نحو بناء مدونة اختبار معيارية عربية لفك الالتباس الدلالي

عمرو الجندي*[1]، سعيد الوكيل**[2]، ياسر حفني***[3]، أحمد عبد الحميد عمر**[4]
* مجمع اللغة العربية بالقاهرة
** كلية الآداب جامعة عين شمس
*** كلية الحاسبات والذكاء الاصطناعي جامعة حلون
[1] amr25@hotmail.com
[2] alwakil@aucegypt.edu
[3] yhifny@yahoo.com
[4] ahmed.omar@art.asu.edu.eg

ملخص

تهدف هذه الورقة البحثية إلى تقديم منهجية للمراحل التي يمكن اتباعها لبناء مدونة اختبار معيارية؛ بهدف استخدامها في أبحاث فك الالتباس الدلالي آليا في نصوص اللغة العربية الفصحى، خاصة وأنه لا يكاد توجد مدونة يحتكم إليها الباحثون في اختبار نماذجهم اللغوية الإحصائية، ولا حتى توجد قائمة موحّدة من الكلمات الملتبسة التي يمكن أن تخضع للاختبار؛ مما يحدو بالخبراء إلى حالة من عدم اليقين العلمي بشأن أفضل النماذج المقترَحة بشأن فك الالتباس الدلالي العربي آليا. وقد عمل البحث على اتباع منهج مقترح بدءا من تحديد الكلمات الملتبسة المستعملة في كل العصور، ومرورا بجمع مدونة كبرى وتهيئتها، وانتهاء بتصنيفها تمهيدا لاستخراج مدونة اختبار تمثل المدونة الكبرى قدر الإمكان.

الكلمات المفتاحية

المدونة اللغوية – تصنيف المدونة اللغوية – مدونة اختبار معيارية – قائمة الكلمات الملتبسة – السياق – فك الالتباس الدلالي.

أولا: المقدمة

وصف الفيلسوف الأمريكي Abraham Kaplan الالتباس بأنه نزلة البرد التي تصيب اللغة [1]، وتمثل المعاني المتعددة للكلمة الواحدة جوهر هذا الالتباس، والذي يؤدي بدوره إلى صعوبات في العديد من تطبيقات المعالجة الآلية للغات الطبيعية، وعلى رأسها الترجمة الآلية [2]. وفيما يتعلق بلغتنا العربية فإن الثراءَ الصرفي واللهجي لكلماتها، والحريةَ النسبية لترتيبها داخل الجمل، وغيابَ علامات الضبط في أغلب نصوصها المكتوبة، فإن كل ذلك يزيد من الالتباس [3].

وتعتمد عملية إزالة الالتباس عن الكلمات التي تحمل أكثر من معنى، تعتمد في الأساس على الكلمات المجاورة في الجملة أو النص، أو ما يعرف بالسياق Context.

وفي الوقت الذي تطورت فيه الأبحاث في هذا المجال بالنسبة للعديد من اللغات منذ عدة عقود؛ فإن الأبحاث التي اهتمت بفك الالتباس الدلالي في اللغة العربية لم تبدأ سوى منذ عقدين فقط من الزمان [4]، ولعل أبرز الأسباب التي أدت إلى ذلك: تركيز الباحثين أولا على البنية التحتية للغة العربية (ويُعنى بذلك مستويات الصوت والصرف والنحو والمعجم) [5] [6]، ونقص الذخائر (المدونات) اللغوية [7]، بالإضافة إلى تأخر ظهور أساليب الذكاء الاصطناعي المتطورة التي يحتاج إليها هذا النوع من الأبحاث [5].

وقد لاحظنا – في حدود ما توصلنا إليه – أنه لا توجد مدونة اختبار عربية موحدة يحتكم إليها الباحثون لتقييم ما تقدمه أبحاثهم من نماذج لغوية إحصائية؛ ومن ثم فقد غدت معظم النتائج التي يعرضونها محل عدم يقين علمي؛ لذلك فإن هدف هذا البحث هو تقديم منهجية تساعد على بناء مدونة اختبار معيارية؛ بحيث تكون ممثلة قدر الإمكان للواقع اللغوي، وفي الوقت نفسه صالحة للتطوير بناء على المنهج المقترح، ومن ثم تستخدم في تحسين

قدرة الحاسوب على فك الالتباس الدلالي، واكتشاف كل من: القرائن اللغوية اللازمة لبناء نموذج لغوي إحصائي دقيق يصلح للتطبيق الفعلي على نصوص العربية الفصحى، والأدوات والطرائق الأكثر كفاءة في تعيين القرائن.

وبعد عرض للمقدمة نستعرض في الصفحات الآتية كلا من هذه العناصر: (ثانيا) الدراسات السابقة، (ثالثا) تكوين قائمة الكلمات الملتبسة، (رابعا) بناء المدونة اللغوية، وأخيرا خلاصة تتضمن أهم ما توصل إليه البحث، وأبرز نقاط العمل المستقبلية إن شاء الله تعالى.

ثانيا: الدراسات السابقة

رغم أن الدراسات التي عُنيت بفك الالتباس الدلالي تعود إلى أربعينيات القرن العشرين؛ فإن الدراسات العربية في هذا الصدد لم تظهر إلا مع بدايات القرن الحادي والعشرين، عندما وضع بعض الباحثين أطرًا نظرية لفك الالتباس الدلالي [8]، فيما تمكن آخرون من تنفيذ دراسات تطبيقية على مجموعات صغيرة ومختلفة من كلمات الفصحى المعاصرة [4] و [9 - 21].

وقد لاحظنا – كما ذكرنا في المقدمة – أنه لا توجد مدونة اختبار عربية موحدة يحتكم إليها الباحثون لتقييم ما تقدمه أبحاثهم من نماذج لغوية إحصائية.

