعية، وبشكل خاص للغات الأوربية واليابانية، وبذلك ظهر الجيل الثاني من برامج الترجمة الآلية فظهرت في الأسواق برامج تجارية تستعمل الحواسيب. وهذا التطور شمل البحوث المعجمية والنحو والصرف والدلالة.

وفي عام ١٩٨٩ كانت بداية الجيل الثالث من برامج الترجمة الآلية التي استندت إلى المعلومات الإحصائية حينما قامت شركة IBM (٢٩)، بمشروع Candide المستند إلى الترجمة بالأمثلة والترجمة المحدودة الموضوع مع تعدد اللغات المترجم منها والمترجم إليها، فازداد الاهتمام بالترجمة الآلية لحد لم يسبق له مثيل. وقد استندت هذه الترجمة إلى الذخيرة اللغوية Corpus Based والتي لا تزال في تطور حتى اليوم مع بعض التكامل مع الأساليب الأخرى للترجمة الآلية (١).

أما الترجمة من اللغة الإنكليزية إلى العربية فقد كان أول ظهورها في نهاية السبعينيات في ولاية يوتا الأمريكية، ولم يكن يحتوي البرنامج على تحرير أولي، لكن كان بالإمكان إجراء تحرير نهائي. وقد احتوى النظام على مرحلتين الأولى لتحليل اللغة المصدر والثانية لتوليد اللغة الهدف. كان تحليل اللغة الإنكليزية

موجهًا لكي يولد العبارات في اللغة الهدف بشكل مباشر باستعمال معجم ثنائي اللغة للعبارات. ولم تكن الألفاظ والجمل الإنكليزية تحلل بعمق، بل فقط بما يكفي لتوليد المقابل باللغة العربية. وهكذا كان النظام باتجاه واحد ولم يكن يغوص بعمق في اللغة المصدر. وقد استعمل البرنامج لترجمة الموسوعة البريطانية (بريتانكا) إلى اللغة العربية (١٤). كما قامت سلطنة عمان بالحصول على امتياز للبرنامج لكي يستعمل في ترجمة الوثائق الإنكليزية إلى العربية.

لم تكن الطريقة المباشرة في الترجمة تغوص في أعماق اللغة المصدر، ومن ثم أدرك مطورو البرمجيات أن الطريقة المباشرة لا تكفي لمعالجة اللغات الطبيعية المعقدة. فقد وجد أن التعامل مع الجمل باستبدال مواقع الفعل بين الإنكليزية والعربية مثل Ali drinks coffee فترجمتها إلى: "يشرب علي القهوة" احتوى على استبدال موقع الفعل والفاعل لكن ذلك لا يكفي في جملة مثل:

The boy whose clothes are dark and carrying a flag in his right hand talked to me.

فمثل هذه الجملة تحتاج إلى تحليل عميق لتحديد الفاعل ومن ثم افتترانه بالفعل حيث بينهما عدد كبير من الكلمات. وعلى ذلك فإن الطريقة المباشرة لا تعطي نتائج دقيقة لمثل هذه الجمل الطويلة، لذا فقد برزت حاجة لتطوير المعرفة بكيفية تحليل الجمل الطويلة بعمق وكفاءة دون غموض، وعند ذلك ظهرت الطريقة التحويلية التي قدمت للترجمة الآلية أمرين: الوصف لتحليل الجملة ومن ثم التقنية الجديدة لتمثيل هذا التحليل العميق، وهو ما طور خلال السبعينيات والثمانينيات من القرن الماضي.

إن استعمال أنظمة الترجمة الآلية قليلة الدقة يدفع إلى -بشكل مطرد- للاستعانة بمرجمين من البشر. وقد وظفت بلدان كثيرة الترجمة الآلية لخدمة متطلباتها الاجتماعية والاقتصادية والعلمية والتقنية في طليعتها الولايات المتحدة وروسيا والصين واليابان (٢٣و٢٤).

لقد أجريت منذ الثمانينيات في القرن الماضي العديد من الأبحاث حول المعالجة الصرفية للغة العربية وكان معظم تلك الدراسات ينصب على عملية فصل السوابق واللاحق في الكلمة العربية المركبة، ومن ثم استخلاص الجذر لمعرفة العمليات الصرفية التي أجريت عليها والتي يمكن للحاسوب القيام بها بعد برمجته بشكل دقيق.

قام كيفن نايت ودانيال ماركو بتحويل الأبحاث العلمية المتعلقة بالترجمة الآلية إحصائياً في عام ٢٠٠٢ إلى منتج للترجمة الآلية من العربية إلى الإنكليزية، وكان ذلك مثلاً جيداً لتكوين ذخيرة لغوية متقابلة بين العربية والإنكليزية والاستفادة منها للترجمة الآلية.

كما وجد أن استكشاف أسماء الأعلام والبيانات المعروفة ببرنامج خاص يمكن أن يحسن من الترجمة، فما أن يكتشف مثل هذا الاسم حتى يمكن القيام بعدة عمليات تفيد في الترجمة، فمثلاً عند اكتشاف أن كلمة مثل "أحمد" هي اسم علم فإنها لا تترجم على أنها I praise

كما ظهرت بعض الأنظمة الهجينية (٣٤) التي تستعمل للترجمة المكتوبة والمنطوقة باستعمال الطريقتين: الطريقة الإحصائية والطريقة المستندة إلى القواعد (٢٤ و ٢٥)، وبعد النظر إلى كل من الطريقتين من ناحيتي الفوائد والمضار يمكن المزاوجة بينهما للحصول على الفوائد من كليهما واستبعاد المضار (١٢).

هناك الآن في الأسواق أعداد كبيرة من برامج الترجمة الآلية بين اللغات الأوروبية، وقد ازداد الطلب عليها بشكل واسع وساهمت شبكة الإنترنت بزيادة الحاجة إلى الترجمة الآلية وسهلت في تبادل المادة المترجمة إلى من يحتاجها. ولا يزال التقدم على أشده في تكامل عمل المترجمين من البشر مع الترجمة الآلية، حيث يحتاج التقدم بمستوى الترجمة الآلية إلى مترجمين أكفاء لتطويرها والارتقاء بها (٧).

## مزايا الترجمة الآلية

يحتاج المجتمع البشري اليوم إلى كمية هائلة من المعلومات التي يراد ترجمتها بما يفوق ما يتوفر من مترجمين من البشر خاصة بعد تفجر ثورة المعلومات وتنوع اللغات التي تساهم في إنتاج المعرفة اليوم وما يحتاجه من لا يتكلمون تلك اللغات. ورغم التأريخ الحافل للغة العربية في مساهمتها في المعرفة البشرية، لكنها اليوم ليست المصدر الأهم في العلوم والتقنيات الحديثة. وهذا يجعل عملية الترجمة من اللغات الأخرى إلى اللغة العربية ذات أهمية بالغة للناطقين بالعربية، فإمّا أن يتعلموا لغة أو لغات أخرى بجانب العربية أو أن يترجم لهم ما يصدر من معرفة باللغات الأخرى. فالتخلف في الترجمة إلى اللغة العربية يفسر الإقبال الشديد على تعلم اللغة الإنكليزية خاصة والتدريس بها في الجامعات والمدارس في كثير من الأقطار العربية (١).

يضاف إلى ما سبق سبب ثان في أهمية الترجمة الآلية هو أن عملية الترجمة مملة للمترجمين وتستهلك الأوقات الطويلة منهم، فهم يحاولون التغلب على المل بتغيير أسلوب الترجمة أو بالراحة. والمترجم حاله حال البشر، ينام ويمرض ويغير عمله ويتقاعد، وهذا يجعل المترجمين من البشر بضاعة نادرة. كما أن غالبية المترجمين يتقنون لغة واحدة مع اللغة الأم. وهذا بخلاف الآلة التي يمكنها أن تعمل الوقت الطويل ويمكن استبدالها بما هو أكفأ منها مع التقدم التقني المستمر إذا ما تقادمت (١).

وهناك أسباب أخرى منها أن المؤسسات الضخمة ترغب في استعمال مصطلحات وعبارات محددة ينبغي أن تعاد كلما مرت تلك العبارات، والمترجمون من البشر عادة يحاولون تغيير أساليبهم ابتعاداً عن الملل، إضافة إلى أسباب اقتصادية وتجارية في استعمال الترجمة المستندة للحاسوب حيث تزيد من حجم ومن سرعة ما يراد ترجمته مما يحقق للمؤسسات والشركات إنتاجية عالية بسرعة هائلة، كما أن حاجة القطاع التجاري قد تكون إلى ترجمة تعطي فكرة عامة عن

المنتجات الصناعية والتجارية دون أن تكون على درجة عالية من الرصانة. يجري هذا مع ازدياد حجم مثل هذه المعلومات المطلوب ترجمتها وتوسيعها لتشمل لغات جديدة في عصر العولمة. وهذا ما يجعل الترجمة الآلية رغم انخفاض كفاءتها مفيدة (٢٢١).

## ٢٥. مستويات برامج الترجمة الآلية وأنواع منها

يمكن تصنيف برامج الترجمة الآلية بحسب مستوياتها التي تعكس مدى تعقيدها ومدى كفاءتها في الترجمة إلى الأصناف الآتية بشكل تقريبي (٢٦٩):

**المستوى الأدنى:** تستدعي الترجمة الآلية البدائية ترجمة كلمة لكلمة دون أية معالجة آلية قبل ذلك، ويمكن بعد ذلك للمترجم البشري أن يفهم من المادة الناتجة المقصود من العبارات فيقوم بإعادة صياغتها بشكل جديد بحيث يعيد ترتيب الكلمات والعبارات ويولد الجملة المطلوبة ويقوم بالتأكد من دقة وسلامة الجملة. وهذا المستوى يحتاج إلى معجم ثنائي اللغة ضخمة ومعجم للكلمات التي تحتاج أكثر من كلمة تقابلها.

**المستوى التالي الأعلى من المستوى الأدنى:** في هذا المستوى تجرى بعض المعالجات الصرفية لغرض التعامل المخفي مع الكلمات. ويمكن اعتبار الترجمة بالأمثلة ضمن هذا المستوى بوجود ذخيرة من جمل وعبارات متوازية.

**المستوى المتوسط الأول:** يحتاج هذا المستوى إلى شجرة إعراب للجملة (محلل نحوي) باللغة المصدر ومن ثم تسقط على اللغة المترجم إليها.

**المستوى المتوسط الثاني:** في هذا المستوى هناك حاجة للتمثيل الدلالي بين اللغات، فمثلاً إذا قلنا "رأسي يؤلمني" بالعربية، فترجمتها بالإنكليزية هي "I have headache" والتي تترجم حرفياً إلى "لدي صداع". وعلى هذا فمن الضروري فهم المعنى وتمثيله بشكل سليم، لكي يمكن إيجاد المرادف له في اللغة الأخرى.

**المستوى الأعلى:** لا يزال البحث في هذا المستوى يجري، وذلك للأخذ بعين الاعتبار الأساليب البلاغية العميقة في اللغتين، وتكاد تعاني كل برامج الترجمة الآلية حالياً من قصور في هذا الجانب (٢٢).

### **وفي كل الأحوال على برنامج الترجمة الآلية أن يقوم بعمليتين رئيسيتين:**

١. **النقل:** أي عملية إيجاد ما يقابل كلمات وعبارات وجمل النص الأصلي، وكذلك إيجاد التراكيب النحوية المقابلة للتراكيب الواردة فيه.

٢. أما العملية الثانية فهي صياغة الجمل الناتجة من عملية النقل السابقة صياغة صحيحة صرفياً ونحوياً وأسلوبياً، مثل صياغة: (عامل) + (جمع) في صورة (عمال) و (معلم) + (جمع) + (حالة النصب أو الجر) في صورة (معلمين)، وكذلك وضع الصفة في العربية بعد الموصوف ومراعاة قواعد المطابقة اللازمة (٢١و٢٠).

وفيما يأتي تفصيل لأهم أساليب الترجمة الآلية المستعملة اليوم. ونشير إلى أن كل هذه الأساليب ما زالت قاصرة إن استعملت لوحدها، لذلك فمن الضروري للارتقاء بالترجمة الآلية اعتماد أكثر من أسلوب بشكل متوازي لرفع مستوى الترجمة خاصة لفك الغموض (٢١و٢٤).

### **الترجمة باستعمال لغة وسيطة**

تستند لغة التشبيك العالمية (٢٧) إلى تحويل اللغة المصدر إلى لغة تمثيل للنص لا يعتمد على أية لغة، ومن ثم يمكن ترجمة هذه الرموز إلى أية لغة أخرى، وهذا يعني أن هذه الطريقة يمكن أن تخدم الترجمة إلى لغات متعددة بالتوازي، بينما تحتاج الطريقة الإحصائية التي سيرد ذكرها فيما بعد إلى جهد وذخيرة لغوية متوازية بين كل لغتين يراد الترجمة بينهما. إن التحدي الذي تواجهه هذه الطريقة هو تكوين لغة تمثل المقصود باللغات كلها دون لبس لكي يمكن الترجمة منها. يحتوي برنامج الترجمة بهذه الطريقة على ثلاثة أجزاء هي:

الوحدات اللغوية والتي تضم المعاجم التي تحتوي على كلمة عالمية تمثل المفردات العالمية وتضم القواعد النحوية لتحويل اللغات الطبيعية إلى هذه اللغة العالمية بما في ذلك من علاقات وأوصاف وملصقات وكل ما تحتاجه اللغات الطبيعية لكي تكون عبارات فصيحة وواضحة ومفهومة في اللغة المراد الترجمة إليها، وهذا ما يستدعي وجود قاعدة معرفة هرمية الشكل للمفاهيم الموجودة في اللغات الطبيعية. إن هذه اللغة هي لغة اصطناعية يمكن بوساطتها العبور نحو اللغة الهدف وتتمثل كفاءتها في دقة تمثيل اللغة المصدر ومدى إمكانية تحويل ذلك التمثيل إلى اللغة الهدف.

إن الفكرة وراء الترجمة الآلية المستندة إلى التحويل (٢٨) هي أن يكون هناك تمثيل وسيط يلتقط معنى الجملة الأصلية بحيث يكون وعاء للمعرفة لا يعتمد على لغة معينة ولا على حقل معين لغرض توليد الترجمة الصحيحة. ويمكن تحسين الترجمة بصياغة المعرفة اللغوية وتزويد الحاسوب بقواعد كافية للتعامل مع الخصائص اللغوية. إن قوة هذه الطريقة الحقيقية هي في إمكانية تمثيل المعرفة بهذه اللغة الوسيطة.

تعتمد الطريقة على تمثيل النص من اللغة المصدر بشكل شبكة نحوية كلية باستعمال لغة تسمى لغة التشبيك الكلية والتي يمكن بعد ذلك تحويلها إلى أية لغة أخرى (٢٧).

وقد جربت هذه الطريقة على نحو ١٠٠٠ صفحة من موسوعة أنظمة دعم الحياة إلى اللغات الستة التي تدعمها وثائق اليونسكو وكانت النتائج لا بأس بها. وعلى هذا فحينما يكون المطلوب ترجمة نص لعدد من اللغات فإن التحويل بهذه الطريقة قد يكون خياراً مفضلاً (٣٠).

