نحو بناء نموذج الذخيرة اللغوية العربية في ماليزيا

إعداد

أسوندي بن لامن ياشيم

المشرف
الأستاذ الدكتور نهاد الموسى

قدمت هذه الرسالة استكمالا لمتطلبات الحصول على درجة الماجستير في اللغة العربية و آدابها

كلية الدراسات العليا
الجامعة الأردنية

آذار ٢٠٠٩

الجامعة الأردنية

نموذج التفويض

أنا الطالب أسوندي بن لامن ياشيم، أفوّض الجامعة الأردنية بتزويد نسخ من رسالتي للمكتبات أو المؤسسات أو الهيئات أو الأشخاص عند طلبهم حسب التعليمات النافذة للجامعة الأردنية.

التوقيع: 

التاريخ: 24-03-2009

قرار لجنة المناقشة

نوقشت هذه الرسالة (نحو بناء نموذج الذخيرة اللغوية العربية في ماليزيا) وأجيزت بتاريخ :
2009/3/17م

أعضاء لجنة المناقشة	التوقيع
الدكتور نهاد الموسى، مشرفاً أستاذ – اللغة و اللسانيات	.....
الدكتور صلاح جرار، عضوا أستاذ – الأدب الأندلسي	.....
الدكتور محمود الحديد، عضواً أستاذ مشارك – فقه اللغة و اللسانيات	.....
الدكتور ، وليد العناتي، عضواً أستاذ مشارك – اللسانيات التطبيقية (جامعة البترا)	.....

الشكر و التقدير

أجزل الشكر و أعظمه إلى أفراد أسرتي الذين عانوا مرارة رحيلي عنهم و لوعة الغياب.

و إلى وزارة التربية الماليزية التي ابتعثتني لأواصل دربا أمضي به في تعلم اللغة العربية خدمة للقرآن العظيم.

و إلى أساتذتي في الجامعة الأردنية الذين غمروني بفيض علمهم و فضلهم وأخص بالذكر الأستاذ الدكتور نهاد الموسى الذي تفضل بقبول الإشراف على هذا البحث.

و إلى إخوتي من الزملاء العرب الذين شَرُفت بأن جعلوني محطّ فضلهم و حسن صحبتهم و أخص بالذكر أخي الأستاذ إياد العسيلي الذي تفضل بمراجعة هذه الدراسة وأحاطني بموفور أدبه و علمه.

فهرس المحتويات

قائمة الجداول

قائمة الأشكال

نحو بناء نموذج الذخيرة اللغوية العربية في ماليزيا

إعداد

أسوندي بن لامن ياشيم

المشرف

الأستاذ الدكتور نهاد الموسى

ملخـــــص

تحاول هذه الدراسة تقديم مشروع مقترح يمكن أن يكون منطلقا تأسيسيا لمشروعات مؤسسية كبيرة في مجال بناء الذخيرة العربية في ماليزيا. و تتلخص دواعي الدراسة في أنه، على الرغم من أهمية الذخيرة اللغوية، تندر الدراسات على مستوى العالم العربي – حسب علم الباحث – التي تتناول هذا الموضوع تعرّف بالذخيرة اللغوية، وتبين طبيعتها و أنواعها، وتوضح أدواتها البرمجية، و تقترح مشروعا لبنائها سواء على المستوى الفردي أو المؤسسي. و ليس ما جاء في هذه الدراسة من توصيف نظري إلا المشروع في شِقّه الأول؛ إذ تترقب هذه الدراسة تمامها في الجانب التطبيقي بالإفادة من نظم البرمجيات المتقدمة في ماليزيا لبناء الذخيرة اللغوية العربية، و هو المشروع الذي أعود إلى بلدي حاملا إنجازه على عاتقي.

المقدمة

إن عملية جمع المخطوطات التراثية، والكتب المؤلفة حديثا و قديما، والصحف العربية المتدفقة لغرض التحليل لعمل شاق. وهذا ما حدث منذ فترة طويلة خاصة للباحثين الأكاديميين. وبفضل التقدم التكنولوجي وتقدم التقنيات في تخزين المعلومات حاسوبيا أصبح هذا العمل الشاق سهلاً، وصار ممكنا الوصول إلى كميات كبيرة من المعلومات بطرق سهلة وميسرة، و إن تكن مسألة استيعاب ما يُسْتحدث تظل مفتوحة، و هو ما يقتضي تدبيرا منهجيا ثابتا للمتابعة، و محصول هذا كله ما يعرف بالذخيرة اللغوية أو متن اللغة، و هي مجموعة من النصوص المقروءة آليا، و يمكن تحليلها تحليلا لسانيا حاسوبيا.

ولا شك فيما لهذه الذخيرة من أهمية بالنسبة للباحثين الأكاديميين و اللسانيين التطبيقيين، لاسيما في مجال تعليم اللغة؛ لأنها تتيح للنتائج والبيانات المجموعة التي ستحتويها أن تقدم اللغة بطريقة واقعية وظيفية مقنعة. ويقدم الاستخدام الفعّال للذخيرة اللغوية فوائد جمة في مجال فرز قواعد اللغة صرفا و نحوا، وتأليف القواميس، والدلالة على وجوه التباين اللغوي في الاختيارات المعجمية و نِسَب التواتر في الظواهر النحوية والصرفية، و هي عدة للسانيات التاريخية. ولعلها أهدى دليل إلى رسم مناهج تعليم اللغة و وضع تآليف تعليمها.

و لعل هذه المقاصد تتعاظم فائدتها حين تتمثل أبعاد استثمار العربية في ماليزيا، من حيث إن العربية تمثل وجها من وجوه الثقافة الماليزية، و من حيث إنها تمثل مطمحا براغماتيا ينتفع بالجانب الثقافي و الاقتصادي، و هو ما ينبغي أن يتجلى ظاهرا في مناهج

تعليم العربية في ماليزيا، و إلا تكن رجعا و صدى جامدا يقصِّر عن بلوغ "وظيفة العربية" في ماليزيا.

تتلخص دواعي الدراسة في أنه، على الرغم من أهمية الذخيرة اللغوية، تندر الدراسات على مستوى العالم العربي – حسب علم الباحث – التي تتناول هذا الموضوع لقلة المراجع العربية في اللسانيات الحاسوبية النظرية و التطبيقية التي تعرف بالذخيرة اللغوية، وتبين طبيعتها و أنواعها، و توضح أدواتها البرمجية، و تقترح مشروعا لبنائها سواء على المستوى الفردي أو المؤسسي.

و على ذلك فإن أهمية هذه الدراسة تنبع من أهمية الذخيرة اللغوية نفسها و خاصة الذخيرة العربية، إذ تقصد إلى تقديم مشروع مقترح يمكن أن يكون منطلقا تأسيسيا لمشروعات مؤسسية كبيرة، و لا سيما مع تنامي أهمية علم حوسبة اللغة العربية في جانبها العملي أو التطبيقي، إذ إن بناء الذخيرة العربية المكتوبة و المنطوقة إنما يعتمد على منجزات معالجة اللغة العربية. و على ذلك تتمثل أهمية الدراسة في:

١- التعريف بالذخيرة اللغوية على أنها ثبت البيانات الضخمة التي يمكن تحليلها لسانيا في أقصر وقت ممكن.

٢- الإنباه على ضرورة تطوير برامج التحليل اللغوي لتخدم اللغة العربية على وجه مرض.

٣- محاولة رسم نموذج لبناء الذخيرة العربية في ماليزيا لتصبح هذه المحاولة نقطة بدء توصل إلى جهود جبارة مؤسسية في بنائها.

و يسعى الباحث – من خلال هذه الدراسة – إلى تحقيق الأهداف الآتية:

١- توصيف البرامج الحاسوبية الخادمة لتحليل الذخيرة اللغوية مثل Monoconc Pro، وWordSmith، و Xaira، و aConcorde، و توضيح إمكاناتها مع بيان قصورها عن خدمة نظام الكتابة العربي.

٢- توصيف أنموذج المجمع اللغوي الماليزي (Dewan Bahasa & Pustaka) في تخطيط ذخيرة DBP الماليزية مع الإنباه على المستفاد منها.

تتميز هذه الدراسة في أنها توظف الذخيرة اللغوية في خدمة اللغة العربية. و لا سيما أنها تساهم مساهمة جادة مع المجمع اللغوي الماليزي DBP في محاولة وضع خطة مشروع بناء الذخيرة اللغوية العربية في ماليزيا، علما بأن هذه المحاولة هي المحاولة الأولى – على حد علم الباحث – في هذا المجال. و تبدأ هذه المحاولة بتناول مفردات العربية في عدة نصوص مختارة: ١) صحف و مجلات ٢) إجابات الطلبة عن الاختبارات ٣) مواقع إنترنت ٤) كتب مدرسية ٥) رسائل أو بحوث جامعية ٦) قصص شعبية مترجمة ٧) نصوص تراثية مخطوطة ٨) قواميس. فهذه الدراسة لها جانبها النظري ولها طموح عملي عريض.

و قد اعتمدت الدراسة مراجع باللغتين الإنجليزية و العربية، و إذا كان الاعتماد على المراجع الإنجليزية أكبر فلأنها تتقدم المراجع العربية بأشواط في مجال التقدم التقني الذي سبقتنا إليه الحضارة الغربية.

و اجتهد الباحث في ترجمة كثير من النصوص الإنجليزية و المصطلحات العلمية في علم اللسانيات الحاسوبية، و استفاد من قاموسين: قاموس مصطلحات المعلوماتية واللغويات الحسابية لنبيل الزهيري، و معجم المصطلحات لمجمع اللغة العربية بالقاهرة، إلا أنهما لم يفيا بعض المصطلحات اللسانية الحاسوبية حقها، مما حدا بالباحث إلى الاجتهاد في ترجمة تلك المصطلحات.

و وجد الباحث نفسه في حاجة إلى الاتكاء على أكثر من منهج لغوي وفق ما اقتضته طبيعة الدرس؛ فاستخدم المناهج: الوصفي و التاريخي و التقابلي. و ما غلب المنهج الوصفي على الدراسة إلا كونها دليلا لإعداد ذخيرة لغوية عربية في ماليزيا.

و قد جاءت الرسالة في مقدمة و تمهيد و ثلاثة فصول و خاتمة؛ بينت المقدمة أهمية الدراسة و أهدافها و تفردها على صعيد الدراسات الماليزية، فيما عرض التمهيد إلى منزلة الذخيرة اللغوية في اللسانيات الحاسوبية، مستعرضا بعض الجهود النظرية في معالجة اللغة الطبيعية و اللسانيات الحاسوبية، مشيرا إلى بعض النشاطات الجادة في هذا المجال على المستويين العربي والأجنبي.

أما الفصل الأول فعرض إلى الذخيرة اللغوية، و جاء في ثلاثة مباحث؛ أشار الأول إلى أصول الذخيرة و معناها، فيما فصل المبحث الثاني في أنواع الذخيرة التي وصلت إلى ١٥ نوعا. و قدم المبحث الثالث بعض نماذج من الذخائر العربية و غير العربية مبينا مسوغات هذا الانتقاء.