ثالثا: تكوين قائمة الكلمات الملتبسة

يمكننا أن نعرّف قائمة الكلمات الملتبسة تعريفا بسيطا بأنها تلك القائمة التي تتضمن عددا من الكلمات ذات الدلالات المعجمية المتعددة. وعلى الرغم من البساطة التي تبدو عليها تلك القائمة فإن حقيقة الأمر أعقد قليلا مما تبدو عليه؛ وذلك أن وجود الكلمة غير المناسبة ضمن تلك القائمة قد يؤدي إلى تدني دقة النموذج اللغوي الإحصائي المأمول. فعلى سبيل المثال ليس من الممكن أن تحتوي هذه القائمة على كلمات مثل (أحوس، ضبيس، مشمعل، ...إلخ) ثم نتوقع أن ينجح النموذج اللغوي المعتمد على مدونة لغوية معاصرة في استنتاج المعنى الصحيح لها. وعليه فإن إعداد قائمة بالكلمات الملتبسة دلاليا هي خطوة رئيسة نحو بناء مدونة اختبار معيارية.

مرت مرحلة تكوين قائمة الكلمات الملتبسة بالخطوات الآتية:

1. جمع الكلمات

لقد جرى جمع هذه القائمة من المكنز الكبير [22]، الذي يحتوي على أكثر من اثنين وعشرين ألف (22,000) كلمة فريدة، فيما تشكل الكلمات التي تحتمل أكثر من معنى دلالي فيها حوالي الثلث (35%) بحوالي أربعة عشر ألف (14,000) كلمة.

2. تهيئة الكلمات

نظرًا لأن أغلب النصوص العربية تُكتب دون ضبط بالشكل؛ فقد حُذفت كل علامات الضبط من القائمة الأولية للكلمات، وهذا أدى – بطبيعة الحال – إلى زيادة الالتباس، وظهور ما يمكن أن نطلق عليه (المشترك الكتابي).

فعلى سبيل المثال كلمة (ثبت) الخالية من علامات الضبط تحتمل المعاني الآتية: (الأمن، التشجيع، التوقف، الثبات، الحق، السكون، الصبر، المقاومة والدفاع، المواظبة، الرزانة، العقل، التأكيد، الترسيخ، التقوية، النصرة، الإثبات، الفهرس)، بينما لو ضُبطت على هذه الأنحاء (ثَبَتَ، ثَبُتَ، ثَبَّتَ، ثَبَتٌ) لغدا لكل صيغة (مشترك لفظي) منها عدد أقل من المعاني السابقة.

3. تصنيف الكلمات

قُسمت الكلمات بحسب التصنيفات الآتية:

(1) التصنيف النوعي

هناك نوعان من الكلمات في اللغة: الأول يُطلق عليه (كلمات المحتوى) والثاني (الكلمات الوظيفية) [23]، وهذا البحث يستهدف (كلمات المحتوى المتعددة الدلالة)، وهي التي تتميز كل كلمة فيها بأكثر من معنى دلالي يتسبب في حدوث اللبس عند عدم فهم الحاسوب للسياق اللغوي؛ ولذا فقد تم تحديد الكلمات الوظيفية بدقة، وجرى استبعادها من المدونة.

والحقيقة أن لاستبعاد الكلمات الوظيفية سببين آخرين: (أولهما) أنها ترد بكثرة كبيرة مقارنة بباقي الكلمات، ولا شك أن تقليل حجم المدونة اللغوية – وبخاصة مدونة التدريب التي يستقرئ الحاسوبُ من خلالها السياقات اللغوية ليتعلم قواعد التقارب الدلالي للكلمات – يعمل على زيادة الأداء، وذلك من خلال تقليل الوقت الذي يستهلكه الحاسوب في معالجة بيانات تلك المدونة، وعليه فإن حذف أدوات مثل: عن – إلى – في – على سيسهم بشكل كبير في تقليل حجم المعالجات غير الضرورية للمهمة المطلوب إنجازها. و(ثانيهما) أنها قليلة الالتباس الدلالي.

وتجدر الإشارة إلى أنه قد جرى أيضا استبعاد الكلمات الشبيهة بالكلمات الوظيفية، مثل: (مِنْ: مَنْ، مَنّ)، (إِلَى: إلَى)، (أَمْ: أَمّ، أُمّ)، (أَنْ: أَنّ)، (أَنْتَ: أَنَتَ، أَنَّتْ)، (أَنْتُنَّ: أَنْتَنَ)، (أَنَّى: أَنَى، أَنِيَ).

(2) التصنيف الدلالي

يحاول البحث الوصول إلى الكلمات المُلبسة في اللغة العربية، والكلمات المُلبسة هي تلك التي يكون لها أكثر من معنى في قائمة الحقول الدلالية. وهنا يبرز سؤال نظري: هل هناك في اللغة كلمة ليس لها سوى معنى دلالي واحد؟ وبعبارة أخرى: هل تنقسم كلمات اللغة فعلا إلى كلمات وحيدة المعنى وأخرى متعددة المعاني؟

تظهر قائمة الحقول الدلالية للدكتور أحمد مختار عمر أن لكلمة مثل (أبٌ) معنى دلاليًّا واحدًا هو (الوالد)؛ في حين يورد المعجم الكبير ثلاثة معان أساسية هي: (الوالد، والجد وإن علا، وصاحب الشيء)، ويورد معجم لغة الشعر العربي معنيين أساسيين هما: (الوالد، والصاحب).

وعلى أية حال فإن الكلمات قد صُنِّفت من حيث دلالاتها صنفين رئيسين: الأول كلمات أحادية الدلالة، والثاني كلمات متعددة الدلالات، ثم قُسِّم الصنف الثاني قسمين فرعيين هما: كلمات ذات دلالات متقاربة، وكلمات ذات دلالات متباعدة، وذلك على النحو الآتي:

(أ) كلمات أحادية الدلالة

الكلمات أحادية الدلالة هي تلك الكلمات التي لم يرد لها سوى معنى واحد في قائمة المكنز الكبير، وقد أضاف إليها البحث الكلمات المشتركة رسما مع مصادرها وليس لهما سوى معنى واحد، مثل الفعل (رَسَفَ) ومصدره (رَسْفٌ)؛ حيث إنهما بعد حذف علامات الضبط يصبحان كأنهما كلمة واحدة ذات معنى واحد في سياقاتها المتعددة، وهذه الكلمات تكاد تنحصر في الأفعال الثلاثية ومصادرها.