### الترجمة الآلية المستندة إلى الإحصاء

تحتاج هذه الطريقة الحديثة لجمع أكبر ما يمكن من ذخيرة لغوية (corpus) والقيام بجهد إحصائي كبير عليها لكي تصبح مناسبة للاستخدام في الترجمة



الآلية. وبالطبع فإن الذخيرة المترجمة هي بالأساس مترجمة من قبل مترجمين من البشر، وهي بذلك تستخلص خبرات البشر للإفادة منها في الترجمة الآلية (١).

تعتبر الذخيرة اللغوية لأية لغة مرتكزاً أساسياً اليوم للباحثين في اللغات الطبيعية. فالخزن على الحواسيب جعل جمع وتصنيف وتحليل الذخيرة اللغوية ميسوراً. وهذه الذخيرة تمكن الباحثين من إطلاق الوصف على خصائص اللغة وعلى النحو فيها ودراساتها تاريخياً وتغييرها مع الزمن (١٣).

تجمع الذخيرة اللغوية المناسبة (يجب أن تكون من الضخامة ما فيه الكفاية) بين محتوياتها غالبية الكلمات الشائعة في اللغة وغالبية التعابير اللغوية والتراكيب النحوية والصرفية. وعلى هذا فيتوقع منها أنها تمثل اللغة بشكل تقريبي. ونظراً لاعتماد هذه الطريقة على الذخيرة ثنائية اللغة (أو متعددة اللغات)، فيجب أن تكون هذه الذخيرة شاملة وواسعة ودقيقة، وهو أمر في غاية الندرة. وتتوفر مثل هذه الذخيرة اليوم في اللغات العالمية الأخرى غير العربية لضخامة ما يتوفر من كتابات محوسبة كالصحف والمجلات والكتب المطبوعة والمقالات المنشورة على الإنترنت والإعلانات ونشرات الشركات التجارية وهي تمثل اللغة، ومن ثم تستفيد الترجمة الآلية من مثل هذه الذخيرة (١٥ و١٦).

إن الحجم الأدنى للذخيرة التي يمكن الاعتماد عليها لغرض الاستعمال في الترجمة يبلغ مئات الملايين من الكلمات. وبهذا الصدد لا تزال الذخيرة اللغوية المتوفرة باللغة العربية مع ترجماتها للغات الأخرى محدودة.

هناك عدة أنواع من الترجمة الإحصائية، فمنها المباشرة التي تستند إلى معجم عبارات لغتين. أما الترجمة الإحصائية غير المباشرة فيجري فيها القيام بعمليات تحويل لغوية عديدة للإفادة من الذخيرة كالتحويلات الصرفية والنحوية والدلالية وحذف السوابق واللواحق والقيام بعمليات إحصائية واستعمال الأنظمة الخبيرة. كما أن هناك الترجمة الإحصائية بالأمثلة التي تحتاج للقيام بإعراب

الجملة بلغتها الأصلية ثم القيام بعدد من التحويلات بإعادة السلسلة الظاهرية في اللغة الأصلية لكي توافق التسلسل المطلوب في اللغة المترجم إليها. وتجرى هذه العمليات على الجزء من الذخيرة المستعمل للتدريب وكذلك على النصوص المراد ترجمتها ويحتاج ذلك لحل مشكلة تقابل عدة كلمات في لغة مقابل كلمة في اللغة الأخرى وبالعكس (١٧).

تستند عملية تهيئة الذخيرة حاسوبياً على استخراج الكلمات المكررة وغير المكررة في الذخيرة (٣١ و ٣٢) ثم بعد ذلك الحصول على قائمة بكل كلمتين متعاقبتين ثم قوائم العبارات المتقابلة بين اللغات المراد الترجمة منها وإليها. وبالطبع فإن هذه الطريقة بأسلوبها المبسط لا تعطي ترجمة دقيقة وهي بحاجة إلى معلومات نحوية وصرفية لزيادة الدقة. وقد استعملت هذه الطريقة في ترجمة محاضر البرلمان الكندي بين اللغتين الفرنسية والإنكليزية فوجدت ناجحة لحد كبير، وذلك بسبب أن موضوع التماثل في أروقة البرلمان ذو نمط معين وبأساليب متعارف عليها بين المتحاورين (١).

إن الخطوة التالية في معظم برمجيات الترجمة الآلية المستندة إلى الإحصاء بعد استكمال الذخيرة اللغوية المتوازنة هي توليد الكلمة المقابلة بين اللغتين وحل مشكلة عدم تساوي عدد الكلمات بين اللغتين (١٨).

تعتمد دقة الترجمة الإحصائية على نوع ودقة الذخيرة المستعملة وعلى البرمجيات المستعملة فيها (٢١). ويجدر بالذكر أن تغير اللغة والمفردات المستعملة بمضي الزمن وضخامة الذخيرة المطلوبة يستوجب وجود معجم تأريخي للغة. كمثال على ذلك فإن كلمة "السيارة" حين ترد في نص مكتوب قبل قرون تعني غير "السيارة" المكتوبة في نص مستعمل اليوم. وكذلك فإن استعمال بعض أسماء الأعلام يحتاج وجود قرائن في البرنامج لكي يقوم بترجمتها بمعناها وعدم اعتبارها أسماء أعلام. فمثلاً كلمة: "أحمد" في جملة "أحمد الله على نعمائه" هي ليست إسم علم. ومن الصعوبة بمكان على الآلة تحديد هل الكلمة ذات معنى أم اسم علم بدون قرائن كافية.

## الترجمة بالأمثلة

الترجمة المعتمدة على الأمثلة تستند إلى ترجمة جمل قصيرة باستعمال ذخيرة لغوية ضخمة مبوبة على مستوى العبارات وقد استفيد من وثائق الأمم المتحدة متعددة اللغات بهذه الطريقة. لا تعتمد هذه الطريقة من الترجمة على الهيكلية، حيث أن الأمثلة تخزن بشكلها السطحي مع تعاملات صرفية وتجزئة للكلام محدودة، حيث تقابل كل جملة مع ترجمتها باستعمال عمليات صرفية محددة. وتجري عملية المقابلة بالرجوع إلى خوارزمية تحدد المعاني التقريبية للكلمات بدل استعمال معاجم ثنائية اللغة. وأهم ما في هذه الطريقة أنها تحاكي الطريقة التي يفكر بها المترجم البشري، فالذخيرة التي تستند إليها هذه الطريقة هي القوائم متعددة اللغات للكلمات والعبارات والجمل المتوازية. وتقوم الترجمة بإيجاد المقابل للنص المطلوب ترجمته من القوائم المتوفرة. وهذه العملية تسمى التحويل، تليها عملية دمج العبارات المختلفة لتوليد النص النهائي باللغة الهدف (١).

إن الترجمة بالأمثلة لا يمكن أن تعتبر كافية لإتمام عملية الترجمة كاملة، لكنها يجب أن تتم مع طرائق أخرى للترجمة بحيث يكمل بعضها بعضاً. وتستند هذه الطريقة إلى إجراء أبحاث على الذخيرة اللغوية لاستخلاص الأمثلة والعبارات الشائعة المتقابلة بين اللغتين المراد الترجمة بينهما، إضافة إلى استخلاص قوالب نحوية بين اللغتين لتعويض الكلمات المقابلة بين اللغتين. هذه الطريقة فعالة في العادة للترجمة بين اللغات التي تعود لأصول لاتينية مثلاً. لكن الأمر يزداد صعوبة بين اللغات التي تعود إلى عوائل مختلفة.

## الترجمة الآلية واللغة العربية

تمتاز اللغة العربية بميزات خاصة مختلفة عن اللغات الأخرى، وفي الوقت نفسه تجمعها مع اللغات الأخرى ميزات عامة، لذلك يمكن الاستفادة من هذه الميزات العامة مع اللغات الأخرى والبناء عليها لتغطية خصوصيات اللغة العربية

سواء في الترجمة من العربية أو الترجمة إليها. فمن الأمور المشتركة بين عدد من اللغات تعدد المعاني للكلمة الواحدة ووجود جمل ليس فيها فعل. كما تحوي اللغة العربية على عوامل صرفية عديدة منها السوابق واللواحق والأواسط وتغيير صيغة الكلمة. هذا يعني أن أية ترجمة آلية يجب أن تحوي تحليلاً قوياً للوصول إلى الجذر.

إن الترجمة الجيدة يجب أن تقتنص المعنى بكفاءة، ونعني بالكفاءة الوصول إلى المعنى المقصود بدقة باستخدام أكبر ما يمكن من قرائن. وهذا ما تفتقده أنظمة الترجمة الآلية بالوقت الحاضر، وهو ما يضع عبئاً على من يقوم بالتصحيح بعد الترجمة، والذي قد يكون أكثر مما يمكن أن يبذل بالترجمة البشرية أحياناً إذا كانت الترجمة سيئة.

أما خصوصيات اللغة العربية ذات العلاقة بالترجمة اليوم فمنها غياب التشكيل رغم أنه جزء رئيسي من النص المنطوق، مما يفرض على القارئ افتراض تشكيل من عنده بما يملكه من ثروة لغوية. وهو ما يقوم به المترجم كذلك، مما يستدعي إيجاد بديل لذلك في الترجمة الآلية (١٠). ويزيد الأمور تعقيداً ندرة استعمال علامات الوقف والفواصل في النصوص العربية.

كما تمتاز اللغة العربية بدمج السوابق واللواحق والإدغام والإقلاّب وغيرها. أما الجمل الطويلة في اللغة العربية فتحتاج إلى عمليات حاسوبية معقدة لتحليلها مما يؤدي إلى أن تصبح المعالجة بطيئة. ومن العمليات التي تساعد في ذلك تقسيم الجملة إلى عبارات المضاف والمضاف إليه والجار والمجرور وأشباه الجمل (١١).

أما من الناحية التاريخية فإن التغييرات التي صاحبت اللغة العربية كانت أقل بكثير من لغات أخرى كالإنكليزية التي شهدت تطوراً هائلاً خلال القرون الأخيرة بالمقارنة مع اللغة العربية التي لا يزال تراثها المكتوب منذ صدر الإسلام ولحد الآن لم يحدث عليه تغيير جوهري في المفاهيم العامة.

ولغرض أن تكون الترجمة الآلية مقبولة يجب القيام بعدة عمليات على النص العربي برمجيًا، منها على سبيل المثال لا الحصر ما يأتي (١٢):

- إضافة الرموز للنص كالقوارز والفواصل والنقاط ومعالجة الهمزات والتفريق بين الهاء والتاء المربوطة والتشكيل للأحرف التي فيها لبس.
- مراعاة وجود السوابق واللاحق وتحديد أجزاء كل منها إن وجدت كألف لام التعريف وحروف العطف وحروف الجر المتصلة والضمائر وعلامات الإعراب وتحديد جذر جذع الكلمة ووزنها الصريح إذا احتيج لذلك.
- مراعاة قواعد تغيير كتابة الكلمة بعد إلحاق بعض الضمائر المتصلة مثل كلمة "مكتبة - مكتبهم" و"أعلى - أعلاه".
- تحديد فيما إذا كان العدد المتضمن ضمن الكلمة مفردًا أو مثنى أو جمعًا وذلك من خلال علامات الجمع الواو والنون والياء والنون وعلامات المثنى من خلال الألف والنون والياء والنون وغيرها.
- تحديد المعنى عندما يكون هناك وجود لأكثر من احتمال في تحديد الكلمة المراد ترجمة معناها مثل الكلمات التي تعطي أكثر من معنى باختلاف التشكيل كأن يكون الفعل مبنياً للمجهول أو للمعلوم أو أن يكون اسماً أو فعلاً مثل: ذَهَبَ وَ ذَهَبُ (١).
- التحليل النحوي لتحديد التركيب الدقيق للعبارة أو الجملة حيث يقوم البرنامج بتطبيق قواعد منطقية بهدف الوصول إلى الترجمة باللغة الهدف.
- الربط بين تركيب الجملة ومركباتها الدلالية لغرض الوصول إلى الترجمة الهدف.
- التركيب المنطقي بتمثيل الجملة الأصلية باللغة الهدف.

- تحسين الجملة باللغة الهدف نحويًا وصرفيًا ودلاليًا.

### الأبحاث في الترجمة من اللغة العربية وإليها

اهتمت في الآونة الأخيرة كثير من الجامعات الأمريكية والأوربية والعربية بمعالجة اللغة العربية حاسوبياً والترجمة منها وإليها. وتتوزع الأبحاث على النحو والصرف والدلالة والترجمة وجمع الذخيرة اللغوية وأساليب تكوين المعجم المحوسب وعملية الاستفسار باللغة العربية وتكوين الخلاصات وغيرها (١).

إن الكلمات العربية المركبة غالباً ما تتكون من سوابق ولواحق، وحيث إن عملية فصل هذه اللواحق فيها كثير من الإبهام، لذلك هناك حاجة للرجوع إلى المعنى بالإضافة إلى القواعد الصرفية والنحوية. كما أن عملية فصل اللواحق تزداد صعوبة في العربية بسبب غياب التشكيل عادة.

تقع عملية فصل اللواحق في ثلاثة أصناف: صنف يستند إلى قواعد وصنف يستند إلى الإحصاء وصنف يعتمد المعجمية كأساس له. الصنف الذي يعتمد القواعد يستند على المعرفة اللغوية المأخوذة من الخبرة البشرية بحيث تحول هذه المعرفة إلى قواعد وهذه القواعد معروفة أن الوصول إليها ليس بالسهولة بمكان، كما إن الكثير من هذه القواعد ليست حدية ولها استثناءات. أما الصنف الذي يعتمد على الإحصاء فيحتاج إلى تدريب وبعض المعرفة البشرية لبنائه.

لقد وجد أن القيام بعملية رصينة في فصل اللواحق ذو تأثير أفضل بكثير من القيام بالعملية بشكل سطحي. ويمكن أن يتحقق تحسن آخر باستعمال أكثر من وسيلة من وسائل فصل اللواحق.

إن العوامل التي تحكم ترجمة التعابير المركبة متعارف عليها لدى اللسانيين، وحيث إن معظم التعابير المركبة تكون مركبة من أسماء لذلك فهي تخضع لتصريف الأسماء في اللغة، أما استعمال حروف الجر في اللغة العربية وتلازم بعضها مع

الأسماء فإنه يجعل التعامل معها غير ممكن بشكل مباشر (١٦ و ١٧). ويلاحظ أن اللغات المختلفة يمكن أن تحتوي كلمات متشابهة التسلسل وكلمات متعكسة التسلسل وكلمات متشابهة التهجئة.

ومما يجب معالجته مشكلة المتلازمات اللفظية التي يمكن تقسيمها إلى أربع فئات حسب درجة التلازم، فالمتلازمات اللفظية المفتوحة الحرة مثل "بادئ الأمر" يمكن استبدالها بـ "أول الأمر"، فاستبدال كلمة "بادئ" بمرادفها "أول" لم يخل ذلك بالمعنى. والمتلازمات اللفظية المقيدة وهي أن ترد لفظتان معاً بحيث تستعمل إحداها بمعناها الحرفي وتستعمل الأخرى بمعناها المتخصص الذي يمكن أن يكون مجازياً مثلاً "يشق طريقه". ففي هذا المثال، لا يعني الفعل "يشق" معناه الحرفي أي "أن يقسم المرء شيئاً إلى شطرين" بل معنىً مجازياً هو "أن يسلك المرء طريقاً صعباً". وهناك العبارات المسكوكة المجازية مثل قولنا "والليل إذا تنفس" والعبارات المسكوكة المحضة مثل "وضعت الحرب أوزارها"، بمعنى انتهت وتوقفت إذ لا يمكن في هذه العبارات إبدال كلمة بكلمة أخرى أو تقديم إحداها أو تأخير أخرى (٧).