و كان الفصل الثاني في تحديات الذخيرة اللغوية. و قد انتظم هذا الفصل أربعة مباحث. الأول منها في تحدي التقنية و الآلية، و الثاني في مراحل تحليل الذخيرة، فيما عرض المبحث الثالث إلى البرامج المقترحة، و خلص هذا الفصل إلى الحديث عن البرامج المنتظرة.

و جاء الفصل الثالث في الخطة المقترحة لذخيرة المجمع اللغوي الماليزي (Dewan Bahasa dan Pustaka) للغة العربية. و قسم هذا الفصل إلى ثلاثة مباحث. عرض الأول إلى توصيف الأنموذج اللغوي الماليزي، فيما عرض المبحث الثاني إلى التصميم المقترح لأنموذج الذخيرة العربية المقترح. وانتهى هذا الفصل بعرض معطيات الذخيرة.

و عرض الباحث في الخاتمة أبرز المشكلات التي واجهت هذه الدراسة و بيّن بعض توصياتها و نتائجها.

التمهيد

منزلة الذخيرة اللغوية في اللسانيات الحاسوبية

اللسانيات الحاسوبية فرع من فروع علم اللسانيات تطبّق فيها مبادئه و أسسه، لدراسة القضايا اللغوية في شتّى ظواهرها. و هي وثيقة الارتباط بمصطلحات أخرى كالذكاء الاصطناعي، و معالجة اللغة الطبيعية، و الترجمة الآلية، و تكنولوجيا اللغة، إذ تهدف إلى وضع نماذج حاسوبية لحوسبة الملكة اللغوية عند الإنسان.[١]

و الواضح أنّها تجمع بين اللسانيات و علم الحاسوب معا في "نظام بينيّ أو حقل بينيّ و تتبصّر في الظاهرة اللغوية من الجوانب التى تقتضيها المعالجة الحاسوبية في منطلقها النظري و مستهدفاتها التطبيقية. وهي على التقائها وافتراقها تمثّل جهودا متكاملة في بناء درس لغوي مضاف للظاهرة اللغوية التى تظلّ على هذا الصعيد مفتوحة لكل نظر مستأنف نحو أفقها المنداح إلى تلك الغاية التى لم يبلغها الإنسان بعد".[٢]

و لعلّ بداية تاريخ علم اللسانيات الحاسوبية تعود إلى عام ١٩٤٩ حين اقترح Warren Pereira أن الترجمة باستخدام آلة تبدو مستحيلة. و قد عقد المؤتمر الأوّل من نوعه في الترجمة الآلية في MIT عام ١٩٥٢[٣] و أدّى إلى نشر أوّل مجلة محكمة عن الترجمة الآلية تحمل اسم Mechanical Translation. و بدأ مصطلح اللسانيات الحاسوبية

---

[1] انظر الموسى، نهاد، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية ص ٥٣.

[2] محاضرة د. نهاد الموسى في مؤسسة شومان عن "حصاد القرن في اللسانيات". صحيفة الرأي الأردنية، العدد ٢٩-٩-٢٠٠٥. انظر http://www.alrai.com/pages.php?news_id=54817

[3] انظر Mitkov, Ruslan: The Oxford Handbook of Computational Linguistic, 2002, Oxford Press Limited, First Published 2003, page xvi.

بالظهور في أوّل استخدام له عام ١٩٦٥ عندما أضيف للمجلة لفظة Computational Linguistics لتصبح Mechanical Translation and Computational Linguistics وذلك بعد إنشاء جمعية الترجمة الآلية و اللسانيات الحاسوبية عام ١٩٦٢ (The Society of Mechanical Translation and Computational Linguistics).[١]

و في عام ١٩٨٠ استقرّ لفظ اللسانيات الحاسوبية دون اقترانه بـــ "الترجمة الآلية" بقرار حسمته الدورية الأمريكية للسانيات الحاسوبية (American Journal for Computational Linguistics) ولعل ذلك – كما يرى David Hayes[٢] يعود إلى آفاق أوسع تَعد بها اللسانيات الحاسوبية في إطار أساس علمي و بحثي في مجال اللغة ومعالجتها حاسوبيا بخلاف الترجمة الآلية.

و أما في السنوات العشرين الأخيرة فثمة تزايد ملحوظ في أعمال اللسانيات الحاسوبية العربية؛ إذ بدأت تظهر مشروعات مثل مشروع MEDAR[٣] و تعقد مؤتمرات و حلقات بحث عمل تخص موضوع معالجة اللغة العربية الطبيعية Arabic Natural Language Processing (ANLP) مع تطبيقاتها الكثيرة مثل الترجمة الآلية، واستخراج المعلومات، ونُظم الإرشاد (Tutoring Systems)، مما يستدعي مناهج مبتكَرة و تقنيات فعالة لأغراض تيسير عمليات تحليل هذه اللغة بما تحمله من غنى في بنيتيها الظاهرة والمقدرة.

---

[1] انظر Mitkov, Ruslan: The Oxford Handbook of Computational Linguistic, 2002, page xvi.

[2] كان عضوا في " Automatic Language Processing Advisory Committee of the National Academy of Sciences.

[3] MEDAR (Mediterranean Arabic Language and Speech Technology). هو مشروع إنشاء شبكة من المراكز الشريكة من أفضل التطبيقات في مجال اللغة العربية ومعالجة الخطابة بهدف إنشاء خريطة الطريق للتعاون على المدى البعيد. و هذا المشروع تدعمه المفوضية الأوروبية ويمتد من ١ فبراير ٢٠٠٨ حتى ١ أغسطس ٢٠١٠. انظر: http://www.medar.info/index.php

ثمة مجالات كثيرة شغلت اللسانيين و المبرمجين المهتمين بالعربية و حوسبتها لإيجاد أفضل الحلول لمواجهة الأسئلة في مجال معالجة اللغة الطبيعية و اللسانيات الحاسوبية، ويذكر منها على سبيل المثال لا الحصر:

- الموارد اللغوية (Language Resources)
- عنونة الكلمات: فرز أقسام الكلمة و وسم كل منها بإشارة مميزة (Part of Speech Tagging)
- التحليل و التوليد الصرفيان (Morphological Analysis and Generation)
- الإعراب الميسّر و العميق (Shallow and Deep Parsing)
- الترجمة الآلية (Machine Translation)
- إزالة الغموض البنيوي أو التركيبي (Word Sense and Syntactic Disambiguation)
- التحليل الدلالي (Semantic Analysis)
- استخلاص المعلومات واسترجاعها (Information Extraction and Retrieval)
- الإجابة عن السؤال (Question Answering)
- تجميع النص و تقسيمه (Text Clustering and Classification)
- تعدين النص ومحتوى الشبكة (Text and Web Content Mining)
- تعرف كيان الاسم (Named Entity Recognition)
- المعالجة المبنية على اللغة العامية (Colloquial-Based Language Processing)

و من النشاطات الجادة في مجال معالجة اللغة العربية الطبيعية:

- الملتقى الرابع للسانيات العربية والإعلامية، (١٩٨٧، مركز الدراسات والأبحاث الاقتصادية والاجتماعية، الجامعة التونسية، تونس).
- المؤتمر الثاني حول اللغويات الحسابية العربية (١٩٨٩، جامعة الكويت، الكويت).
- ندوة استخدام اللغة العربية في تقنية المعلومات (١٩٩٣، الرياض، مكتبة الملك عبد العزيز العامة، المملكة العربية السعودية).
- سلسلة ورشات العمل عن المناهج الحاسوبية للغات السامية (١٩٩٨، معالجة اللغة العربية الطبيعية ACL (Arabic Language Processing) ، Montreal، كندا).
- معالجة اللغة العربية الطبيعية – ورشة عمل (٢٠٠١، Toulouse، فرنسا).
- حلقة العمل في اللغة العربية، الموارد والتقييم (٢٠٠١، مصادر العربية و تقييمها LREC (Language Resources and Evaluation Conference)، Las Palmas، Canary Island).
- الدورة الخاصة حول العربية في التجهيز والتشغيل الآلي (٢٠٠٤، معالجة العربية بالجهاز التلقائي – TALN[1]، فاس، المغرب).
- مؤتمر حضارة الأمة وتحدي المعلوماتية (٢٠٠٤، كلية الآداب بجامعة الزرقاء الأهلية، الأردن).

[1] معالجة العربية بالجهاز التلقائي (TALN) أي Arabic Processing In Traitment Automatique Du Language Naturel

- مؤتمر NEMLAR [1] العالمي للعربية: مواردها و أدواتها (٢٠٠٤، القاهرة، مصر) وسيعقد المؤتمر الثاني في مارس عام ٢٠٠٩ المقبل.
- معالجة اللغة العربية الطبيعية – ALC (٢٠٠٥، Ann Arbor ، الولايات المتحدة الأمريكية).
- المؤتمر الدولي الخامس في اللغة والترجمة: قضايا معاصرة في الترجمة والتعريب (٢٠٠٥، مركز أطلس العالمي للدراسات و الأبحاث، عمان، الأردن).
- مؤتمر اللغة العربية في عصر المعلوماتية (٢٠٠٦، مجمع اللغة العربية بدمشق، سوريا).
- تحديات حوسبة العربية و معالجتها NLP و الترجمة الآلية MT، (٢٠٠٦ ، لندن، المملكة المتحدة).
- معالجة اللغة العربية الطبيعية – ALC (٢٠٠٧، براغ ، الجمهورية التشيكية).
- ندوة تقنية المعلومات و العلوم الشرعية و العربية (٢٠٠٧، كلية علوم الحاسب والمعلومات بجامعة الإمام محمد بن سعود، السعودية).
- الندوة العالميّة الأولى عن الحاسب واللغة العربيّة (٢٠٠٧، مدينة الملك عبد العزيز للعلوم والتقنية بالتعاون مع جمعية الحاسبات السعودية، السعودية).
- مؤتمر المعلوماتية وأنظمة حوسبة اللغة (٢٠٠٨ NLP-INFOS ، جامعة القاهرة، مصر).

[1] NEMLAR (Network for Euro-Mediterranean Language Resources) و هو مشروع بدأ من أجل المساعدة في تمهيد الطريق لجهد تعاوني للموارد اللغة العربية في منطقة البحر الأبيض المتوسط. و كان المشروع بدعم من الإتحاد الأوروبي من برامج تابع لـ UNCO-MED أي برنامج تعاوني بين الإتحاد الأوروبي و البلدان في منطقة البحر الأبيض منذ عام ٢٠٠٣ لغاية الآن. و لها ١٤ شريكا من مصر، و الأردن، و لبنان، و المغرب، و تونس، و الضفة الغربية و قطاع غزة، و الدنمارك، و فرنسا، و اليونان، و هولندا. انظر http://www.medar.info/index.php

- حلقة عمل عن حوسبة اللغة العربية في العالم – "اللغة العربية و معالجة اللغات المحلية"، مركز التحديثات والتوقعات، (LREC ٢٠٠٨).
- ملتقى لسانيات اللغة العربية: الإطار [المفاهيمي] والأبعاد المنهجية (مارس ٢٠٠٩، الأغواط، الجزائر).