(ب) كلمات متعددة الدلالات

لوحظ أن معاني بعض الكلمات المتعددة الدلالات تكاد تكون متطابقة، أو على أقل تقدير ذات سياقات متقاربة للغاية، وهذا النوع من الكلمات ربما يحتاج إلى تدقيق أكثر في المعالجة والتعلم الآلي؛ حتى يتمكن الحاسوب من اكتشاف عوامل الاختلاف الدقيق بين معانيها، ومن ثم يقوم بفك التباسها بأقل درجة ممكن من الخطأ.

لذلك عمد البحث إلى تصنيف الكلمات المتعددة الدلالات إلى صنفين على النحو الآتي:

أ) كلمات ذات دلالات متقاربة

يمكن أن نمثل للكلمات ذات الدلالات المتقاربة بالكلمة (ساجل) التي تدل على (المجادلة) و(المسابقة)، وعلى الرغم من الفرق الواضح بين الدلالتين فإنهما متقاربتان. وقد يكون مرد هذا التقارب علاقات دلالية متعددة: كالتضاد مثلا في (خبث الرائحة، طيب الرائحة)، والنوعية كما في (التفاؤل، التنبؤ)، والتناظر كما في (الطحن، الفتة، الدق) و(الانطفاء، التكاسل، السكون، الضعف، الكسل، المرض)، وأشباه الترادف كما في (التجويف، الثغرة، الفجوة). وهذا سيتم تناوله ضمن مرحلة متقدمة للغاية.

ب) كلمات ذات دلالات متباعدة

من أمثلة الكلمات ذات الدلالات المتباعدة كلمة (فدفد) التي تعني (الصحراء، أو اللبن الرائب)، ويتضح ما بينهما من تباين شديد في المعنى.

وتجدر الإشارة إلى أن هذا الصنف قد تتحول بعض كلماته إلى كلمات أحادية الدلالة، وذلك إذا ما أدخلنا في الاعتبار التصنيف الاستعمالي؛ حيث إن بعض الكلمات زوجية الدلالة قد تكون إحدى دلالاتها مقصورة على العربية التراثية، بينما الأخرى ممتدة حتى العربية المعاصرة؛ وحينها يتم اعتبار هذا النوع من الكلمات أحادية الدلالة فيما يتصل بالعربية المعاصرة، وهكذا. وهذا القسم هو ما سيعتمد عليه هذا البحث في بناء مدونة اختباره المقترحة.

(3) التصنيف الاستعمالي

يميز المكنز الكبير بين الرصيد الإيجابي الذي يمكن استخدامه في لغة العصر الحديث، والرصيد السلبي الذي فقد وجوده في اللغة الحية بمستوييها التراثي والحديث، ولم ينتقل من جيل إلى جيل إلا من خلال المعاجم. كما يميز بين الرصيد الإيجابي المعاصر الذي يمثل اللغة الحية السائدة، وبين الرصيد الإيجابي التراثي الذي لا يصادفه الباحث إلا في النصوص القديمة. ولا يعني وصف اللفظ بأنه من الرصيد المعاصر أنه استجد في العصر الحديث، وإنما يعني أنه مستعمل في العصر الحديث حتى لو كان قديما [22].

وعلى الرغم من أن هذا التمييز محتاج إلى مراجعة دقيقة بعرضه على مدونات العصور المختلفة للتأكد من دقته؛ فقد تقرر العمل به في هذا البحث، على أن تستدرك مسألة المراجعة والتدقيق للأعمال المستقبلية.

رابعا: بناء المدونة اللغوية

المدونة اللغوية في مجال اللسانيات الحاسوبية هي مجموعة كبيرة نسبيًّا من البيانات العامة أو المتخصصة في مجال معين، وقد تكون هذه البيانات نصوصا أو صورا أو مرئيات أو صوتيات، وتُخزّن إلكترونيا لأهداف البحث والتطوير المختلفة، كاستكشاف الظواهر والتنبؤ بها وتحليلها واستخراج المعلومات وما إلى ذلك. وقد يجرى توسيم هذه البيانات Data Annotation\Labeling بمعلومات تصنفها أو تصف وحداتها [24 - 26].

وفي هذا البحث فإن الهدف المباشر من بناء المدونة هو **توفير سياقات للمشترك الكتابي المتعدد المعنى عبر تاريخ العربية الفصحى وتشعّب المجالات المعرفية التي كُتبت بها؛** بحيث يسهم جزء من هذه السياقات – مدونة التدريب Training set – في بناء نموذج لغوي إحصائي Statistical Language Model من خلال تقنيات التعلم الآلي يعمل على إنشاء الحقول الدلالية آليا واستنتاج المعاني السياقية للكلمات الملتبسة دلاليا بدقة عالية، بينما يسهم جزء ثانٍ – مدونة التطوير Dev set – في تحسين دقة هذا النموذج، وأخيرا يسهم الجزء الثالث – مدونة الاختبار Test set – في اختباره.

هذا وتمر عملية بناء المدونة اللغوية Corpus Building بعدد من المراحل، تنتظم كل واحدة منها مجموعة من المعايير الدقيقة التي تعمل على تحقيق الأهداف الرئيسة لبناء المدونة، وذلك على النحو الآتي:

1. جمع المدونة

بالإضافة إلى الدقة الإملائية فقد روعي في جمع نصوص المدونة أن تكون ممثلة للغة العربية الفصحى وفق المعايير الثلاثة الآتية:

أولا: أن تكون مستوعبة لجميع عصور اللغة العربية الفصحى، ابتداء من الإسلامي وانتهاء بالحديث.

