بناء ذخيرة لُغوية قياسية معاصرة للغة العربية لأغراض استرجاع المعلومات

د. إبراهيم حسن أبو الخير

قسم علم المعلومات، كلية العلوم الاجتماعية

جامعة أم القرى

قسم المكتبات والمعلومات ، كلية الآداب – جامعة المنيا

مستخلص

الذخيرة اللغوية هي مجموعة من النصوص المكتوبة بلغة معينة، أو بأكثر من لغة، والتي يتم جمعها، وتخزينها، ومعالجتها، آلياً بشكل منظم على الحاسب الآلي؛ وفقاً لطريقة استخدامها والعمل عليها، وهي تعد أحد أهم المصادر في مجال البحث في استرجاع المعلومات، والمعالجة الطبيعية للغة، وكذلك اللغويات الحاسوبية، وقد أصبح بناء الذخائر اللغوية أمراً شائعاً ومألوفاً في هذه المجالات منذ سنوات، كما أن أحجام الذخائر قد زادت بشكل كبير في الآونة الأخيرة؛ بسبب التطور الهائل في التكنولوجيا المستخدمة في بنائها. إن هذه الدراسة هي محاولة من قبل الباحث لبناء ذخيرة لُغوية قياسية معاصرة للغة العربية. والذخيرة الناتجة، هي ذخيرة نصية مكتوبة مكونة مما يزيد على خمسة ملايين مقال وتحقيق صحفي، بإجمالي عدد كلمات يزيد على مليار ونصف المليار كلمة، منها حوالي أكثر من ثلاث ملايين كلمة فريدة لم تتكرر، وقد تم جمعها من المقالات الصحفية في عشرة مصادر من ثمانِ دول عربية، على مدار أربع عشرة سنة، وقد تم تشفير الذخيرة بنوعين من التشفير هما: الكود الموحد UTF-8، وكود ويندوز للغة العربية Windows cp-1256، كما تم توسيمها بلغة SGML، ولغة XML.

الكلمات الدالة: الذخائر اللغوية Corpus, Corpora، استرجاع المعلومات Information Retrieval، بناء الذخائر اللغوية Corpus Creation، المعالجة الطبيعية للغة Natural Language Processing، اللغويات الحاسوبية Computational Linguistics.

تمهيد

تعتمد كفاءة نظم استرجاع المعلومات بكل أنواعها، بشكل كبير على التجارب التي تجريها الشركات التي تنتج هذه النظم والتى يجريها مصمموها؛ هذه التجارب في الأساس هي تجارب لمحاكاة الواقع من حيث الاستفسارات المقدمة لنظام المعلومات، وما يحتويه هذا النظام من وثائق تجيب على تلك الاستفسارات، وعادة ما تكون هذه التجارب تجارب معملية مغلقة Ad-hoc retrieval experiments يتحكم الباحثون فيها، وفي كل عنصر من عناصر عملية الاسترجاع؛ ومن ثم يتمكنون من تحديد أسباب النجاح أو الخلل في تلك العملية، والعمل على إصلاح ومعالجة الأسباب التي أدت لذلك.

وتعد الذخيرة اللغوية أحد أهم العناصر في هذا النوع من التجارب في نظم استرجاع المعلومات، والتي لا غنى عنها للباحثين في هذا المجال، وفي مجال المعالجة الطبيعية للغات بشكل عام، والسبب في ذلك هو أن تلك الذخائر تمثل نموذجاً للاستخدام الفعلي والعملي للغة في كافة المجالات وشتى المواقف، ويمكن من خلالها دراسة نماذج واقعية للغة والتراكيب اللغوية. وقد استقر استخدام الذخائر اللغوية في عدد كبير من لغات العالم تقريباً، وهناك تقدم ملحوظ في هذا المجال للغات الأوروبية خاصة الإنجليزية والألمانية والفرنسية. لكن لا زال استخدام الذخائر اللغوية في اللغة العربية حديثاً نسبياً بالمقارنة بتلك اللغات.

ويحاول الباحث في هذه الدراسة بناء ذخيرة لغوية قياسية معاصرة للغة العربية؛ لكي يتم استخدامها في التجارب العلمية في مجال استرجاع المعلومات، والمعالجة الطبيعية للغة، وكذلك اللغويات الحاسوبية، على أن تكون هذه الذخيرة متاحة للباحثين؛ فالذخائر الحالية إما أنها صغيرة الحجم، ولا تعطي نتائج علمية قابلة للتطبيق، أو أنها باهظة الثمن لا يستطيع عدد كبير من الباحثين الحصول عليها.

ونعني هنا، بمفهوم اللغة القياسية المعاصرة Modern Standard Language، اللغة العربية الفصحى المستخدمة في الكتابات الرسمية، وكتابة المقالات والتحقيقات الصحفية، وليس اللغة الكلاسيكية Classical Language؛ لغة الكتب التراثية القديمة، ولا اللغة الدارجة أو العامية colloquial Language.

مشكلة الدراسة وأهميتها

إن اللغة العربية هي لغة القرآن الكريم، ويستخدمها ما يربو على المليار ونصف المليار مسلم حول العالم في عبادتهم، وهي اللغة الأولى أو اللغة الأم لما يقارب ٢٥٠ مليون شخص حول العالم، وهي اللغة الرسمية لـ ٢٢ دولة عربية، وإحدى اللغات الرسمية لدول

غير عربية مثل: تشاد وإريتريا ومالي وتركيا (Encyclopedia Britannica Almanac, 2009)، بالإضافة إلى هذا فإنها تعد إحدى اللغات الست الرئيسة الرسمية للأمم المتحدة (الأمم المتحدة، ٢٠١٥)، وذلك منذ السبعينات من القرن الماضي بالقرار رقم ٣١٩٠(د-٢٨) لعام ١٩٧٣ (الأمم المتحدة، ١٩٧٣).

وعلى الرغم من كل ما سبق فإنها في حاجة أكثر من غيرها للمزيد من الأبحاث والدراسات لسبر أغوارها، خاصة في المصادر الإلكترونية المتاحة للاستخدامات البحثية، والتي من ضمنها الذخائر اللغوية؛ إذ مازالت اللغة العربية فقيرة نسبياً في هذا المجال، ومن هنا تنبع أهمية الدراسة، إذ إن هناك حاجة ملحة ومتزايدة لوجود ذخائر لغوية باللغة العربية، تكون كبيرة الحجم، ولها دلالة معبرة، وذلك لكي تقدم رصداً دقيقاً وشاملاً لاستخدامات اللغة؛ كما أن هناك حاجة لذخيرة عربية تمثل أكثر من دولة، وتمثل أكثر من أسلوب في الكتابة، وتمثل أكثر من مصدر، وتكون موزعة على عدد من السنوات؛ على أن تكون هذه الذخيرة متاحة لخدمة البحث العلمي، خاصة في مجال استرجاع المعلومات، والمعالجة الطبيعية، واللغويات، أو اللسانيات الحاسوبية وهو المنهج القائم على دراسة اللغة في ضوء النصوص اللغوية المخزنة على الحاسب.

ولعل المشكلة الحقيقية في ذخائر اللغة العربية الموجودة حالياً هي صغر حجمها من حيث عدد النصوص المستخدمة في إنتاج الذخيرة، وبالتالي قلة عدد الكلمات المتاحة فيها؛ مما قد يكون له دلالة سلبية على نتائج البحوث، كما سبق وأن أشرنا.

بالإضافة إلى ذلك؛ فإن الذخائر الجيدة كبيرة الحجم، غالباً ما تكون ذخائر تجارية، وغلاء ثمن هذه الذخائر بشكل كبير، يشكل عبئاً إضافيا على الباحثين، الذين لا يتوفر لهم التمويل الملائم لشراء مثل هذا النوع من المصادر لاستخدامه في تجاربهم وأبحاثهم.

أهداف الدراسة

تهدف هذه الدراسة في الأساس إلى تحديد مفهوم الذخيرة بوجه عام، وأنواعها وسبب الحاجة إليها، والتعرف على الذخائر اللغوية العربية، وتحديد مبادئ تصميم وبناء الذخائر اللغوية، فضلاً عن الهدف الرئيس للدراسة، وهو إنتاج ذخيرة عربية متكاملة، ويمكن بلورة أهداف الدراسة فيما يأتي:

- تحديد مفهوم الذخيرة اللغوية.
- معرفة أنواع الذخائر اللغوية والفروق بينها.
- تحديد أسباب الحاجة للذخائر اللغوية وأهميتها.

- إلقاء الضوء على التجارب السابقة في بناء الذخائر اللغوية باللغة العربية، إذ إن جميع الدراسات التي تناولت الذخائر اللغوية العربية كانت كلها تقريباً باللغة الإنجليزية!.
- تحديد المبادئ الأساسية لتصميم وبناء الذخائر اللغوية.
- بناء ذخيرة لغوية قياسية حديثة للغة العربية، وإتاحتها للباحثين في هذا المجال؛ وذلك للعمل بها، وهو الهدف الرئيس لهذه الدراسة.
- المساهمة في إثراء العمل في استرجاع المعلومات باللغة العربية من خلال بناء هذه الذخيرة.

مفهوم الذخيرة اللغوية

يشير المعنى اللغوي (معجم المعاني الجامع، ٢٠١٠)، لكلمة **ذَخيرة**، وجمعها **ذخيرات** أو **ذخائرُ**، إلى أن **الذخِيرَةُ** قد تعني: عُدَّة الحرب من رصَاص و قذائف، أو ما يُعدّ للآخرةِ من الأعمال الحسنة، أو الذُخْر، أو المَئُونة، أو ما يُدّخر من القوت.

ونرى أن هذه المعاني جميعاً، تشير إلى أن الذخيرة هى تجميع لشىء معين، وهذا يقترب من المعنى الاصطلاحي للكلمة. وتشير الذخيرة corpus، وجمعها بالإنجليزية corpora (Corpus Linguistics, 2011)، إلى مجموعة من النصوص المكتوبة بلغة معينة أو بأكثر من لغة، والتي يتم جمعها، وتخزينها، ومعالجتها، آلياً بشكل منظم على الحاسب الآلي؛ وفقاً لطريقة استخدامها والعمل عليها.

وقد تم صياغة مصطلح "الذخيرة اللغوية" في عام ١٩٩٩ على يد الباحث اللغوي ورائد العمل العربي في هذا المجال الدكتور عبدالرحمن صالح (صالح، ١٩٩٩)، وهو المصطلح الأكثر شيوعاً واستخداماً للتعبير عن المفهوم السابق، لكن البعض يطلق على مجموعة النصوص التي يتم تجميعها أياً كان نوعها مصطلح مدونة (المدونة العربية، ٢٠١٣؛ صالح، محمود، ٢٠١٤)، ويرى الباحث أن مصطلح الذخيرة اللغوية يعبر بشكل أفضل عن المفهوم، خاصة أن كلمة مدونة قد تتسبب في الخلط والبلبلة عند البعض، لأنها تشير إلى المواقع الإلكترونية التي تجمع تدوينات أصحابها Blogs، والتي عادة ما تكون كالمفكرة، أو ساحة لطرح الآراء والأفكار، ناهيك عن أن مصطلح المدونة غير شائع نسبياً.

أنواع الذخائر اللغوية

هناك عدد كبير من أنواع الذخائر اللغوية، نوجزها فيما يأتي (Zaghouani, 2014; Corpus Linguistics, 2011):

- الذخيرة النصية الخام raw text corpus، وتضم الذخائر أحادية اللغة corpus، وكما يشير الاسم فهي مجموعة من النصوص الرقمية بلغة واحدة monolingual، والذخائر متعددة اللغات multilingual corpus، وعادة ما تكون لغتين فقط، لكن يوجد عدد قليل من الذخائر بأكثر من لغتين، ولا يشترط أن تكون النصوص لها علاقة ببعضها، أو أحدهم ترجمة للآخر، والذخائر المشكلة diacritical corpus، وذخائر الويب web corpus.
- الذخيرة المشروحة annotated corpus، وتعني الذخيرة التي يكون فيها توضيح لكل كلمة أو جملة بداخلها، و تسمى أحياناً، بنك الكلمات الشجري Treebank، و بنك الكلمات عبارة عن ذخيرة لغوية تحدد فيها أقسام الكلام parts of speech (POS)، أو تعالج فيها الجمل؛ ليتم توضيح نوع كل كلمة سواء أكانت فعلاً أم اسماً.. إلخ. وغالباً ما تكون في شكل كلمات منفصلة، وأمام كل كلمة نوعها، وذلك في شكل شجري، وتضم بنوك الكلمات ذخائر الأسماء name entity corpus، ذخائر تصحيح الأخطاء؛ الذخيرة التعليمية أو الإملائية orthographic corpus، وهي ذخيرة تستخدم لملاحظة الطرق المختلفة لكتابة الكلمات وتحديد الأخطاء اللغوية الشائعة في الكتابة.
- المعاجم اللغوية lexicon corpus وقوائم الكلمات word lists.
- ذخائر الكلام أو الذخائر الصوتية speech corpus، وهي الذخائر المسجلة من الأخبار من الراديو أو التلفزيون، وقد تصاحبها كتابة كاملة لكل ما هو مسموع transcription.
- ذخائر التعرف على خط اليد hand recognition corpus، وتضم النصوص الممسوحة رقمياً digitally scanned texts والمشروحة annotated.
- ذخائر الأسئلة والأجوبة questions and answers corpus.
- الذخائر المقارنة comparative corpus.
- ذخائر الملخصات abstracts corpus.
- ذخائر كشف الانتحال plagiarism corpus.
- ذخائر البريد الإلكتروني لكشف البريد غير المهم spam email.
- ذخائر الشبكات الاجتماعية مثل تويتر وفيس بوك وهي ما تسمى بذخائر كشف الحالة النفسية أو العاطفة sentiment analysis.
- الذخيرة الموازية parallel corpus، وهي ذخيرة ثنائية اللغة، وتختلف عن الذخيرة متعددة اللغات في أن الذخيرة تحتوي على نصوص بلغة معينة وترجمتها بلغة أخرى.