و لا شك في مدى الفائدة المتحصلة من نتائج هذه المؤتمرات في مجال مشروع بناء الذخيرة اللغوية العربية؛ فالاستعانة بتجارب الآخرين العملية منها و النظرية تغني أي مشروع في هذا المجال.

و أما لسانيات الذخيرة (Corpus Linguistics) فهي مجال جديد نسبيا في اللسانيات الحاسوبية؛ وفرع حادث عليها؛ إذ يعود تاريخ نشأتها إلى تدوين ذخيرة Brown لـــ Kucera & Francis ما بين عامي ١٩٦١ و ١٩٦٤ في جامعة Brown في الولايات المتحدة. و قد اُشتهر هذا التدوين باسم "ذخيرة Brown" الذي قام على أساس جهد ثنائي قبل أن يقوم Kucera & Francis بإنشاء مؤسسة تُعنى ببناء ذخيرة أكبر حجما، وتحليلها حاسوبيا.[1]

و لذخيرة Brown خمسمئة عينة من النصوص المكتوبة في عام ١٩٦١ و تحتوي على مليون كلمة. و تقسم هذه العينات إلى ١٥ قسما أهمها: التقرير الإخباري، والتحرير الصَحَفي، والإعلانات الصحفية، والدين، والمهارات والهوايات، والأعراف و التقاليد،

[1] انظر David Crystal, The Cambridge Encyclopedia Of The English Language, 2nd Edition, 2003, page 448. و انظر Sandra Kubler: Introduction to Corpus Linguistics, Seminar fur Sprachwissenschaft, University of Tubingen, page 12.

والمنوعات، والعلم، والأدب: شعرا و نثرا، و الأحاجي و الألغاز. وعلى الرغم من باكورة العمل للسانيات الذخيرة آنذاك و خصوصا في شكوك تحليل الإحصاء اللغوي فإن ذخيرة Brown قد حققت مفاد بنائها فكانت نموذجا أو معيارا متبعا في وصف متن اللغة الإنجليزية و لغات أخرى كما تمنى صاحباها.[١]

و كانت ذخيرة Brown في بدئها مجرد رصد للكلمات نفسها إضافة إلى إمكانية العثور عليها، مع بعض سمات إحصائية بسيطة كنسب التواتر للكلمة و شيوعها. وعلى مدى عدة سنوا ت تطورت هذه الذخيرة تقنيا و طُبّق عليها تقسيمات الكلام (Part of Speech Tagged) فقد تطورت هذه التقسيمات إلى ما يناهز ٨٠ تقسيما مع رصد الكلمات الأجنبية الوافدة على الإنجليزية و بعض ظواهر لغوية أخرى. و على النهج نفسه تقوم كثير من الذخائر الإنجليزية كما هو واضح في ذخيرة Lancaster-Oslo-Bergen Corpus (LOB).[٢]

و مصطلح "لسانيات الذخيرة" قد استخدم مرارا و تكرارا بفضل Aarts & Meijs عام ١٩٩٠[٣] وفي الوقت نفسه دونت ذخيرة Lancaster-Oslo-Bergen Corpus (LOB) وهي مماثلة لذخيرة Brown غير أنها بريطانية المنشأ و فيها مليون كلمة. و ما بين عامي

[1] انظر Paul Baker, Andrew Hardie and Tony McEnery, A Glossary of Corpus Linguistics, 2006, Edinburgh University Press Ltd, page 25 & 26.
[2] انظر http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
[3] انظر Jacqueline Léon, Claimed and Unclaimed Sources of Corpus Linguistics, Henry Sweet Society Bulletin, Issue No. 44, May 2005, Page 36.

١٩٩١ إلى ١٩٩٤ دونت جامعة أكسفورد مع ٦ مؤسسات أخرى ذخيرة BNC (British National Corpus) و فيها ٩٠ مليون كلمة مكتوبة و ١٠ ملايين كلمة منطوقة و معنونة.[١]

و قد تقدّمت أوروبا في هذا المجال؛ إذ ظهرت عدة ذخائر أخرى مثل BoE (The Bank of English) من إعداد جامعة Birmingham في ١٩٩١-١٩٩٥ مع ٢٠٠ مليون كلمة حيث دونت لغرض تأليف القاموس و تعلم اللغة الإنجليزية. و ظهرت بعد ذلك ذخيرة ANC (American National Corpus). و الجدير بالإضافة أن جميع هذه الذخائر مُعلمة وفق أقسام الكلام (Part of Speech Tagged).[٢]

و في الذخائر المتكاثرة حجما لا بد من وجود برامج تمكّن من تناولها و هذه البرامج تسمى المُفَهرِسات (concordancers). و دورها هو رسم منهج البحث في الذخيرة و فحصها واستخراج بعض المعطيات و إتاحتها للتحليل المعجمي أو النحوي وفقا لمستويات لغوية أخرى دلالية و خطابية إلخ. و من اللافت للنظر أن بعض هذه البرامج مجاني و بعضها زهيد الثمن مما يضعها في متناول المهتمين بدراسة اللغة.[٣] و لا ريب في حجم الإفادة التى تقدمها مثل هذه البرامج للدارسين والباحثين، و ذلك بفضل ما تتيحه من تقدم متسارع في مجال معالجة المعلومات، و ما تسنح به من فرص بناء الذخائر الخاصة بالمستخدمين جراء الإفادة من برامج الفهرسة المجانية.

---

[1] و يقصد بـ "معنونة" أن النصوص قد وضعت لها علامات لتمييزها من حيث نوع الكلمة

[2] انظر Tony McEnery, Richard Xiao And Yukio Tono, Corpus-Based Language Studies, An Advanced Resource Book, Routledge Taylor & Francis Group, 2006, Page 4.

[3] للمزيد يراجع http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm

الفصل الأول

الذخيرة اللغوية

المبحث الأول

أصولها و معناها

"Corpus" لفظة لاتينية تعني الجسد (body) و جمعها (corpuses) أو (corpora)،[١] أما في العربية فقد وضعت لها عدة مقابلات منها: المدونة، والمفهرَسات، والمكتنزات النصّية، و الذخيرة اللغوية، و المادة اللغوية المسجلة، و متن اللغة، و المجموعة النصية، والمخزون النصي. و على اختلاف هذه الإطلاقات فإنها تشترك جميعا في الدلالة على مجموعات النصوص المكتوبة و المنطوقة لغرض محدد.

و عرّفها David Crystal[٢] بأنها مجموعة من البيانات اللغوية، أكانت نصوصا مكتوبة أم خطابات منطوقة يمكن اعتمادها نقطة انطلاق في وصف اللسانيات أو إثبات فرضيات لغوية. و أضاف John Sinclair[٣] أنها مجموعة نصوص لغوية واقعية و طبيعية اختيرت لوصف حالة أو حالات من اللغة.

---

[1] انظر Michael Pearce, The Routledge Dictionary of English Language Studies, 1st Published 2007. page 45.

[2] انظر David Crystal, *A* Dictionary of Linguistics and Phonetics, Blackwell*,* 3rd Edition, 1991. *Page 86.* و انظر http://www.mml.cam.ac.uk/call/cert/14/

[3] انظر John Sinclair, Corpus Concordance, Collocation*,* OUP, 1991. و انظر http://www.mml.cam.ac.uk/call/cert/14/

و واضح أن هذين التعريفين يفتقران إلى عنصر ثَبَت البيانات الإلكترونية (electronic database) و هي إحدى السمات المميزة لمفهوم الذخيرة. و قد أشار Mc Arthur & Mc Arthur إلى هذه السمة حين تحدثا عن الذخيرة اللغوية التى تخزن كمية ضخمة من النصوص قد تتجاوز ملايين الكلمات. و هذه البيانات محللة و مصنفة ومقسمة حسب أبنية الكلمة و تركيبات أخرى ذات علاقة، و ذلك باستخدام أدوات أو برامج الفهرسة (concordancing programs).[١]

و يرى نبيل الزهيري أن تُعرّفَ كلمة "Corpus" في جانبين اثنين هما: اللسانيات البنيوية و اللغويات الحسابية. أما في الجانب الأول فيقول: "الذخيرة مجموع البيانات (data) اللغوية التى يسجلها الباحثون اللغويون في مذكرات أو تسجيلات صوتية للكلام المدون على الطبيعة كما يلفظه أهل السليقة بلغة معينة، لتكون المادة (material) التى يستشهد الباحثون بها لوصف هذه اللغة علميا".[٢]

و أما في الجانب الثاني فيقول: "الذخيرة كم كبير من النصوص يختزن في ذاكرة الكمبيوتر لأغراض استخراج المعلومات و الرد على الاستفسارات و ما شابه ذلك. و قد يلحق بكلمات النصوص تعاليق و شروح (برموز الشِفرة الآلية) [عن] أقسام الكلمة و خصائصها الدلالية و اقتراناتها مع غيرها من الكلمات الأخرى. و قد يكون المخزون عاما أو متخصصا في مجال معين أو واسطة معينة كالكتب و المراجع أو الصحف و الدوريات".[٣]

---

[1] انظر Thomas Burns McArthur and Tom McArthur, The Oxford Companion to the English Language (1992), Oxford University Press.

[2] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ص ٧٨.

[3] المرجع نفسه

لماذا كلمة "الذخيرة"؟

يرى عبد الرحمن الحاج صالح أن لفظة الذخيرة "اسم أطلق على المدونة الخاصة التى هي هدف المشروع العربي المسمى بمشروع الذخيرة العربية.[١] فكلمة "ذخيرة" هي خاصة بهذا المشروع لأنها تتصف بأنها محوسبة: يمكن أن يلقى عليها أسئلة من أي نوع كان ولها موقع في الإنترنت وتسمى أيضا بالإنترنت العربي. و ثمة مشروع مماثل يجري حاليا بمصر من عمل مركز تسجيل التراث الثقافي و الطبيعي التابع لجامعة الإنسكندرية المسمى بـــ ICA أي International Corpus of Arabic.[٢]

وأهم من هذا كله أنها مدونة نصوص استخرجت من الاستعمال القديم والحديث (التراث وما يكتب بالعربية أو باللغات الأجنبية منقولا إلى العربية). فالذخيرة في رأيه ليست مدونة مفردات أو مدونة نصوص إذا قورنت بالمكنز العربي المعاصر[٣] أو المكنز الكبير[٤] أو المكتبات الإلكترونية، مجانية و مدفوعة الثمن في أقراص مدمجة أو على الشبكة الإلكترونية. لقد صارت الذخيرة تتجاوز المدونة اللغوية إلى مدونة ثقافية علمية شاملة الآن. فالذخيرة تستجيب لأي طلب معلومات في جميع المجالات.[٥]

---

[1] إن مشروع الذخيرة اللغوية العربية مشروع عربي أشرفت عليه المنظمة العربية للتربية والثقافة والعلوم. وقد عرض، لأول مرة، على مجلسها التنفيذي في ديسمبر ١٩٨٨ فوافق أعضاؤه على تبنيه. و نشأ من فكرة الاستعانة بالحاسوب واستغلال سرعته الهائلة في علاج المعطيات وقدرته العجيبة في تخزين الملايين من هذه المعطيات في ذاكرته، لإنشاء بنك آلي من المعطيات يحتوي على أهم ما حرّر بالعربية مما سينتَج على مرّ السنين. راجع http://www.isesco.org.ma/arabe/publications/Langue_arabe/p12.php

[2] راجع جريدة الرياض، العدد ١٤٥٩٩، ١٣ يونيو ٢٠٠٨. و انظر http://www.alriyadh.com/2008/06/13/article350422.html

[3] المكنز هو ذخيرة للكلمات و هي مرشد للباحث عن الكلمات المرتبطة بمفهوم ما (يمثلها المدخل). و الهدف العام منه لا يختلف عن معاجم المعاني التقليدية غير أن المكنز يتصف بأن تنظيمه مبني على الألفاظ بوصفها تمثيلا لمعاني المختلفة إذ لا يحتاج الباحث إلى البحث في الفهارس و رؤوس الموضوعات بل كل ما عليه هو أن يذكر كلمة شائعة تتعلق بفكرة ما ثم يبحث عنها في مكانها وفق الترتيب الألفبائي. راجع صفحتي "م" و "ن" في مقدمة المكنز.