ثانيا: أن تكون مشتملة على أكبر قدر ممكن من المجالات المعرفية، كالفلسفة والدين واللغات وغيرها.

ثالثا: أن تكون متضمنة أكبر تنوع ممكن من أوعية النشر، كالكتب والدوريات والمجلات والصحف وغيرها.

لذلك فقد اعتُمد على عدد من المدونات لتحقيق المعايير الثلاثة الآنفة الذكر، وهذه المدونات هي: (1) المكتبة الشاملة (2) مدونة ويكيبيديا العربية (3) مدونة صحيفة الخليج (4) مدونة أبو الخير. وتجدر الإشارة إلى أن كل هذه المدونات متاحة على الإنترنت.

2. تهيئة المدونة

إن عملية تهيئة نصوص المدونة أشبه ما تكون بما يقوم به واضعو المقررات الدراسية، الذين يعملون على أن يلبي كلُّ مقرر مجموعةً من الأهداف التفصيلية، وذلك بما يتناسب مع الفئة المستهدفة من المتعلمين. وفي هذا البحث سيكون الحاسوب هو ذلك المتعلم الآلي الذي يُبتغى تعليمُه تعرفَ الكلمات المتقاربة دلاليا ومن ثم بناءَ الحقول الدلالية واستنتاجَ المعنى السياقي للكلمة بشكل مجرد عن أية تفصيلات أخرى أو عقبات إضافية، وهذا ما حدا إلى أن تُهيأ النصوص بشكل يتناسب مع ذلك؛ ومن ثم كانت الإجراءات الآتية:

(1) استبعاد النصوص المقدسة والشِّعر

إن القرآن الكريم والشعر العربي يتميزان بطبيعة مختلفة عن بقية النصوص العربية؛ فالنص المقدس ليس من كلام البشر، والشِّعر من كلام البشر لكنه يخضع لعوامل شكلية كثيرة تجعله مختلفا عن الكلام العادي. ومن ثم فهما مختلفان بوضوح في ألفاظهما ومعانيهما وتراكيبهما عن باقي أوعية النصوص الأخرى، وقد افتُرض أن هذا الاختلاف قد يشكّل عبئًا على نحو ما على عملية التعلم الآلي التي سيقوم بها الحاسب الآلي.

(2) حذف علامات الضبط

بالرغم من أن حذف علامات الضبط يزيد الكلمات التباسًا فوق التباس؛ فإن الغالبية العظمى من النصوص العربية تُكتب خاليةً من علامات الضبط؛ لذا وفي ضوء أن المدونة يُفترض أن تكون ممثلة للواقع اللغوي، فقد

تخلى البحث عن جميع علامات الضبط سواء على مستوى البنية أو الإعراب، قاصدًا أن يتعلم الحاسب الآلي استنتاجَ المعنى السياقي دون الاعتماد على ضبط الكلمة.

وبطبيعة الحال فإن البحث في هذه الحالة لا يتعامل فقط مع (المشترك اللفظي)، بل مع ما سبق أن أطلقنا عليه (المشترك الكتابي).

(3) التحديد المبدئي للسياقات

ويقصد به وضع حد لسياق الكلمة؛ فهل هو الجملة؟ أم الفقرة؟ وإذا كان الجملة فكيف يمكن تحديد بدايتها ونهايتها؟ وإذا كانت الفقرة فأليس من الواضح أنه سياق طويل جدا؟! على أية حال تم الاستقرار على التسديد والمقاربة، فتم جعل بداية السطر الجديد هو بداية السياق، كما جُعلت النقطة وعلامتي الاستفهام والتعجب نهاية للسياق. واستُبعدت الفاصلة من كونها حدا للسياق لأنها لا تستخدم دائما للفصل بين الجمل المكتملة، وكذلك الحال مع النقطتين، وغيرهما من علامات الترقيم التي قد تستخدم بكثرة بين الكلمات، وقد يُتغافل عنها تماما في بعض النصوص.

في عام 1949 ناقش Warren Weaver مسألة معالجة الالتباس الدلالي للكلمات في اللغة الإنجليزية، وأنه ينبغي أن يُعتمد في ذلك على السياق [19] [27]. وفي عام 1950 أجرى Abraham Kaplan عدة تجارب بهدف الوقوف على الحجم الذي ينبغي أن يكون عليه هذا السياق من أجل فك التباس الكلمة، وتوصل في النهاية إلى أن سياقا مكونا من الكلمة الغامضة دلاليا وكلمتين قبلها وكلمتين بعدها كافٍ لفك الغموض الدلالي لتلك الكلمة [10]. وقد اعتمدت Masterman نتائج Abraham Kaplan في اللغة الروسية كما اعتمدها كل من Choueka و Lusignan في اللغة الفرنسية.

يرى الفيلسوف الأمريكي Abraham Kaplan أن السياق الذي يتكون من كلمة أو كلمتين على كل جانب من الكلمة الغامضة الدلالة يرى أن هذا السياق لديه فعالية في الكشف عن معنى الكلمة الغامضة تشبه فعالية الجملة كاملة [1].

وتجدر الإشارة إلى أمرين مهمين: (الأول) أن هذا التحديد موصوف بالمبدئي؛ وذلك لأنه في مرحلة تحسين نتائج النموذج اللغوي الإحصائي قد يُلجأ إلى تحديد جديد للسياق عبر عدد الكلمات المحيطة بالكلمة الملتبسة. (الثاني) أن اتخاذ بعض علامات الترقيم وسيلة لتحديد السياق ليس سوى جهد المقل؛ وذلك لأن علامات الترقيم تُستخدم بدرجة كبيرة من العشوائية في النصوص العربية على اختلاف الزمان والموضوع والمستعمِل.