ومما يجب تكوينه المعجم العربي المحوسب الذي ينبغي أن يحصر جميع المعاني للكلمة ودراسة مدى شيوع الكلمات ودراسة تكرار الكلمة كمادة وكفئة معجمية ودراسة استخدام المرادفات وشكل الكلمة وفق حالتها الإعرابية ودراسة الكلمة وفق اشتقاقها الصريفي ومعرفة مدى السلامة والصحة اللغوية ومعرفة المتلازمات اللغوية (٧). كما أن المعجم المحوسب يجب أن يحوي رموزاً خاصة لتصريف الكلمة ومعلومات أخرى عنها تدرج فيها مفردات اللغة بالتفصيل بحيث يكون بالإمكان الإفادة منها حاسوبياً.

### محاولات الترجمة الآلية من اللغة العربية وإليها

قامت محاولات عديدة للبدء بترجمة آلية من اللغة العربية وإليها. وقد أثمر

- بعضها بتكوين أنظمة ترجمة آلية، بينما أصبح البعض الآخر طي النسيان. وفيما يأتي بعض من هذه المحاولات (١):
- نظام "المترجم العربي" الذي طورته شركة ATA في لندن، وبرنامجاً مصغراً أسمته "الوايفي" (٦).
  - نظام "الناقل العربي" الذي طورته شركة سيموس في باريس للترجمة بين العربية وكل من الإنكليزية والفرنسية (٣٥).
  - نظام شركة أبتك Apptek (٣٦).
  - نظام سيستران Systran ويترجم النظام هذا الآن بين ١٥ لغة. لقد ابتداءً تطوير سيستران لمترجم من العربية إلى الإنكليزية في عام ٢٠٠٢ وله موقع للترجمة الآن (٣٧)، وقد امتاز سيستران نتيجة استعماله المكثف للمعاجم بقابليته لترميز مكونات اللغة التفصيلية بميزات نحوية وصرفية دقيقة. لكن دراسة لفحص المتلازمات اللفظية (التي تشكل عائقاً أمام المترجم والمتعلم) على نظام سيستران وجدت أن النظام أخفق في ترجمة معظم المتلازمات اللفظية (٧).
  - شركة ألبس Alps: لديها برامج للترجمة بين عدد من اللغات، وتطبق مبدأ الترجمة التحوارية (٣٨).
  - موقع المسبار وهو موقع يهتم أيضاً بالترجمة الآلية من الإنكليزية إلى العربية وبالعكس، ويمتاز بالسهولة والمرونة عند استخدامه إذا كانت العبارات بسيطة (٣٩). ويرتبط موقعه مع المترجم العربي والوايفي.
  - موقع freetranslation ويترجم بين ٤٣ لغة (٤٠) بينها اللغة العربية.
  - لقد وظفت شركة صخر لبرامج الحاسب محرراً للترجمة الآلية الخاص بها (٨) في دعم موقع الترجمة (٤١)، والموقع متوقف الآن.



- برنامج شركة جوجل: وهو برنامج مجاني يستند إلى الترجمة الإحصائية من ذخيرة لغوية مأخوذة من الإنترنت ويتعلم من أخطائه، فإذا ما ترجم جملة خاطئة وأخبره المستخدم أن الترجمة خاطئة وأن المفروض أن تكون بشكل آخر، فإنه يخزن هذه المعلومات ويستعملها في المستقبل بشكل أصح. عدد اللغات التي يترجم منها وإليها تبلغ ٨٠ لغة (٤٢) بينها اللغة العربية.

- ترجمة بينج: من شركة مايكروسوفت. وهي تُستخدم من قبل فيسبوك وياهو لترجمة التعليقات. تستخدم بينج "الترجمة الآلية الإحصائية"، غير أنها تختلف عن ترجمة جوجل في طريقة تنفيذها أو البيانات التي تعتمد عليها. وتشمل ترجمة بينج ٤٥ لغة بينها اللغة العربية، وتسمح خاصية الترجمة التعاونية للمستخدمين بأن يصححوا أو يحسنوا الترجمات (٤٣).

### مقارنة بين بعض برامج الترجمة الآلية من اللغة العربية وإليها

سنقارن بين ثلاث برمجيات للترجمة الآلية هي جوجل وسيستران وبينج المذكورة مواقعها أعلاه. وقد اخترنا جملة عربية لترجمتها ومقارنتها إلى الإنكليزية. ثم ترجمنا الجملة إلى الإنكليزية وأدخلت تلك الترجمة للأنظمة الثلاثة للحصول على الترجمة العربية. وقد أعيدت التجربة مرتين في شهري أيلول (أشير لها بالرمز (×)) وكانون الأول (أشير لها بالرمز (××)) من عام ٢٠١٤ ويلاحظ حدوث اختلاف في كل الترجمات من الإنكليزية إلى العربية في المواقع الثلاثة، بينما الترجمة من العربية إلى الإنكليزية اختلفت في ترجمة جوجل وبقيت نفسها في سيستران وبينج. ويلاحظ عدم الدقة والركاكة في الترجمات الآلية كلها تقريباً. كما يلاحظ أن الترجمة قد تختلف بمضي الزمن.

النص الأصلي بالعربية	حصل تقدم هائل للترجمة الآلية خلال السنوات الماضية خاصة بين اللغات الأوروبية ذات الأصول اللاتينية لما بينها من مشتركات
ترجمة جوجل إلى الإنكليزية	Got tremendous progress of the translation mechanism during the past years, especially among European languages with Latinos because of the participants, including (*) Progress has been made tremendous machine translation during the past years, especially among European languages with Latin asset to them from participants (**)
ترجمة سيستران إلى الإنكليزية	Huge progress for the automatic translation during the last years collected especially between the European languages self of the Latin origins for what between her from joint (*) & (**)
ترجمة بينغ إلى الإنكليزية	Advances of machine translation over the past years, especially between the European languages with Latin origins to their subscribers (*) & (**)
ترجمة النص بالإنكليزية	A tremendous progress occurred during recent years between European languages due to their common Latin origin.
ترجمة جوجل إلى العربية	حدث تقدم هائل خلال السنوات الأخيرة بين اللغات الأوروبية بسبب الأصل اللاتيني المشترك (×) حدث هناك تقدماً هائلاً خلال السنوات الأخيرة بين اللغات الأوروبية بسبب الأصل اللاتيني المشترك (××)
ترجمة سيستران إلى العربية	تقدم هائل وقع أثناء سنوات الأخيرة بين لغات أوربي واجب إلى أصلهم لاتيني عادي (×) تقدم هائل وقع أثناء سنوات الأخيرة بين لغة أوربي واجب إلى أصلهم عادي لاتيني (××)

ترجمة بينغ إلى العربية	التقدم هائل الذي حدث خلال السنوات الأخيرة بين اللغات الأورو بية بسبب أصلها اللاتيني الشائعة (×)
	التقدم هائل الذي حدث خلال السنوات الأخيرة بين اللغات الأورو بية بسبب أصلهم اللاتينية الشائعة (××)

في حال الترجمة من العربية إلى الإنكليزية، ينصح (٤٤) باستخدام العربية الفصحى البسيطة الشبيهة بما قد يوجد في المواقع الإخبارية، والابتعاد عن التعبيرات البلاغية. كما ينصح باستخدام بنية الجملة التي تبدأ بالاسم ثم الفعل كما في بنية الجملة الإنكليزية، مثلاً "الرجل جاء" بدلاً من "جاء الرجل". فحين تكون الجمل العربية طويلة قد يتبع الفعل أكثر من عشر كلمات قبل ورود الفاعل، مما يجعل من الصعب على الترجمة الآلية أن تنقل الفعل إلى مكانه الصحيح في الجملة الإنكليزية، بل إن كثيراً من أنظمة الترجمة الآلية تحذف الفعل بالكامل.

### برامج مساعدة للترجمة

تخدم المعالجة الآلية للغة العربية بالإضافة إلى الترجمة الآلية: التلخيص الآلي والتوليد الآلي للغة واستخلاص المعلومات واسترجاع المعلومات والإجابة على الأسئلة والتنقيب في النصوص وتحويل الكلام المنطوق إلى مكتوب وتحويل النص إلى كلام منطوق والتعرف الضوئي على الحروف ... وهناك أبحاث جارية لكل هذه المواضيع. ومن البرامج المتوفرة الآن: برنامج التشكيل والتصحيح الآلي (٤٥) Arab Diac ومجموعة برامج صخر للتدقيق الإملائي Sakhr Corrector وبرنامج التشكيل الآلي (٤٦) Automatic Diacritizer وبرنامج صخر للتحليل الصرفي (٤٧) وبرنامج استخراج الكلمات المفتاحية (٤٨) Sakhr Keywords Extractor ومحرك صخر للتصنيف الشجري "سراج" ومحرك صخر للتلخيص الآلي وبرامج تنقيب ووصف الكلام (٤٩) Part of Speech Tagging. وهناك برامج التحليل الصرفي لتحديد جذور الكلمات والمعلومات الصرفية والنحوية ومن هذه البرامج:

برنامج (٥٠) ArabMorpho حيث يستخدم ذخيرة للألفاظ وتحليل إحصائي لفك الغموض ومحلل (٥١) Beesley Xerox وهو مولد ومحلل صريفي مساعد في التعليم وبرنامج باك وولتر للتحليل الصريفي (٥٢) وبرنامج التنقيب في النصوص للوصول إلى الأفكار الأساسية وبرنامج محرك البحث النصي العربي (٥٣) Swift وبرنامج Arab Dictions الذي يقوم بتحليل الكلمات العربية إلى وحداتها الصرفية ثم يربط كل الأشكال الصرفية بمادة المعجم المقابل لها وفق الجذر (٥٤)، وبرنامج معالجة الكلام المنطوق Arab Talk الذي يحول النصوص المكتوبة إلى كلام منطوق معتمداً على تشكيل آلي (٥٥)، وبرنامج إبصار للمعاقرين بصرياً والمكفوفين (٣٣).

### ترجمة الكلام المنطوق

تتوفر في الأسواق بعض القواميس الصوتية لترجمة عبارات بين لغات متعددة بينها اللغة العربية. وكثيراً ما تحتوي عبارات باللهجات العامية. ويمكن استعمال بعض الأنظمة الصوتية في حجز الفنادق والحجز على الخطوط الجوية والمشاركة في المؤتمرات وطلب الطعام من المطاعم والاستفسار عن اتجاه السير والحجز لدى العيادات الطبية والمستشفيات واستئجار السيارات وغيرها (١٩). إن مشاكل ترجمة الكلام المنطوق تزيد على مشاكل النص المكتوب، وذلك لأن الكثير من الكلام المتداول يحوي أخطاء نحوية أو جملاً ناقصة. ولغرض تكوين ترجمة آلية فورية، يجب وضع قواعد عامة ذات مرونة لا تلتزم بقواعد اللغة بصرامة، وذلك للأخذ بعين الاعتبار تحديد نطاق الكلمات التي يستعملها المتكلم وطريقة نطقها وسرعة النطق وطول الجملة وصيغ التوقف بين الجمل. وهناك محاولة لاستعمال الرسائل القصيرة على الهاتف النقال للترجمة وتمتاز بأنها قصيرة ومحدودة المعجم، وبأنها قابلة للتطور والتوسع (٢٠). ولا تزال الأبحاث لمثل هذه الأنظمة في مراحل التطوير.

## الاستنتاجات والمقترحات

الترجمة الآلية من اللغة العربية وإليها إحدى الوسائل التي تكمن فيها خدمة كبيرة في نهضة الأمة وأجيالها المستقبلية وذلك لأن اللغة العربية اليوم ليست هي اللغة العالمية للعلوم والتقنية، فما يصدر في اللغات الأخرى وخاصة الإنكليزية من أبحاث ودراسات وكتب ومؤلفات أخرى يحتاج للترجمة للغة العربية. كما أن اللغات التي يتكلم بها المسلمون بحاجة إلى التراث العربي الإسلامي غير المتوفر في تلك اللغات وهناك حاجة ماسة لترجمته. لذلك فالترجمة الآلية من اللغة العربية وإليها ليست من الكماليات بل هي من الضرورات القصوى لنهضة الأمة وعودتها لمكانتها بين الأمم. وأول ما تحتاجه الترجمة الآلية هو وجود معجم عربي محوسب وتكوين ذخيرة متعددة اللغات مع العربية لتساعد في الترجمة الآلية المستندة إلى الإحصاء ودعم الأبحاث اللغوية المتعلقة بالترجمة الآلية من اللغة العربية وإليها. ولغرض الوصول إلى ذلك هناك حاجة إلى القيام بحملة توعية للقيادات السياسية والعلمية والجهات الداعمة للبحث العلمي على أهمية البحث العلمي في حوسبة اللغة العربية من قبل الجامعات اللغوية العربية وأقسام الحاسوب واللغة العربية واللسانيات في الجامعات العربية كما ينبغي توجيه الأبحاث نحو التطبيق العملي وتكوين قيادات بحثية في أقسام اللغة العربية ذات خلفية حاسوبية جيدة وفي أقسام الحاسوب ذات خلفيات جيدة باللغة العربية لكي يكون التواصل والبحث العلمي على أتم وجه (١).

## المصادر

### المصادر العربية

١. محمد زكي خضر، اللغة العربية والترجمة الآلية - المشاكل والحلول - المؤتمر الحادي عشر للتعريب - عمان / الأردن ١٢-١٦ تشرين الأول ٢٠٠٨م.
٢. محمود اسماعيل صالح (الصيني) - الحاسوب في خدمة الترجمة والتعريب: - <http://www.wata.cc/forums/showthread.php?5421> - الحاسوب-في-خدمة-الترجمة-والتعريب-
٣. محمد بن محمد الغزالي أبو حامد - المستنصر في علم الأصول- ١٨
٤. محمد الصرايرة - اللغة العربية والترجمة الآلية - محاضرة في مجمع اللغة العربية الأردني - الموسم الثقافي التاسع عشر ٢٠٠١م
٥. أبو الحجاج محمد بشير - المعالجة الآلية للغة العربية جهود الحاضر وتحديات المستقبل - مجلة لغة العصر المصرية - ٢٠٠٩
٦. عيدان، عدنان، طارق إبراهيم، الترجمة الآلية من اللغة الإنكليزية إلى اللغة العربية: تجربة شركة أي-تي-أي . لتقنية البرامج، مركز دراسات الوحدة العربية، ص ٢٨٩
٧. أمّنة فاطمة الزهراء - إشكالية حدود الترجمة الآلية: ترجمة نظام "سيستران" للمتلازمات اللفظية - رسالة الماجستير كلية الآداب و اللغات - قسم الترجمة - جامعة منتوري -قسنطينة-الجزائر

٨. عبد الحميد بن العزلان، برنامج "عجيب" في الترجمة الآلية من اللغة الإنجليزية إلى اللغة العربية - رسالة ماجستير - الجامعة الإسلامية العالمية - ماليزيا - أغسطس ٢٠٠٥

### المصادر الأجنبية

1. Thepchai Supnithi. Virach Sornlertlamvanich. Thatsanee Charoenporn. A Cross System Machine Translation COL-ING-02 on Machine Translation in Asia - Vol 16. Sep. 2002
2. Moustafa Elshafei. Husni Al-Muhtaseb. and Mansour Al-ghamdi. Machine Generation of Arabic Diacritical Marks. The 2006 International Conference on Machine Learning; Models, Technologies & Applications (MLMTA'06).
3. Habash. Nizar. Bonnie Dorr and Christof Monz.. Challenges in Building an Arabic-English GHMT System with SMT components. Proceedings of the Association for Machine Translation in the Americas (AMTA-2006). Boston, MA. 2006.
4. Fatiha Sadat. Nizar Habash. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. pages 1-8. Sydney. July 2006.
5. Ahmed Abdelali. Jim Cowie and Hamdy S. Soliman. Building A Modern Standard Arabic Corpus. Workshop on computational Modeling of Lexical acquisition. The split meeting Croatia 25<sup>th</sup> - 28<sup>th</sup> July 2005.