[4] لقد ضم هذا المعجم بين دفتيه معجما للموضوعات أو المعاني أو المجالات، و معجما ثانيا للمترادفات و المتضادات، و معجما ثالثا لمعاني الكلمات، و معجما رابعا للألفاظ أو الكلمات. راجع صفحة ٧ في مقدمة المعجم.

[5] مراسلة مع الدكتور عبد الرحمن الحاج صالح بالبريد الإلكتروني في ١٦ أبريل ٢٠٠٨.

و أضاف عبد الرحمن الحاج صالح أن الذخيرة – ذخيرة المشروع العربي– تقوم بإنجازها ١٨ دولة عربية وليست عملا خاصا بمؤسسة واحدة، وذلك نظرًا لضخامة ما ستحتوي عليه من نصوص، ولأنه لابد أن يسهم في إنجازها كل البلدان الناطقة بالعربية.[١]

و لعل كلمة الذخيرة هي أقرب لروح هذا المشروع اللغوي بما تدور عليه هذه اللفظة من المعاني اللغوية حفظا و استبقاء و استزادة و استدعاء و استحضارا. و هي بهذا منسجمة مع أصل المراد بها في اللغة، إذ إن الذخيرة – في لسان العرب– هي واحدة الذخائر و معناها هو ما اُدّخر و استُبقي و اختِير و اتُخذ. و منه قول العرب: فرس مُذَّخَرٌ أي المُبَقَّى لِحُضْرِه.[٢] و قد يضاف إلى ما سبق شيوع هذا المصطلح بين الدارسين.

[1] مراسلة مع الدكتور عبد الرحمن الحاج صالح بالبريد الإلكتروني في ١٦ أبريل ٢٠٠٨.

[2] انظر ابن منظور، جمال الدين أبو الفضل محمد بن مكرم (ت ٧١١هـ)، لسان العرب، دار صادر، بيروت، ١٩٩٧. مادة "ذخر"

فوائد الذخيرة

و لعل من أهم فوائد الذخيرة في علم اللغة أنها تجعل الباحثين في موقف يقيني من نتائج بحثهم؛ لأن المنهج المطبق في البحث و دراسة اللغة منهج تجريبي غير ظني ولا بَدَهي، معتمد على التجربة العلمية و العملية؛ و لأن النصوص في الذخيرة كثيرة متكاثرة قد تصل إلى مئات الملايين من الكلمات تحلل تحليلا لسانيا في أقصى سرعة.

و هذه بعض الأمثلة المعجمية و النحوية التى تبرهن على جدوى الذخيرة و وجوه استثمارها و الانتفاع بها:[1]

١- المراد في المشترك اللفظي

- مصر في **قلب** الأحداث
- أجرى عملية **قلب** مفتوح
- كانت الشمس في **قلب** السماء
- تسببتْ الريح في **قلب** القارب

[1] راجع:

- مهديوي، عمر، و حمادة، سلوى، نحو بناء قاعدة بيانات معجمية للعلاقات الدلالية بين الكلمات. انظر http://www.arabcin.net/arabiaall/3-2006/3.html
- ياغي، حسين محمد، المدونات و تعليم النحو و الصرف و المفردات، المؤتمر الدولي الأول لتعليم اللغة العربية للناطقين بغيرها، ٢٠٠٨، الجامعة الأردنية.
- Aswandi, Shariman & Zaimuddin, (2008), Kamus Arab-Melayu Berasaskan Korpus http://prokamus.wordpress.com/2008/12/20/contoh-analisis-kombinasi-kata/
- Sameh Alansary, Magdy Nagi And Noha Adly: Building An International Corpus Of Arabic (ICA): Progress Of Compilation Stage. Page 3-6.

٢- نِسب تردد الكلمة

- للفصل بين الشائع و الغريب مثل السيف و الحسام
- لبيان اللفظ الأكثر انتماء للغة من الآخر مثل تواليت و مرحاض
- لتوضيح اللفظ الأرقى في المستوى من الآخر مثل هانم و ست
- لميز اللفظ الأكثر تخصصا من الآخر مثل حكم ذاتي و استقلال

٣- العثور على تنوع المعنى

- جامعة **عين** شمس
- **عين** الماء
- **عين** الإنسان
- **عين** وزيرا للخارجية
- **عين** السلطان

٤- البحث عن المتلازمات اللفظية

- لا علاقة – علاقة ودية – علاقات عامة – انقطاع علاقة – تطبيع علاقة – علاقة حب – ربط علاقة – ذات علاقة – على علاقة بـــ – على علاقة مع

(يتبع كمنهج جديد في تأليف القواميس)

٥- أشكال الكلمة في السياق

- في ميز "المصريون" من "المصريين"

٦- التحليل الإعرابي

- في الفصل بين السوابق و اللواحق بين:

علمية – علّمتْنا – علمه – علماء – تعليم – علوم – العلم

- في تحديد مواقع حروف الجر و العطف

ماذا يأتي في الغالب بعد/ قبل: من – على – في – أو – ثم

المبحث الثاني

أنواع الذخيرة

الذخيرة إذن، هي مجموعة من النصوص اللغوية مخزونة ومحللة و منظمة حاسوبيا، أيا كان نوعها، و هي لا تضم بالضرورة أية معلومات جديدة عن اللغة، ولكنها تعطي آفاقا جديدة للأبحاث اللغوية، وتساعد في تطوير عمليات مختلفة في تعلم اللغة والترجمة و تحليل المفردات.

و ثمة أنواع مختلفة و تقسيمات متعددة للذخيرة؛ فهناك من يقسمها إلى: الذخيرة مع التعاليق (Annotated Corpus) و دون التعاليق (Unannotated Corpus)،[١] و ثمة تقسيم ثان بحسب اللغات المستخدمة، يسلكها في ضربين: ذخيرة أحادية اللغة (Monolingual Corpus) و ذخيرة متعددة اللغات (Corpus Multilingual)، و تقسيم ثالث إلى نصوص دون عنونة (Raw Text) و نصوص منتقاة (Marked-Up Text)، ولا سيما أن هناك من يعد شبكة الإنترنت ذخيرة. و مهما يكن من تنوع الذخيرة فإن أسهل تقسيم لها يعود إلى الميز بينها طبقا لجدواها و فحواها.[٢] و لنا أن نعود إلى كتاب A Glossary of Corpus Linguistics لـــ Paul Baker و Andrew Hardie و Tony McEnery لنتبين تعريفا للذخائر بأنواعها المختلفة:[٣]

---

[1] انظر Tony McEnery and Andrew Wilson, Corpus Linguistics, Edinburgh University Press, 1996. Page 25.

[2] انظر Maryam Mohammadi, Specialized Monolingual Corpora in Translation, Translation Journal, Volume 11, No. 2. April 2007. http://translationjournal.net/journal/40corpus.htm

[3] ترجم الباحث هذه المصطلحات إلى العربية و أبان مقصود كل منها

١- الذخيرة مع التعاليق (Annotated Corpus)

و قد عرفت كذلك بـــ (Treebank) و هي نوع من الذخيرة التى تتضاف عليها معلومات لسانية عديدة صرفا و نحوا و تركيبا أو ما يسمى (tagged corpus). و من أشهر هذه الذخائر ذخيرتا LDC [1] و ELRA.[٢]

٢- الذخيرة دون التعاليق (Unannotated/ Raw Corpus)

تقابل هذه الذخيرة الذخيرة مع التعاليق؛ على أن نصوصها غير محللة، لكنها تفيد الدراسات اللغوية أيضا. و لا شك في أنها تصبح أوسع امتدادا و فائدة، إذا ما انضافت إليها بعض التعاليق. و هي تعد من أكثر الذخائر وجودا؛ لأن من السهل تجميع الذخيرة لكن عملية التحليل تواجه تحديات جد شائكة.

٣- الذخيرة المتخصصة (Specialized Corpus)

و هي تضم نصوصا لغوية معينة مكتوبة كانت أو منطوقة؛ إذ ليس لها حدود معينة من التخصيص، إلا أن بعض المعايير قد تحدد نوع النصوص كالزمن والموضوع. و من أمثلة الذخيرة ذات الأغراض الخاصة، ذخيرة Cambridge

[1] Language Data Consortium جمعية مختبرات البحوث لغرض البحث و التطوير؛ المفتوحة للجامعات و الشركات والحكومات. و تُعنى بإنشاء ثبت بيانات الكلمة و القواميس و موارد لغوية أخرى و جمعها و نشرها على صعيد المنطوق و المكتوب. تترأسها جامعة Pennsylvania الأمريكية منذ إنشاء الجمعية عام ١٩٩٢. و للمزيد يراجع http://www.ldc.upenn.edu/

[2] European Language Resources Association اتحاد الموارد اللغوية الأروبي الذي مقره باريس، فرنسا. تأسس عام ١٩٩٥. و مهامه توفير الموارد اللغوية لأغراض هندسة اللغة و تقييم تقنياتها. ولتحقيق هذا الهدف قد قام هذا الاتحاد بتحديد مصادر لغوية معينة و نشرها و جمعها و تثبيتها و بنائها. للمزيد يراجع http://www.elra.info/

وNottingham للخطاب الإنجليزي (CANCODE)[1] التى تحتوي على ٥ ملايين كلمة، و ذخيرة Michigan التعليمية للإنجليزية المنطوقة (MICASE).[2]

٤- الذخيرة العامة (General Corpus)

هذا نوع من الذخيرة يتضمن أنواعا متعددة من النصوص في موضوعات شتى مكتوبة و منطوقة. و تسمى أحيانا الذخيرة المرجعية باعتبارها مواد مرجعية لتعلم اللغات والترجمة. و من نماذج هذه الذخيرة، الذخيرة البريطانية الوطنية (BNC)[3] في ١٠٠ مليون كلمة و بنك الكلمات الإنجليزية (BoE)[4] الذي يتضمن ٤٠٠ مليون كلمة.