- الذخيرة التجميعية collocation، وتشير إلى الذخائر التي تضم الكلمات التتابعية وتواتر ورودها معاً.
- ذخيرة السياق concordance، وتعني الذخيرة التي يتم فيها تجميع الكلمات وفقاً للسياق وتستخدم لتحليل الاستخدامات المختلفة للكلمات في لغة ما.
- ذخيرة المفردات، وهي ذخائر للكلمات فقط lemmas، ومتغيراتها سواء تغير معناها عند إضافة زوائد لها tokens، أو تم تجميعها؛ لأن لها نفس المعنى، رغم الزوائد التي أضيفت عليها، وتستخدم أيضاً، في ملاحظة طرق الكتابة، وتحديد الأخطاء اللغوية الشائعة، وكذلك في إنشاء بنوك الكلمات.

كما يمكن تقسيم الذخائر اللغوية من حيث الغرض الذي تستخدم من أجله الذخيرة (Rogati & Young, 2004)، إلى ذخيرة عامة، وذخيرة متخصصة، الذخيرة العامة عادة ما تحتوي على موضوعات مختلفة مثل: الصحف بأنواعها، والموسوعات، والدوريات، بينما الذخيرة المتخصصة فغالباً ما تكون متخصصة في موضوع واحد فقط، وتكون قليلة نسبياً بالمقارنة بالذخائر العامة، وغالباً ما تكون تجميعاً للعناوين و/أو المستخلصات فقط، وليس كامل الوثائق أو المقالات.

ويمكن تقسيمها أيضاً (Rogati & Young, 2004)، إلى ذخيرة متزامنة، وذخيرة غير متزامنة، وتعني الذخيرة المتزامنة أن جميع النصوص جمعت في نفس الحقبة الزمنية، وتستخدم غالباً لمقارنة التغيرات اللغوية من إقليم لآخر، أما الذخيرة غير المتزامنة فيتم تجميعها لفترات زمنية مختلفة، ويتم مقارنة استخدام اللغة في تلك الفترات.

الحاجة للذخائر اللغوية

إن للذخائر اللغوية فوائد عدة لا يمكن إغفالها للكثير من المجالات منها؛ استرجاع المعلومات، والمعالجة الطبيعية للغة، واللغويات الحاسوبية، ولا تقتصر فوائدها على مجال الحاسب الآلي، بل تتعداها للمتخصصين في اللغويات وتاريخ اللغات وعلم اللغة بشكل عام. ويمكن أن نوجز فوائد الذخائر اللغوية فيما يأتي (Atkins, Clear, & Ostler, 1992; Dash, 2008; الربيعة،2012 ; مشروع الذخيرة العربية، 2008):

- رصد دقيق وشامل لاستعمال اللغة في إقليم خاص في عصر من العصور.
- رصد منظم للاستعمال الحقيقي لمصطلحات مجال فني معين.
- تصفح لمعاني الكلمات من خلال سياقها عبر الزمان.
- تحليل لغة كاتب أو شاعر معين، وكذلك الأساليب اللغوية لمؤلف ما أو عدة مؤلفين.

- بناء المعاجم اللغوية، سواء أكانت أحادية اللغة، أم متعددة اللغات، أم معاجم دلالية.
- دراسة اللغة والبناء اللغوي في وقت أو عصر معين.
- دراسة الكلمات بشكل مفرد، والعلاقات بينها، والجذور والتصريفات المختلفة للكلمات، والفروق بينها، ومدى ارتباطها باستخدامات لغوية محددة.
- دراسة الفروق بين النصوص اللغوية المختلفة.
- تعليم اللغة.
- تحسين أداء نظم استرجاع المعلومات Information Retrieval Systems، وتحسين أداء نظم الأسئلة والأجوبة Question & Answer Systems.

مبادئ تصميم وبناء الذخائر اللغوية :

عند التخطيط لإنشاء ذخيرة لغوية فإنه يجب أن نأخذ في الاعتبار عدة عوامل أو عناصر عند إنشائها (Mansour, M., 2013):

أولاً– *الحجم Size:* يجب أن يكون حجم الذخيرة كبيراً نسبياً، بغض النظر عن حجم وحداتها، ويرجع السبب في ذلك إلى أنه عند إجراء البحوث باستخدام الذخيرة الصغيرة لا تكون النتائج موثوقاً بها إلى حد كبير، ولا يمكن تعميمها، وكقاعدة عامة فإنه كلما زاد حجم الذخيرة كان ذلك أفضل.

ثانياً– *تحديد الغرض Purpose:* ويعني ببساطة نوعية البحوث والدراسات التي سوف تجرى على الذخيرة وفي ماذا سوف تستخدم؛ فالذخيرة في حد ذاتها لا فائدة منها دون استخدامها في أحد البحوث العلمية سواء أكانت لغوية أم في استرجاع المعلومات والمعالجة الطبيعية للمعلومات.

ثالثاً– *تنوع الموضوعات Diversity:* حيث يجب أن لا تقتصر الذخيرة اللغوية على موضوع واحد إلا إذا كان ذلك الموضوع في ذاته هو الدراسة، وكلما تنوعت الموضوعات في الذخيرة كان ذلك أفضل.

رابعاً– *تمثيل اللغة Representativeness:* ويقصد بها تحقيق الذخيرة للتوازن بين الدول المختلفة التي تتحدث تلك اللغة؛ بمعنى أن تحتوي الذخيرة على جزء من أكثر من دولة.

خامساً– *التوازن بين المصادر Balance :* بمعنى أن لا يطغى مصدر على الآخر وأن تكون جميع المصادر متوازنة التمثيل.

وتتفق هذه المبادئ الخمس مع المبادئ التي وضعها آتكينز وكلير وأوستلر عام ١٩٩٢ (Atkins, Clear, & Ostler, 1992)، وهي التمثيل، والتنوع، والتوازن والحجم؛

مع زيادة الغرض من بناء الذخيرة وهو أمر يجب وضعه في الاعتبار من البداية؛ إذ إنه يحدد وبشكل كبير كيفية معالجة الذخيرة وشكلها النهائي.

الدراسات والذخائر السابقة والمثيلة

في هذا الجزء من الدراسة نستعرض عدداً من الدراسات والمحاولات السابقة؛ لإنشاء الذخائر اللغوية العربية، وتجدر الإشارة إلى أن هذا الجزء يقتصر على دراسات إنشاء الذخائر باللغة العربية فقط دون غيرها من اللغات، كما يقتصر العرض على الذخائر النصية أحادية اللغة فقط دون غيرها من أنواع الذخائر، كذخائر الكلمات، والذخائر المعجمية، والذخائر المسموعة، وذخائر الآراء؛ بسبب صغر حجمها، وأنها تقاس بعدد الجمل وليس النصوص وحجمها وعدد الكلمات. وتم تقسيم هذا الجزء إلى قسمين؛ أولهما يختص بعرض الذخائر، ودراسات إنشاء الذخائر المجانية، وثانيهما يختص بعرض الذخائر، ودراسات إنشاء الذخائر التجارية.

أولاً- الذخائر المجانية

أ- ذخيرة ملخصات المؤتمر السعودي لعلوم الحاسب الآلي Saudi SACS Arabian National Computer Science Conf. (Abu Salem 92, Hmeidi, Kanaan, & Evens, 1997)، وتتكون من ٢٤٢ (مائتان واثنان وأربعون) مستخلصاً من المؤتمر المذكور تضم ٤٦٩٦٨(ست وأربعون ألفاً وتسعمائة وثمانية وستون) كلمة، وقد بدأت الذخيرة عام ١٩٩٢ بعناوين البحوث فقط، ثم أضيف إليها المستخلصات، ويضم كل مستخلص ٣٦ حقلاً، منها: العنوان، المؤلفون، المصادر، المستخلص نفسه وهي نصية تماماً دون أي تكويد، فقط رمز لكل حقل في بدايته.

ب- ذخيرة جريدة الراية القطرية (Hasnah 96)، وهي ذخيرة نصية لجريدة الراية القطرية تضم ١٨٧(مائة وسبعة وثمانون) مقالاً، وحوالي ٢١٩٩٧٨(مائتان وتسعة عشر وتسعمائة وثمانية وسبعون) كلمة، منها٣٠٠٩٦(ثلاثون ألف وست وتسعون) كلمة فريدة غير مكررة.

ولعل السبب في ذكر هاتين الذخيرتين، رغم صغر حجمهما، يرجع إلى أنهما من أوائل الذخائر اللغوية التي تم إنشاؤها في بداية العمل على استرجاع المعلومات باللغة العربية بشكله الجديد، في معهد إلينويز للتكنولوجيا Illinois Institute of Technology، والسبب في صغر الحجم يعود للإمكانات الحاسوبية الضعيفة في هذه الأيام. ويحسب

لمنشئي هاتين الذخيرتين أنهم كانوا أصحاب السبق والريادة في هذا النوع من الدراسات في العالم العربي، وكلتا الذخيرتين متاحتان بالاتصال الشخصي بالمؤلفين.

ج- مشروع الذخيرة العربية (صالح، ١٩٩٩؛ مشروع الذخيرة العربية، ٢٠٠٨) الذي تتبناه المنظمة العربية للتربية والثقافة والعلوم، يعد أحد أهم المشروعات التي تحاول بناء ذخيرة لغوية تتناول الاستعمال الحقيقي للغة العربية منذ أقدم العصور حتى العصر الحالي وهي عبارة عن تجميع لكل ما أنتجه الفكر العربي منذ الجاهلية إلى العصر الحديث، ومن المفترض أن يضم المشروع:

١- المؤلفات ذات القيمة الكبيرة في الآداب والعلوم والتكنولوجيا.
٢- المحاضرات القيمة المنشورة في المجلات الأدبية والعلمية.
٣- جميع المعاجم العربية والمزدوجة اللغة قديماً، وحديثاً.
٤- المقالات ذات القيمة المنشورة في المجلات الأدبية والعلمية.

والهدف الرئيس للمشروع كان هو إنشاء بنك إلكتروني للغة العربية المستعملة فعلياً، وعمل معجم إلكتروني للغة العربية والكلمات المرادفة في اللغة الإنجليزية والفرنسية.

وقد ظهرت فكرة المشروع إلى الوجود في عام ٢٠٠٤؛ حيث طرحه الخبير اللغوي عبدالرحمن الحاج صالح رئيس المعجم الجزائري للغة العربية؛ بهدف تمكين الباحثين العرب من إيجاد ما يبحثون عنه من معلومات بشأن أي موضوع في أي مجال وفي أي تخصص بسهولة على الإنترنت بهدف توفير بنك آلي للمعلومات، وأقر المشروع في عام ٢٠١٠ من قبل جامعة الدول العربية؛ ليصبح مؤسسة قائمة بذاتها (مشروع الذخيرة العربية، ٢٠٠٨).

ومن أسف فإنه لا يوجد شيء ملموس أو نتاج واضح لهذا المشروع حتى الآن، وكل ما ظهر هو تجارب فردية لوضع بعض الكتب على الإنترنت لا تتجاوز الأربعمائة كتاب (صالح، عبدالرحمن؛ ٢٠١٤).

د- ذخيرة اللغة العربية المعاصرة للطيفة السليطي وإيريك آتويل (Al-Sulaiti & Atwell, 2005; Al-Sulaiti & Atwell, 2006)، وهي في الأصل رسالة الماجستير للباحثة لطيفة السليطي في جامعة ليدز (Al-Sulaiti, 2004) ومصدرها الصحف من بداية التسعينات وحتى ٢٠٠٤، وهي متاحة مجاناً للتحميل؛ صحف، ومواقع إنترنت، وإيميلات، وقد تم تجميعها يدوياً، وتتكون الذخيرة من ٨٤٢٦٨٤ (ثمانمائة واثنان وأربعون ألفاً وستمائة وأربع وثمانون) كلمة، وتضم الذخيرة خمس فئات موضوعية رئيسة هي: الأدب، الفنون، العلوم، إدارة الأعمال، ومتفرقات تضم تحتها ٤١ فئة موضوعية صغرى في ٤١٦

ملفاً، وكان الغرض الأساسي من تطوير هذه الذخيرة هو استخدامها في البحث العلمي في مجال استرجاع المعلومات، وتعليم اللغة العربية لغير الناطقين بها، والذخيرة تعد من الذخائر الجيدة جداً؛ نظراً لتنوع موضوعاتها وتوازنها، بالإضافة إلى أنه تم توسيم الذخيرة بلغة XML.