3. تصنيف المدونة

التصنيف لغة: هو جعل الأشياء أنواعًا وأضرابًا، وتمييز بعضها عن بعض [28]. واصطلاحا Classification: هو "تقسيم الأشياء أو المعاني وترتيبها في نظام خاص وعلى أساس معين؛ بحيث تبدو صلة بعضها ببعض، ومنه تصنيف الكائنات وتصنيف العلوم" [29]. أو هو "تجميع أفراد لا حصر لها تحت عدد محصور من الأنواع العامة Class تبعًا لخصائص معينة مشتركة بينها يمكن استنتاج أنها تصدق على أفراد النوع ... والمصنِّف ينتقل من الأدنى إلى الأعلى بالتصاعد بعكس القسمة Division" [24].

من المعروف أن بعض أصناف التباين بين مدونتي التدريب والاختبار يؤدي إلى ارتفاع معدل أخطاء أغلب تقنيات معالجة اللغات الطبيعية، وبخاصة ما يرتبط منها بالمستوى الدلالي؛ ذلك أن الدلالة – بطبيعتها المعنوية – أكثر تأثرا بعوامل التباين من المستويين الصرفي والتركيبي اللذين يتسمان – لطبيعتهما الشكلية – بالثبات النسبي.

وتعتبر مرحلة تصنيف المدونة امتدادًا للمرحلة السابقة (تهيئة المدونة)، وتسعى هذه المرحلة إلى تحقيق العديد من الأهداف على النحو الآتي:

أولا: المساعدة على تنظيم المدونة وتوصيفها نوعيًّا وإحصائيًّا بدقة.

ثانيا: المساعدة على بناء مدونة اختبار أو عدة مدونات اختبار؛ بحيث تكون ممثِّلة بدقة للمدونة العامة ومن ثم للواقع اللغوي.

هذا وسوف يكون تصنيف المدونة على النحو الآتي:

(1) التصنيف التاريخي المعرفي

يُقصد بالتصنيف التاريخي المعرفي أن تصنَّف النصوص المختلفة للمدونة بحسب العصور التاريخية التي أُلِّفت فيها، ثم تُصنف نصوص كل عصر وفق المجال المعرفي الذي تنتمي إليه؛ لتصبح المحصلة النهائية مجموعات من النصوص المصنَّفة تاريخيا ومعرفيا في الوقت نفسه، وذلك على النحو الآتي:

(أ) التصنيف التاريخي

اتبع البحث – بشكل تقريبي – التقسيم الخماسي الذي سار عليه المعجم التاريخي للغة العربية [30]، والذي يعتمد بشكل رئيس على ما أقره اتحاد المجامع اللغوية العلمية العربية [31]، وذلك على النحو الآتي:

1- **عصر ما قبل الإسلام**: الممتد من (...) إلى (1ق.ه = 622م).

2- **العصر الإسلامي**: الممتد من (1ه = 622م) إلى (131ه = 749م).

3- **العصر العباسي**: الممتد من (132ه = 750م) إلى (655ه = 1257م).

4- **عصر الدول والإمارات:** الممتد من (656ه = 1258م) إلى (1219ه = 1804م).

5- **العصر الحديث:** العصر من (1214ه = 1798م) إلى وقتنا الحالي.

وتجدر الإشارة إلى خلو المدونة – بحسب تواريخ وفيات مؤلفي كتبها – من أية نصوص تنتمي إلى عصر ما قبل الإسلام؛ وذلك لسببين: (الأول) أنه جرى استبعاد الدواوين الشعرية التي تنتمي بوضوح إلى ذلك العصر؛ لما سبقت الإشارة إليه من استبعاد الشِّعر. (الثاني) أن كل مصادر النصوص المتبقية ينتمي مؤلفوها إلى العصر الإسلامي وما تلاه من عصور.

(ب) التصنيف المعرفي

هو التصنيف على أساس الموضوع؛ حيث تكتسب العديد من الكلمات معانيَها من المجال المعرفي الذي تُستعمل فيه. وسوف نتناول هذا التصنيف من خلال النقاط الآتية:

أ) مستويات التصنيف المعرفي

هناك عدة مستويات للتصنيف المعرفي؛ بدءًا من المستويات العامة التي تشمل عددًا من العلوم الأساسية، ومرورًا بالمستويات الفرعية التي تتضمن فروع كل علم، وانتهاء بالمستويات الأكثر تفرّعًا التي يتضمن كل واحد منها موضوعات متخصصة في أفرع العلوم المختلفة.

وقد قرر البحث الاكتفاء بتصنيف المدونة وفق المستوى الأول (التصنيف المعرفي العام أو ما يُطلق عليه تصنيف ديوي العشري) لتحقيق الهدف، كما هو موضح بالجدول (1).

جدول (1): التصنيف المعرفي العام (تصنيف ديوي العشري)

م	التصنيف	الكود
1	المعارف العامة	099 - 000
2	الفلسفة وعلم النفس	199 - 100
3	الديانات	299 - 200
4	العلوم الاجتماعية	399 – 300
5	اللّغات	499 - 400
6	العلوم البحتة	599 - 500
7	العلوم التطبيقية	699 - 600
8	الفنون والاستجمام والديكور	799 - 700
9	الآداب	899 - 800
10	التاريخ، الجغرافيا والتراجم	999 - 900