٥- الذخيرة المرجعية أو العِيارية (Comparable/ Reference Corpus)

و هي الذخيرة التى تتكون من نصوص متشابهة في نوعها و محتواها للمقارنة بين لغة و لغة أخرى، أو بين لغة و لغات أخرى كثيرة مثل ذخيرة "العقود القانونية في الإنجليزية و الفرنسية" ومثلها الذخيرة الإنجليزية الدولية (ICE)[5] التى تقارن ١٠٠ مليون كلمة إنجليزية باللهجات الإنجليزية المختلفة.

1 Cambridge and Nottingham Corpus of Discourse In English
2 The Michigan of Academic Spoken English
3 British National Corpus
4 Bank Of English
5 International Corpus of English

٦- الذخيرة المتوازية (Parallel/ Aligned Corpus)

و هي تتكون من نصوص مترجمة من لغة إلى لغة أو إلى لغتين أو أكثر. نضرب لذلك مثلا المادة الطبية المترجمة من الإنجليزية إلى الإسبانية والفنلندية والفرنسية. و كذلك ذخيرة EUROPARL[1] و هي سجلات رسمية للبرلمان الأوروبي تحوي ٢٠ مليون كلمة من جميع اللغات الأوروبية، و يفيد هذا النوع من الذخيرة في البحث عن المصطلحات و التعبيرات المتقابلة بين كل من اللغات المترجمة. و تفيد كذلك في رصد ما يقع من إشكالات في لغة المترجمين و بونها عن لغة المتعلمين.

٧- ذخيرة المتعلمين (Learner Corpus)

مجموعة المقالات التى ألفها متعلمو لغة ما. و هذه الذخيرة مُعَدّة للعثور على الاختلافات بين النصوص الصادرة من متعلمي اللغة و الناطقين بها. و منها، الذخيرة الدولية لمتعلمي الإنجليزية ICLE[2] و تحوي مليوني كلمة، و ذخيرة Louvain لمقالات الناطقين بالإنجليزية و تعرف بـــ LOCNESS.[3]

[1] European Parliament Proceedings Parallel Corpus 1996-2006
[2] International Corpus of Learner English
[3] The Louvain Corpus of Native English Essays

٨- الذخيرة التربوية (Pedagogic Corpus)

هي الذخيرة التى لها علاقة مباشرة بالتربية كما يدل اسمها. و هي مجموعة النصوص و الوسائل المعينة في العملية التعليمية و التعلمية مكتوبة و غير مكتوبة، وتتسع هذه الذخيرة لتشمل العملية التعليمية و التعلمية في كافة مستوياتها من المدرسة إلى الجامعة. وهذه الذخيرة تخدم المتعلمين بوجه خاص لتحسين معارفهم في اللغة و إغناء مهاراتهم التعلمية و برامجهم التعليمية.

٩- الذخيرة التاريخية و التعاقبية (Historical and Diachronic Corpus)

هي مجموعة النصوص التى تنتمي إلى حقب زمنية معينة بغرض إظهار رحلة تطور الكلمات على مدى زمن محدد. و من أشهرها ذخيرة Helsinki التى تحتوي على مليون و نصف مليون كلمة. و ذخيرة ARCHER[1] عن تاريخ الإنجليز.

١٠- الذخيرة المرصودة (Monitor Corpus)

هذه الذخيرة هي امتداد للذخيرة التاريخية و التعاقبية في البحث عن أثر التغيرات و التطورات و التحولات التى قد تطرأ على كلمة ما بالإضافة إلى رصدها سنويا، أو شهريا، أو حتى يوميا. و منها Longman Written American Corpus.

[1] Representative Corpus of Historical English Register

١١- الذخيرة متعددة اللغات (Multilingual Corpus)

تمثل هذه الذخيرة نصوصا صغيرة من ذخائر فردية أحادية اللغة يطبق فيها نظام واحد أو طريقة واحدة في أخذ العينات و تصنيف فئات الكلمة، بيد أن نصوصها متنوعة المصادر و متعددة اللغات.

١٢- الذخيرة النفعية (Opportunistic Corpus)

تتصف بأنها قليلة التكلفة لأن نصوصها تدون من مصادر شبكية إلكترونية مجانية أو شبه مجانية، يسهل الحصول عليها و تحويلها. و يغلب أن يكون هذا النوع من الذخيرة غير مكتمل باعتبارها جهدا فرديا أو شخصيا.

١٣- الذخيرة المثالية (Sampled Corpus)

مجموعة النصوص المختارة، اختيرت بعناية و درست بدقة و جِدّ. وتتصف هذه الذخيرة بالثبوت، إذ لا يطرأ عليها أي تغيير بعد اكتمالها.

١٤- الذخيرة المشبَعة (Saturated Corpus)

نوع جديد نسبيا للذخيرة و تخص موضوع المعجميات بحيث يتوقف معدل نمو المفردات لتصبح ثابتة أو مشبعة.

١٥- الذخيرة المنطوقة (Spoken Corpus)

الذخيرة التى تحتوي على نصوص مكتوبة من كلام منطوق و متداول. وتحلل أصوات الكلمات و تدون باستخدام الكتابة الصوتية، و مثالها ذخيرة LLC اختصارا لـــ (London-Lund Corpus of Spoken English)

و يأتي الاستعراض لأنواع الذخيرة السابقة في محاولة لتخير النوع الأكثر مواءمة حسب اعتقاد الباحث في بناء الذخيرة العربية في ماليزيا. و يراها الباحث في بعض الأنواع دون غيرها مثل الذخيرة دون التعاليق أو الذخيرة الموازية وذخيرة المتعلمين. و مردّ ذلك يعود إلى أن مشروع الذخيرة المقترح ما يزال في بداياته و أن الأنواع السابقة قد تكون أسهل في التطبيق و التنفيذ.

أما عند الشروع الفعلي في تنفيذ المشروع مستقبلا فإن بقية الأنواع التى تحتاج إلى تقنيات عالية و آليات مطوّرة تتعلق بالأنظمة الصوتية و النحوية والدلالية ... تصبح أكثر قابلية للاستخدام و التنفيذ؛ مثل الذخيرة مع التعاليق التى تحتاج إلى معلومات لسانية عديدة، والذخيرة المنطوقة التى تحتاج إلى تحليل أصوات الكلام باستخدام الكتابة الصوتية، و الذخيرة المشبعة التى تتطلب معرفة بمعدل نمو المفردات بحيث يمكن وصفها بالثابتة أو المشبعة.

المبحث الثالث

نماذج من الذخائر اللغوية، عربية و غير عربية

أ- العربية:

إن المدارس اللغوية الأمريكية و الأوروبية قد تفوّقت على المدارس اللغوية العربية في بناء الذخيرة وتطويرها لتقدمهم في مجال التكنولوجيا المعلوماتية. و يرى الباحث أن التعريف العام ببعض تجارب الذخائر اللغوية غير العربية مهم في محاكاة هذه التجارب وبخاصة لدى البدء في إعداد ذخيرة جديدة مثل الذخيرة العربية المقترحة في ماليزيا. علما بأن معظم هذه الذخائر متوافر على الشبكة الإلكترونية العالمية.

إلا أن تفوق النماذج الغربية السابقة لم يمنع من وجود ذخائر للغة العربية أنشأتها المؤسسات العلمية من جامعات ومعاهد ومجامع وحكومات، وهي في معظمها جهود لمؤسسات و حكومات غير عربية. نذكر منها على سبيل المثال لا الحصر:[١]

١- الذخيرة العربية لـــBuckwalter[٢]

تجربة المعجمي Buckwalter في عام ١٩٨٦ لجمع المفردات بنسخها نسخا من جريدة الشرق الأوسط. و تحتوي ذخيرته هذه على ٤٠،٠٠٠ كلمة و هي أول تجربة

[1] انظر
http://www.globalwordnet.org/AWN/Resources.html#0.0.4.2%20Arabic%20Dependency%20Treebank%20|outline
[2] راجع http://www.qamus.org/

لبناء ذخيرة عربية، و اعتمدت على النصوص الورقية إذ لم تكن النصوص الإلكترونية موجودة.

٢- ذخيرة Leuven[1]

جهد فردي لـــ Mark van Mol من جامعة Leuven الكاثوليكية ببلجيكا عام ٢٠٠٠. وقد بدأ عمله في ١٩٩٠ بغرض تأليف معجم تعليمي جديد (عربي – هولندي – عربي). و كانت مصادره هي البرامج الإذاعية و التلفازية من أخبار و مقابلات ومحاورات و خطابات و مسرحيات من الجزائر و مصر، و السعودية. وأدخل في ذخيرته كذلك ٥٠ كتابا مدرسيا في تعليم العربية لتسع دول، و نصوصا عديدة من مجلات الأخبار و نشراتها في مواقع الإنترنت.

٣- ذخيرة جمعية المعطيات اللسانية (The Linguistic Data Consortium- LDC)[2]

مشروع ضخم لجامعة بنسلفانيا الأمريكية عام ١٩٩٤ لبناء الذخيرة العربية وتطويرها مكتوبة ومنطوقة. و على الصعيد المكتوب هناك ذخيرتان: Arabic Newswire Corpus و Arabic Gigaword تحتويان على ٨٠ مليون كلمة من نصوص إخبارية لـــ AFP، و وكالة أخبار الجمهورية الإسلامية (Islamic Republic News Agency)، و وكالة الأخبار XinHua، و صحافة الأمة

[1] راجع http://ilt.kuleuven.be/arabic/ARAB/indexARAB.php
[2] راجع http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T02

(Press Ummah). أما على الصعيد المعطيات المنطوقة فهي ١١٠ سلسلات من تسجيلات الأخبار الإذاعية في إذاعة صوت أمريكا (The Voice Of America) و١٢٠ مكالمة هاتفية بين المصريين المقيمين في أمريكا و كندا، و تدرس هذه الذخيرة المنطوقة ظواهر صوتية كالنبر و التنغيم و التواتر و التردد في كل كلمة ملفوظة.

٤- ذخيرة Nijmegen[١]

طورتها جامعة Nijmegen الهولندية لبناء معجم (عربي- هولندي – عربي) مستمد من الذخيرة التى اعتمد في إنشائها على مجلتي الوسط السعودية و العربي الكويتية، و بعض القصص القصيرة من صحيفتي الحياة و القدس. و لعل تأليف هذا المعجم يأتي ردا على عدم وجود معجم عربي هولندي على نمطه الخاص بعيدا عن الترجمة المباشرة من المعاجم العربية.

٥- ذخيرة CLARA (Corpus Linguae Arabicae)[٢]

الذخيرة العربية (الحديثة الفصيحة) و تحتوي على ٣٧ مليون كلمة مأخوذة من النصوص العربية المنشورة منذ سنة ١٩٧٥ في شبه الجزيرة العربية و سوريا ومصر و تونس و المغرب لتأليف قاموس (عربي- تشيكي) بدعم مالي من وزارة التربية التشيكية.

---

[1] راجع http://www.let.kun.nl/wba/Content2/1.4.5_Nijmegen_Corpus.htm
[2] راجع http://enlil.ff.cuni.cz/veda/projekty/clara.htm

٦- الذخيرة المصرية[1]

الذخيرة المسماة بـــ Egypt طورها مركز John Hopkins لمعالجة اللغة والكلام (The Center For Language and Speech Processing) في جامعة John Hopkins الأمريكية، و هي آلة ترجمة شبه أوتوماتيكية لترجمة القرآن إلى اللغة الإنجليزية.