6. (14) Ali Farghaly. Arabic Machine Translation: A Developmental Perspective. International Journal on Information and Communication Technologies. Vol. 3. No. 3. June 2010
7. (15) Chris Callison-Burch Philipp Koehn Miles Osborne. Improved Statistical Machine Translation Using Paraphrases. Proceedings NAACL-2006
8. (16) Preslav Nakov Hwee Tou Ng. Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. Journal of Artificial Intelligence Research 44 (2012) 179-222
9. (17) Zouhair MAALEJ. Guidelines for the translation of English nominal compounds into Arabic: A computational discourse model. The Tunisian Review of Modern Languages 01/2003; 11:139-167.
10. (18) Saab Mansour and Hermann Ney. Arabic-Segmentation Combination Strategies for Statistical Machine Translation LREC 2012
11. Ying Zhang. Survey of Current Speech Translation Research. Language Technologies Institute. Carnegie Mellon University. [citeseer.ist.psu.edu/752690.htm](http://citeseer.ist.psu.edu/752690.htm)
12. Avinash J. Agrawal. Manoj B. Chandak Mobile Interface for Domain Specific Machine Translation Using Short Messaging Service [ieeexplore.ieee.org/iel5/4151644/4151645/04151821.pdf?tp=&isnumber=&arnumber=4151821](http://ieeexplore.ieee.org/iel5/4151644/4151645/04151821.pdf?tp=&isnumber=&arnumber=4151821)



13. Nicola Ueffing, Hermann Ney. Word-Level Confidence Estimation for Machine Translation. RWTH Aachen University. [portal.acm.org/citation.cfm?id=1220671](http://portal.acm.org/citation.cfm?id=1220671)
14. Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz, John Hutchins. Computer linguistik: was geht, was kommt? Computational linguistics: achievements and perspectives. Festschrift für Winfried Lenders (Sankt Augustin: Gardez Verlag, 2002), p. 159-162] Machine translation today and tomorrow: [www.hutchinsweb.me.uk/Lenders-](http://www.hutchinsweb.me.uk/Lenders-HYPERLINK) HYPERLINK "http://www.hutchinsweb.me.uk/Lenders-2002.pdf"2002
15. Zhuang Xinglai. The Emerging Role of Translation Experts in the Coming MT Era. Translation Journal. Volume 6. No. 4 Oct. 2002
16. Fuji Ren and Hongchi Shi. Parallel Machine Translation: Principles and Practice. [doi.ieeecomputersociety.org/10.1109/ICECCS.2001.930184](http://doi.ieeecomputersociety.org/10.1109/ICECCS.2001.930184)
17. Yasser Salem. Generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model. M.Sc. Thesis. Institute of Technology Blanchards town, Dublin, Ireland. April. 2009
18. Mostafa Aref, Muhammed Al-Mulhem & Husni Al-Muhtaseb. English to Arabic Machine Translation: A Critical Review and Suggestions for Development. The Fourth Saudi Engineering Conference, 1995 Vol. 3, pp421-4

19. Sameh Alansary. Magdy Nagi. Noha Adly. The Arabic Universal Networking Language System. 7th International Computing Conference in Arabic (ICCA). Riyadh. Saudi Arabia. May 31 – June 2 2011
20. Omar Shirko. Nazlia Omar. Haslina Arshad and Mohammed Albared. Machine Translation of Noun Phrases from Arabic to English Using Transfer-Based Approach. Journal of Computer Science 6 (3): 350-356. 2010
21. W. John Hutchins. Machine Translation. A Brief History. Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by: E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press. 1995. Pages 431-445
22. Yasser Salem. Arnold Hensman. Brian Nolan. Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model. Proceedings of the Eighth International Conference on Information Technology and Telecommunication (ITT 2008). Galway. Ireland pp 103-110
23. Abu Shquier. M. and T. Sembok. Word Agreement and ordering in English Arabic machine translation. Proceeding of the International Symposium on Information Technology. Aug. 2008. IEEE Xplore. pp: 1-10
24. Thuy Nguyen & Stephan Vogel. Context-based Arabic Morphological Analysis for Machine Translation. Proceedings of the Twelfth Conference on Computational Natural Language Learning. 2008/8/16 pp 135-142

25. [http://www.ebsarfoundation.org/?page\\_\\_id=151](http://www.ebsarfoundation.org/?page__id=151)
26. Sabine Hunsicker, Chen Yu & Christian Federmann. Machine Learning for Hybrid Machine Translation. Proceedings of the 7th Workshop on Statistical Machine Translation. pp. 312–316. Montreal, Canada. June 7–8. 2012.
27. <http://taln-arabe.blogspot.com/2010/03/mlts.html>
28. <http://www.apptek.com/products/omnifluent-translate/index.html>
29. [www.systranet.com](http://www.systranet.com)
30. Patrick Corness. The ALPS computer-assisted translation system in an academic environment. Proceedings of a conference on Translating and the Computer 14–15 November. London: Aslib. 1986
31. <http://www.almisbar.com/>
32. [www.freetranslation.com](http://www.freetranslation.com)
33. <http://www.tarjim.com> HYPERLINK "<http://www.tarjim.com/>"
34. <http://translate.google.com>
35. <http://www.bing.com/translator/>
36. <https://www.facebook.com/journalismwork/posts/613446758684466>

37. [http://www.rdi-eg.com/rdi/technologies/arabic\\_\\_nlp\\_\\_pg.htm](http://www.rdi-eg.com/rdi/technologies/arabic__nlp__pg.htm)
38. [http://www.rdi-eg.com/rdi/technologies/arabic\\_\\_nlp\\_\\_pg.htm](http://www.rdi-eg.com/rdi/technologies/arabic__nlp__pg.htm)
39. <http://sakhr.software.informer.com/>(48) Samhaa El-Beltagy & Ahmed Rafea. KP-Miner: A keyphrase extraction system for English and Arabic documents. Journal of Information System. Vol 34 No. 1. March 2009 pp 132-144(49) <http://cl.indiana.edu/~skuebler/papers/arab.pdf>
40. <http://www.aclweb.org/anthology/W10-3808.pdf>
41. [www.helsinki.fi/esslli/evening/.../beesley-helsinki.pp](http://www.helsinki.fi/esslli/evening/.../beesley-helsinki.pp)
42. <http://www.bibalex.org/ica/ar/about.aspx>
43. <https://www.academia.edu/2547637/>
44. <http://www.rdi-eg.com/technologies/ArabDiaction.aspx>
45. <http://arab-talk.com/>



## تقنيات التعرف الآلي على الكلام المنطوق وتطبيقاتها في القرآن الكريم: واقع وطموح

د. يحيى محمد الحاج (\*)

yelhadj@ariscam.org

(\*)

أستاذ مساعد في جامعة الإمام محمد بن سعود الإسلامية

حصل على درجة البكالوريوس في علوم الحاسب من جامعة القاضي عياض بالمغرب سنة ١٩٩٦ ثم نال درجة الماجستير والدكتوراه في المعالجة المتوازية والموزعة ضمن برامج أوروبية مشتركة في سنتي ١٩٩٨ و ٢٠٠١. التحق بهيئة التدريس بقسم علوم الحاسب في كلية علوم الحاسب والمعلومات بجامعة الإمام بالمملكة العربية السعودية سنة ٢٠٠٢، وتم تكليفه مؤخراً بالعمل في مركز الحوسبة في المجالات الشرعية والعربية الذي أنشئ حديثاً بالجامعة والمجموعة البحثية حول الحوسبة في المجالات العربية والشرعية ARISCOM.



## ملخص

تمثل هذه الورقة بحثاً استطلاعياً حول تقنيات التعرف الآلي على الكلام المنطوق وواقع استخداماتها في مجال القرآن الكريم. يتم في البداية تقديم عرض مفصل حول تقنيات التعرف الآلي على الكلام، مجالات تطبيقها بشكل عام ومفاهيمها الأساسية وأدواتها التقنية وطرق بناء تطبيقاتها. ومن ثم يتم استعراض أهم الأعمال والجهود المبذولة في سبيل تطوير هذه التقنيات لخدمة القرآن الكريم كرسد للواقع الفعلي لمدى الاستفادة منها في مجال تعلم وتعليم القرآن الكريم. وتختتم الورقة بلفت أنظار المهتمين والمعنيين إلى ضرورة تكثيف الجهود وتضافرها للاستفادة من هذه التقنيات واستغلالها بشكل أمثل لتسهيل تعلم وتعليم القرآن الكريم؛ وتشير الورقة إلى نماذج يمكن أن يكونا نواه لأنظمة فعلية وبيئات تفاعلية للتعلم الذاتي للقرآن الكريم.





## مقدمة

يعتبر مجال التعرف الآلي على الأصوات المنطوقة أو ما يعرف اختصاراً بالتعرف الآلي على الكلام من أهم وأقدم المجالات الحاسوبية التي حظيت بجهود بحثية معتبرة لكون تواصل الإنسان مع الآلة بلغته الطبيعية يمثل حلماً طالما قد راوده [١]. ونظراً لأهمية الكلام لدى الإنسان والأريحية في التعاطي معه، تزايدت بشكل ملحوظ الاستخدامات المتعددة لتقنيات التعرف الآلي على الكلام وظهرت تطبيقاته في مجالات شتى، منها التحكم عن بعد باستخدام الهاتف، ومساعدة المعوقين وذوي الاحتياجات الخاصة، والتعرف على المتحدث، والتعرف على اللغة، وإدخال النصوص، والأرشفة ثم البحث والاسترجاع [١]. وقد برزت في السنوات الأخيرة توجهات جديدة تسعى إلى تطبيق هذه التقنيات في الأغراض التعليمية وبالذات في تعليم القراءة وتصحيح النطق. وبالرغم من حداثة هذه التوجهات البحثية، إلا إنها حظيت باهتمام كبير ومتسارع في لغات عدة؛ حيث أطلقت مشاريع عملاقة في بلدان مختلفة، نذكر من بينها: مشروع "ليس" [٢-٣] و"كوليت" [٤-٥] لتطوير برمجيات للتعرف على قراءة الأطفال وتحسين مستواها، ومشروع "سبيس" [٦-٧] لتطوير تقنيات للتعرف على الكلام المنطوق بشكل يتناسب مع التطبيقات التعليمية والطبية، ومشروع "القارئ الرفيق" [٨] لمساعدة الأطفال والكبار على تعلم القراءة، ومشروع "القارئ المساعد" [٩] لمساعدة الطلاب على تحسين قدراتهم على القراءة الصحيحة وزيادة رصيدهم اللغوي ورفع مستوى فهمهم.

أما على مستوى اللغة العربية فالجهود لا تزال محدودة، ولم تنتشر بعد الاستخدامات الفعلية لتقنيات التعرف الآلي على الكلام في تعليم اللغة العربية بشكل عام سواء للناطقين بها أو لغير الناطقين بها على غرار ما يوجد في لغات أخرى. ولعل من أهم المجالات التطبيقية لتقنيات التعرف الآلي على الكلام

العربي المنطوق، تعليم النظام الصوتي للعربية الفصحى لمساعدة المسلمين (عرباً وعجماً) على تلاوة القرآن بشكل صحيح سواء لحفظه أو لأداء المناسك والشعائر الإسلامية. فالقرآن هو معجزة الإسلام الخالدة، أنزله الله ليتلى ويتدبر ويعمل به؛ ونبه الرسول صلى الله عليه وسلم على فضله وفضل تعلمه وتعليمه، حيث قال "خيركم من تعلم القرآن وعلمه". وقد ظهر الحرص على حفظ القرآن الكريم وضبط تلاوته منذ نزوله على الرسول صلى الله عليه وسلم [١٠]، حيث تولاه الصحابة رضوان الله عليهم مهمة حفظه وتحفيظه لجيل التابعين رضي الله عنهم، وانتقل متواتراً عبر الأجيال يتعلمه كل جيل مشافهة عن الجيل السابق. وقد ظهرت الدراسات المتعلقة بوصف أصوات العربية من حيث مخارجها وطرق إخراجها في القرن الثاني الهجري على يد الخليل بن أحمد الفراهيدي وتلميذه سيبويه [١١]؛ وتبع ذلك في القرن الثالث وما بعده تقعيد علم التجويد وتدوينه كعلم يهتم بتلاوة القرآن الكريم التلاوة الصحيحة [١٢]. وقد أبدع سيبويه ومن تبعه من علماء اللغة في وضع توصيفات دقيقة لأصوات العربية، وقد سبقوا غيرهم من الأمم في ذلك الوقت. ومع التطور التقني الذي شهده العالم ظهرت أجهزة ومعدات تقنية مكنت من معرفة الجهاز الصوتي لدى الإنسان وما يصدر عنه من موجات صوتية أثناء الكلام واستفادت منها لغات أخرى؛ ولم يتوقف الأمر عند هذا الحد بل برزت تقنيات حاسوبية تسعى إلى تحديد الصوت المنطوق والتأكد من سلامة نطقه؛ إلا إن هذه الوسائل التقنية لم تستغل بما يكفي لخدمة اللغة العربية بشكل عام والقرآن الكريم بشكل خاص.

نهدف في هذا البحث إلى استعراض تقنيات التعرف الآلي على الكلام لتقريبها من القارئ العربي، واستطلاع تطبيقاتها الحالية في القرآن الكريم؛ ونوجه الراغب في الاستزادة من المعلومات المتعلقة بالأجهزة الصوتية الحديثة وكيفية الاستفادة منها في ضبط مخارج الحروف وتعلم أحكام التجويد، نوجهه إلى بحث الدكتورين منصور الغامدي وعبد الله الأنصاري [١٣] المقدم في ندوة القرآن

الكريم والتقنيات المعاصرة "تقنية المعلومات" بمجمع الملك فهد لطباعة المصحف الشريف سنة ٢٠٠٩، وكذلك بحث الدكتور غانم قدوري [١٤] وأحمد راغب [١٥].

تضم هذه الورقة - غير المقدمة والخاتمة - قسمين أساسيين: قسمًا نستعرض فيه بنوع من التفصيل الخلفية العلمية لتقنيات التعرف الآلي على الكلام وأدواته ومنهجية بناء تطبيقاته، في حين نقدم في القسم الآخر أهم الأعمال التي اهتمت باستخدام تقنيات التعرف الآلي على الكلام في مجال القرآن الكريم كرسد للواقع الفعلي لاستغلال هذه التقنيات في خدمة القرآن.