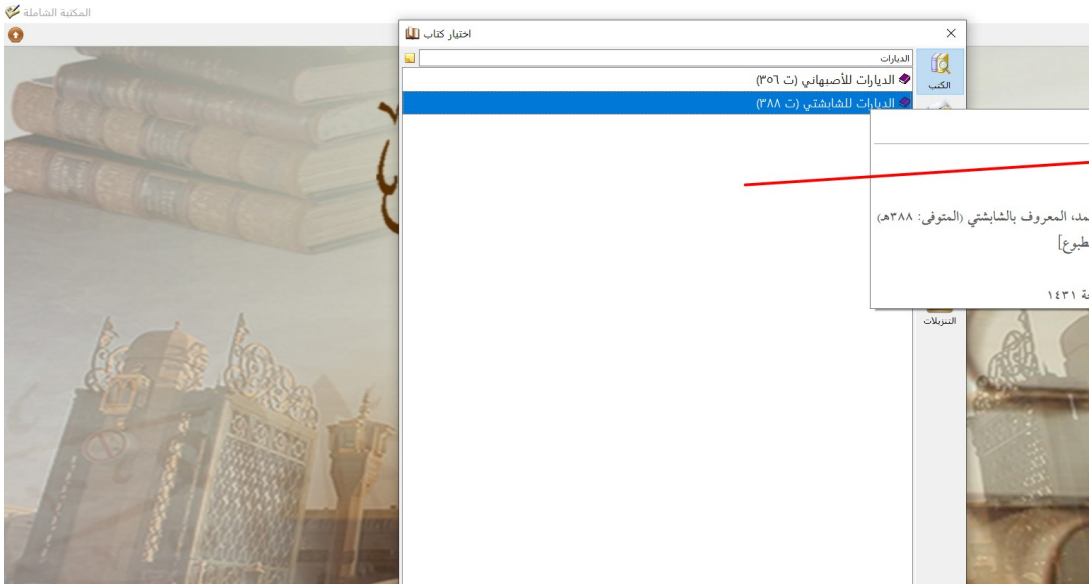
ب) وسائل التصنيف المعرفي

هناك العديد من وسائل التصنيف المعرفي، وقد لجأ البحث إلى الأدوات الآتية على الترتيب:

أولا: المكتبة الشاملة.

ثانيا: اتحاد مكتبات الجامعات المصرية.

ثالثا: مشروع Classify.

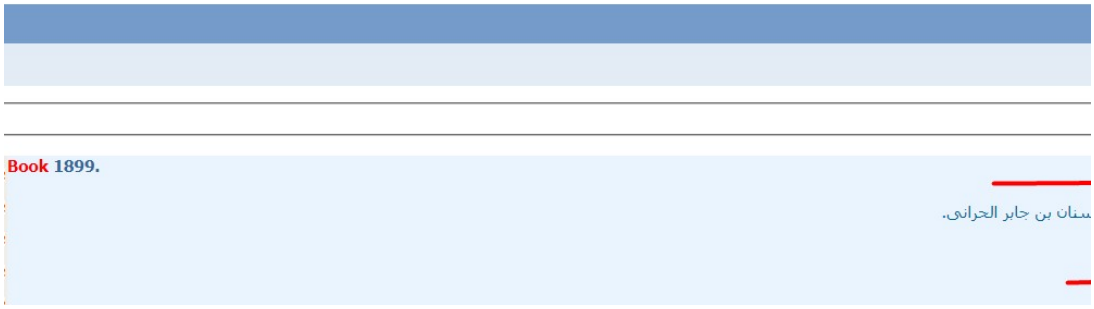


شكل (1): تحديد المجال المعرفي لأحد الكتب في المكتبة الشاملة

اعتمد البحث في البداية على المكتبة الشاملة في تعرف المجالات المعرفية لكتب المدونة؛ وذلك من خلال تحديد عنوان الكتاب، وعندها يظهر المجال المعرفي الذي ينتمي إليه، كما هو مبين في الشكل (1).

لم يكن تصنيف المكتبة الشاملة دقيقًا إلى حد كبير؛ حيث صنّفت مؤلفات من علوم ومجالات مختلفة تحت تصنيف الكتب الأدبية، ومن أمثلة ذلك: (العقل والهوى للحكيم الترمذي)، و(تاريخ الرسل والملوك)، و(المطر والسحاب لابن دريد)، و(التنبيه والإشراف للمسعودي)، و(الديارات للأصبهاني)، و(الجليس الصالح والأنيس الناصح)؛ لذلك تم اللجوء إلى بعض الأدوات الأخرى التي تستخدم تصنيفًا أكثر دقة، وهو تصنيف ديوي العشري.

وقد تمت الاستعانة بموقعين إلكترونيين للمساعدة في تصنيف الكتب وفق تصنيف ديوي العشري؛ الموقع الأول هو الخاص باتحاد مكتبات الجامعات المصرية والموضح بالشكل (2)، والموقع الثاني هو الخاص بمشروع Classify التابع لمركز المكتبة الرقمية على الإنترنت والموضح بالشكل (3).



شكل (2): تحديد المجال المعرفي لأحد الكتب باستخدام محرك بحث اتحاد مكتبات الجامعات المصرية

شكل (3): تحديد المجال المعرفي لأحد الكتب باستخدام Classify

إذا كان الكتاب مصنَّفًا مرة ضمن تصنيف موضوعي عام ومرة ضمن تصنيف موضوعي خاص – كما هو في الشكل (4) – فإننا نختار التصنيف المعرفي الخاص.

شكل (4): مستويان من التصنيف المعرفي لكتاب واحد على محرك بحث اتحاد مكتبات الجامعات المصرية

أحيانا يكون الكتاب مصنفًا تحت أكثر من تصنيف موضوعي؛ وحينها يقوم الباحث باختيار التصنيف الأكثر ترددًا مع الكتاب؛ فعلى سبيل المثال كتاب (المعاني الكبير لابن قتيبة الدينوري) صُنِّف تحت التصنيفات الموضوعية الآتية: البلاغة العربية (مرتين)، والشعر العربي (ثلاث مرات)، والدين الإسلامي (مرة واحدة)؛ فاختار الباحث أن يكون ضمن تصنيف الشعر العربي.

وفي حال تنازعَ تصنيفان كتابًا ما بنسبة متطابقة عمد الباحث إلى الكلمات المفتاحية ليرجّح تصنيفًا على آخر؛ كما في كتاب (الأعضاء والنفس والعقل والهوى للحكيم الترمذي) الذي صُنّف مرة ضمن الفلسفة الإسلامية

مصحوبًا بهذه الكلمات المفتاحية: (الفلسفة الإسلامية، العقل، الإسلام والطب) وأخرى ضمن علم النفس مصحوبًا بكلمة مفتاحية واحدة هي: (العقل)؛ ومن ثم رجّح الباحث أن ينسب الكتاب إلى الفلسفة الإسلامية.