٧- ذخيرة دينار (DIINAR Corpus)[2]

مشروع المعجم العربي الآلي متعدد اللغات و هو مشروع يرأسه J. Dichy من جامعة Lumiere Lyon II و هدفه الرئيس إنتاج معجم (عربي- إنجليزي – فرنسي) باستخدام برامج التحليل اللغوي الحديثة و يحتوي على ١٠ ملايين كلمة.

٨- ذخيرة الاتحاد الأوروبي للموارد اللغوية (ELRA-European Language Resources Association)[3]

صدر عن الاتحاد مشروعا الذخيرة:

[1] راجع http://www.clsp.jhu.edu/ws99/projects/mt/
[2] راجع http://Medar.Info/The_Nemlar_Project/Publications/Nemlar-Report-Survey-Final_Web.Pdf
[3] راجع http://www.elra.info/

أ) ذخيرة جريدة النهار

جمع ١٤٠ مليون كلمة من الجريدة التى صدرت ما بين ١٩٩٥- ٢٠٠٠ وتحفظ الكلمات في قرص مدمج على نظام الملف الـ HTML. و في كل سنة هناك ٤٥،٠٠٠ مقالة تحوي ٢٤ مليون كلمة.

ب) ذخيرة جريدة الحياة

مشروع مشترك بين جامعة Essex والجامعة المفتوحة (Open University) البريطانيتين لخدمة الهندسة اللغوية و استرجاع المعلومات (information retrieval) وهما عملان حاسوبيان خالصان. و تقسم هذه الذخيرة إلى عدة أقسام منها: العام، والأخبار، والاقتصاد، والعلوم، والرياضة.

٩- ذخيرة مزدوجة عربي – إنجليزي[1]

مشروع قامت به جامعة الكويت بدعم من المؤسسة الكويتية للتقدم العلمي يهدف إلى تحسين معجم ثنائي (إنجليزي – عربي) و تطويره في الترجمة المقابلة بين اللغتين الإنجليزية و العربية. و هذه الذخيرة تحوي ٣ ملايين كلمة منتقاة من سلسلة "عالم المعرفة" الكويتية.

[1] راجع http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

١٠- ذخيرة متعددة اللغات[١]

أنشأها ستّار عزويني عام ٢٠٠٣ من جامعة مؤسسة العلوم و التكنولوجيا بـــ Manchester (UMIST). و درس فيها ترجمة مصطلحات التكنولوجيا المعلوماتية من العربية إلى السويدية و بالعكس. و قد جمع ذخيرته من الكتب و البحوث و بعض المواقع الإلكترونية المتعلقة بأنظمة الحاسوب و برامجه.

١١- الذخيرة العربية للعلوم العلمية العامة (General Scientific Arabic Corpus-GSAC)[٢]

جهد قام به أمين المهنا من الجامعة نفسها إذ جمع في ذخيرته نصوصا من مجلة كويتية هي"العلوم و التكنولوجيا" و درس فيها بناء المصطلحات العلمية والتكنولوجية المترجمة إلى العربية.

[1] راجع http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

[2] المرجع نفسه

١٢- الذخيرة العربية الكلاسيكية/ الفصحى التراثية (CAC-Classical Arabic Corpus)[1]

طورها عبد الحميد علوي كذلك من جامعة مؤسسة العلوم و التكنولوجيا Manchester (UMIST). تحوي ٥ ملايين كلمة من الأشعار من عصر صدر الإسلام إلى القرن الحادى عشر الميلادي و تدرس هذه الأشعار دراسة معجمية.

١٣- ذخيرة ArabiCorpus[2]

من عمل Parkinson من جامعة Brigham Young للبحث في مدونة غير معنونة أو مصنفة عن كلمات وتعابير عربية. وهي تذكر درجة شيوع المبحوث عنه، وتدرج السياقات النصية لكل كلمة ومصاحباتها اللفظية. و تحتوي على ٦٨ مليون كلمة ونيف.

[1] راجع http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm
[2] راجع http://arabicorpus.byu.edu/

ب- غير العربية:

- الإنجليزية[1]

١- ذخيرة Brown تعود إلى عام ١٩٦١، و تعتبر من أوائل الذخائر و هي تجمع مليون كلمة من نصوص شتى للإنجليزية الأمريكية.

٢- ذخيرة Lancaster-Oslo-Bergen و عرف كذلك بـــ LOB التى تأتي في العام نفسه ١٩٦١ و مثلها كمثل ذخيرة Brown سوى أنها تجمع الإنجليزية البريطانية.

٣- الذخيرة البريطانية الوطنية (BNC)، وجمعت فيها العينات الكلاسيكية للإنجليزية البريطانية، و تحوي ١٠٠ مليون كلمة.

٤- ذخيرة الإنجليزية الدولية (International Corpus of English) أو ICE، وتحوي مليون كلمة إنجليزية حية و معاصرة مكتوبة و منطوقة تسجل تنوع استخدام الإنجليزية في بريطانيا و أمريكا، و أستراليا، و كندا، و جنوب أفريقيا، و الهند، وهونغ كونغ، و ماليزيا، و سنغافورة و بعض دول الكومنولث.

٥- ذخيرة وكالة Reuters، و جمع فيها ٩٠ مليون كلمة. و قد ذكرت Reuters أن مدة جمعها استغرقت عاما واحد من ١٩٩٦-٠٨-٢٠ إلى ١٩٩٧-٠٨-١٩.

٦- ذخيرة الأخبار البريطانية: و تتضمن ٢٠٠ مليون كلمة من مجموعة الأخبار التى جمعت منذ ٢٠٠٤ لأكبر صحف في بريطانيا؛ و هي Guardian، و Observer، وIndependent، و Daily Telegraph، و Times.

[1] انظر http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm

٧- ذخيرة الإنترنت الإنجليزية: و هي مجموعة من الكلمات تشكلت من ١١٠ ملايين كلمة تم تجميعها تلقائيا من شبكة الإنترنت عام ٢٠٠٥ مع بقية ذخائر الإنترنت بالصينية، و الفرنسية، و الألمانية، و الإيطالية، و الأسبانية، و الروسية، والبولندية.

- **الروسية**[١]

١- الذخيرة الوطنية الروسية: و هي مجموعة من النصوص المماثلة في التصميم للذخيرة البريطانية الوطنية (BNC)، و النسخة التجريبية تتضمن ١٠٠ مليون عبارة.

٢- ذخيرة الإنترنت الروسية: وشكّلت نحو ٩٠ مليون كلمة تم تجميعها تلقائيا من شبكة الإنترنت في فبراير إلى أبريل عام ٢٠٠٥.

٣- ذخيرة الصحف الروسية: و هي صحيفة Izvestia، و Trud، و Strana و تشتمل على ٧٨ مليون كلمة.

٤- الذخيرة الروسية الموحدة/ الفصيحة: و تشكلت من مجموعة القصص الخيالية الروسية الحديثة مع تبيان لغموض الفئات الصرفية للغة الروسية، و تحتوي على ١,٦ مليون كلمة.

[1] انظر http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm

- **الصينية**[1]

١- جزء من ذخيرة LDC Gigaword الصينية: حيث تشمل ٣٥ مليون كلمة صينية، حُللت نصوصها باستخدام أدوات NEUCSP التحليلية من مختبر حوسبة اللغة من جامعة الشمال الشرقي الصينية (North-Eastern University). و تجمع هذه النصوص الأخبار لمدة سنة كاملة (٢٠٠١) و هي تماثل نموذج ذخيرة Reuters.

- **لغات أخرى (مرتبة كميا)**[2]

١- الذخيرة البولندية (IPI Pan Polish Corpus) و تحوي ٣٠٠ مليون كلمة و نيفا.

٢- الذخيرة DBP الماليزية و تحوي ١٣٥ مليون كلمة و نيفا.

٣- الذخيرة التشيكية الوطنية (Czech National Corpus) و تحوي ١٠٠ مليون كلمة و نيفا.

٤- الذخيرة الهنغارية الوطنية (Hungarian National Corpus) و تحوي ٨٠ مليون كلمة و نيفا.

٥- الذخيرة الكرواتية الوطنية (Croatian National Corpus) و تحوي ٣٠ مليون كلمة و نيفا.

٦- الذخيرة METU التركية (METU Turkish Corpus) و تحوي ١٠ مليون كلمة.

---

[1] انظر http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm

[2] انظر Sandra Kubler*: Introduction to Corpus Linguistic*, Seminar fur Sprachwissenschaft University of Tubingen, page 16.

الفصل الثاني

تحديات الذخيرة اللغوية العربية

المبحث الأول

تحديات الآلية و التقنية

إن الحياة في مجتمع المعلوماتية في عصر العولمة و الرقمية ترتبط ارتباطا قهريا بشكل أو بآخر بتكنولوجيا الحاسوب، و لا شك أن الحاسوب يستخدم في معظم نواحي الحياة الإنسانية من اتصالات و اقتصاد كما أنه يساعد كثيرا في إدارة الحكومة الإلكترونية في كثير من بلدان العالم لخدمة المجتمع. و المطلوب إذن أن يتمكن الحاسوب من تقديم المعلومات تقديما فعالا و دقيقا، و يفرز المعلومات مكتوبة و منطوقة ناهيك عن ترجمتها.

و هنا يأتي دور تكنولوجيا اللغة البشرية (Human Language Technologies)، التى تمكن المستخدمين من الاتصال بالحاسوب و استخدامه بطريقة طبيعية لا تخضع لنظم الحاسوب الرمزية. و لا شك أن انقطاع بعض اللغات عن تكنولوجيا اللغة البشرية يحرم مستخدمي هذه اللغات من الآليات الهائلة التى توفرها هذه التكنولوجيا لخدمة اللغة.