هـ- ذخيرة جريدة أخبار الخليج وقد بناها الدكتور مراد عباس (Abbas and Samaili, 2005) خصيصاً لأغراض المعالجة الطبيعية للمعلومات، ومتاحة مجاناً للباحثين، وهي تجميع لمقالات جريدة الخليج البحرينية، ويبلغ حجمها ٤ ميجابايت، وثلاثة ملايين كلمة وتغطي البحرين. وقد أضيف إليها ذخيرة جريدة الوطن ٢٠٠٤ (Abbas, Samaili, & Berkani, 2011) وهي أيضاً، من جمع الدكتور مراد عباس، ومتاحة مجاناً للباحثين، ومن جريدة الوطن العمانية فقط، ويبلغ حجمها منفرداً ١٠ ميجابايت، وعدد الكلمات بها عشرة ملايين كلمة، لسنة ٢٠٠٤ فقط، وهي أيضاً متاحة مجاناً للتحميل. وقد تم استخدام هذه الذخيرة لإنشاء ذخيرة كلمات KALIMAT، وهي أيضاً، ذخيرة مجانية متعددة الاستخدامات بها ٢٠٢٩١ (عشرون ألفاً ومئتان وواحد وتسعون) مقالاً، وإجمالي ١٨١٦٧٨٣ كلمة مشروحة، ومحددة النوع annotated جميعها من جريدة الوطن العمانية (El-Haj, Koulali, 2013).

وقد استخدم الباحث ذخيرة الوطن السابق ذكرها، وقام باجراء عدد من تجارب عليها للخروج بعدة ذخائر هي: ذخيرة نصية من خلال تحويل ملفات html لنص دون أية تهيئة، وجمع المصادر في ٢٠٥٧ ملفاً تجميعياً للنصوص، وكذلك عمل على تحديد الأسماء والمعالم فيها، بالإضافة إلى تحديد أجزاء الكلام POS، وأخيراً التحليل الصرفي الكامل للذخيرة.

و- الذخيرة العربية الحديثة المعاصرة Arabic Modern Standard Corpus لأحمد عبدالعلي وآخرين (Abdalali et al 2005)، وهى ذخيرة بنيت خصيصاً لأغراض استرجاع المعلومات والمعالجة الطبيعية للمعلومات، وهي عبارة عن تجميع لمقالات الصحف في إحدى عشرة دولة عربية هي: مصر والكويت وعُمان والجزائر ولبنان والسعودية والمغرب والأردن وقطر وسوريا والعراق، وهي متاحة مجاناً للتحميل، وتحتوي على ١٠٢١٣٤(مائة واثنان ألفاً ومائة وأربعة وثلاثون) مقالاً، بها ١١٣(مائة وثلاثة عشر) مليون كلمة بحجم ٨٠٠ ميجابايت، وهي ذخيرة جيدة جداً تم بناؤها بشكل منظم، لكن يعيبها عدم الاتزان؛ فمثلاً جريدة الراية القطرية لها ٢٧٠ مقالاً فقط، وجريدة المغرب بها ١٧١٩٦ مقالاً.

ز- الذخيرة العربية مفتوحة المصدر Open Source Arabic Corpus (OSAC) (Saad, Ashour 2010)، وهي ذخيرة مفتوحة المصدر جمعها الدكتور معتز سعد من الصحف العربية اليومية، من ثلاثة مصادر رئيسة، الأول موقع هيئة الإذاعة البريطانية بالعربية، وتم جمع ٤٧٦٣ (أربعة آلاف وسبعمائة وثلاثة وستون) مقالاً من سبع فئات مختلفة بإجمالي ١,٨ مليون وثمانمائة كلمة؛ منها حوالي ١٠٦٧٣٣(مائة وستة آلاف وسبعمائة وثلاثة وثلاثون) كلمة فريدة من نوعها؛ وكان المصدر الثاني هو موقع السي إن إن العربي، وتم جمع ٥٠٧٠ (خمسة آلاف وسبعون) مقالاً، في عدة موضوعات بما يزيد عن المليوني كلمة، وتحديداً ٢.٢ مليون، منها١٤٤٤٦٠ (مائة وأربعة وأربعون ألفاً وأربعمائة وستون) كلمة فريدة، وكان المصدر الأخير هو عدة مواقع متنوعة، وتم جمع ٢٢٤٢٩ (اثنان وعشرون ألفاً وأربعمائة وتسعة وعشرون) مقالاً نصياً، في عشر فئات موضوعية بإجمالي ١٨ مليون كلمة، وحوالي ٤٤٩٦٠٠ (أربعمائة وتسع وأربعون ألفاً وستمائة) كلمة فريدة. وقد تم تجميع المصادر الثلاث في ذخيرة واحدة يبلغ حجمها تقريباً ٢٢ مليون كلمة، وتم تحويل الذخيرة بكاملها إلى نص بصيغة UTF-8.

ح- الذخيرة الدولية للغة العربية The International Corpus of Arabic (ICA) ، (Alansary, & Nagi, 2014; Alansary, Nagi, & Adly, 2007)، بدأ العمل فيها عام ٢٠٠٦ من قبل فريق بحثي في مكتبة الإسكندرية وجامعة الإسكندرية؛ بهدف مساندة الباحثين في مجال اللغة العربية، وقد كان مخططاً لها أن يكون بها ١٠٠ (مائة) مليون كلمة، وهي ذخيرة قياسية معاصرة للغة العربية Modern Standard Arabic، وقد تم اختيار عدد كبير من المصادر؛ لتكون ممثلة لقطاع كبير من اللغة العربية، وكيف تستخدم اللغة في شتى المجالات؛ وقد اعتمدت هذه الذخيرة على تصنيف ديوي العشري في تصنيف النصوص الموجودة بها مع بعض التعديلات الطفيفة، حيث تضم الذخيرة أحد عشر تقسيماً موضوعياً، هي العلوم الاستراتيجية، العلوم الاجتماعية، الرياضة، الديانات، الأدب، الإنسانيات، والعلوم الطبيعية، والعلوم التطبيقية، والفنون، والببليوجرافيا، بالإضافة إلى بعض المتفرقات، والتي تضم بدورها ٢٤ فئة موضوعية، وكانت أكبر الفئات الموضوعية تمثيلاً هي المتفرقات بنسبة ٢٥%، وأقلها تمثيلاً هي العلوم الطبيعية، والعلوم التطبيقية بنسبة ١% لكل منها.

وقد وصل عدد الكلمات بها حتى الآن إلى ٨٠ مليون كلمة تقريباً، وكان المصدر الأول هو الصحف مثل: الأهرام المصرية، الدستور الأردنية، الحياة اللبنانية، والمصدر

الثاني كان بعض المقالات المنشورة على الإنترنت كالمدونات والمنتديات، وكانت الكتب الإلكترونية هي المصدر الثالث، وأخيراً بعض البحوث الأكاديمية ورسائل الدكتوراه.

وكان إجمالي النصوص ٧٠٠٢٢(سبعون ألفاً واثنان وعشرون) مقالاً ومصدراً، وإجمالي ٧٩٥٦٩٣٨٤(تسع وسبعون مليوناً وخمسمائة وتسع وستون ألفاً وثلاثمائة وأربعة وثمانون) كلمة، منها ١٢٧٢٧٦٦(مليون ومائتان واثنان وسبعون ألفاً وسبعمائة وست وستون) كلمة فريدة على مدار ٢٢ سنة من ١٩٩٣ – ٢٠١٤، والذخيرة حجمها ١٠٠ ميجا متاحة مجاناً بالطلب.

ط- ذخيرة جامعة الملك سعود للنصوص العربية الكلاسيكية King Saud University Corpus of Classical Arabic (KSUCCA)، وهي ذخيرة نصية تحتوي على أكثر من خمسين مليون كلمة من العربية الفصحى، قامت بإنشائها مها سليمان الربيعة كجزء من بحثها في رسالة الدكتوراه؛ لبناء نموذج للدلالات المعجمية العربية واستخدامات أخرى، وهي متاحة مجاناً للباحثين للاستخدام الشخصي والأكاديمي وهي في الأساس، عبارة عن كتب مكتبة الشاملة للتراث الإسلامي، وهي مقسمة إلى ستة فروع موضوعية هي: الدين، والأدب، واللغويات، والعلوم، والسير والتراجم، وعلم الاجتماع، وتمتد الفترة الزمنية لهذه الذخيرة من القرن السابع وحتى القرن الحادي عشر (الربيعة ، ٢٠١٢؛ Alrabiah, Al-Salman, & Atwell, 2013؛ الربيعة والسلمان وأتويل، ٢٠١٤)، والذخيرة لم يكن فيها أي نوع من التهيئة كانت نصاً خاماً فقط، وهي متاحة مجاناً للتحميل من موقعها.

ي- الذخيرة العربية بجامعة الأردن (UJAC) University of Jordan Arabic Corpus (Hammo et al., 2013)، وهي ذخيرة حديثة وصغيرة نسبياً، قام بإنشائها أربعة باحثين بجامعة الأردن، ضمن مشروع لتطوير عدة أدوات ومصادر للمعالجة الآلية للغة العربية، وقام الباحثون بجمع عدة مقالات من خمس عشرة صحيفة عربية، ومصادر أخرى لنظم تسع عشرة دولة عربية وتضم الذخيرة ٦١٠٣٧ (واحد وستون ألفاً وسبع وثلاثون) ملفاً، وعدد ٧٥٢٢٩٤١ (سبعة ملايين وخمسمائة واثنان وعشرون ألفاً وتسعمائة وواحد وأربعون) كلمة، ومنها ٧٠٧٣٨٥(سبعمائة وسبعة آلاف وثلاثمائة وخمس وثمانون) كلمة فريدة بحجم ٤١,٥ ميجابايت. وكانت الذخيرة في شكل نصوص XML وبتشفير UTF-8.

ك- ذخيرة مدينة الملك عبدالعزيز للعلوم والتقنية King Abdulaziz City for Science and Technology (KACST)، أو المدونة العربية، أو المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية (المدونة العربية، ٢٠١٣؛ Al-Thubaity,

2014)، هي إحدى المشاريع الاستراتيجية لمبادرة الملك عبدالله للمحتوى العربي، الذي يهدف إلى بناء ذخيرة لغوية عربية تحوي سبعمائة مليون كلمة مما دوِن بالعربية ابتداءً من العصر الجاهلي وحتى العصر الحديث، ومن مختلف المناطق والبلدان، مع الأخذ في الاعتبار طبيعة وحجم النشاط الفكري لكل فترة، و تنوع أوعية النشر فيها (مخطوطات، صحف، كتب، مجلات، دوريات علمية،...إلخ) و السائد من المجالات العلمية والفكرية المختلفة (المعتقدات، علوم العربية، العلوم الطبيعية، الأدب،...إلخ). كما يشمل المشروع– بالإضافة إلى المادة اللغوية المصنفة– إنشاء موقع للمدونة على الإنترنت فضلاً عن أدوات للبحث والتحليل اللغوي والإحصائي تعزز الاستفادة من مواد الذخيرة.

وقد روعي في تصميم الذخيرة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية، عدة معايير خارجية لاختيار نصوص الذخيرة تعتمد على خمس ركائز أساسية هي البعد الزمني، والبعد الجغرافي، والوعاء المعلوماتي، و المجال المعرفي والتصنيف الموضوعي، ويمكن تناول هذه الأبعاد على النحو الآتي[1]:

١– البعد الزمني أو البعد التاريخي، ويمتد من عصر ما قبل الإسلام وحتى عصرنا هذا والذي أثر على الوعاء أو الصورة التي ظهر فيها النص وكذلك على حجم النصوص المطلوب جمعها لكل فترة زمنية.

٢– والبعد الجغرافي، ويقصد به المكان الذي صدر منه النص. ولأن الذخيرة تعنى باللغة العربية بمجموعها وتحاول أن تكون ممثلة للغة ومتغيراتها فإنه لم يتم تحديد بلد عربي بعينه لجمع النصوص بل أن المطلوب هو تنوع البلدان والكتّاب.

٣– الوعاء المعلوماتي، تم فيه اختيار ما يناسب كل فترة زمنية وما كان سائداً فيها من علوم ومعارف وما كان أكثر انتشاراً وتداولاً بين الناس من أوعية للنشر وما تكون لغته مناسبة ورصينة. فمثلاً تم استبعاد المنتديات الحوارية و صفحات الإنترنت الخاصة والتي يغلب عليها هذا الوقت اللهجات الدارجة ولا تتقيد باللغة العربية الفصحى. وتم اختيار عشرة أوعية للنشر وهي المخطوطات المحققة، والصحف، والمجلات، والكتب، والرسائل الجامعية، والدوريات المحكمة، والإصدارات الرسمية، ووكالات الأنباء، والإنترنت والمناهج الدراسية. وتم اختيار هذه الأوعية بناء على انتشارها وتأثيرها ورصانة لغتها.

[1]. نقلاً عن موقع الذخيرة "http://www.kacstac.org.sa/Pages/Default.aspx"

٤- المجال المعرفي، وهو يشير إلى المجالات التي تتدرج تحت كل وعاء من الأوعية المختارة، وهي تحدد مجال النص وسمته العامة. ففي الصحف على سبيل المثال، هناك مجالان رئيسان هما: الأخبار والمقالات. وفي المخطوطات المحققة، وفي الفترة التي كتبت فيها هذه المخطوطات كانت هناك مجالات عامة مثل: العقائد والفقه وأصوله وعلوم اللغة، وغيرها بما يناسب كل فترة. وينطبق هذا على كل وعاء من الأوعية. وهذا البعد يعطي الذخيرة فرصة أكبر لإيضاح الاختلافات بين كل مجال وآخر، وفترة وأخرى، كما يوضح أيضاً تنوعها وتمثيلها للغة بشكل أكبر.