### تقنيات التعرف الآلي على الكلام وأدواتها

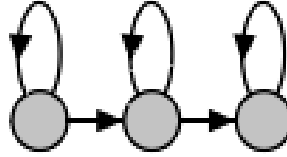
لقد أفضت الجهود البحثية في مجال التعرف الآلي على الكلام إلى نجاحات معتبرة وتجسدت بشكل كبير في العقود الثلاثة الأخيرة مع استخدام الطرق الإحصائية، وخصوصًا ما يعتمد منها على نماذج مركوف الخفية (Hidden Markov Models - HMM) [١٦-١٨]. وسنستعرض بتفصيل، في الأجزاء التالية، منهجية التعاطي مع تقنيات التعرف الآلي على الكلام المعتمد على هذا النوع من النمذجة.

### منهج بناء أنظمة التعرف الآلي على الكلام المعتمد على نماذج ماركوف الخفية

يعتمد مبدأ استخدام نماذج مركوف الخفية (HMMs) [١٩] في التعرف الآلي على الكلام، على تقسيم مفردات المعجم اللغوي المراد التعرف عليها إلى سلسلة من الوحدات الصوتية التي يتم تمثيل كل واحدة منها بنموذج صوتي (Acoustic Model) مكون من مقاطع متوالية (Sequence of States)، حيث يمثل كل مقطع في النموذج قسمًا متجانسًا (أو أجزاء متجانسة) من الإشارة الصوتية للصوت الممثل.

### البنية النمطية لنماذج ماركوف الخفية في تمثيل الأصوات المنطوقة

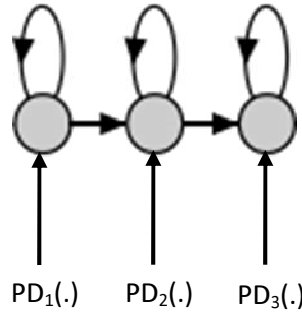
تبين أغلب الدراسات الحديثة أنه يكفي استخدام نماذج ماركوف خفية ذات ثلاث مقاطع رئيسية (Three Emitting States) لتمثيل الأصوات في بنيتها الأساسية [١]، بحيث يمثل المقطعان الأول والأخير الأجزاء الانتقالية مع الوحدات الصوتية المجاورة في حين يمثل المقطع الأوسط الجزء الثابت والمستقر من الصوت نفسه؛ الشكل ١ يقدم تمثيلاً تصورياً لهذا النموذج.



الشكل ١ : البنية الأساسية لتمثيل الوحدات الصوتية

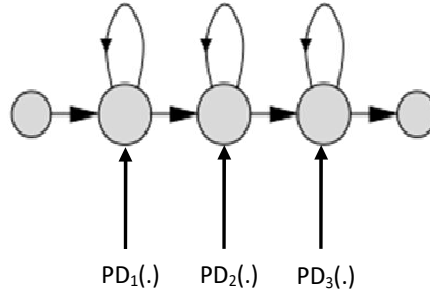
وفي هذا النموذج، يلاحظ أن الأجزاء المتشابهة (أو المتجانسة) في بداية الإشارة الصوتية يتم تثبيتها على المقطع الأول في النموذج من خلال السهم الذي يخرج من المقطع ويعود إليه؛ ومع التقدم الزمني في الإشارة الصوتية وتغير خصائصها يتم الانتقال إلى المقطع الثاني (الأوسط) الذي بدوره تثبت عليه الأجزاء المتشابهة من الإشارة الممثلة للقسم الثابت والمستقر من الصوت عبر السهم الذي يخرج من المقطع ويعود إليه؛ بعد ذلك يتم الانتقال إلى المقطع الثالث لتثبت عليه الأجزاء المتبقية من الإشارة الصوتية كذلك بواسطة السهم الذي يخرج من المقطع ويعود إليه.

ولضبط عمليات مطابقة أجزاء الإشارة على مقاطع النموذج الثلاثة، يتم استخدام توزيعات احتمالية (Probability Distributions-PDs) ترفق بالمقاطع لتمثيل الجزء الصوتي المصاحب، (أنظر الشكل ٢).



الشكل ٢: البنية الأساسية لتمثيل الوحدات الصوتية مع التوزيعات الاحتمالية

ومن الناحية الإجرائية، يضاف عادة إلى المقاطع الثلاثة (المذكورة آنفاً) مقطعان استثنائيان، واحد لتمثيل البداية والآخر لتمثيل النهاية، وذلك لربط عناصر السلسلة الصوتية بعضها ببعض (انظر الشكل ٣).



الشكل ٣: البنية الإجمالية لتمثيل الوحدات الصوتية

وبناء على ما سبق فإن كل وحدة صوتية تمثل بنموذج مركوف خفي (HMM) (انظر الشكل ٤) يرمز له بـ  $\lambda = (\Pi, A, B)$  ،

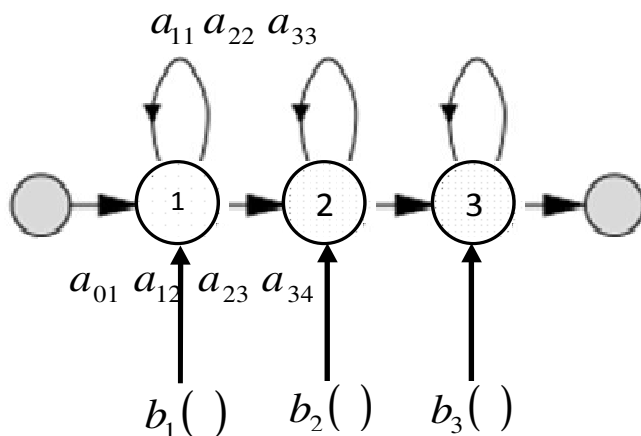
حيث:

أ.  $\Pi = (\pi_i)_{1 \leq i \leq N}$  : تمثل توزيع القيم الاحتمالات الأولية لمقاطع النموذج، وهي دائماً تساوي "واحد" عند المقطع الذي يمثل مدخل النموذج ومنطلقه (  $\pi_1 = 1$  )، و"صفر" لبقية المقاطع (  $\pi_i = 0, i = 2, \dots, N$  )؛ أما  $N$  فترمز لعدد المقاطع في النموذج.

ب.  $A = (a_{ij})_{1 \leq i, j \leq N}$  : مصفوفة تضم قيماً عددية تمثل احتمالات الانتقال الممكنة والمسموح بها بين مقاطع النموذج، وتحدد الهيكل أو البنية العامة للنموذج (Model Topology).

ت.  $B = (b_i(.))_{1 \leq i \leq N}$  : مجموعة التوزيعات الاحتمالية المرفقة بمقاطع النموذج؛ حيث يرمز  $b_i(.)$  للتوزيع الاحتمالي المرفق بالمقطع رقم  $i$  في النموذج. ومن خلال هذه التوزيعات يتم التعرف على الأجزاء الصوتية أثناء مطابقة الإشارة على مقاطع النموذج.

نشير إلى أن التوزيعات الاحتمالية  $b_i(.)$  عادة ما تكون من نوع (Gauss-ian)؛ وقد تكون هذه التوزيعات منفردة بواقع واحد لكل مقطع (Single Gaussian) بالنسبة للتطبيقات البسيطة، أو خليط من التوزيعات (Gaussian Mixture Models - GMMs) في كل مقطع من مقاطع النموذج بالنسبة للتطبيقات الأكثر تعقيداً وذلك لزيادة القدرة على تمييز الخصائص الصوتية وضبطها بشكل أدق.



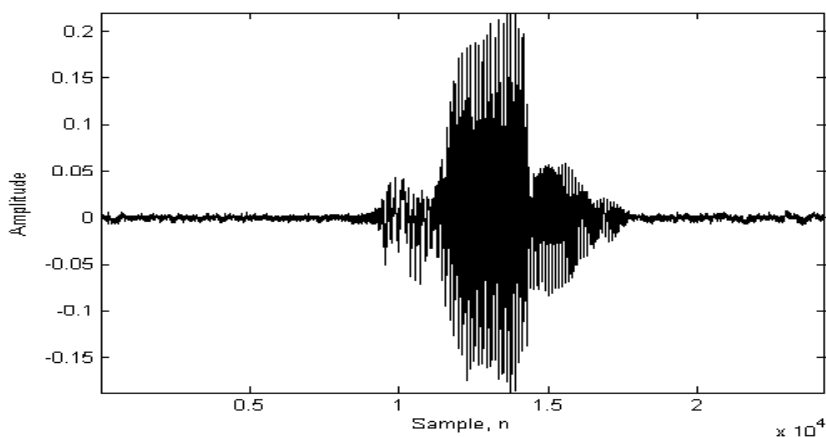
الشكل ٤ : بنية نمطية لنموذج ماركوف ذو ثلاث مقاطع أساسية

وتحتاج نماذج ماركوف الخفية بشكل عام إلى عمليات تدريب مكثفة على نسخ متعددة للوحدات الصوتية التي تتم نمذجتها لتتمكن التوزيعات الاحتمالية من تجميع الخصائص الصوتية المختلفة حسب النطق. ومن أجل ذلك لا بد من وجود قاعدة بيانات صوتية معدة سلفاً وفقاً لمواصفات ومميزات مناسبة.

### استخلاص الخصائص الصوتية من الإشارات لتدريب النماذج

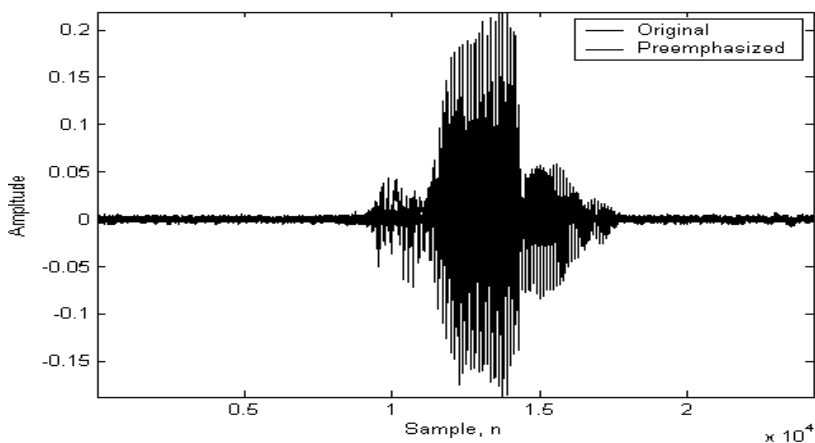
يبدأ التحضير لعملية التدريب، بمعالجة الإشارات الصوتية في قاعدة البيانات واستخراج الخصائص منها (Feature Extraction)، وهو ما يقتضي تقسيم الإشارة إلى مقاطع أو نوافذ زمنية قصيرة (تتراوح عادة بين ١٠ و ٢٠ ملсекند) يتم تمثيل كل منها بمتجهات تضم مجموعة من الخصائص المميزة للمقطع. وهناك أنواع من الخصائص وطرق متعددة لاستخلاصها، لكن أكثرها استخداماً وشيوعاً في مجال التعرف الآلي على الكلام هو ما يعرف بـ (Mel Frequency Cepstral Coefficients). وللتحضير لاستخلاص هذه الخصائص، لا بد من تحويل الإشارة من صيغتها العادية على شكل موجة إلى صيغة رقمية قابلة للتخزين والمعالجة (أنظر الشكل ٥). وهنا ينبغي تحديد معدل العينات (Sampling Rate) التي تقيس عدد العينات المأخوذة في الثانية من الإشارة الصوتية إضافة إلى سعة تمثيلها (عدد البتات المستخدمة). وبما أن أغلب المعلومات في الكلام البشري موجودة في ترددات أقل من عشرة آلاف هرتز، فإن عشرين ألف عينة في الثانية كافية للتعرف على الكلام. فالكلام المسجل مثلاً بواسطة الهاتف لا يحتاج إلا إلى أربعة آلاف هرتز، أي أن ثمانية آلاف عينة في الثانية كافية له.





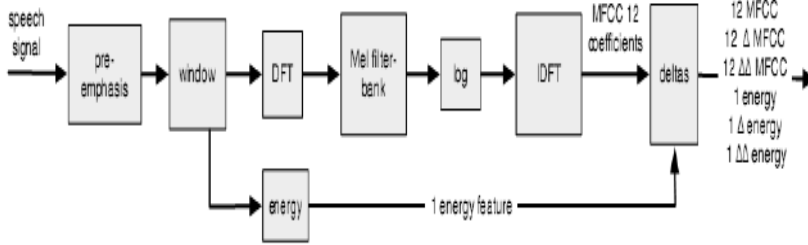
الشكل ٥ : إشارة صوتية مقسمة إلى عينات

وتمر عملية استخلاص الخصائص أو المؤثرات الصوتية بعدة مراحل كما يبينها الشكل ٧. وتبدأ هذه المراحل بمعالجة الترددات العالية والمنخفضة لتكون في مستوى متناسب (preemphasis) (الشكل ٦).



الشكل ٦ : معالجة الترددات العالية والمنخفضة

يتم بعد ذلك تقسيم الإشارة إلى نوافذ زمنية صغيرة يتم تمثيل كل جزء منها بمتجه يتكون من مجموعة من القيم (عادة ما تكون ٣٩ قيمة) تمثل الإثنا عشر عنصراً الأولى من قيم MFCC مضافاً إليها قيمة الطاقة للنافذة الزمنية لتمثيل الخصائص الثابتة في المقطع الصوتي، ثم يضاف إلى ذلك قيمة الاشتقاق الأول والثاني لهذه العناصر لتمثيل الخصائص الديناميكية للمقطع. نشير هنا إلى أن عدد القيم في المتجه الذي يمثل المقطع الصوتي يمكن أن يكون أقل أو أكثر حسب أنواع التطبيقات، إلا إن القيم المذكورة هنا (٣٩ قيمة) تعتبر متوسطة وهي كثيرة الاستخدام في تطبيقات التعرف الآلي على الكلام.



الشكل ٧: خطوات استخلاص الخصائص الصوتية (مستقاه من [١])

### تدريب النماذج الصوتية

يتم التدريب بإدخال الخصائص الصوتية المستخرجة من الإشارات الصوتية إلى النماذج بشكل تكراري حتى تصل النماذج إلى القيم المثلى (Optimal Values). وهذا يعني أننا نقوم بتغيير قيم النموذج الصوتي التي رأيناها سابقاً، أي مجموع قيم مصفوفة الانتقال بين مقاطع النموذج  $(a_{ij})_{1 \leq i, j \leq N}$  وقيم التوزيعات الاحتمالية المرفقة بمقاطع النموذج  $(b_i(\cdot))_{1 \leq i \leq N}$ . يتم ذلك من خلال قاعدة بيانات التدريب التي يفترض أن تضم مجموعة من التكرارات (نطق متكرر) لكل وحدة صوتية. ويمكن أن يتم التدريب بطريقتين، تدريب منفصل للوحدات الصوتية يعرف بـ (Isolated-Unit Training) أو تدريب مدمج (Embedded-Unit Training).