وأما العربية فقد بدأت العناية بمعالجتها على مستويين؛ المستوى الأول: و يختص بما قدمه الباحثون العرب أنفسهم منذ مطلع العام ١٩٦١ حين عقد المؤتمر الأول للتعريب في

الرباط، حتى الندوة العالمية الأولى عن الحاسب و اللغة العربية التى عقدت في العام ٢٠٠٧ في السعودية. و نوقشت و اقترحت دراسات و مشاريع كثيرة بشأن معالجة اللغة العربية الطبيعية من جوانب عدة. سيكتفي الباحث بتعداد بعضها لأخذ فكرة عامة عن تلك الجهود وبخاصة في الجوانب التطبيقية حيث لاحظ الباحث للأسف قصورا في هذا الجانب إذا ما قورن بالعرض النظري. و فيما يأتي بعض البحوث التى اطلع عليها الباحث و يظن أنها تفيد الدارس في مجال بناء الذخيرة اللغوية العربية. و منها:[1]

- منطق النحو العربي و العلاج الحاسوبي – عبد الرحمن الحاج صالح
- الحاسب الآلي و صناعة المعجم العربي – محمود فهمي حجازي
- اللسانيات و برمجة اللغة العربية في الحاسوب – محمد علي الزركان
- معالجة اللغة العربية بالحاسوب – محمد عبد المنعم حشيش
- القراءة الآلية للنص العربي – حازم يوسف عبد العظيم
- نظام تصحيح الهجاء – حسام الدين حسن محجوب
- الفعل العربي و طرق معالجته – صرح الدين صالح حسين
- نظرية حاسوبية لسانية لبناء المعاجم الآلية – محمد الحناش
- نحو معجم عربي للتطبيقات الحاسوبية – محمود إسماعيل صيني
- الاسترجاع الموضوعي بواسطة كلمات العنوان – ناصر محمد السويدان
- البحث من العنوان في قواعد البيانات العربية – بخيت سليمان بخيت

[1] راجع http://www.voiceofarabic.net/index.php?option=com_content&view=article&id=107:2008-06-28-16-11-23&catid=16:2008-06-07-09-45-13&Itemid=27. و راجع http://www.mghamdi.com/

- التشريح البنائي لمشكّلٍ آليّ عربي لتوظيفه في نظام تخليق آلي للصوت المنطوق من النص العربي المكتوب – محمد عطية محمد العربي
- المصاحبة الصوتية وأثرها الدلالي في القرآن الكريم – دراسة فونولوجية حاسوبية - أحمد راغب أحمد
- ألفاظ الأعداد العربية وخوارزميات تركيبها - عبد الله الزامل
- الحروف العربية والحاسوب – محمد زكي محمد خضر
- جهود معهد بحوث الحاسب والإلكترونيات في بحوث اللغة العربية – منصور الغامدي و آخرون
- استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة – عبد المحسن بن عبيد الثبيتي
- الحوسبة التوليدية للصرف العربي – بوشعيب راغين
- المتطلبات للغوية لمعالجة التعابير الاصطلاحية – وفاء كامل فايد
- المعجم العربي في ضوء اللسانيات الحاسوبية – عمر مهديوي
- ترميز اللغة العربية – هند بنت سليمان الخليفة
- نظام ترميزي جديد لكتابة أصوات اللغة العربية – منصور الغامدي و آخرون
- اللغة العربية و المعالجة الآلية، برامج صخر نموذجا – عبد الغني أبو العزم
- نحو معجم حاسوبي أحادي للناطقين بغير العربية – وليد العناتي
- الدليل نحو بناء قاعدة بيانات للِّسانيّات الحاسوبيّة العربية – وليد العناتي
- نظام حاسوبي لتشكيل النص العربي، التقرير الفني النهائي – منصور الغامدي و آخرون

أما المستوى الثاني في خدمة اللغة العربية تكنولوجيا فيعنى بما قدمه الباحثون الأجانب في هذا المجال و قد أشرنا في أول بحثنا إلى أهم ما قدموه من جهود.[١]

و يرى الباحث أن التعاون بين أصحاب الجهود السابقة و الجهود العربية المبذولة في معالجة اللغة العربية الطبيعية ما يزال دون مستوى الطموح المرجو. و حتى على مستوى التعاون بين الباحثين العرب أنفسهم.

بيد أن هذه الجهود بما قدمته من آليات و أدوات في سبيل تطوير تكنولوجيا اللغة العربية ما زال أمامها رحلة شاقة و طويلة لتصل إلى ما وصلت إليه الإنجليزية، والروسية، والفرنسية، و اليابانية، و الصينية من تقدم في تكنولوجيا اللغة البشرية.

و يعود تحدي آليات تكنولوجيا اللغة العربية و أدواتها إلى قلة مصادرها و مواردها اللغوية (language resources) مع عظمة مصادرها التراثية مدونة و مخزونة في الذخيرة. و إن تكنولوجيا اللغة العربية يمكن بناؤها بعدما جمعت عددا ضخما معقولا من ذخيرة، كما أن هذه المصادر الإلكترونية المحوسبة ضرورية في اللغة و صناعتها وترجمتها. و لعل بعض الشركات الضخمة تبني على نفقتها الخاصة مصادر لغوية خاصة بها لأغراض تجارية مثل Treebank و لكن الشركات محدودة الميزانيات و الإمكانات لا تطيق ذلك. و غالبا ما تحتفظ هذه الشركات لنفسها بذخائرها و لا تتيح لغيرها استعمالها.

[1] انظر التمهيد.

و من البدهي أن تكنولوجيا اللغة البشرية في معالجة اللغة الطبيعية التى تطبق على لغات أوروبية لا تتماشى مع العربية في نظامها الكتابي إذ يجري من اليمين إلى اليسار، كما لا تظهر صوائتها القصيرة في الكتابة إلا في القرآن الكريم و كتب الأطفال. وهي لغة اشتقاقية تتعدّد فيها المعاني بتعدد المباني.[1] و لا شك أن التحدي الذي يواجه الذخيرة اللغوية العربية في مستوياتها النحوية و الصرفية و الدلالية يتطلب جهودا كبيرة و تقنيات مبتكرة.

[1] انظر Mark Van Mol, Exploring Annotated Arabic Corpora, Preliminary Results, page 1. http://ilt.kuleuven.be/arabic/ENG/onderzoek/index.php

المبحث الثاني

مراحل تحليل الذخيرة

و قبل الحديث عن برامج تحليل الذخيرة أو الفهرسة في المبحث التالي، يجدر بنا الحديث عن مراحل قياسية متبعة في تحليل الذخيرة؛[١] لأن التحديات الكبرى التى يواجهها أنصار العربية في موضوع معالجة اللغة الطبيعية تكمن في هذه المراحل و هي:

١- الوسم/ التمييز بالعلامات (Tokenisation)

و هي مرحلة أساسية في التعامل مع النص و يعني بها عملية كسر تسلسل الحروف في النص بالعثور على حدود الكلمات كلمة كلمة و تحديد نقطة بدايتها و نقطة انتهائها. ويبدو هذا يسيرا في الإنجليزية و ما شابهها من أنظمة الكتابة الرومانية. و لكن العربية تواجه تحديات أكبر من حيث تنوع شكلها الكتابي في الوصل و الفصل، و الضبط عن طريق الحركات.[٢]

٢- عنونة الكلمات بأقسامها (Part-of-Speech Tagging)

في هذه المرحلة يتم فرز أقسام الكلمة و وسم كل منها بإشارة مميزة. و قد خُدمت الإنجليزية في هذا المجال؛ فالمفردات تم تقسيمها مثلا إلى مفرد و جمع، و مذكر و مؤنث.

---

[1] انظر Geoffrey Leech, Developing Linguistic Corpora: a Guide to Good Practice, Adding Linguistic Annotation, 2004. و انظر http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm

[2] راجع الموسى، نهاد، قضايا اللغة العربية المعاصرة، (فصل الكتابة العربية)

وكذلك تمت عنونة الأفعال وفق صيغها الزمنية. و تختلف برامج العنونة من حيث القدرة إذ هناك برامج لا تزيد عنوناتها على ٢٠ عنونة و أخرى متقدمة و مطورة قد تصل إلى ٤٠٠.[١]

و تطمح الذخيرة العربية الشاملة التى قد يبزغ فجرها يوما ما إلى عنونة صرفية ونحوية و معجمية شاملة متقدمة و متطورة. لكن هذا الطموح يصطدم بحاجز صعوبة تنفيذ هذه العنونة و على الأخص الصرفية الاشتقاقية منها.[٢]

٣- الإعراب الجزئي (Partial Parsing)

و في الإعراب الجزئي تأتي وظيفة المعرب الجزئي (Partial Parser) ليطبق قواعد نحوية عليها و ذلك بإنتاج شجرة أو جدول الإعراب من الجملة. و يتضمن الجدول مجموعة الأعاريب الجزئية (Partial Parse) التى ينتجها برنامج الفهرسة أثناء عملية الإعراب و التى يناظر كل إعراب جزئي منها جزءا من الجملة (Substring).[٣]

[1] انظر A Survey of Machine Learning Approaches to Analysis of Large Corpora, Xunlei Rose Hu and Eric Atwell. Page 46.

[2] راجع بعض الأمثلة التطبيقية على هذه المشكلات في الموسى، نهاد، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية. ص ٢٤٤-٢٤٥.

[3] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ص ٤١٣.

و تبدو صعوبة هذه المرحلة في أن "العربية هي اللغة الوحيدة التى قيل عنها إن عليك أن تفهمها قبل أن تقرأها"[١] و أن إمكانية الإعراب الجزئي فيها صعبة للغاية لأنه يعتمد على المعنى و ليس على مجرد مواقع الكلمات.

٤- التحليل الدلالي (Semantic Analysis)

و يقصد به الكشف عن معنى الكلام بطريقة منهجية، و هو الهدف من تحليل الجملة. و يتلخص في تجريد البنية المنطقية و العلاقات المنطقية للتركيب النحوي للجملة بعد عملية التحليل و التركيب ثم ترجمة هذه البنية المنطقية بالاعتماد على التعريفات القاموسية لمعاني لكلمات إلى صيغ منطقية (logical form) يمكن إجراء عمليات الاستنتاج المنطقي عليها وتحليلها حسابيا.[٢]

إن التعليقة الدلالية (Semantic Annotation) يقصد منها إلى زيادة البيانات لتسهيل تعرف مضمون الدلالات الكامنة وراء الجملة أو السياق. و الطريقة المتبعة الشائعة في هذه المهام تحويل التعريفات القاموسية لمعاني الكلمات إلى صيغ منطقية. والجدير بالذكر، أنه ليس هناك قياس واحد متفق عليه للسمات الدلالية التى ينبغي أن يؤخذ بها سوى عمليات الاستنتاج المنطقي.[٣]

[1] انظر البعلبكي، روحي، الترجمة الإلكترونية، آفاق الحاضر و المستقبل.
http://www.alarabimag.com/common/book/afaq013_1.htm
[2] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ص ٣٣٧.
[3] المرجع نفسه

و تبرز في هذه المرحلة مشكلة تتعلق بازدياد احتمالية اللبس في النص العربي و فيما قد يكون بعض اللبس قابلا للمعالجة بأدلة الاعتماد المتبادل، فإن وقوع الكلمة في دائرة ألفاظ من مجال دلالي خاص قد لا يكون كفيلا بكشف المقصود و دفع اللبس.[١]

٥- الحاشية/ التعليقة الخطابية (Discourse Annotation)

تختلف هذه المرحلة عما سبق من تحليل لساني للذخيرة بأنها لا تحلل الجملة كلمة كلمة، و لا هي تفك شفرات معاني الجمل جملة جملة فحسب. بل إنها تقوم بالكشف عن بنية الخطاب (discourse structure) التى تحددها العلاقات اللغوية و المنطقية الصريحة والضمنية التى تربط بين أجزاء الكلام. و توضح التعليقة علاقة الكلام بالسياق الذي يستعمل فيه لفهم معناه الإجمالي و رفع ما قد يوجد من لبس في معاني الجمل.[٢]

و ليس ثمة قياس عالمي في وضع حاشية أو تعليقة للخطاب في التحليل الدلالي (Semantic Analysis)، و يؤمل من الذخيرة المنطوقة الآخذة بالنمو و الازدياد في الحجم أن تساهم في كشف الغموض و رفع اللبس.[٣]

[1] راجع مزيدا من الأمثلة و التوضيح لهذه الآراء في الموسى، ٢٠٠٠، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية ص ٢٨١-٢٨٥.