٥- التصنيف الموضوعي، إذ يندرج تحت كل مجال من المجالات المخصصة للأوعية عدة مواضيع تفصل المجال، وتوضح تنوعاته الأدق، وتظهر الثراء والتنوع في كل مجال ووعاء، ففي وعاء الصحف، وتحت مجال الأخبار، هناك عدة مواضيع مثل: الأخبار الاجتماعية، الأخبار السياسية، الأخبار الرياضية، الأخبار الاقتصادية،...إلخ.

وتحتوي الذخيرة على ٧٣٢٧٨٠٥٠٩ (سبعمائة واثنان وثلاثون مليوناً وسبعمائة وثمانون ألفاً وخمسمائة وتسع) كلمات، منها ٧٤٦٤٣٩٦ (سبعة ملايين وأربعمائة وأربعة وستون ألفاً وثلاثمائة وست وتسعون) كلمة فريدة، وذلك لـحوالي ٨٦٩٨٠٠ (ثمانمائة وتسعة وستون ألفاً وثمانمائة) مصدر نصي ومقال.

ثانياً- الذخائر التجارية

أ- ذخيرة جريدة الحياة اللبنانية (Al-Hayat Arabic Corpus, 2001)، وهي ذخيرة نصية لمقالات جريدة الحياة اللبنانية، تم إنشاؤها ضمن مشروع بحثي مشترك بين جامعة إسكس University of Essex والجامعة المفتوحة؛ للاستخدام ضمن أبحاث استرجاع المعلومات وهندسة اللغة، وتضم ٤٢٥٩١ (اثنان وأربعون ألفاً وخمسمائة وواحد وتسعون) مقالاً بإجمالي ١٨٦٣٩٢٦٤ (ثمانمائة عشر مليون وستمائة وتسع وثلاثون ألفاً ومئتان وأربع وستون) كلمة، في ٢٦٨ ميجابايت في ٧ قطاعات موضوعية هي: عام، وسيارات، وحاسب آلي، وأخبار، واقتصاد، وعلوم، ورياضة، وقد اقتصرت الذخيرة على عام ١٩٩٨.

ب- ذخيرة جريدة النهار اللبنانية (An-Nahar Newspaper Text Corpus, 2001)، وهي ذخيرة نصية لجريدة النهار اللبنانية في الفترة من ١٩٩٥ وحتى ٢٠٠٠، وتضم الذخيرة خمساً وأربعين ألف مقال لكل سنة من السنوات الست، بإجمالي ٢٧٠ (مائتان وسبعون ألف)

مقال، وعدد أربع وعشرين مليون كلمة لكل سنة بإجمالي ١٤٤ (مائة وأربعة وأربعون) مليون كلمة. والذخيرة متاحة بصيغة html

ج- ذخيرة الأخبار العربية (الجزء الأول) Arabic Newswire- Part 1 (Graff, & Walker, 2001; Abu El-Khair, 2007)، وكانت هي البداية في إنشاء الذخائر اللغوية كبيرة الحجم، وظهرت في عام ٢٠٠١ بالتزامن مع بداية استخدام اللغة العربية في مؤتمر استرجاع المعلومات النصية (TREC) Text Retrieval Conference، وقد قام بتجميعها ديفيد جراف وكيفين ووكر بجامعة بنسيلفانيا الأمريكية، واتحاد البيانات اللغوية Linguistic Data Consortium(LDC). وتتكون الذخيرة من مقالات صحفية من وكالة الأخبار الفرنسية AFP (Agence France Press) في الفترة من مايو ١٩٩٤ وحتى ديسمبر ٢٠٠٠، وهي مصاغة في صيغة SGML وتشفير UTF-8 وتضم ٣٨٣٨٧٢ (ثلاثمائة وثلاث وثمانون ألفاً وثمانمائة واثنان وسبعون) مقالاً، وتضم ستاً وسبعين مليون كلمة وعدد ٦٦٦٠٩٤ (ستمائة وست وستون ألفاً وأربع وتسعون) كلمة فريدة، وربما تكون هذه الذخيرة هي الأكثر استخداماً من قبل الباحثين في مجال استرجاع المعلومات باللغة العربية؛ وذلك لأنها تحتوي على مجموعة من الأسئلة، ٧٥ سؤالاً تحديداً وإجاباتها وأحكام الصلة المرتبطة بها Relevance Judgment.

د- ذخيرة اللغة العربية المليونية بطبعاتها المختلفة Arabic Gigaword Corpus، وهي ذخيرة شاملة مجمعة للأخبار العربية من بعض الصحف ووكالات الأنباء العربية، وقد صدر منها خمسة إصدارات حتى الآن؛ الإصدارة الأولى (Graff, 2003)، تم فيها جمع ذخيرة الأخبار العربية السابق ذكرها، والتي تحتوي على مقالات من وكالة الأخبار الفرنسية، وذخيرة صحيفة النهار اللبنانية، وذخيرة صحيفة الحياة اللبنانية، بالإضافة إلى مقالات من مصدر رابع هو وكالة الأخبار الصينية بالعربية Xinhua. وقد صدرت الذخيرة في صيغة SGML بتشفير UTF-8 عام ٢٠٠٣ من LDC وتحتوي الذخيرة على ١٢٥٦٧١٩ (مليون ومائتان وست وخمسون ألفاً وسبعمائة وتسع عشرة) مقالاً، وإجمالي ٣٩١٦١٩ (ثلاثمائة ألف وواحد وتسعون وستمائة وتسع عشرة) كلمة فريدة.

وفي الإصدارة الثانية (Graff et al., 2006)، تم زيادة عدد من المقالات من كل مصدر، وإضافة مصدراً آخر وهو جريدة الأمة، وتحتوي الذخيرة على ١٥٩١٩٨٧ (مليون وخمسمائة وواحد وتسعون ألفاً وتسعمائة وسبعة وثمانون) مقالاً، وإجمالي ٤٨١٩٠٦ (أربعمائة وواحد وثمانون ألفاً وتسعمائة وست) كلمات. وبنفس الطريقة في الإصدارة الثالثة (Graff, 2007)،

تم زيادة عدد المقالات من كل مصدر، وإضافة مصدر سادس هو جريدة الصباح، وتحتوي الذخيرة على ١٩٩٤٧٣٥ (مليون وتسعمائة وأربع وتسعون ألفاً وسبعمائة وخمسة وثلاثون) مقالاً، وإجمالي ٥٧٦٧٩٩ (خمسمائة وست وسبعون ألفاً وسبعمائة وتسع وتسعون) كلمة فريدة.

وفي الإصدارة الرابعة (Parker et al., 2009)، تم زيادة جريدة الأهرام، وجريدة الشرق الأوسط، وجريدة القدس العربي، وأصبح عدد المقالات ٢٧١٦٩٩٥ (مائتا مليون وسبعمائة وستة عشر ألفاً وتسعمائة وخمسة وتسعون) مقالاً، وإجمالي ٨٤٨٤٦٩ (ثمانمائة وثمان وأربعون ألفاً وأربعمائة وتسع وستون) كلمة فريدة، أما الإصدارة الخامسة (Parker et al., 2011) والأخيرة، فقد تمت الزيادة على نفس المصادر السابقة بنفس الطريقة، وأصبحت الذخيرة تغطي تسع صحف، ووكالات أخبار من ست دول وهي: إنجلترا، وفرنسا، والصين، ومصر، وتونس، ولبنان، للفترة من ١٩٩٤ وحتى ٢٠١٠، وبلغ إجمالى المقالات الصحفية والأخبار ٣٣٤٦١٦٧ (ثلاثة ملايين وثلاثمائة وست وأربعون ألفاً ومائة وواحد وستون) مقالاً، وإجمالي ١٠٧٧٣٨٢٠٠٠ (مليار وسبع وسبعون مليون وثلاثمائة واثنان وثمانون ألف) كلمة، وجميع الإصدارات من هذه الذخيرة يمكن الحصول عليها بمقابل من اتحاد البيانات اللغوية LDC.

هـ- ذخيرة مشروع نملار أو شبكة المصادر اللغوية الأورومتوسطية NEMLAR (Network for Euro-Mediterranean LAnguage Resources)، وهي ذخيرة صغيرة نسبياً مكونة من نصف مليون كلمة مشروحة ومقسمة لثلاث عشرة فئة، وكانت هي إحدى نتائج مشروع الشبكة والذي نفذ بين عامي ٢٠٠٣ – ٢٠٠٥، وكان يهدف لإنتاج وتوفير ذخائر لغوية وأدوات تقنية أساسية للباحثين في اللغة العربية والتنبيه والتوعية بهذا المجال، ودعم التعاون والمشاركة بين الأطراف العربية بعضها البعض، والأطراف الأوروبية العاملة في هذا المجال (NEMLAR Project. 2010) ، والذخيرة متاحة في موقع جمعية المصادر اللغوية الأوروبية European Language Resources Association (ELRA) (NEMLAR Written Corpus, 2003).

وتلا ذلك المشروع مشروع تكميلي له بين ٢٠٠٨ – ٢٠١٠، وهو مشروع ميدار Medar (Mediterranean Arabic Language and Speech Technology)، أو المشروع المتوسطي لتقنيات اللغة العربية المكتوبة والمنطوقة، وقد كان أحد أهداف مشروع ميدار هو تحديث الذخيرة اللغوية السابقة، ولكن كان الاهتمام بشكل أكبر بمشروعات نملار الأخرى؛

لذا لم يتم الإضافة للذخيرة بشكل ملحوظ، وإنما كانت تعديلات عليها في التحليل الصرفي والتشكيل وتحديد أجزاء الكلام، بالإضافة إلى ذلك فقد تم إضافة بعض البرامج في مشروع ميدار التكميلي (ميدار، ٢٠١٠)، والذخيرة متاحة في موقع جمعية المصادر اللغوية الأوروبية (ELRA)، (MEDAR Evaluation Package, 2010).

منهج الدراسة وأدواتها

أولاً–المنهج:

تهدف الدراسة في الأساس إلى إنتاج وبناء ذخيرة لغوية للغة العربية، ولهذا الغرض تنتهج الدراسة المنهج البنائي Constructive Research Method، والمنهج البنائي هو أحد مناهج البحث المستحدثة في بداية التسعينات من القرن الماضي، وظهر في البداية كجزء من منهج دراسة الحالة (Lukka, 2000)، وهو مستخدم بكثرة في مجال إدارة الأعمال، وقد شاع استخدام المنهج فيما بعد في مجال العلوم الاجتماعية بشكل عام، وكذلك مجال الحاسب الآلي والبرمجيات، وعلوم الصيدلة.

ويستخدم هذا المنهج في الأساس لإنتاج وبناء معارف جديدة بشكل عملي تطبيقي، بناءً على المعرفة الموجودة مسبقاً، لكن بشكل جديد أو طريقة مبتكرة في محاولة لإيجاد الحلقة المفقودة، أو سد فجوة ما في جزء معين من موضوع البحث، ويكون ذلك عن طريق بناء تطبيقات مفصلة خصيصاً لسد تلك الفجوة، وقد تكون هذه التطبيقات نماذج عملية Models، مخططات Diagrams، خطط Plans، خرائط تنظيمية Organizational Charts، تصميمات نظم System Desgins، خوارزميات Algorithms، لغات اصطناعية Artificial Languages، أو برمجيات Software (Crnkovic, 2010). وعادة ما يكون نتاج البحوث البنائية هو تطوير لشيء معين موجود بالفعل وليس اكتشافاً جديداً.

وهذا ما يحاول الباحث القيام به في هذه الدراسة، إنتاج وبناء معارف جديدة بشكل عملي تطبيقي، وبالتحديد بناء ذخيرة لغوية قياسية معاصرة للغة العربية.

ثانياً– أدوات جمع البيانات:

في محاولة الباحث لجمع النصوص المنشورة في مواقع الصحف العربية، كان لابد من الاستعانة بأحد برامج استيراد المواقع، أو استخلاص النصوص Website copying or scraping وهي كثيرة ومتنوعة، وتم تجربة الكثير منها في هذه الدراسة للوصول لمعرفة

أفضلها، ثم استخدامه، وبدأ الباحث ببرنامج "دبليو جيت" wget[2] المستخدم من قبل اتحاد البيانات اللغوية لتجميع البيانات، لكن كان يعيبه البطء الشديد، وربما يعود السبب إلى انه مصمم في الاساس للعمل في بيئة يونكس UNIX، لذا تم استبعاده، وتم تجربة برنامج "أتش تي تي تراك" htttrack[3]، لكن أيضاً، تم استبعاده بسبب البطء الشديد، كما تمت محاولة برنامج "إنترنت دانلوود مانجر" Internet Download Manager[4]، وبرنامج "كايوتك ويب كوبي" cyotek webcopy[5]. لكن تم استبعادهما؛ لأنهما لا يتعمقان في الموقع بالشكل الكافي، ويتوقفا دون سبب، بالإضافة للبطء. واستقر الباحث على برنامجين للعمل هما: MetaProducts Offline Explorer Pro [6] و Visual Web Ripper[7]، ويمتازان بالسرعة، وإمكانية تنقية ما يتم استيراده من أي شيء لا قيمة له بالنسبة للبحث مثل ملفات الصور والفيديو، والملفات المساعدة في المواقع مثل: ملفات الجافاسكريبت Javascript، والأنماط الإنسيابية CSS (Cascading Style Sheets).