فالنوع الأول يقتضي أن يكون كل ملف صوتي في قاعدة بيانات التدريب مقسماً وممرزاً بحيث تحدد البداية والنهاية الزمنية لكل وحدة صوتية فيه؛ ويتم تجميع الأجزاء الخاصة بكل وحدة صوتية من مختلف الملفات وتستخدم مع بعض لتدريب النموذج الذي يمثل تلك الوحدة الصوتية. وتعتمد جودة التدريب هنا على دقة تحديد الحدود الزمنية للأصوات. ونظراً لما يتطلبه ذلك من وقت وجهد فإنه قد يكون من الصعب تجهيز قواعد بيانات صوتية كبيرة وفقاً لهذه الطريقة وبالتالي يتم اللجوء إلى الطريقة الثانية التي تعرف بالتدريب المدمج. وفيها يتم الاكتفاء بتحديد قائمة الوحدات الصوتية لكل ملف صوتي ثم يترك للخوارزميات مهمة تقسيم الإشارة وتحديد الحدود الزمنية بشكل آلي من خلال مطابقة الإشارة الصوتية على تلك القائمة. ومع التقدم في مراحل التدريب تعمل الخوارزميات على ضبط تلك الحدود بشكل أفضل تدريجياً حتى تصل إلى الأمثل. ويبقى وجود بيانات صوتية مقسمة زمنياً على مستوى الوحدات أمراً بالغ الأهمية في تسريع التدريب وضبطه بشكل أفضل، إضافة إلى أهميته في عدة أنواع من التطبيقات.

وقبل الشروع في التدريب بأي من الطريقتين، يتم عادة تحديد قيم ابتدائية للنموذج تكون منطلقاً له؛ ويتم ذلك في الطريقة الأولى بتقسيم الإشارة بشكل متساو بين مقاطع النموذج ثم يتم استخدام إحدى صور الخوارزمية المعروفة باسم صاحبها Viterbi [٢٠-٢١] لتحديد التسلسل الأمثل؛ وخوارزمية Viterbi تعتمد على مبدأ البرمجة الديناميكية، وقد ظهرت لأول مرة في منتصف الستينيات من القرن الماضي. أما في الطريقة الثانية فيتم حساب متوسط وانحراف عام على كل بيانات التدريب ويتم استخدامهما كقيم ابتدائية لكل التوزيعات الاحتمالية في النموذج (وهذا النمط يعرف بـ "الإنطلاقة المنبسطة أو المستوية" Flat Start)؛ وبالنسبة لاحتمالات الانتقال بين مقاطع النموذج فيكفي أن توضع بشكل متساو في البداية على أن تتغير أثناء التدريب. بعد تلك التحضيرات الأولية يتم التدريب الفعلي باستخدام خوارزمية تعرف هي الأخرى بأسماء أصحابها Baum-Welch

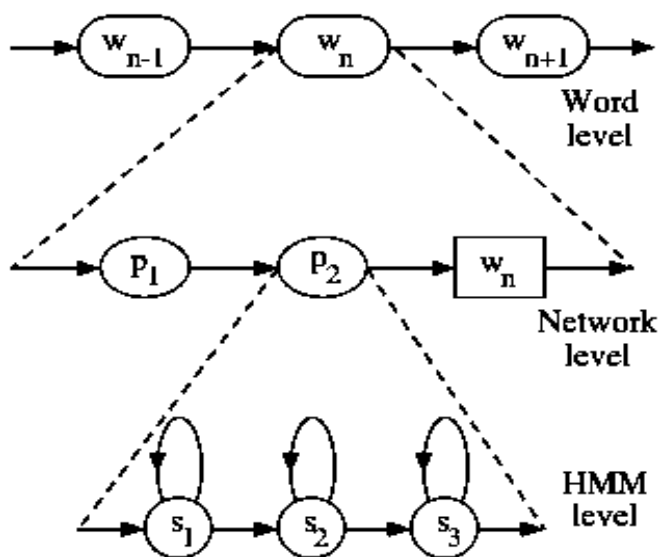
[٢٢]، وتعمل على حساب القيم المثلى لنماذج ماركوف الخفية بإدخال الخصائص الصوتية بشكل تكراري.

### اختبار وتجريب النماذج

بعد الانتهاء من تدريب النماذج الصوتية، يتم عادة العمل على قياس مدى نجاعتها من خلال عرض عينات صوتية عليها ومحاولة تحديد ما نُطق ثم قياس مدى صحته. ويحتاج هذا الأمر إلى وضع نموذج لغوي (language model) يحدد التتابعات المسموح بها حسب نوع التطبيق؛ فقد يكون التطبيق عبارة عن كلمات أو وحدات منفردة يتم نطق كل منها على حدة والتعرف عليه بشكل منفصل (ويعرف هذا النوع بـ Isolated-Words Recognition أي التعرف على كلمات منفردة)، ويكون النموذج اللغوي في هذه الحالة بسيطاً؛ وقد يكون التطبيق عبارة عن جمل تضم سلسلة من الكلمات وفقاً لقواعد يتم تحديدها وتشكل طبيعة النموذج اللغوي (ويعرف هذا النوع بـ Connected-Words Recognition أي التعرف على كلمات متصلة أو مترابطة)؛ وقد يكون التطبيق يسعى إلى التعرف الآلي على الكلام في صيغته الطبيعية (وهو ما يعرف بـ Continuous Speech Recognition أي التعرف على الكلام المتواصل)، ويبني النموذج اللغوي في هذه الحالة وفقاً للقواعد اللغوية التي تحكم بنية الجملة. ويستخدم الجانب النصي في ذخيرة التدريب لحساب احتمالات التتابعات الممكنة وفقاً لنماذج إحصائية تعرف بـ "N-Gram". ويلعب النموذج اللغوي دوراً مهماً في تحسين نسبة التعرف لأنه يساعد على تحديد الاختيار الأمثل من بين الاحتمالات الممكنة بناءً على التسلسلات التي تدرب عليها.

ففي مرحلة التعرف يتم أولاً بناء فضاء للبحث (Search Graph) على شكل شبكة تربط بين الوحدات الصوتية وفقاً لما يحدده النموذج اللغوي؛ فإذا كان النموذج اللغوي معرّفاً على مستوى الجمل، فإن الشبكة تبنى على ثلاثة مستويات، الأول يمثل كيفية توالي الكلمات في الجملة، أما الثاني ففيه يتم تعويض كل كلمة

بحالاتها النطقية وفقاً لما يحدده القاموس النطقي (Pronunciation Lexicon)، وفي المستوى الثالث يتم تعويض كل وحدة في القاموس النطقي بالنموذج الصوتي الذي يمثلها (انظر الشكل ٨). أما إن كان النموذج اللغوي معرّفاً على مستوى الوحدات اللغوية الأصغر من الكلمة، فيتم تعويض كل وحدة بنموذجها الصوتي مباشرة للحصول على الشبكة.



الشكل ٨: توليد فضاء أو شبكة البحث في مرحلة التعرف

يتم البحث في هذه الشبكة -أيضاً- باستخدام خوارزمية البرمجة الديناميكية Viterbi التي تعمل على إيجاد أفضل مسار في الشبكة، وهو عبارة عن سلسلة من الوحدات اللغوية، حيث يتم تحويلها بواسطة القاموس النطقي إلى كلمات ثم إلى جمل عبر النموذج اللغوي؛ وذلك في مسار عكسي لبناء الشبكة.

### قياس أداء أنظمة التعرف

لقياس أداء نظام التعرف، يتم مقارنة ما تم التعرف عليه بواسطة خوارزمية

Viterbi مع ما تم نطقه أصلاً، حيث ينبغي أن تكون هناك عينة مجهزة سلفاً وموضوعة في صيغة مماثلة لما يخرجها نظام التعرف؛ وتعرف هذه العينة بعينة الاختبار، وهي جزء من قاعدة البيانات الصوتية التي أشرنا إلى ضرورة وجودها خلال الحديث عن التدريب.

وكوحدة لقياس الأداء، يتم حساب عدد الكلمات أو الوحدات الصوتية التي تم التعرف عليها بشكل صحيح (دون النظر إلى مكانها وترتيبها حسب المدخلات) وتقسم على عدد الكلمات المدخلة؛ ويعرف هذا المقياس بـ "Word Recognition Rate" أي نسبة التعرف على الكلمات. ويوجد مقياس أكثر صرامة لقياس دقة نظام التعرف (يسمى "Recognition Accuracy" أي دقة التعرف)، حيث تحسب فيه نسب الاستبدال والإضافة والحذف، بمعنى أن الكلمات التي يتم التعرف عليها بشكل صحيح لا تحسب إلا إذا كانت في مكانها المناسب حسب ترتيب المدخلات.

### أدوات بناء أنظمة التعرف الآلي على الكلام

يعتبر بناء أنظمة التعرف الآلي على الكلام بشكل عام أمراً صعباً لكونه يحتاج إلى جهود كبيرة وتخصصات مختلفة (صوتية ولغوية وحاسوبية وغيرها). ولتسهيل وتسريع بناء تلك الأنظمة تم التوجه في الأوساط الأكاديمية إلى إنشاء بيئات تطويرية توفر جملة من الأدوات الأساسية يمكن للمطور الاستعانة بها في بناء تطبيقاته المختلفة. ولعل من أشهر تلك البيئات وأكثرها استخداماً نظام Sphinx [٢٣] وحزمة أدوات HTK (Hidden Markov Model Toolkit) [٢٤] اللذان يمكنان من بناء أنظمة التعرف الآلي على الكلام بمختلف أنواعها، وذلك من حيث:

١. الحجم، أي عدد الوحدات التي يمكن للنظام التعرف عليها "vocabulary"؛ ويمكن أن يكون هذا العدد صغيراً أو متوسطاً أو كبيراً،

٢. طريقة النطق، أي كيفية التعاطي مع المدخلات في النظام؛ هل يتعامل النظام مع كلمات أو وحدات منفصلة، أم يتعامل مع كلمات متصلة تشكل جملاً قصيرة مبنية وفقاً لقواعد محددة، أم يتعاطى مع الكلام الطبيعي في مجمله،

٣. عدد المستخدمين، أي عدد المتحدثين الذين يمكن للنظام التعرف عليهم. فقد يكون للنظام مستخدماً واحداً تدرب على صوته ولا يتعرف إلا عليه، ويدعى النظام في هذه الحالة "speaker dependent"؛ وقد يكون النظام قادراً على التعرف على عدد معين من المستخدمين تدرب على أصواتهم، ويدعى في هذه الحالة "multi-speakers"؛ وقد يكون النظام مستقلاً عن المتحدث، بمعنى أنه قادر على التعرف على صوت أي متحدث لكونه تدرب على عدد كبير من المتحدثين، ويدعى في هذه الحالة "speaker in-dependent".

تم تطوير نظام Sphinx في جامعة كارنيجي ميلون بالولايات المتحدة الأمريكية، وهو يستخدم نماذج ماركوف الخفية في نمذجة الأصوات ويضم قسمين رئيسيين، قسم للتدريب وآخر للتعرف. وقد ظهرت منه إصدارات مختلفة تستخدم طرق مختلفة في التعرف آخرها Sphinx4 [٢٥] الذي تم تطويره باستخدام لغة الجافا وهو يحتوي على ثلاثة أجزاء أساسية أساسية هي:

١. الواجهة (FrontEnd): تقوم باستقبال الإشارة وتحويلها إلى مجموعة من الخصائص الرقمية وفقاً لآليات استخلاص الخصائص التي شرحت في الجزء السابق.
٢. اللغوي (Linguist): يقوم بترجمة النموذج اللغوي للنظام ودمجه مع معلومات القاموس اللفظي الذي يضم الحالات النطقية الممكنة والنماذج الصوتية لتكوين مجال وشبكة البحث.
٣. المتعرف (Decoder): يقوم باستخدام الخصائص الرقمية ومجال البحث الذي تم إنشاؤه لأداء عملية التعرف الفعلية.

أما حزمة أدوات HTK التي تعتبر الأشهر والأكثر انتشاراً، فقد تم تطويرها في جامعة كامبريدج ببريطانيا. وفيها يتم تمثيل كل وحدة صوتية باستخدام نموذج

ماركوف خفي واحد، وكل كلمة كسلسلة من النماذج. ويتكون HTK من أربعة أجزاء رئيسية هي:

١. أدوات تجهيز البيانات: وتقوم باستقبال الإشارة الصوتية ومعالجتها ثم استخلاص الخصائص الرقمية منها وفقاً لأحدى الصيغ المناسبة (MFCC، LPC، إلخ.)؛ كما تقوم هذه الأدوات بتجهيز البيانات للمرحلة المقبلة.
٢. أدوات التدريب: وتمكن من تعريف الهيكل العام للنماذج الصوتية وتحديد قيمها الابتدائية ثم تدريبها بالآلية والخوارزميات التي ذكرناها في الجزء السابق.
٣. أدوات التعرف: وتمكن من استخدام النماذج الصوتية والنموذج اللغوي والقاموس النطقي مع خوارزمية Viterbi لمحاولة التعرف على ما تم نطقه.
٤. أدوات التحليل: تستخدم لتحليل النتائج وتقييم الأداء.

ولعل هذين النظامين يعتبران من أنضج الوسائل المعتمدة على الطرق الإحصائية والمستخدمه اليوم في بناء أنظمة التعرف الآلي على الكلام. ولا توجد حسب علمنا دراسات مقارنة تبين أهمية واحد على الآخر حسب نوع التطبيق المراد عمله.

### واقع استخدام تقنيات التعرف الآلي على الكلام في خدمة القرآن الكريم

إن الأعمال البحثية في مجال التعرف الآلي على الكلام العربي بشكل عام لا تزال محدودة إذا ما قورنت بما تم في اللغات الأخرى كالإنكليزية مثلاً، على الرغم من أن اللغة العربية تصنف من حيث عدد المتحدثين بها في المرتبة الثانية عالمياً بأكثر من ٤٢٢ مليون متحدث في حين تأتي اللغة الإنكليزية في المرتبة الرابعة وذلك وفقاً لما أوردته موسوعة أنكرتا الشهيرة وما جاء في هذه الإحصائية [٢٦]. ولعل من بين الأسباب التي قد تكون وراء ذلك النقص غياب البنى التحتية اللازمة لبناء



أنظمة التعرف الآلي على الكلام، سواء ما تعلق منها بالصوتيات العربية وقواعد بياناتها [٢٧] أو تقنيات معالجتها كلغة طبيعية [٢٨] أو غير ذلك. ومع هذا فقد بدأت تلوح بوادر لتزايد الاهتمام في السنوات الأخيرة، حيث انطلقت أبحاث متعددة، نذكر من بينها [٢٧-٣٤].

أما الأعمال التي اهتمت بتوظيف تقنيات التعرف الآلي على الكلام لخدمة القرآن الكريم فهي قليلة على حد علمنا، وبعضها عبارة عن محاولات متواضعة لا تعكس أهمية القرآن الكريم في حياة المسلمين.

ومن بين هذه الأعمال التي اهتمت بتوظيف تقنيات التعرف الآلي على الكلام لخدمة القرآن الكريم، ما قام به مرتجى وزملاؤه في جامعة عمان، حيث سعوا لبناء نظام للتعرف الآلي على التلاوة القرآنية مرتبط بالمتحدث (Speaker-Dependent) واستخدموا في ذلك نماذج ماركوف الخفية لتمثيل الوحدات الصوتية حسب السياق (tri-phone HMM model) [٣٥]. وقد استخدموا عينة من ٢٥ آية للتدريب والاختبار، ووصلت نسبة التعرف عندهم إلى ٨٠٪ حيث اعتبروها مناسبة ومرضية جداً إذا ما قورنت مع قلة الآيات المستخدمة. وقد قام نفس الفريق، لاحقاً، بتطوير النظام ليصبح مستقلاً عن المتحدث (Speaker-Independent) واستخدموا تقنية MLLR (Maximum Likelihood Linear Regression) لملاءمة النموذج [٣٦]. وقد أخذوا الجزء الأخير من القرآن الكريم (جزء "عم") بتلاوات خمسة قراء للتدريب والاختبار ولكنهم لم يحددوا ما الحجم المستخدم في التدريب والاختبار كل على حدة. كما أنهم لم يبينوا ما إذا كان الاختبار وقع على جزء مما استخدم في التدريب أم لا؟ ولم يذكروا كذلك ما إذا كانوا قد استخدموا نموذجاً لغوياً أم لا؟ مع أن لكل ذلك علاقة وطيدة بأداء النظام. وقد وصلوا إلى نتائج تتراوح بين ٦٨٪ و ٨٥٪ حسب القارئ، أي بمعدل حوالي ٧٧٪. وأوردوا في بحثهم أن هذه النتائج تم الحصول عليها باستخدام نماذج ماركوف مكونة من ١٣

مقطع لكل واحد على حدة، ونظرًا لذلك فقد ذكروا أن مرحلة التعرف في نظامهم تعتبر بطيئة. واستخدموا في بناء تلك الأنظمة أدوات HTK.