وأحيانًا تتنازع عدة تصنيفات مؤلفًا واحدًا ولا نتمكن من الترجيح بينهما؛ فنعود – على سبيل المثال – مرة أخرى إلى المكتبة الشاملة، كما في كتاب (طبائع النساء لابن عبد ربه)، الذي صنفته المكتبة الشاملة ضمن الكتب الأدبية.

وقد يلجأ البحث إلى مطالعة الكتاب لتحديد الحقل الموضوعي الدقيق الذي ينتمي إليه الكتاب، كما في كتاب (اللغات في القرآن لابن حسنون)، والذي بدا له أنه أقرب لتصنيف (الألفاظ القرآنية).

(2) التصنيف الالتباسي الدلالي

كما سبق فإن هذا البحث معني بالكلمات متعددة الدلالات المتقاربة، وعليه فقد جرى تكوين قائمة الكلمات المطلوبة، وعليه أيضا جرى تصنيف سياقات المدونة؛ بغية استخراج السياقات التي تحتوي على واحدة أو أكثر من كلمات القائمة باعتبارها الخطوة الثانية نحو تصنيف المدونة من حيث الالتباس الدلالي.

ومما لا شك فيه أن مدونة الاختبار لا بد أن تكون جميع سياقاتها محتوية على كلمات ملتبسة؛ حتى يؤدي الاختبار الدور المنوط به. لذلك كان من الضروري استخراج جميع السياقات التي تحتوي على كلمات ملتبسة دلاليا، ومن ثم تُستخرج من هذه السياقات نسبة مرضية لتكوين مدونة الاختبار.

على أية حال جرى تقسيم السياقات المحتوية على قائمة الكلمات المُلبسة إلى: (أ) السياقات المكونة من ثلاث كلمات فأقل. (ب) السياقات المكونة من أربع كلمات فأكثر. ويعود السبب في ذلك التقسيم إلى أن السياقات المكونة من ثلاث كلمات فأقل ربما تصعّب من عملية التعلم الآلي؛ لكن ذلك لا يعني إهمال هذه المسألة، حيث سيتم معالجتها ضمن الأعمال المستقبلية.

الخلاصة

تمكنا في هذه الورقة البحثية من عرض منهج مقترح لبناء مدونة اختبار معيارية لفك الالتباس الدلالي في اللغة العربية الفصحى؛ ويبقى في أعمالنا المستقبلية أن تُستخرج هذه المدونة من مدونتها الأم، ثم يجري عنونتها دلاليا تمهيدا لبدء عملية تدريب الحاسوب واختبار النموذج اللغوي الإحصائي المتولد عن ذلك، فضلا عن تطوير هذه المدونة وفقا لأغراض البحث في موضوع فك الالتباس الدلالي آليا في اللغة العربية.

الشكر

نود أن نتوجه بأصدق الشكر إلى الشركة الهندسية لتطوير الأنظمة الرقمية RDI التي أمدتنا بقائمة إلكترونية لكلمات المكنز الكبير، وإلى الدكتور وليد نزيه – عضو هيئة التدريس بقسم علوم الحاسب بكلية هندسة وعلوم الحاسب جامعة الأمير سَطّام بن عبد العزيز – على معالجته لأجزاء كبيرة من المدونة.

المراجع

[1] Abraham Kaplan, “an experimental study of ambiguity and context”, Mechanical Translation, vol.2 no.2, November (1955), pp. 45.

[2] أحمد شفيق الخطيب، قراءات في علم اللغة، دار النشر للجامعات، ط 1، (2006)، ص 112.

[3] David Pinto et al, "Word sense induction in the Arabic language: A self-term expansion-based approach", Paper presented at the 7th Conference on Language Engineering of the Egyptian Society of Language Engineering- ESOLE, (2007), pp. 235 – 245, Cairo, Egypt.

[4] Laroussi Merhbene, et al, "A Semi-Supervised Method for Arabic Word Sense Disambiguation Using a Weighted Directed Graph", International Joint Conference on Natural Language Processing, (2013), pages 1027–1031, Nagoya, Japan.

[5] نبيل علي، اللغة العربية والحاسوب، دار تعريب، (1988)، ص 531 - 532.

[6] Bassem Haddad, "Semantic representation of Arabic: A logical approach towards compositionality and generalized Arabic quantifiers". International Journal of computer processing of oriental languages, Vol 20, No. 1, (2007), pp. 37-52.

[7] Bakhouche Abdelaali, et al, "Ant Colony Algorithm for Arabic Word Sense Disambiguation through English lexical information", International Journal of Metadata, Semantics and Ontologies, Vol. 10, (2015), p 202–211.

[8] حسين البسومي، اللبس الدلالي في المعالجة الآلية للغة العربية المعاصرة المكتوبة، رسالة دكتوراه، (2011)، جامعة القاهرة.

[9] Anis Zouaghi, et al, "Contribution to Semantic Analysis of Arabic Language", Hindawi Publishing Corporation, (2012).

[10] Anis Zouaghi, et al, "A Hybrid Approach for Arabic Word Sense Disambiguation", International Journal of Computer Processing of Languages, Vol. 24, No. 2, (2012), pp. 133-151.

[11] Laroussi Merhben, et al, "Lexical disambiguation of Arabic language: An experimental study", Polibits, Vol. 46, (2012), pp. 49–54.

[12] Madeeh El-Gedawy, "Using Fuzzifiers to Solve Word Sense Ambiguation in Arabic Language", International Journal of Computer Applications (0975 – 8887) Vol. 79, No. 2, (2013), pp. 1-8.

[13] Marwah Alian, et al, "Word sense disambiguation for Arabic text using Wikipedia and Vector Space Model", International Journal of Speech Technology, Vol 19, No. 4, (2016), pp. 857–867.

[14] Meryeme Hadni et al, "Word Sense Disambiguation for Arabic Text Categorization", The International Arab Journal of Information Technology, Vol. 13, No. 1A, (2016), pp. 215 –

222.