ثالثاً–حدود الدراسة:

الحدود الموضوعية:

شملت الذخيرة المقالات في كل الموضوعات دون التقيد بموضوع معين، أو استبعاد موضوع معين، وشملت على سبيل المثال: السياسة، الرياضة، الفن، العلوم والتكنولوجيا، واختلفت نسبة تمثيل كل موضوع وفقاً للمصادر.

الحدود النوعية:

اشتملت الذخيرة على المقالات الصحفية والتقارير الصحفية النصية فقط، دون الصور، والمقالات المسموعة (متاحة في المصري اليوم فقط)، والتقارير الصحفية المرئية.

الحدود اللغوية:

على الرغم من وجود مقالات صحفية بالإنجليزية في معظم مصادر الذخيرة فأنها اقتصرت على اللغة العربية فقط.

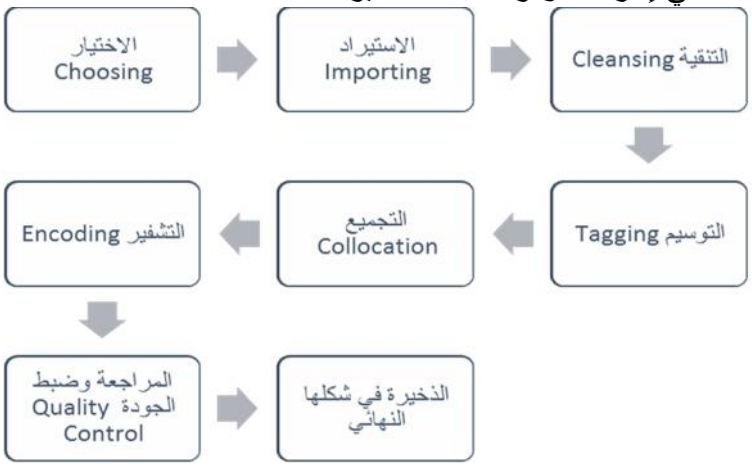
---

2. https://www.gnu.org/software/wget
3. https://www.httrack.com
4. https://www.internetdownloadmanager.com
5. http://www.cyotek.com/cyotek-webcopy
6. http://www.metaproducts.com/mp/offline_explorer_pro.htm
7. http://www.visualwebripper.com

النتائج

أولاً- إجراءات ومراحل البناء:

يوضح الشكل التالي إجراءات ومراحل بناء الذخيرة:



شكل رقم (١): إجراءات ومراحل بناء الذخيرة

- المرحلة الأولى- كانت مرحلة اختيار المصادر التي سوف يتم العمل عليها لبناء الذخيرة، وفي الفقرة التالية شرح مفصل لأسباب ومعايير الاختيار.
- المرحلة الثانية- هي مرحلة التجميع والاستيراد لصفحات كل مصدر، وتم استخدام البرنامجين السابق ذكرهما للعمل في هذه المرحلة، وهما؛ MetaProducts Offline Explorer Pro، وتم استخدام نسخة تجريبية من هذا البرنامج، حيث أنها كاملة الوظائف ولا فرق بينها وبين النسخة الكاملة سوى تاريخ الاستخدام إذ أنها تنتهي بمدة معينة؛ وVisual Web Ripper وتم استخدام نسخة كاملة خاصة بالباحث من هذا البرنامج، إذ أن النسخة التجريبية لا تستورد سوى ألف صفحة فقط.
- المرحلة الثالثة- هي المرحلة التي تم فيها تنقية Cleansing، واستبعاد كل تيجان لغة تهيئة المواقع سواء كانت html ،أو aspx ،أو غيرها، وكذلك استبعاد الإعلانات؛ ليكون النص صافياً دون أية علامات ترميز لا قيمة لها، وقد تم الاستعانة ببرنامج كتبه الباحث بلغة بيرل Perl للقيام بذلك.
- المرحلة الرابعة- وهي لصيقة بمرحلة التنقية، هي وضع الوسوم الخاصة بالذخيرة، وسيتم شرح هذه المرحلة بالتفصيل لاحقاً.

- المرحلة الخامسة– هي مرحلة التجميع، فيها تم تجميع كل المقالات المستوردة من كل مصدر معاً؛ ليكون لكل منهم ملف واحد كبير، وقد تم الاستعانة ببرنامج كتبه الباحث بلغة بيرل Perl للقيام بذلك.
- المرحلة السادسة– هي مرحلة التشفير encoding، والمقصود هنا، توحيد تهيئة النص، وسيتم شرح هذه المرحلة بالتفصيل لاحقاً.
- المرحلة السابعة والأخيرة، هي مرحلة ضبط الجودة، وفيها تم انتقاء عينات عشوائية من الذخيرة لمراجعتها، والتأكد من عدم وجود أي أخطاء فيها.

ثانياً– مصادر الذخيرة:

هناك الكثير من المصادر الصحفية التي يمكن استخدامها لإنتاج ذخيرة لغوية، لكن وقع اختيار الباحث على عشرة مصادر صحفية؛ لتكون هي الأساس الذي ستبنى به الذخيرة، وقد تم تجربة عدد كبير من مواقع الصحف في بداية الدراسة وذلك للاختيار من الصحف الموجودة، مع ملاحظة أن شهرة الموقع أو الصحيفة، أو حتى كثرة عدد القراء لم تكن معياراً للاختيار، وإنما كان هناك عدد من الأسباب والمعايير الفنية الأخري التي تم على أساسها اختيار هذه المصادر، وهي:

- أن لا يكون هناك تداخل بينها وبين أي من الذخائر السابقة، فعلى سبيل المثال تم استبعاد جريدة الأهرام المصرية رغم أنها أقدم الصحف العربية على الإنترنت، ويوجد لها أرشيف رقمي كبير، لأنها موجودة في الذخيرة المليونية لاتحاد البيانات اللغوية LDC.
- أن يكون المصدر موجوداً على الخط المباشر منذ فترة، وذلك لكي يكون به عدد كبير من المقالات؛ إذ إن حجم المقالات المنشورة وعددها يعد من الأسباب الرئيسة للاختيار، ويرتبط ذلك بوقت ظهور الصحيفة على الإنترنت، وربما كان ذلك أحد العوائق الكبرى للدراسة؛ لعدم التمكن من معرفة هذه المعلومة بسهولة، وعدم وجود مصدر واحد يوضح بداية وجود كل صحيفة على الإنترنت، رغم وجود الكثير من المواقع التي تحصر الصحف الإلكترونية العربية، لكن أياً منها لم يذكر بداية تلك الصحف على الإنترنت.
- التنوع في الدول، وتمثيل مختلف المناطق في العالم العربي.
- وجود النص في صورة ممكنة التحرير، والعمل عليها إن جاز التعبير.

أن يكون موقع الصحيفة يسمح لبرامج الاستيراد بالزحف داخله، واستيراد المقالات محلياً؛ حيث قام الباحث باستبعاد عدد من المصادر بسبب التأمين المشدد على الموقع، والذي يمنع برامج الاستيراد من العمل فيه.

وتجدر الإشارة إلى أنه تم جمع البيانات في الفترة من ديسمبر ٢٠١٣ وحتى يونيو ٢٠١٤، كما أنه تم إعادة استيراد مقالات جريدة المستقبل اللبنانية، ووكالة الأخبار اليمنية، بسبب خطأ بسيط في إعدادات برنامج الاستيراد أدى لعدم استيراد تاريخ المقال في كلا المصدرين، وتم اكتشاف ذلك أثناء عملية المراجعة وضبط الجودة.

ويشير الجدول رقم (١) إلى المصادر التي تم اختيارها لبناء الذخيرة، والفترة التي تم تغطيتها لكل مصدر منهم.

جدول رقم (١): مصادر الذخيرة والفترة الزمنية المغطاة لكل منها

المصدر	الدولة	من	إلى	الموقع
الاتحاد الإماراتية	الإمارات	يناير ٢٠٠٨	يونيو ٢٠١٤	http://www.alittihad.ae
الشروق الجزائرية	الجزائر	فبراير ٢٠٠٨	مايو ٢٠١٤	http://www.echoroukonline.com/ara
الرياض السعودية	السعودية	أكتوبر ٢٠٠٠	ديسمبر ٢٠١٣	http://www.alriyadh.com
اليوم السعودية	السعودية	يوليو ٢٠٠٢	ديسمبر ٢٠١٣	http://www.alyaum.com
تشرين السورية	سوريا	يناير ٢٠٠٤	مايو ٢٠١٤	http://www.tishreen.news.sy
القبس	الكويت	يناير ٢٠٠٦	أبريل ٢٠١٤	http://www.alqabas.com.kw
المستقبل اللبنانية	لبنان	سبتمبر ٢٠٠٣	أبريل ٢٠١٤	http://www.almustaqbal.com
المصري اليوم	مصر	ديسمبر ٢٠٠٥	يناير ٢٠١٤	http://www.almasryalyoum.com
اليوم السابع	مصر	يناير ٢٠٠٨	مايو ٢٠١٣	http://www.youm7.com
وكالة أنباء سبأ اليمنية	اليمن	ديسمبر ٢٠٠٩	مايو ٢٠١٤	http://www.sabanews.net

ويوضح الجدول مصادر الذخيرة، ودولة كل مصدر، والفترة الزمنية التي تم تغطيتها، والموقع الإلكتروني. ونرى أنه تم تجميع الذخيرة من عشرة مصادر هي: الاتحاد الإماراتية، والشروق الجزائرية، والرياض السعودية، واليوم السعودية، وتشرين السورية، والقبس الكويتية، والمستقبل اللبنانية، والمصري اليوم، واليوم السابع المصرية، ووكالة أنباء سبأ اليمنية؛ وكانت هذه المصادر موزعة على ثمان دول هي، الإمارات، والجزائر، والسعودية، وسوريا، والكويت، ولبنان، ومصر، واليمن. وقد مثلت مصر والسعودية بصحيفتين لكل منهما؛ وذلك لكونهما أكثر الدول التي يوجد بها صحف إلكترونية، وأقدمها في مجال الصحافة الإلكترونية

ويشير الجدول أيضاً إلى تفاوت فترة التغطية من مصدر لآخر، ويرجع ذلك لسببين؛ أولهما، فترة تجميع الذخيرة السابق ذكرها، وثانيهما، موقع المصدر نفسه، والذي أحياناً ما يمنع استيراد الأخبار الحديثة، لكن يسمح باستيراد أرشيف الأخبار مثل المصري اليوم، واليوم السعودية. وقد كانت أكثر الصحف التي تم الاستيراد منها هي: الرياض السعودية بأربع عشرة سنة تقريباً، والمستقبل اللبنانية، والقبس الكويتية باثنتي عشرة سنة، وكانت أقلهم من حيث عدد السنوات، وكالة أنباء سبأ اليمنية واليوم السابع المصرية بست سنوات فقط؛ وذلك بسبب حداثة الموقع لكل منهما.

ثالثاً– الميتاداتا:

كما سبق وأشرنا، توسيم الذخيرة Tagging، أو وضع واصفات البيانات Metadata، أو التهيئة Formatting، هي إحدى المراحل المهمة في بناء الذخيرة؛ وذلك لكى يتم التعامل مع الذخيرة من قبل البرامج المختلفة بشكل صحيح، حيث كانت بعض الذخائر السابقة، وخاصة المجانية، عبارة عن نصوص فقط، دون أي نوع من التهيئة، وهو ما كان يسبب بعض الأخطاء عند التعامل مع برامج تطلب تهيئة معينة، ولذلك كان الباحثون يعيدون تهيئة الذخيرة وهو ما كان يشكل عبئاً إضافياً عليهم من حيث الوقت والجهد، وتجدر الإشارة إلى أن هذه المشكلة غالباً ما يواجهها الباحثون عند استخدام الذخائر المجانية، إلا أنها غير موجودة في الذخائر التجارية المدفوعة الثمن، لانها جميعها تحتوي على واصفات بيانات.

وقد تم توسيم كل مقال في الذخيرة بطريقتين حتى يمكن العمل عليها بأكثر من برنامج وفقاً لما يرغبه الباحثون، الطريقة الأولى من التهيئة تم فيها استخدام لغة SGML (Standard Generalized Markup Language) للترميز وذلك وفقا للطريقة المعتمدة في مؤتمر استرجاع المعلومات النصية TREC، أما الطريقة الثانية من التهيئة فقد تم فيها استخدام لغة XML (Extensible Markup Language)، وهي نفس طريقة التوسيم المستخدمة في ذخائر اتحاد البيانات اللغوية LDC، ولعل الفروق بسيطة لكن لكل منها استخداماته.

ويوضح جدول رقم(٢) الوسوم المستخدمة في بناء الذخيرة بكلتا الطريقتين، مع وصف بسيط لكل وسم منهم.