وقد أهتم طبال وزملاؤه في الجامعة اللبنانية باكتشاف الآيات القرآنية في الملفات الصوتية بغية تحديدها واستخلاصها [٢٧]. وقد استخدموا نظام Sphinx في ذلك، ولكنهم اقتصروا فقط على سورة الإخلاص بمقرئين متعددين (عشرين قارئًا) ولم يبينوا ما إذا كان الاختبار تم على نفس التلاوات التي استخدمت في التدريب. وعلى اعتبار أنهم يسعون فقط إلى تمييز الآيات القرآنية عن غيرها، فقد استخدموا نموذجًا لغويًا للآيات الموجودة في سورة الإخلاص. ويذكرون أنهم قاموا بتجارب على نوعين من التلاوات، تلاوات مرتلة وتلاوات مجودة، وحصلوا في المعدل على نسبة ٩٢٪ في النوع الأول و ٩٠٪ في النوع الثاني مع العلم أنهم استخدموا ٢٠ قارئًا في كل نوع. وحسبما أوردوا فإن التلاوات المرتلة أسرع من تلك المجودة وأقل تركيزًا منها على استيفاء المدد الزمنية للأصوات القرآنية.

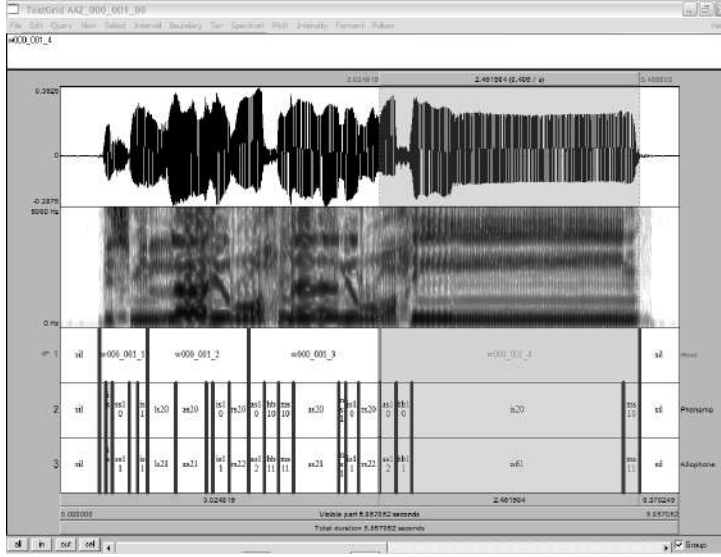
ومن بين الأعمال الهامة في المجال، مشروع "J-QAF" [٢٨] الذي أطلقته الحكومة الماليزية سنة ٢٠٠٤م ضمن خطتها لتطوير الرأس المال البشري "En-hancing Human Capital" ورصدت له مبالغ مالية كبيرة؛ وهو يسعى إلى بناء أنظمة حاسوبية لتعليم تلاوة القرآن الكريم واللغة العربية والواجبات الدينية في المدارس الابتدائية إضافة إلى تعلم الخط الملاوي العربي. هذا المشروع يتكون من سبع وحدات وأقسام تعليمية (Teaching Modules) بنيت بشكل تدريجي حتى نهاية الفترة الرئيسية للمشروع سنة ٢٠١٠م؛ ومن بين هذه الوحدات، أقسام لتحفيظ القرآن الكريم وتعليم قواعده التجويدية إضافة إلى قسم لتعليم اللغة العربية. والأقسام التي تعنى بالقرآن الكريم في هذا المشروع بدأت بتقنيات الملتيميديا التفاعلية [٣٩-٤٠] ثم تدرجت لتصل في المراحل الأخيرة إلى إدخال تقنيات التعرف الآلي على الكلام حسب ما ذكر [٤١]؛ ولم يتسنّ لنا إيجاد تفاصيل

حول آليات التعرف الآلي المستخدمة ولا مستوى الدقة الذي وصلت له. وقد قام ظريف وزملاؤه مؤخراً بدراسة لتقييم فاعلية أقسام تعليم القرآن في هذا المشروع وقياس أثرها على الطلاب [٤٢]، وأظهرت الدراسة مستوى عالياً في ضبط التلاوة القرآنية (النطق وأحكام التجويد) وانعكاسات ذلك على تحسن مستويات الطلاب في تعلم اللغة العربية.

ومن الأعمال البارزة في مجال توظيف تقنيات التعرف الآلي على الكلام لخدمة القرآن الكريم، مشروع حفص [٤٣] الذي أنتجته الشركة الهندسية لتطوير نظم الحاسبات المصرية RDI في صورة برنامج كمبيوتر تفاعلي يسعى للتمكين من التعلم الذاتي لأحكام تجويد القرآن الكريم. تقوم الفكرة على محاولة البرنامج محاكاة المقرئ في الكتابات من خلال إسماع المتعلم مثلاً لتلاوة نموذجية ثم يُطلب منه أن يحاكيها قدر استطاعته؛ يقوم البرنامج بتحليل ما نطقه المتعلم وتوليد تقرير صوتي وكتابي بأخطاء المتعلم ويطلب منه التركيز على تصحيح نطقه للمواضيع التي أخطأ فيها ثم إعادة التسميع؛ وهكذا إلى أن يجيز البرنامج قراءة المتعلم. وقد نشرت أوراق بحثية تبين الجوانب التقنية لآلية التعرف الآلي المتبعة في تحديد الأخطاء النطقية خلال تلاوة المتعلم للأمثلة المضمنة في البرنامج والتي تدور حول مواطن وجود أحكام تجويدية [٤٤-٤٦]. وقد ذكر أن تقييم البرنامج يتم من خلال قاعدة بيانات تضم تلاوات للأمثلة التي يتعامل معها البرنامج لقراء مجازين؛ حيث قرئ كل مثال عدة مرات، بعضها بشكل صحيح وبعضها أدخلت فيه أخطاء نطقية شائعة. ثم طلب من خبراء لغويين تحليل مختلف النطق المسجل وكتابته صوتياً مع توصيف للأخطاء ليستخدم كل ذلك في آليات التعرف الآلي والمقارنات أثناء استخدام البرنامج. وتشير النتائج إلى أن البرنامج يعطي القرار الصحيح حول القراءة (صحة القراءة أو رسالة تشخيص الأخطاء) بنسبة ٨٤٪، وهي نسبة جيدة. وتذكر الشركة أنها تعمل على تحسين أداء البرنامج بشكل

مستمر، وتسعى لتطويره مستقبلا ليتجاوز حدود التعامل مع حالات نطقية محددة ويصبح قادراً على تصحيح التلاوة والاستخدام في التحفيظ الآلي بشكل عام.

ومن ضمن الأعمال البارزة في المجال أيضاً، مشروع التعليم الآلي للقرآن الكريم [٤٧] الذي مولته مدينة الملك عبد العزيز للعلوم والتقنية، والذي سعى إلى وضع نموذج لبيئة حاسوبية للتحفيظ الآلي للقرآن الكريم تستخدم فيها تقنيات التعرف الآلي على الكلام لمساعدة القارئ على تصحيح نطقه وضبط حفظه. وقد ركز المشروع على إعداد بنى تحتية قوية تساعد على الوصول إلى مصحح آلي دقيق يمكن استخدامه في نظام حلقات افتراضية لتعليم القرآن الكريم. ومن أهم البنى التحتية التي ركز عليها المشروع، بناء قاعدة بيانات صوتية لتلاوات قرآنية وتجهيزها بشكل مناسب لبناء أنظمة للتعرف الآلي على الأصوات القرآنية بمختلف صورها وكذلك أصوات اللغة العربية الفصحى حيث هي لغة القرآن. وقد استخدم جزء "عم" كعينة وتم تسجيله بتلاوات عشرة قراء تم اختيارهم من طلاب الجمعية الخيرية لتحفيظ القرآن الكريم بالرياض؛ وتمت التلاوات والتسجيلات تحت إشراف مختص في القرآن وعلومه. هذه التلاوات تم تقطيعها وترميزها صوتياً على ثلاثة مستويات هي الكلمة والأصوات الأصول (الفونيمات) ثم الأصوات الفروع (الآلوفونات) التي تمثل مختلف الحالات النطقية مثل الغنة والقلقلة والتفخيم والترقيق ودرجات المدود وغير ذلك من الظواهر الصوتية في التلاوة القلائية [٤٨-٤٩]؛ والشكل ٩ يوضح مستويات التقطيع وآلية الترميز المستخدمة.

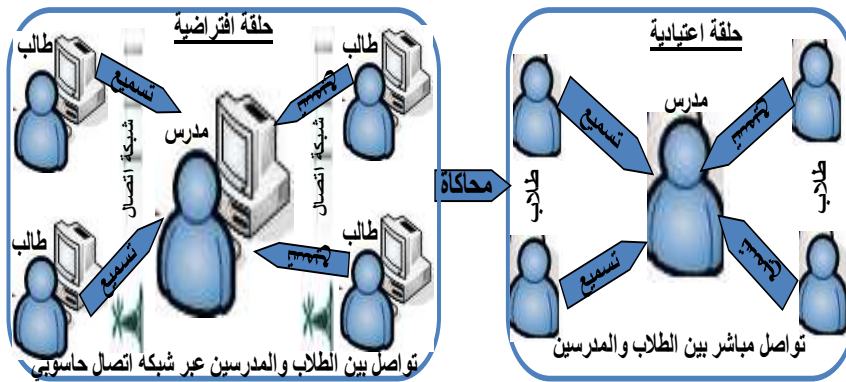


**الشكل ٩ : مستويات التقطيع في قاعدة البيانات الصوتية للقرآن الكريم**

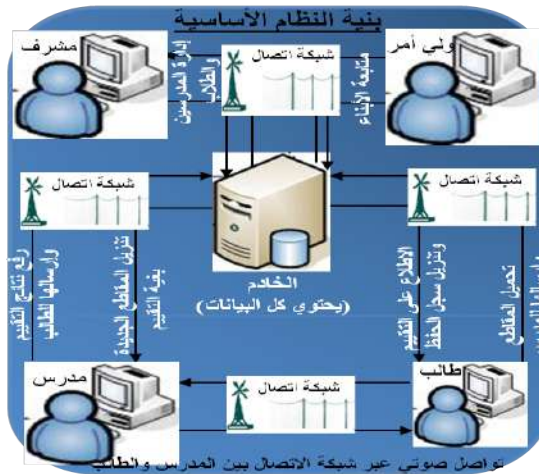
بالاعتماد على هذه القاعدة الصوتية تم بناء نموذجين أوليين للتعرف الآلي على الأصوات، واحد على مستوى الأصوات الأصول [٥٠] والآخر على مستوى الأصوات الفروع [٥١]. وقد استخدمت نماذج ماركوف الخفية مع خليط من التوزيعات الاحتمالية؛ وتم قياس الأداء من خلال سلسلة من التجارب، تم في كل واحدة منها الاختبار على عينة من قارئ محدد لم تستخدم في التدريب الذي يتم على بقية تلاوة هذا القارئ وتلاوات الآخرين، ثم يحسب معدل نسب التعرف الإجمالية لكل القراء. وقد وصلت النتائج إلى ٩٢٪ بالنسبة للأصوات الأصول و٨٨٪ بالنسبة للأصوات الفروع؛ وهي نسب جيدة وتوحي بإمكانية الوصول إلى مستويات عالية من الدقة في التعرف على مختلف الأصوات القرآنية أثناء التلاوة.

وفي المشروع تم وضع تصور لنظام حلقات إفتراضية [٥٢-٥٣] تحاكي نظام التحفيظ الاعتيادي (الأشكال ١٠ و ١١)، وتسهل التواصل بين المعلمين والمتعلمين

خصوصاً لمن هم في بلدان غير إسلامية. وقد وضعت آلية لربط المصحح الآلي المعتمد على تقنيات التعرف الآلي على الكلام مع نظام الحلقات الافتراضية لمساعدة المتعلم على تصحيح نطقه وتجريب حفظه قبل عرضه على المدرس من خلال آليات التواصل.



شكل ١٠ : محاكاة حلقات التحفيظ (مستقاه من ٣٥)



شكل ١١ : معمارية نظام الحلقات الافتراضية (مستقاه من [٥٣])

وقد اهتم المشروع في جانب آخر من جوانبه بدراسة التشابه اللفظي وتحديد مواطنه لأهمية ذلك في التحفيظ. وقد استخدمت تقنيات في مجال تحليل النصوص لقياس درجة التشابه بين الكلمات ووضع معايير ومقاييس لترتيب الآيات حسب مستويات التشابه [٥٤]. ووضعت آلية لربط التشابه مع التحفيظ لمساعدة الدارس في إتقان حفظه. كما أضيف في نظام التحفيظ مكتبة تضم أهم كتب علوم القرآن الأخرى (الإعراب، التفسير، أسباب النزول) وتم ربطها مع النص القرآني لتسهيل الوصول إليها.

وبشكل عام، فإن مشروع التعليم الآلي للقرآن الكريم وضع أسساً قوية لبناء بيئة حاسوبية متكاملة للتحفيظ الآلي للقرآن الكريم، لكنه يحتاج إلى جهة تتبناه وتحوله من نموذج بحثي إلى منتج فعلي يخدم الراغبين في حفظ القرآن الكريم وتعلم علومه.

## الخاتمة

لقد قدمنا في هذا البحث بسطاً لأشهر تقنيات التعرف الآلي على الكلام ومنهجية بناء نظمته وأدواته؛ ثم استعرضنا بنوع من التفصيل واقع استخدام هذه التقنيات في الوقت الحالي لخدمة القرآن الكريم. ولعل القارئ والمتتبع لهذه الأعمال يشعر بأنها لا تزال محدودة وتحتاج إلى تراكم في الخبرة وتضافر في الجهود لنصل إلى مستوى من التحكم في تلك التقنيات يمكن من تطويعها بما يتماشى مع خصوصية القرآن الكريم وقدسيته.