[15] Mohamed M. El-Gamal, "Arabic Word Sense Disambiguation", Master Thesis, Arab Academy for Science, Technology and Maritime Transport, College of Engineering and Technology, (2012).

[16] Mohamed M. El-Gamal, et al, "A Comparative Study for Arabic Word Sense Disambiguation Using Document Preprocessing and Machine Learning Techniques", ALTIC, Alexandria, Egypt, (2011).

[17] M. E. B. Menai and W. Alsaeedan, "Genetic Algorithm for Arabic Word Sense Disambiguation", 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Kyoto, Japan, (2012), pp. 195-200.

[18] Mona Diab, "Word sense disambiguation within a multilingual framework", PhD dissertation, University of Maryland, (2003).

[19] Nadia Bouhriz, et al, "Word Sense Disambiguation Approach for Arabic Text", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, (2016), p. 381-385.

[20] Samir Elmougy, et al, "Naïve Bayes Classifier for Arabic Word Sense Disambiguation", in Proceeding of the Sixth International Conference on Informatics and Systems, (2008), pp: 16-21.

[21] Soha M. Eid, et al, "A Comparative Study of Rocchio Classifier Applied to supervised WSD Using Arabic Lexical Samples", Tenth Conference on Language Engineering, (ESOLEC'10), (2010), Cairo, Egypt.

[22] أحمد مختار عمر، المكنز الكبير، دار سطور، الرياض، ط 1، (2000)، ص 17 - 18.

[23] محمد علي الخولي، مدخل إلى علم اللغة، دار الفلاح للنشر والتوزيع، (2010)، ص 77.

[24] نبيل الزهيري، قاموس مصطلحات المعلوماتية واللغويات الحسابية، مكتبة لبنان ناشرون، ط 1، (2003)، ص 78، 49.

[25] https://research.aimultiple.com/data-annotation

[26] https://research.aimultiple.com/data-labeling

[27] Warren Weaver, "Translation. In Machine Translation of Languages", MIT Press, Cambridge, MA, (1949), pp. 9.

[28] مجمع اللغة العربية، المعجم الوسيط، القاهرة، ط 5، (2021)، ج 1، ص 836.

[29] مجمع اللغة العربية، المعجم الفلسفي، دار المعارف، القاهرة، ط 3 مصورة عن ط 1، (2022)، ص 45.

[30] https://www.almojam.org/page-2-7

[31] محمد حسن عبد العزيز، المعجم التاريخي للغة العربية وثائق ونماذج، دار السلام، القاهرة، ط1، (2008)، ص 175، 196.

المؤلفون

عمرو الجندي

باحث بمجمع اللغة العربية في القاهرة، وعضو لجنتي التحرير والتنسيق بمشروع المعجم التاريخي التابع لاتحاد المجامع العلمية واللغوية العربية. كانت رسالته للماجستير حول التشكيل الآلي للنصوص العربية، أما رسالة الدكتوراه فتدور حول فك الالتباس الدلالي آليا. شارك في العديد من المشروعات البحثية، وفي تحرير عدد من المعاجم، وفي تطوير عدد من محركات البحث المعجمية.

أ.د. سعيد الوكيل

ناقد وشاعر وكاتب. وأستاذ النقد والأدب العربي الحديث بكلية الآداب جامعة عين شمس، ووكيل الكلية (السابق) للدراسات العليا والبحوث. له عدة مؤلفات حول السرد، وعدد من الدواوين الشعرية، والكتابات الشعرية والقصصية للأطفال، والكثير من الأبحاث في المجلات العلمية المحكمة، والكثير من المشاركات في المؤتمرات الدولية.

أ.د. ياسر حفني

أستاذ تكنولوجيا المعلومات بكلية الحاسبات والذكاء الإصطناعي جامعة حلوان. حاصل على الدكتوراه من جامعة شيفيلد – بريطانيا. تشمل اهتماماته البحثية معالجة الكلام والإشارات، والتعلم الآلي، وهندسة اللغة. أشرف على العديد من الرسائل العلمية، وله العديد من الأوراق العلمية المنشورة في المجلات المصنفة.

د. أحمد عبد الحميد عمر

مدرس علم اللغة بكلية الآداب جامعة عين شمس. حاصل على الدكتوراه من جامعة أمستردام. مدير تحرير مجلة "فصول" منذ 2020 إلى الآن. ترجم من الإنجليزية كتاب (المناورة الاستراتيجية في الخطاب الحجاجي) Strategic Maneuvering in Argumentative Discourse الصادر عن المركز القومي للترجمة، وصدرت له مقالات في المجلات العلمية المحكمة.

Towards building A Standard Arabic Test Corpus for Word Sense Disambiguation

Amr El-Gendy*[1], Said Al-Wakil**[2], Yaser Hifny***[3], Ahmed Abdul-Hamid Omar**[4]

**Academy of the Arabic Language in Cairo*
*** Faculty of Arts, Ain Shams University*
****Faculty of Computers and Artificial Intelligence, Helwan University*

[1] amr25@hotmail.com
[2] alwakil@aucegypt.edu
[3] yhifny@yahoo.com
[4] ahmed.omar@art.asu.edu.eg

Abstract: ***This paper aims to provide a methodology for the stages that can be followed to build a Standard Arabic Test Corpus. The aim is to use it to automatically Word Sense Disambiguation in the texts of the classical Arabic language, especially since there is hardly a corpus that researchers can use to test their statistical linguistic models. Even there is no unified list of ambiguous words that can be subject to testing; which leads experts to a state of scientific uncertainty about the best proposed models for Arabic Word Sense Disambiguation automatically. The research followed a proposed approach, starting from defining the ambiguous words used in all ages, and passing through the collection and preparation of a major corpus, and ending with its classification in preparation for extracting a test set that represents the major corpus as much as possible.***

Keywords: ***Linguistic Corpus – Linguistic Corpus Classifying – Standard Test Corpus – List of Ambiguous Words – Context – Word Sense Disambiguation.***