جدول رقم (٢): الوسوم المستخدمة في بناء الذخيرة

وصف الوسوم	SGML (TREC Format)	XML
وسم المستوى العلوي، بداية المقال	<DOC>	<Alittihad>
رقم مسلسل للمقال	<DOCNO> </DOCNO>	<ID> </ID>
موقع المقال على الإنترنت	غير موجود في هذه الطريقة	<URL> </URL>
عنوان المقال	<HEADLINE></HEADLINE>	<headline></headline>
تاريخ نشر المقال	<dateline></dateline>	<dateline></dateline>
نص المقال	<TEXT></TEXT>	<text></text>
وسم المستوى العلوي، نهاية المقال	</DOC>	</Alittihad>

ويدلنا الجدول رقم(٣)، على الرموز المستخدمة لمصادر الذخيرة، حيث تم تحديد رمز للدلالة على كل مصدر من مصادر الذخيرة؛ ليكون هو وسم المستوى العلوي في لغة XML، مأخوذ من اسم الصحيفة باللغة الإنجليزية بشكل مباشر دون تعديل. كما أنه قد عين لكل صحيفة أو مصدر إختصار من ثلاثة حروف باللغة الإنجليزية لكي يستخدم في حقل الرقم المعرف لكل مقال في كلتا اللغتين XML & SGML، على أن يكتب الرقم المعرف بهذا التسلسل، رمز المصدر متبوعاً برمز للغة العربية، ثم رقم مسلسل، <ID> RYD_ARB_0000001 </ID>.

جدول رقم (٣): الرموز المستخدمة لمصادر الذخيرة

المصدر	الاختصار	المصدر بالإنجليزية
الاتحاد الإماراتية	ETD	Alittihad
الشروق الجزائرية	SHG	Echoroukonline
الرياض السعودية	RYD	Ryiadh
اليوم السعودية	YMS	SaudiYoum
تشرين السورية	TRN	Techreen
القبس	QBS	Alqabas
المستقبل اللبنانية	MTL	Almustaqbal
المصري اليوم	MSY	Almasryalyoum
اليوم السابع	YM7	Youm7
وكالة أنباء سبأ اليمنية	SBN	Sabanews

رابعاً– التشفير:

تعد مرحلة التشفير encoding، من المراحل المهمة في بناء الذخائر اللغوية، إذ يجب توحيد تهيئة النص في كل أجزاء الذخيرة، وذلك لأنه ليس من الضروري أن تكون كل المواقع تستخدم نفس نوع التشفير.

وقد تم تشفير ملفات الذخيرة بنوعين من التشفير لخدمة الباحثين؛ الأول تشفير ويندوز للغة العربية windows-cp1256[8]؛ وذلك لمن يريد استخدام الذخيرة مع برنامج Lemur Toolkit[9] لاسترجاع المعلومات؛ لأنه التشفير المعتمد في هذا البرنامج للتعامل مع اللغة العربية.

أما النوع الثاني من التشفير فهو تشفير الكود الموحد UTF-8 [10]؛ وذلك لمن يريد استخدام الذخيرة مع مجموعة برامج جامعة ستانفورد الأمريكية؛ لمعالجة اللغة الطبيعية [11]، وكذلك برنامج Python NLP Toolkit [12]، ولعل هذه البرامج هي الأشهر في معالجة اللغة الطبيعية واسترجاع المعلومات، وهناك الكثير من البرامج التي لا مجال لذكرها، لكن وجود هذين النوعين من التشفير لذخيرة واحدة، بالإضافة لطريقتين من التوسيم، سوف يكون مساعداً للباحثين في استرجاع المعلومات باللغة العربية بشكل كبير.

خامساً– إحصاءات الذخيرة:

إن الذخائر اللغوية يجب أن يكون حجمها كبيراً، وأن الذخيرة كلما زاد حجمها كان ذلك أفضل، خاصة للباحثين؛ فالذخيرة اللغوية الكبيرة تحتوي على الكثير من الاستخدامات العملية أو الفعلية للغة، وهو ماتحققه هذه الذخيرة. ويدلنا الجدول رقم (٤)، والذي يضم إحصائية عامة للذخيرة، على ذلك، فقد بلغ عدد المصادر عشرة مصادر من ثمانِ دول على مدار أربع عشرة سنة، وبلغ إجمالي المقالات التي تم تجميعها ما يزيد على الخمس ملايين ومئتي ألف مقال، تحتوي على أكثر من المليار ونصف المليار كلمة، منها ثلاثة ملايين وثلاثمائة ألف كلمة فريدة غير مكررة، كما بلغ حجم الذخيرة المادي حوالي عشرة جيجابايت بتشفير ويندوز للغة العربية، وستة عشر جيجابايت بتشفير الكود الموحد.

---

8. https://msdn.microsoft.com/en-us/goglobal/cc305149.aspx
9. www.lemurproject.org
10 http://unicode.org/resources/utf8.html
11 http://nlp.stanford.edu
12 www.nltk.org

جدول رقم (٤): إحصائية عامة للذخيرة

عدد المصادر	تسع صحف، ووكالة أنباء
عدد الدول المغطاة	ثمان دول
عدد النصوص	٥٢٢٢٩٧٣ مقالاً
حجم القاعدة	٩.٦٩ جيجابايت / ١٦.٧ جيجابايت
عدد السنوات المغطاة	١٤ سنة
إجمالي عدد الكلمات	١٥٢٥٧٢٢٢٥٢ كلمة
عدد الكلمات الفريدة	٣٣٠٣٧٢٣ كلمة

ولكي يتضح الأمر بشكل أكبر، نورد في الجدول رقم (٥)، إحصاءً مفصلاً لما تم تجميعه من كل مصدر من مصادر الذخيرة العشر، وذلك وفقاً لعدد المقالات التي تم استيرادها من كل مصدر، ونسبتها لإجمالي المقالات بالذخيرة، كذلك يوضح الجدول إجمالي عدد الكلمات لكل مصدر، ونسبتها لإجمالي الكلمات بالذخيرة، و يضم الجدول أيضـــــــاً،

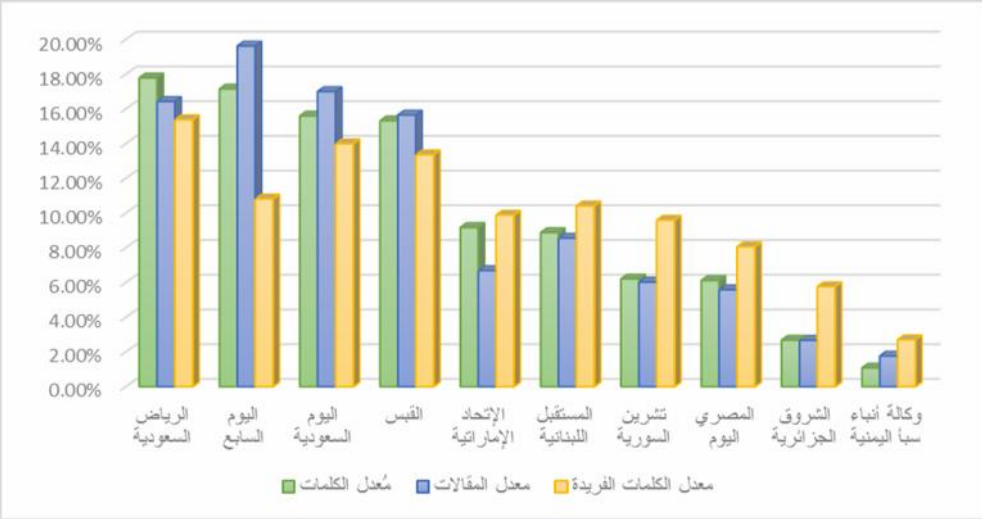
جدول رقم (٥): عدد المقالات والكلمات وفقاً لكل مصدر في الذخيرة

المصدر	المقالات		إجمالي عدد الكلمات		الكلمات الفريدة	
	عدد	نسبة	عدد	نسبة	عدد	نسبة
الرياض السعودية	**858188**	**16.43%**	**271353697**	**17.79%**	**1451320**	**15.39%**
اليوم السابع	**1025027**	**19.63%**	**261700304**	**17.15%**	**1020444**	**10.82%**
اليوم السعودية	**888068**	**17.00%**	**237914494**	**15.59%**	**1319996**	**13.99%**
القبس	**817274**	**15.65%**	**233741575**	**15.32%**	**1260511**	**13.36%**
الاتحاد الإماراتية	**349342**	**6.69%**	**139962699**	**9.17%**	**932628**	**9.89%**
المستقبل اللبنانية	**446873**	**8.56%**	**135446906**	**8.88%**	**982765**	**10.42%**
تشرين السورية	**314597**	**6.02%**	**94695378**	**6.21%**	**905169**	**9.60%**
المصري اليوم	**291723**	**5.59%**	**93398135**	**6.12%**	**760511**	**8.06%**
الشروق الجزائرية	**139732**	**2.68%**	**40978911**	**2.69%**	**543799**	**5.77%**
وكالة أنباء سبأ اليمنية	**92149**	**1.76%**	**16530153**	**1.08%**	**255098**	**2.70%**
الإجمالي	**5222973**	**100.00%**	**1525722252**	**100.00%**	**3303723**	

إحصاء بعدد الكلمات الفريدة وفقاً لكل مصدر. وقد تم ترتيب الجدول بناءً على عدد الكلمات؛ لأنها هي التي تحدد قيمة كل مصدر بالنسبة للذخيرة، ويجب التنويه إلى أن خانة الإجمالي في عمود "عدد الكلمات الفريدة" لا تساوي الجمع المباشر للقيم الموجودة في العمود؛ لأنه قد تم استبعاد الكلمات المكررة المشتركة بين المصادر.

ونلاحظ هنا، أن أكثر المصادر إسهاماً في الذخيرة كانت هي صحيفة الرياض السعودية بما يزيد عن مئتين وواحد وسبعين مليون كلمة، يليها اليوم السابع بما يزيد عن مئتين وواحد وستين مليون كلمة رغم أنها حديثة الظهور، ثم اليوم السعودية بما يزيد عن مئتين وسبع وثلاثين مليون كلمة، وكانت أقل المصادر إسهاماً في الذخيرة هي وكالة أنباء سبأ اليمنية بإجمالي ست عشرة مليون كلمة تقريباً، وصحيفة الشروق الجزائرية بأربعين مليون كلمة تقريباً.

شكل رقم (٢): نسبة المقالات والكلمات، وفقاً لكل مصدر في الذخيرة



ويوضح لنا الشكل رقم (٢) نسبة مساهمة كل مصدر في الذخيرة من حيث نسبة المقالات (العمود الأوسط)، ونسبة الكلمات ككل (العمود الأيسر)، ونسبة الكلمات الفريدة (العمود الأيمن)، ويعطينا هذا الشكل دلالة أو مؤشراً سريعاً عن حجم المقالات وعددها ونسبة التجديد في الكتابة؛ فعلى سبيل المثال تسهم صحيفة اليوم السابع بأكبر عدد من المقالات في الذخيرة، لكن عند النظر لنسبة الكلمات القليلة فيها، فإن ذلك يدل على صغر حجم مقالاتها بالمقارنة بالقبس الكويتية مثلاً، أو الرياض السعودية. ومثال آخر من الشكل؛

فقد يرى البعض معدل الكلمات الفريدة في كل من الشروق الجزائرية، وتشرين السورية مثلاً، مؤشراً إيجابياً يدل على فصاحة الكتاب فيهما، وقد يراه البعض مؤشراً سلبياً يدل على وجود الكثير من الأخطاء؛ الأمر الذي أدى لأن يحسب الحاسب الآلي الكلمات الخطأ على أنها كلمات فريدة لم تتكرر.

الخلاصة

حاول الباحث في هذه الدراسة بناء ذخيرة لغوية معاصرة للغة العربية تتوفر فيها العديد من السمات، لعل أهمها أنها تتناول الاستعمال الحقيقي للغة في العصر الحديث من خلال تجميع عدد كبير من المقالات الصحفية، وقد قام الباحث بإنتاج أربع نسخ من الذخيرة لكي تستعمل مباشرة دون تجهيز ودون أن يحتاج أيٌ من الباحثين الذين سيعملون عليها أن يعملوا على تغيير التشفير أو لغة التوسيم، وهذه النسخ الأربعة هي نتاج تبادل نوعي التشفير والتوسيم المستخدمين وهى:

- XML & CP-1256
- XML & UTF-8
- SGML & CP-1256
- SGML & UTF-8

وفيما يأتي تقويم للذخيرة التي انتجناها، بناء على العوامل التي يجب مراعاتها عند بناء أي ذخيرة لغوية، لكي تكون الذخيرة نموذجاً ممثلاً للغة واستخداماتها، والتي تم ذكرها في بداية الدراسة؛ مع شرح مبسط لواقع الدراسة الحالية وفقاً لكل عامل من تلك العوامل:

١- **الحجم:** حجم الذخيرة كبير نسبياً بالمقارنة مع غيرها من الذخائر؛ فذخيرة مدينة الملك عبد العزيز للعلوم والتكنولوجيا والتي يقول منشؤها أنها أكبر ذخيرة عربية- وهو كلام يجافي الحقيقة- بها ٧٠٠ مليون كلمة فقط من حوالي المليون مقال أو نص، أضف لذلك أنه عمل مؤسسي لم يقم به فرد واحد بل فريق عمل من المدينة، بل وكما ذكر القائم على الذخيرة (Al-Thubaity, 2014) فقد عهد لمتعهد خارجي بتجميع ٢٥% من الذخيرة، كما أن أكبر ذخيرة عربية موجودة بالفعل وهي ذخيرة اللغة العربية المليونية في إصدارتها الخامسة بها قد بلغ عدد المقالات بها ثلاثة ملايين وثلاثمائة ألف مقال تقريباً، وإجمالي مليار وسبع وسبعين مليون كلمة، هذا ويبلغ ثمن هذه الذخيرة ستة آلاف دولار أمريكي، في حين يزيد عدد حجم المقالات بالذخيرة الحالية عن الخمس ملايين مقال، وبها أكثر من مليار ونصف المليار كلمة.