ومن أهم وأنضج ما استعرضناه من أعمال مشروع حفص ومشروع التعليم الآلي للقرآن الكريم؛ فالأول منتج تجاري والثاني نموذج بحثي. فمشروع حفص يركز على تعليم وتطبيق أحكام التجويد من خلال التعامل مع مقاطع محددة من القرآن الكريم، حيث يسمع المتعلم تلاوتها بقراءة نموذجية ثم يطلب منه محاولة محاكاتها ومن ثم يقوم بتوجيهه بناء على مستوى نطقه؛ وهو عمل جاد ومفيد. أما مشروع التعليم الآلي للقرآن الكريم فله أهداف أوسع وأشمل، تضم في صلبها بيئة حاسوبية للتخفيف الآلي للقرآن الكريم مشتملة على مختلف علومه. وقد وضع فيه نموذج لحلقات افتراضية تحاكي نظام حلقات التخفيف الاعتيادية وتسهل التواصل مع المقرئين. وتم وضع تصور لربطه مع نظام التصحيح الآلي للنطق والحفظ ليكون ذلك عوناً للدارس من جهة، ومساعداً للمقرئ من جهة أخرى في تسريع التعااطي مع المتعلمين. والنظام يعتمد في رؤيته على عدم إلزام المتعلمين والمعلمين بالتزامن أثناء العمل مع إتاحة خيار للتواصل المباشر سواء لتقديم توجيهات وتعليقات عند الحاجة أو لإجراء اختبارات لمنح إجازات أو كذلك للقيام بمسابقات قرآنية.





## المراجع

1. D. Jurafsky, J. H. Martin. (2008). Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Prentice Hall, ISBN: 10: 0131873210.
2. <http://www.cs.cmu.edu/~listen/>
3. J. E. Beck, K. Chang, J. Mostow, A. Corbett. (2005). Using a student model to improve a computer tutor's speech recognition. Proceedings of the AIED 05 Workshop on Student Modeling for Language Tutors, 12<sup>th</sup> International Conference on Artificial Intelligence in Education, Amsterdam, 18-22 July.
4. <http://www.colit.org/>
5. R. Col, S. Vuuren, B. Pellom, et. Al. (2003). Perceptive Animated Interfaces: First Steps Towards a New Paradigm for Human Computer Interaction. Proceedings of the IEEE, vol.91, no9, pp. 1391-1405.
6. <http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/>
7. W. Mattheyses, L. Latacz, W. Verhelst. (2009). On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. EURASIP Journal on Audio, Speech, and Music Processing.

8. <http://www.readingcompanion.org/>

9. <http://www.scilearn.com/products/reading-assistant>

١٠. غانم قدوري الحمد. (١٤٠٦هـ). الدراسات الصوتية عند علماء التجويد. مطبعة الخلود. بغداد.

١١. أبوبشر عمرو بن عثمان "سيبويه". (١٩٧٥م). الكتاب. تحقيق: عبد السلام محمد هارون، دار الكتاب العربي. القاهرة.

١٢. محمد بن شحادة الفول. (١٤١١هـ). بغية عباد الرحمن لتحقيق تجويد القرآن. دار بن القيم. الدمام.

١٣. منصور الغامدي، عبد الله الأنصاري. (٢٠٠٩م). التقنيات المعاصرة في خدمة القرآن الكريم. ندوة القرآن الكريم والتقنيات المعاصرة "تقنية المعلومات"، المدينة المنورة - المملكة العربية السعودية، ١٣-١٥ أكتوبر.

١٤. غانم قدوري الحمد. (٢٠٠٩م). استخدام صورة آلة النطق ومخارج الحروف في تعليم قواعد التلاوة "تأصيل وتحليل". ندوة القرآن الكريم والتقنيات المعاصرة "تقنية المعلومات"، المدينة المنورة - المملكة العربية السعودية، ١٣-١٥ أكتوبر.

١٥. أحمد راغب أحمد. (٢٠٠٤م). فنولوجيا القرآن: دراسة لأحكام التجويد في ضوء علم الأصوات الحديث. أطروحة لنيل درجة الماجستير. جامعة عين شمس.

16. L. Rabiner, B.H. Juang. (1993). Fundamentals of Speech Recognition. Prentice Hall.

17. F. Jelinek. (1998). Statistical Methods for Speech Recognition. Cambridge, MA: MIT Press.

18. X. Huang, A. Acero, H. Hon. (2001). Spoken Language Processing. Prentice Hall PTR.
19. L. Rabiner. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2).
20. T. K. Vintsyuk. (1968). Speech discrimination by dynamic programming. Cybernetics, 4(1), 52–57. Russian Kibernetika, 4(1), pp. 81–88.
21. A. J. Viterbi. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13 (2): 260–269. doi:10.1109/TIT.1967.1054010. April.
22. F. Jelinek, L. Bahl, R. Mercer. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Transactions of Information Theory 21 (3): 250–6. doi:10.1109/tit.1975.1055384. May.
23. <http://cmusphinx.sourceforge.net/>
24. <http://htk.eng.cam.ac.uk/>
25. <http://cmusphinx.sourceforge.net/wiki/tutorialsphinx4>
26. <http://www.scribd.com/doc/14436546/Languages-Spoken-by-More-Than-10-Million-People>.
27. M. Algamdi. (2003). KACST Arabic Phonetics Database. The Fifteenth International Congress of Phonetics Science. Barcelona. 3109–3112.
28. M. Elshafei, H. Al-Muhtaseb, M. Alghamdi. (2006). Statis-

- tical Methods for Automatic Diacritization of Arabic text. Proceedings 18th National computer Conference NCC'18. Riyadh. March 26-29.
29. F. A. H. Al-Otaibi. (2001). Speaker-dependant continuous Arabic speech recognition. M.Sc. Thesis. King Saud University.
  30. M. Alghamdi. M. Elshafei. H. Almuhtaseb. (2009). Arabic Broadcast News Transcription System. International Journal of Speech Technology. 10(4). pp. 1572-8110.
  31. H. Soltau. G. Saon. B. Kingsbury. J. Kuo. L. Mangu. D. Povey. G. Zweig. (2007). The IBM 2006 GALE Arabic ASR system. ICASSP (pp. IV-349 – IV-352).
  32. K. Kirchhofl. J. Bilmes. S. Das. N. Duta. M. Egan. G. Ji. F. He. J. Henderson. D. Liu. M. Noamany. P. Schoner. R. Schwartz. D. Vergyri. (2003). Novel approaches to Arabic speech recognition: report from the 2002 John-Hopkins summer workshop. ICASSP (pp. I-344–I-347).
  33. H. Hyassat. R.A. Zitar. (2006). Arabic speech recognition using SPHINX engine. International Journal of Speech Technology. 9(3-4). pp. 1381-2416.
  34. R.A. Haraty. O.E. Ariss. (2007). CASRA+: A Colloquial Arabic Speech Recognition Application. American Journal of Applied Sciences 4(1). pp. 23-32.
  35. E. Mourtaga. M. Abdallah. A. Sharieh. S. Serahn. (2005). Quranic Based Speaker-Dependent Recognition Using Tri-phone/HMM Model. Advanced in Modeling Series B: Signal Processing and Pattern Recognition. 48(5). pp. 43-58.

36. E. Mourtaga. M. Abdallah. A. Sharieh. S. Serahn. (2007). Speaker-Independent Quranic Recognizer Based on Maximum Likelihood Linear Regression. PWASET. Vol. 20. pp. 376-382.
37. H. Tabbal W. Al-Falou. B. Monla. (2007). Analysis and Implementation of an Automated Delimiter of "Quranic" Verses in Audio Files using Speech Recognition Techniques. Chapter of the Book "Robust Speech Recognition and Understanding". edited by: Michael Grimm and Kristian Kroschel. pp.460.
38. <http://www.j-qaf.net>
39. A.M. Aziz. (2009). Developing the j-QAF e-Learning Application for Children on Islam Obligatory Duties (Fardhu Ain) under the Topic 'Ibadat'. Master's thesis. University Utara Malaysia.
40. T. Mssraty. Q. Faryadi. (2012). Teaching the Qur'anic Recitation with Harakatt: A multimedia-based Interactive Learning Method. International Journal of Scientific & Engineering Research. Volume 3. Issue 8.
41. Z. Razak. N.J. Ibrahim. et al. 2008. Quranic Verse Recitation Recognition Module for Support in j-QAF Learning: A Review. IJCSNS. 8(8). pp. 207-216.
42. M. Zarif. N. Mohamad. B. Bakar. (2014). Assessing Quranic Reading Proficiency in the j-QAF Programme. International Education Studies Journal. Vol. 7. No. 6; Published by Canadian Center of Science and Education.
43. <http://hafss.net/>
44. S. Abdou. S. Hamid. M. Rashwan. A. Samir. O. AbdelHamid. M. Shatin. W. Nazih. (2006). Computer Aided Pronunciation

- Learning System Using Speech Recognition Techniques. Proceedings of the INTERSPEECH 2006-ICSLP. Pittsburgh. PA. USA.
45. A. Samir. S. M. Abdou. A. H. Khalil. M. Rashwan. (2007). Enhancing usability of CAPL system for Qur'an recitation learning. Proceedings of the INTERSPEECH 2007 - ICSLP. Antwerp. Belgium.
46. S. M. Abdou. M. Rashwan. (2014). A Computer Aided Pronunciation Learning System For Teaching The Holy Quran Recitation Rules. Proceedings of the ACS/IEEE AICCSA 2014. 10-13 November. Doha. Qatar.
٤٧. يحيى الحاج، عماد الصغير، منصور الغامدي، محمد الكنهل، عبدالله الأنصاري. (٢٠١٢ م). التعليم الآلي للقرآن الكريم. تقرير فني نهائي، مدينة الملك عبدالعزيز للعلوم والتقنية، الرياض، السعودية.
48. M. AlGhamdi. Y.O.M. Elhadj. M. AlKanhil. (2007). A manual system to segment and transcribe Arabic Speech. Proceedings of IEEE ICSPC 2007. Dubai. UAE. pp. 233-236. ISBN 1-4244-1236-6.
٤٩. يحيى الحاج، منصور الغامدي، محمد الكنهل، عبدالله الأنصاري. (٢٠٠٩ م). ذخيرة صوتية لجزء من القرآن الكريم. ندوة القرآن الكريم والتقنيات المعاصرة "تقنية المعلومات"، المدينة المنورة - المملكة العربية السعودية، ١٣ - ١٥ أكتوبر.
50. Y.O.M. Elhadj. M. Alghamdi. M. Alkanhal. (2013). Phoneme-Based Recognizer to Assist Reading the Holy Quran. Recent Advances in Intelligent Informatics. Advances in Intelligent Systems and Computing. Vol. 235. pp 141-152. Springer.

51. Y.O.M. Elhadj. M. Alghamdi. M. Alkanhal. (2013). Approach for Recognizing Allophonic Sounds of the Classical Arabic Based on Quran Recitations. Theory and Practice of Natural Computing. Lecture Notes in Computer Science. Vol. 8273. pp 57-67. Springer.
52. Y.O.M. Elhadj. (2010). E-Halagat: an E-Learning System for Teaching the Holy Quran. Turkish Online Journal of Educational Technology (TOJET). Vol. 9. No 1. pp: 54-61. (ISSN: 2146-7242).

٥٣. يحيى الحاج. (٢٠١٠ م). التحفيز الآلي للقرآن الكريم. مجلة اتصالات جمعية الحاسبات العربية، المجلد ٣، العدد ١.

٥٤. عماد الصغير، عبد الله الأنصاري، احمد خرصي، يوسف العوهلي. (٢٠٠٩ م). محرك بحث المتشابه اللفظي في القرآن. ندوة القرآن الكريم والتقنيات المعاصرة "تقنية المعلومات"، المدينة المنورة - المملكة العربية السعودية، ١٣-١٥ أكتوبر.



## الخاتمة

استقصى هذا الكتاب ما وصلت إليه العديد من المجالات التي يتدخل فيها الحاسوب اليوم لخدمة اللغة العربية، وإن كانت ثمة نتيجة مستخلصة في نهايته، فهي أن الحاجة لتطويرها - جميعاً - ما زالت كبيرة. فللوهلة الأولى - مثلاً - قد يبدو أن نظام تمثيل الكتابة العربية الحالي خالٍ من المشكلات. ولكن ما أن نقرأ البحث الأول من هذا الكتاب حتى نرى ما أخفاه الاعتياد من مشكلات. وفي البحث الثاني، تتجلى غزارة الأبواب المفتحة للتطوير. ففي أساس محدود - وهو أساس المدونات اللغوية - ومن مكان محدود أيضاً - وهو جامعة ليدز - تطل علينا بحوث متضافرة تتسع مجالاتها بتعددتها وتطورها. أما البحث الثالث فقد استقصى واقع قواعد البيانات الرقمية للوثائق التراثية العربية والإسلامية وبين أن الدراسات في هذا المضمار - حالياً - تنشط على قدم وساق ولكن أكثرها لا يفي بحاجات الحوسبة بعد كما أنه لا يتوفر للباحث العربي بلغته.

واستقصى البحث الرابع من الكتاب كثيراً من الأدوات التي تؤدي وظائف تشترك في احتياجها العديد من التطبيقات اللغوية العربية، وذكر مراجع نظرية وعملية لها، كما أشار إلى بعض جوانب النقص فيها. ثم أخلص البحث الخامس لجانب التشكيل الآلي للنصوص العربية، وللقواعد التعليمية الإحصائية منه تحديداً ونسب صحة أدائها من خطئها.

ثم ناقش البحثان الأخيران تطبيقين يمثلان أبرز الغايات للباحثين في حوسبة اللغة العربية؛ وهما: تطبيق الترجمة الآلية من العربية وإليها وتطبيق التعرف على قراءة القرآن الكريم منطوقاً. وقد خلص الباحثان إلى ميسر الحاجة إلى بنى تحتية ومزيد أدوات وإلى تضافر الجهود المؤسسية والرسمية لدعم ذلك كله.

ونحن نأمل - في ختام هذا الكتاب من باكورة إصدارات مركز الملك عبد الله بن عبدالعزيز الدولي لخدمة اللغة العربية - أن يتبنى هذا المركز الدولي وأمثاله المبادرات لتطوير حوسبة اللغة العربية ابتداءً من مستوى البنى التحتية ومروراً بمستوى الأدوات ووصولاً لمستوى التطبيقات. والله نسأل التوفيق لما يحبه ويرضاه.



## فهرس الموضوعات

الصفحة	الموضوع
٥	كلمة المركز
٧	مقدمة
٩	حول نظام تمثيل الحرف العربي أ. مأمون صبحي الحطاب
٢٥	أبحاث جامعة ليدز في مجال لسانيات المدونات العربية د. إيريك أتويل (Eric Atwell) / أ. عبد الله بن يحيى الفيضي
٦٧	قواعد البيانات الإلكترونية للمخطوطات التراثية العربية والإسلامية: الحاضر والمستقبل د. سامح عويضة
١٢٣	استقصاء تقنيات معالجة اللغات الطبيعية وتطبيقاتها في اللغة العربية د. أمجد أبو جبارة
١٨١	استخدام القواعد الاستنباطية في تحسين أداء التشكيل الآلي أ. عمر السيد شعبان
٢٠٧	الترجمة الآلية من العربية وإليها أ.د. محمد زكي خضر
٢٤٥	تقنيات التعرف الآلي على الكلام المنطوق وتطبيقاتها في القرآن الكريم: واقع وطموح د. يحيى محمد الحاج
٢٨٠	الخاتمة
٢٨٣	الفهرس

## هذا الكتاب

يُصدر مجمع الملك سلمان العالمي للغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المجمع حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجهد في استكتاب نخبة من المحررين والمؤلفين للهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المجمع من وراء ذلك إلى تنشيط العمل في المجالات التي تُنبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المجمع الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسلة.

والشكر والتقدير الوافر لسمو وزير الثقافة رئيس مجلس أمناء المجمع، ، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين للتواصل مع المجمع؛ لبناء المشروعات العلمية، وتكثيف الجهود، والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