جدول رقم (٦): مقارنة بين الذخيرة الحالية والذخائر السابقة

م	اسم الذخيرة	عدد النصوص	عدد الكلمات	عدد الكلمات الفريدة	الترخيص	نوعية البيانات
١	الذخيرة الحالية	٥٢٢٢٩٧٣	١٥٢٥٧٢٢٢٥٢	٣٣٠٣٧٢٣	مجاني	مقالات صحفية
٢	ذخيرة اللغة العربية المليونية، ط٥	٣٣٤٦١٦٧	١٠٧٧٣٨٢٠٠٠	غير متاح	٦٠٠٠ دولار	مقالات صحفية
٣	ذخيرة اللغة العربية المليونية، ط٤	٢٧١٦٩٩٥	غير متاح	٨٤٨٤٦٩	٥٠٠٠ دولار	مقالات صحفية
٤	ذخيرة اللغة العربية المليونية، ط	١٩٩٤٧٣٥	غير متاح	٥٧٦٧٩٩	٤٠٠٠ دولار	مقالات صحفية
٥	ذخيرة اللغة العربية المليونية، ط٢	١٥٩١٩٨٧	غير متاح	٤٨١٩٠٦	٣٠٠٠ دولار	مقالات صحفية
٦	ذخيرة اللغة العربية المليونية، ط ١	١٢٥٦٧١٩	غير متاح	٣٩١٦١٩	٣٠٠٠ دولار	مقالات صحفية
٧	ذخيرة مدينة الملك عبدالعزيز للعلوم والتقنية	٨٦٩٨٠٠	٧٣٢٧٨٠٥٠٩	٧٤٦٤٣٩٦	مجاني	متعدد
٨	ذخيرة جريدة النهار اللبنانية	٢٧٠٠٠٠	١٤٤ مليون	غير متاح	٥٠٤ يورو	مقالات صحفية
٩	الذخيرة العربية الحديثة المعاصرة	١٠٢١٣٤	١١٣ مليون	غير متاح	مجاني	مقالات صحفية
١٠	الذخيرة الدولية للغة العربية	٧٠٠٢٢	٧٩٥٦٩٣٨٤	١٢٧٢٧٦٦	مجاني	مقالات صحفية، كتب، رسائل
١١	ذخيرة الأخبار العربية (الجزء الأول)	٣٨٣٨٧٢	٧٦ مليون	٦٦٦٠٩٤	١٢٠٠ دولار	مقالات صحفية
١٢	ذخيرة جامعة الملك سعود للنصوص العربية الكلاسيكية	غير متاح	٥٠ مليون	غير متاح	مجاني	كتب كلاسيكية
١٣	الذخيرة العربية مفتوحة المصدر	٣٢٢٦٢	٢٢ مليون	غير متاح	مجاني	متعدد
١٤	ذخيرة جريدة الحياة اللبنانية	٤٢٥٩١	١٨٦٣٩٢٦٤	غير متاح	٧٢٠ يورو	مقالات صحفية
١٥	ذخيرة جريدة أخبار الخليج (٢)	غير متاح	١٠ مليون	غير متاح	مجاني	مقالات صحفية
١٦	الذخيرة العربية بجامعة الأردن	٦١٠٣٧	٧٥٢٢٩٤١	٧٠٧٣٨٥	مجاني	مقالات صحفية
١٧	ذخيرة جريدة أخبار الخليج (١)	غير متاح	٣ مليون	غير متاح	مجاني	مقالات صحفية
١٨	ذخيرة اللغة العربية المعاصرة	٤١٦ ملف	٨٤٢٦٨٤	غير متاح	مجاني	مقالات صحفية، مواقع انترنت، ايميلات
١٩	ذخيرة مشروع نملار	غير متاح	٥٠٠٠٠٠	غير متاح	٣٠٠ يورو	متعدد
٢٠	ذخيرة جريدة الراية القطرية	١٨٧	٢١٩٩٧٨	٣٠٠٩٦	مجاني	مقالات صحفية
٢١	ذخيرة ملخصات المؤتمر السعودي لعلوم الحاسب الآلي	٢٤٢	٤٦٩٦٨	غير متاح	مجاني	مستخلصات ابحاث
٢٢	مشروع الذخيرة العربية	٤٠٠	غير متاح	غير متاح	مجاني	كتب

ويوضح جدول رقم (٦)، الفرق بين الذخيرة الحالية والذخائر السابقة من حيث، عدد النصوص أو المقالات، وأجمالي عدد الكلمات، وعدد الكلمات الفريدة، و نوعية الترخيص، وأخيراً نوعية البيانات التي تتكون منها الذخيرة، ونلاحظ هنا تميز هذه الذخيرة عن مثيلاتها السابقات، من حيث كمية المقالات وأجمالي عدد الكلمات، وعدد الكلمات الفريدة، بالإضافة إلى كونها ذخيرة مجانية.

٢- ***تحديد الغرض:*** الذخيرة في حد ذاتها لا فائدة منها دون استخدامها في أحد البحوث العلمية، والذخيرة الحالية التي نحن بصدد بنائها هي ذخيرة لغوية عربية قياسية حديثة جمعت لأغراض البحث العلمي بشكل خاص في استرجاع المعلومات والمعالجة الطبيعية للمعلومات، ويجري العمل عليها حالياً، في هذا الإطار من قبل الباحث، بالإضافة لذلك سوف يتيحها الباحث بالطلب لكل من يرغب في الحصول عليها.

٣- ***تنوع الموضوعات:*** لا تقتصر الذخيرة على موضوع واحد، إذ تضم في جنباتها الكثير من الموضوعات كالسياسة، والأدب، والفن، والتكنولوجيا، والرياضة، والاقتصاد، والثقافة، وغيرها، عكس بعض الذخائر السابقة التي اقتصرت على موضوع واحد.

٤- ***تمثيل اللغة:*** تحقق الذخيرة التوازن بين الدول المختلفة التي تتحدث اللغة العربية؛ إذ انها تحتوي على جزء من أكثر من دولة، وبشكل أدق ثمان دول بالتحديد بها أكثر من نصف المتحدثين باللغة العربية.

٥- ***التوازن بين المصادر:*** لم يطغ مصدر على الآخر في الذخيرة، وقد كانت جميع المصادر متوازنة التمثيل، ولكى نكون موضوعيين فإنه ربما كانت قلة المقالات التي أسهمت به وكالة الأنباء اليمنية سبأ تحتاج إلى إضافة مقالات أخرى من اليمن.

التوصيات

لقد تطور العمل على اللغة العربية واستخداماتها في الحاسب الآلي في الفترة الأخيرة تطوراً كبيراً، وبشكل خاص في مجال معالجة اللغة الطبيعية، ولعل أبرز مناحي هذا التطور هو بناء الذخائر اللغوية، إذ أصبحت الذخائر اللغوية أحد الوسائل الرئيسية لتحليل استخدام اللغة، ودراستها والعمل عليها لتطوير نظم استرجاع المعلومات التي تستخدمها.

والذخيرة الحالية هي محاولة متواضعة من قبل الباحث لإثراء البحث في مجال المعالجة الطبيعية للمعلومات ، واسترجاع المعلومات بشكل عام، والذخائر اللغوية بشكل خاص، فقد تم تجميعها خصيصاً لأغراض البحث العلمي، ويمكن ان تستخدم في أكثر من مجال وبأكثر من طريقة، منها:

- يوصي الباحث بأن تستمر عملية تجميع المقالات من دول أخرى، ومصادر أخرى لزيادة حجم الذخيرة وتنوعها.
- يمكن استخدام الذخيرة في عدد من المجالات مثل: علم اللغويات، واللغويات الحاسوبية، بل ويمكن استخدامها في مجالات أخرى مثل: الدراسات اللغوية والأدبية.
- أيضاً، يمكن استخدام الذخيرة في معرفة التراكيب اللغوية في فترة تجميعها، ومقارنة استخدام اللغة بين أكثر من دولة من الدول التي تضمها الذخيرة، أو استخدام اللهجات المختلفة في الكتابة الرسمية.
- كما يمكن استخدامها في بناء معاجم أحادي اللغة للغة العربية، من واقع استخدام اللغة، وقياس مدى شيوع الكلمات أو عدم شيوعها والدقة في المعاني المرادة منها وتحليلها.
- العمل على إنشاء بنك معلومات شجري متكامل للغة العربية باستخدام هذه الذخيرة وإضافة وسوم أجزاء الكلام لها.
- العمل على بناء محلل اشتقاقي بناءً على إحصاءات المتتابعات الحرفية في الكلمات العربية، وهو ما قد يكون له بالغ الأثر في تحسين عملية إسترجاع المعلومات باللغة العربية.

المراجع

الأمم المتحدة. (٢٠١٥). الأسئلة التي يتردد طرحها، اللغات الرسمية للأمم المتحدة .تم الاسترجاع في: ٢٠١٥/٢/٢٥، من http://www.un.org/ar/hq/dgacm/faqs.shtml:

الأمم المتحدة. (١٩٧٣)قرار الأمم المتحدة رقم ٣١٩٠(د-٢٨) .تم الاسترجاع في: ٢٠١٥/٢/٢٥، من:

http://daccess-dds-ny.un.org/doc/RESOLUTION/GEN/NR0/279/60/IMG/NR027960.pdf?OpenElement

الربيعة ، مها سليمان. (٢٠١٢). الذخيرة النصية الفصحى لجامعة الملك سعود .تم الاسترجاع في: ٢٠١٥/٢/٢٥، من: http://ksucorpus.ksu.edu.sa/ar

الربيعة ، مها سليمان و عبدالملك سلمان السلمان وإريك أتويل. (٢٠١٤). الذخيرة النصية الفصحى حجر الأساس لدراسة معاني كلمات القرآن الكريم في ضوء نماذج الدلالة المتوزعة ورقة عمل مقدمة إلى: مؤتمر جامعة طيبة الدولي في توظيف تقنية المعلومات لخدمة القرآن الكريم

وعلومه، المدينة المنورة، المملكة العربية السعودية .

صالح، عبدالرحمن الحاج. (١٩٩٩). مشروع الذخيرة اللغوية. مجلة اللسان العربي, ٤٨. **تم الاسترجاع في:** ٢٠١٥/٢/٢٥، **من:** http://www.atinternational.org/forums/showthread.php?t=288

صالح، عبدالرحمن الحاج. (٢٠١٤). هذه أبعاد مشروع الذخيرة العربية.. وهذا موقع الجزائر منه . أخبار اليوم الجزائرية، ٢٠١٤/٦/٢٥. **تم الاسترجاع في: ٢٠١٥/٢/٢٥، من:** http://www.akhbarelyoum.dz/ar/200243/200256/109357

صالح، محمود إسماعيل. (٢٠١٤) .لسانيات المدونات اللغوية : **مقدمة للقارئ العربي.** تم الاسترجاع في: *25/2/2015*، من: http://dr-mahmoud-ismail-saleh.blogspot.com/2014/04/blog-post_5.html

المدونة العربية: (المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية) .(2013) **تم الاسترجاع في: ٢٠١٥/٢/٢٥، من:** http://www.kacstac.org.sa/Pages/Default.aspx

مشروع الذخيرة العربية. (٢٠٠٨). موسوعة المعرفة تم الاسترجاع في: ٢٠١٥/٢/٢٥، من: http://www.marefa.org/index.php/مشروع_الذخيرة_العربية

معجم المعاني الجامع. (٢٠١٠). ذخيرة **في معجم المعاني الجامع - معجم عربي عربي .تم الاسترجاع في: ٢٠١٥/٢/٢٥، من:** http://www.almaany.com/ar/dict/ar-ar/ذخيرة

ميدار(٢٠١٠). المشروع المتوسطي لتقنيات اللغة العربية المكتوبة والمنطوقة .تم الاسترجاع في: ٢٠١٥/٢/٢٥، من:
http://www.medar.info/Breif_Arabic/MEDAR_Arabic-brief-June2009.pdf

Abbas, M., & Smaili, K. (2005). *Comparison of topic identification methods for arabic language*. Paper presented at the Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP.

Abbas, M., Smaïli, K., & Berkani, D. (2011). Evaluation of Topic Identification Methods on Arabic Corpora. *JDIM, 9*(5), 185-192.

Abdelali, A., Cowie, J., & Soliman, H. (2005). *Building a modern standard Arabic corpus.* Paper presented at the workshop on computational modeling of

lexical acquisition, The split meeting. Croatia, 25-28 July.

Abu El-Khair, I. (2007). Arabic information retrieval. *Annual review of information science and technology, 41*(1), 505-533.

Abu Salem, H. (1992). *A microcomputer based Arabic bibliographic information retrieval system with relational thesauri (Arabic-IRS). Ph. D.* Dissertation, Illinois Institute of Technology.

Alansary, S., & Nagi, M. (2014). The International Corpus of Arabic: Compilation, Analysis and Evaluation. *ANLP 2014*, 8.

Alansary, S., Nagi, M., & Adly, N. (2007). *Building an International Corpus of Arabic (ICA): progress of compilation stage.* Paper presented at the 7th International Conference on Language Engineering, Cairo, Egypt.

Al-Hayat Arabic Corpus. (2001). *European Language Resources Association, ELRA Catalog number ELRA-W0030 Retrieved 25/2/2015, from: http://catalog.elra.info/product_info.php?products_id=632*

Alrabiah, M., Al-Salman, A., & Atwell, E. (2013). *The design and construction of the 50 million words KSUCCA.* Paper presented at the Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics.

Al-Sulaiti, L. (2004). *Designing and developing a corpus of contemporary Arabic.* (Master), University of Leeds.

Al-Sulaiti, L., & Atwell, E. (2005). *Extending the corpus of contemporary Arabic.* Paper presented at the Proceedings of the CL'2005 Corpus Linguistics Conference.

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, *11*(2), 135-171.

Al-Thubaity, A. O. (2014). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 1-31.

An-Nahar Newspaper Text Corpus. (2001). *European Language Resources Association, ELRA Catalog number ELRA-W0027 Retrieved 25/2/2015,*

from: http://catalog.elra.info/product_info.php?products_id=767

Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A., & Suchomel, V. (2014). arTenTen: Arabic Corpus and Word Sketches. *Journal of King Saud University-Computer and Information Sciences, 26*(4), 357-371.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing, 7*(1), 1-16.

Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., & Suchomel, V. (2013). arTenTen: a new, vast corpus for Arabic.

Corpus Linguistics. (2011). *Corpus Linguistics Terms and Their Meanings. Retrieved 25/2/2015,* from: http://www.cl2011.org.uk/corpus-linguistics-terms-and-their-meanings.html

Crnkovic, G. D. (2010). Constructive research and info-computational knowledge generation *Model-Based Reasoning in Science and Technology* (pp. 359-380): Springer.

Dash, N. S. (2008). *Corpus linguistics: An introduction*: Pearson Education India.

El-Haj, M., & Koulali, R. (2013). *KALIMAT a multipurpose Arabic Corpus.* Paper presented at the Second Workshop on Arabic Corpus Linguistics (WACL-2), UK.

Encyclopaedia Britannica Inc., & Time Inc. (2009). *Encyclopædia Britannica almanac 2010*. Chicago: Encyclopædia Britannica.

Goweder, A., & De Roeck, A. (2001). *Assessment of a significant Arabic corpus.* Paper presented at the Arabic NLP Workshop at ACL/EACL.

Graff, D. (2003). Arabic Gigaword. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2003T12. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2003T12*

Graff, D. (2007). Arabic Gigaword Third Edition. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2007T40. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2007T40*

Graff, D., Chen, K., Kong, J., & Maeda, K. (2006). Arabic Gigaword Second Edition. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2006T02. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2006T02*

Graff, D., & Walker, K. (2001). Arabic newswire part 1. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2001T55. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2001T55*

Hammo, B., Al-Shargi, F., Yagi, S., & Obeid, N. (2013). *Developing Tools for Arabic Corpus for Researchers*. Paper presented at the Second Workshop on Arabic Corpus Linguistics (WACL-2), UK.

Hasnah, A. (1996). *Full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval for Arabic Documents. Ph. D.* Dissertation, Illinois Institute of Technology.

Hmeidi, I., Kanaan, G., & Evens, M. (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. *JASIS, 48*(10), 867-881.

Lukka, K. (2000). The Constructive Research Approach. Retrieved 25/2/2015, from: http://www.metodix.com/en/sisallys/01_menetelmat/02_metodiartikkelit/lukka_const_research_app/kooste

Mansour, M. (2013). The absence of Arabic corpus linguistics: a call for creating an Arabic national corpus. *International Journal of Humanities and Social Science, 3*(12).

MEDAR Evaluation Package. (2010). *European Language Resources Association, ELRA Catalog number ELRA-E0040 Retrieved 25/2/2015, from: http://catalog.elra.info/product_info.php?products_id=1166*

NEMLAR Written Corpus. (2003). *European Language Resources Association, ELRA Catalog number ELRA-W0042 Retrieved 25/2/2015, from:*

http://catalog.elra.info/product_info.php?products_id=873

NEMLAR Project. (2010). The NEMLAR project and its results. Retrieved 25/2/2015, from http://www.medar.info/The_Nemlar_Project

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2009). Arabic Gigaword Fourth Edition. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2009T30. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2009T30*

Parker, R., Graff, D., Chen, K., Kong, J., & Maeda, K. (2011). Arabic Gigaword Fifth Edition. *Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2011T11. Retrieved 25/2/2015, from: https://catalog.ldc.upenn.edu/LDC2011T11*

Rogati, M., & Yang, Y. (2004). Resource selection for domain-specific cross-lingual IR. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). ACM.

Saad, M. K., & Ashour, W. (2010). *OSAC: Open Source Arabic Corpora.* Paper presented at the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, Cyprus.

Zaghouani, W. (2014). *Critical survey of the freely available Arabic corpora.* Paper presented at the Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme.

ملحق رقم (١)، نموذج لإحدى المقالات بالذخيرة في شكله النهائي بتنسيق XML

```
<Alittihad>
<ID>ETD_ARB_0000002</ID>
>
URL>http://www.alittihad.ae/details.php?id=11&y=2008&article=full</URL
<
<Headline>الإمارات ونيبال توقعان مذكرة تفاهم للنقل الجوي</Headline>
<dateline>تاريخ النشر: الثلاثاء ٠١ يناير ٢٠٠٨</dateline>
<Text>
وقعت الإمارات مذكرة تفاهم للنقل الجوي مع جمهورية نيبال في كاتماندو مؤخرا، تقضي بزيادة عدد الرحلات الجوية للناقلات الوطنية لكلا البلدين.
وقع المذكرة من الجانب الإماراتي الكابتن علائشة الهاملي رئيس قسم النقل الجوي في الهيئة العامة للطيران المدني في الدولة، ومن الجانب النيبالي ليلا ماني بوديال، وكيل وزارة الثقافة والسياحة والطيران لحكومة نيبال.
وقد اتفق الطرفان على تعيين طيران الإمارات والاتحاد للطيران والعربية للطيران وطيران رأس الخيمة كناقلات وطنية لدولة الإمارات العربية المتحدة، وكذلك تعيين الطيران النيبالي وشركة الطيران الدولية يتي كناقلات وطنية لحكومة نيبال. كما اتفق الطرفان على زيادة عدد الرحلات الجوية بين البلدين لتصل إلى ٢٨ رحلة أسبوعياً لجميع الناقلات الوطنية. وتتضمن المذكرة تعديل بعض أحكام اتفاقية النقل الجوي الموقعة بين الطرفين بالأحرف الأولى في تاريخ ١٣ أكتوبر ١٩٩٩ المتعلقة بأحكام التعيين والضرائب والسلامة الجوية وتحويل الايرادات والنشاطات التجارية، كما تتضمن الاتفاقية ممارسة الحرية الخامسة لناقلاتنا الوطنية على ثلاث نقاط متوسطة أو ما وراء حدود البلدين ما عدا نقاط ضمن حدود دولتي الهند وتايلاند.
</Text>
</Alittihad>
```

ملحق رقم (٢)، نموذج لإحدى المقالات بالذخيرة في شكله النهائي بتنسيق SGML

<DOC>
<DOCNO>ETD_ARB_0000002</DOCNO>
<Headline>الإمارات ونيبال توقعان مذكرة تفاهم للنقل الجوي</Headline>
<dateline>تاريخ النشر: الثلاثاء ٠١ يناير ٢٠٠٨</dateline>
<Text>
وقعت الإمارات مذكرة تفاهم للنقل الجوي مع جمهورية نيبال في كاتماندو مؤخرا، تقضي بزيادة عدد الرحلات الجوية للناقلات الوطنية لكلا البلدين.
وقع المذكرة من الجانب الإماراتي الكابتن علائشة الهاملي رئيس قسم النقل الجوي في الهيئة العامة للطيران المدني في الدولة، ومن الجانب النيبالي لبيلا ماني بوديال، وكيل وزارة الثقافة والسياحة والطيران لحكومة نيبال.
وقد اتفق الطرفان على تعيين طيران الإمارات والاتحاد للطيران والعربية للطيران وطيران رأس الخيمة كناقلات وطنية لدولة الإمارات العربية المتحدة، وكذلك تعيين الطيران النيبالي وشركة الطيران الدولية يتي كناقلات وطنية لحكومة نيبال. كما اتفق الطرفان على زيادة عدد الرحلات الجوية بين البلدين لتصل إلى ٢٨ رحلة أسبوعياً لجميع الناقلات الوطنية. وتتضمن المذكرة تعديل بعض أحكام اتفاقية النقل الجوي الموقعة بين الطرفين بالأحرف الأولى في تاريخ ١٣ أكتوبر ١٩٩٩ المتعلقة بأحكام التعيين والضرائب والسلامة الجوية وتحويل الايرادات والنشاطات التجارية، كما تتضمن الاتفاقية ممارسة الحرية الخامسة لناقلاتنا الوطنية على ثلاث نقاط متوسطة أو ما وراء حدود البلدين ما عدا نقاط ضمن حدود دولتي الهند وتايلاند.
</Text>
</DOC>

ملحق رقم (٣)، الكلمات الأكثر وروداً في كل مصدر

الاتحاد الإماراتية		الفبس		اليوم السعودية		اليوم السابع		الرياض السعودية	
الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها
في	5018241	في	8754044	في	7171136	فى	7430996	في	8308797
من	3486674	من	5760700	من	5971677	من	6329563	من	6768686
على	2172409	على	3582405	على	3419419	على	4327193	على	3854987
أن	1577316	ان	2529701	أن	1580474	أن	3495390	أن	2347992
إلى	1461127	الى	1854765	ان	1565342	إلى	2627973	إلى	1907478
التي	990055	التي	1529911	التي	1460179	عن	1597982	التي	1701509
عن	709618	عن	1312262	عن	1228542	التى	1510443	عن	1431794
مع	633649	أن	1171935	الى	1143625	مع	1094113	بن	1240842
الذي	616123	تم	1019695	إلى	1077843	الذى	924695	مع	1034834
خلال	488192	مع	989593	الذي	942992	مصر	775638	الذي	1013898
ما	428234	إلى	980154	مع	941340	اليوم	774134	ما	980214
هذا	374033	الذي	929035	بن	902045	ما	766260	هذا	980200
هذه	371818	ما	892285	ما	844493	بعد	747375	ان	939772
تاريخ	371639	لا	854498	هذا	819711	لا	739017	هذه	920760
لا	353815	النشر	823436	الموافق	813405	خلال	671987	لا	897279
النشر	352356	هذا	775243	لا	793024	أو	667734	أو	837878
بين	337326	هذه	733054	هذه	770040	رئيس	650100	العدد	812026
بعد	328732	الكويت	664724	بعد	600476	بين	641132	محمد	785070
إن	327841	خلال	595885	خلال	570358	أنه	635525	الرياض	718126
حيث	323847	بعد	529705	الله	543099	هذا	615353	الله	687530
أو	323501	بين	527476	محمد	538206	إن	594230	خلال	649740
أبوظبي	317821	كل	487299	حيث	521350	محمد	565470	بعد	607913
بن	307500	محمد	447292	بين	504496	هذه	552921	ذلك	587464
وقال	304668	كان	444238	كل	491929	عبد	538860	كما	585686
كما	284131	حيث	439503	أو	483496	كما	538787	بين	578112

وكالة أنباء سبأ اليمنية		الشروق الجزائرية		المصري اليوم		تشرين السورية		المستقبل اللبنانية	
الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها	الكلمة	ترددها
في	573664	في	1312452	من	2198252	في	3137371	في	5054728
من	372874	من	1065643	في	1866488	من	2344531	من	3025019
على	226987	على	621753	أن	1343356	على	1413525	على	2114090
إلى	156808	أن	525342	على	1025795	أن	860724	الى	1264453
أن	116689	إلى	425537	في	994338	إلى	728081	ان	1074891
التي	108448	التي	333566	إلى	608181	التي	663093	أن	1009881
سبأ	103390	عن	257804	عن	592974	عن	563414	التي	847882
سبأنت	89633	الذي	229328	علي	522957	مع	405215	عن	744027
اليوم	84766	ما	180604	مع	396107	الذي	382933	مع	611216
عن	80949	مع	167300	لا	379845	الى	379789	الذي	603219
مع	61564	لا	150134	التى	372140	ما	379079	إلى	555557
الى	56083	هذا	140705	إلي	344387	هذه	362606	لا	531593
خلال	55716	بعد	140101	ما	343558	هذا	345967	ما	517354
الذي	54625	هذه	127566	أو	326766	ان	315892	لبنان	516125
رئيس	52759	الجزائر	122577	بعد	284557	لا	295575	هذا	475619
ان	48081	حيث	116113	مصر	276189	سورية	283859	العدد	454064
اليمن	46244	لم	109889	الذى	257728	أو	261042	صفحة	453791
بين	40316	كان	100807	هذا	255841	بين	244476	هذه	417133
العلم	37824	كل	100212	إن	254342	كل	238141	بين	365966
صنعاء	36611	أو	99921	كان	243410	خلال	207218	بعد	344736
هذه	32799	بين	97277	لم	238898	بعد	202753	كل	327674
بعد	31843	الوطني	91178	بين	226221	كما	198670	رئيس	320715
علم	31690	كما	88822	أنه	217076	لم	187244	كان	284330
ما	31319	خلال	88584	هذه	216148	إن	180235	العلم	273384
وزير	31310	قبل	81127	رئيس	215643	ذلك	178850	الرئيس	266856