[2] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ٢٠٠٣، ص ١٠١.

[3] انظر A Survey of Machine Learning Approaches to Analysis of Large Corpora, Xunlei Rose Hu and Eric Atwell. Page 47.

المبحث الثالث

البرامج المقترحة

لا مراء في أن أهم حديث عن الذخيرة اللغوية، حديث عن الفهرسة (Concordancing)؛ إذ إنها تعد من الإجراءات الرتيبة اللازمة في البحث، و العد، والتنظيم، و العرض للحصول على نتائج تجريبية دقيقة في ظاهرة لغوية ما. و إن المراد من هذه الإجراءات كلها تحليل الذخيرة تحليلا يجيب عن أسئلة اللغة في صوتها، و صرفها، ونحوها، و معجمها و ما إلى ذلك.

و قد اقترح Daniel Wiechmann و Stefan Fuhs[1] عشرة من برامج الفهرسة لتحليل الذخيرة، ثلاثة منها تجارية أي غير مجانية و هي: MonoConc Pro 2.2، و4 Wordsmith Tools، و Concordance. و أما الباقية، فهي برامج مجانية و هي: Multi Language Corpus Tool، و ConcApp 4، و AntConc 1.3، وAconcorde، و Simple Concordance Program، و Concordancer For Windows 2.0، و TextStat 2.6.

و من الملاحظ أن الاتجاه العام في لسانيات الذخيرة و ما بني عليها من برامج الفهرسة هو التركيز على اللغات الأوروبية على الرغم من محاولات أخرى في تطوير اللغات

[1] Daniel Wiechmann and Stefan Fuhs, Corpus Linguistic Theory. Volume 2, Issue 1 (2006), page 107, ISSN. http://www.reference-global.com/doi/abs/10.1515/CLLT.2006.006?cookieSet=1

غير الأوروبية كالصينية و اليابانية و الكورية، لكنها تبدو بدائية و ستقطع مسافة بعيدة لتصل إلى مستوى اللغات الأوروبية.

و على أن أبناء العربية يربو عددهم على ٣٠٠ مليون و المهتمين بها من المسلمين يزيد عددهم على خمس سكان العالم إلا أن الاهتمام بالذخيرة اللغوية العربية لا يزال متواضعا بالنسبة إلى اللغات المذكورة آنفا.

و الجدير بالذكر أن برامج الفهرسة المذكورة لا تتفاعل تفاعلا شاملا مع نظام الكتابة العربية سوى aConcorde. و من هذا المنطلق صمم Andrew Roberts، و Eric Atwel و لطيفة السليطي،[1] برنامجا للعربية سلسلا ميسرا سهل الاستخدام هو برنامج aConcorde ، و اقترحوا ثلاثة برامج أخرى و هي MonoConc، و WordSmith وXiara التى يمكن الاستفادة منها أكثر من بقية البرامج المذكورة. و فيما يأتي تفصيل لهذه البرامج و على الأخص برنامج aConcorde مع بيان محاسنها ووجوه قصورها:

١- MONOCONC PRO 2.2[2]

أصدرت شركة Athelstan برنامجي فهرسة هما: Monoconc، و Paraconc مع عدة نسخ لهما. و يعد Monoconc الذي بناه Barlow[1] من أفضل أحد عشر برنامجا في

[1] انظر aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic, Andrew Roberts, Latifa Al-Sulaiti and Eric Atwell, Corpora Vol. 1 (1): page 41.
[2] راجع http://www.athel.com/mono.html

الفهرسة مع ما يمتاز به من واجهته الجذابة.[٢] و يستخدم لتحليل نصوص اللغة الإنجليزية تحليلا لسانيا بغرض تعليمها و تعلمها. و يقرأ لغات أخرى بجانب اللغة الإنجليزية كالإسبانية، و الفرنسية، و اليابانية، و الصينية، و العربية و غيرها.[٣]

و هو يعطي نتيجة فهرسة KWIC،[٤] و ثبت الكلمات و كذلك معلومات عن التلازم اللفظي. و في مرحلة البحث المتقدم يمكن الاستفادة من بعض ميزات البرنامج – خاصة لنظام الكتابة اللاتينية كالإنجليزية و غيرها – للحصول على نتائج البحث في النصوص حسب السياق، و التعبير العام، و عنونة الكلمات بأقسامها (POS tag)، و المقارنة بين الذخائر.

و إن Monoconc برنامج شبكيّ الصفة؛ و يقصد بذلك أنه قابل للتشغيل المتوازي والبناء الترابطي مما يمكنه من أن يسخر في آن واحد كل المعلومات التى تدخل إليه من أجل حل المشكلة المعروضة عليه.[٥]

---

[1] أستاذ في قسم الدراسات اللغوية التطبيقية و اللسانيات بجامعة أوكلاند، نيوزيلندا

[2] سبق أن ذكرت أن هذا البرنامج برنامج تجاري (commercial software)؛ إذ يباع بـ ٨٥ دولارا أمريكيا برخصة واحدة، أي لحاسوب واحد فقط و ٥٥٠ دولارا أمريكيا برخصة واحدة لـ١٥ حاسوبا و لمدة عامين. و للاستزادة عن البرنامج يمكن تصفح موقع: http://www.athel.com/mono.html

[3] أعد مايكل برلاو الآن مع فريقه نسخة مونوكنك الخاصة للجامعة العربية المفتوحة. و للمزيد راجع www.michaelbarlow.com

[4] و هو نوع من الفهرسة التى تضع نتيجة الكلمة المبحوثة عنها في و سط الجملة فضلا عن مراعاة ترتيبها في السياق.

[5] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ٢٠٠٣، ص ٢٤٣.

٢- WORDSMITH TOOLS 4[١]

المُفهرِس الذي أصدرته مطبعة جامعة أكسفورد (Oxford University Press) من تصميم Mike Scott يقدم برنامجا واسع النطاق لتلبية متطلبات دراسة لسانيات الذخيرة و لا سيما وظائفها المتقدمة للفهرسة. و هو نسخة مطورة من نسخة MicroConcord عام ١٩٩٣، يليه الإصدار الثالث. و قد ظهر حديثا الإصدار الخامس استكمالا للإصدار الرابع القابل للتحديث.

و هو مشابه لـــ Monoconc Pro 2.2 من حيث التقنية و الوظيفة في الفهرسة، إلا أنه يتميز بخاصية تسمى بـــ WebGetter؛ إذ يقصد بها إمكانية عالية لـــ WordSmith في جلب نصوص كثيرة على شبكة الإنترنت. و هو أسرع بقليل من Monoconc Pro في أداء الفهرسة غير أنه لا يتفوق على نظيره في إدارة الموارد؛ لأنه يعجز عن إصدار نتائج كاملة في الفهرسة لكلمة متوسطة التردد (mid-frequency) مثل "of" في الذخيرة البريطانية الوطنية التى تحتوي على نحو ١٠٠ مليون كلمة. و يبطل تشغيله إذا انشغل بالكلمات الأكثر ترددا مثل "the".[٢]

---

[1] انظر Randi Reppen, Review Of Monoconc Pro And Wordsmith Tools, Language Learning & Technology, Vol. 5, No. 3, May 2001, pp. 32-36 http://llt.msu.edu/vol5num3/review4/default.html

[2] و يتوفر الإصدار الرابع من WordSmith بثمن ٨٥ دولارا أمريكيا للحاسوب الواحد و ٥٥٠ دولارا أمريكيا لكل ١٥ حاسوبا. وللمزيد من المعلومات عن انظر: http://www.lexically.net/wordsmith/index.html

٣- XAIRA[١]

إن برنامج XAIRA[٢] صُمم بديلا لبرنامج SARA[٣] لتحليل الذخيرة البريطانية الوطنية (British National Corpus). و استفاد من تكنولوجيا Unicode[٤] و XML،[٥] حيث يمكنه تقديم البحث الدقيق و إعطاء نتائج جيدة وفق عنونة XML. و إذا كانت الذخيرة المنطوقة، تميز صوت الرجل من المرأة فكذلك XAIRA؛ إذ بإمكانه ميز صوت الرجل من صوت المرأة. و يعطي XAIRA فهرسة الكلمة ترتيبا صحيحا للعربية ولكن يبقى أن الحصول على هذا الترتيب الصحيح يتطلب إجراءات طويلة نسبيا.[٦]

٤- ACONCORDE[٧]

هو برنامج ليس تجاريا و لا نتاج مشروع بحثي. و من حيث المواصفات فإنه يبدو سبيطا و غير معقد مقارنة مع البرامج الأربعة المذكورة آنفا سوى أن الهدف من تصميمه هو جعله برنامجا متعدد اللغات قابلا تماما للتفاعل مع العربية. فلا غبار على أنه يتعامل جيدا مع الكتابة من اليمين إلى اليسار أي أنه يمكنه ترتيب فهرسة الكلمات ترتيبا صحيحا، و تعد هذه

---

[1] انظر http://www.oucs.ox.ac.uk/rts/xaira/

[2] XAIRA (XML Aware Indexing and Retrieval Architecture)

[3] SARA (SGML Aware Retrieval Application) Standard Generalized Markup Language

[4] Unicode هي مجموعة رموز عالمية (٦٥٠٠٠ رمزا و نيف) تستخدم لتعريف جميع الرموز والحروف المستخدمة في أغلب لغات العالم وتجميعاً في ترميز واحد لستهيل عرض وإرسال المعلومات بغض النظر عن اللغة المستخدمة. هذا الترميز العالمي يستخدم من ١ إلى ٤ بايت (البايت=٨ بت) لترميز الحروف، ولم يستخدم حتى هذه اللحظة سوى ثلث العدد المتاح في Unicode لترميز حروف هذه اللغات. و هناك ثلاثة أنواع رئيسة تستخدم حاليا لترميز يونيكود و هي: UTF-8، و UTF-16، و UTF-32. راجع http://www.almashroo.com/articles/unicode-utf

[5] و يقصد بها Extensible Markup Language أي لغة التوصيف القابلة للتوسع. راجع:
http://www.almashroo.com/articles/لقاء-مع-العضو-الخبير-في-منظمة-w3c-مخت-وأحد/

[6] انظر Andrew Roberts, Latifa Al-Sulaiti and Eric Atwell, aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic, Corpora Vol. 1 (1): page ٤٤.

[7] انظر http://www.andy-roberts.net/software/aConCorde/

ميزة له على سائر البرامج الأخرى. و من الملاحظ أيضا أنه تفرد في إمكانيته تغيير الواجهة من الكتابة الإنجليزية إلى العربية.[١]

يمكن استخلاص مميزات aConcorde التالية:[٢]

- القابلية لكتابة العربية؛ إذ لا يحتاج إلى ترجمة العربية إلى ألفبائية لاتينية قبل فهرسة الكلمة.
- له واجهتان: إنجليزية و عربية
- مصمم بلغة البرمجة جافا (Java Script)[٣]
- يتماشى مع كثير من نظم التشغيل (operating system) و يقرأ ترميزات (encodings) مثل Unicode، و UTF-16، و UTF-18، وWindows العربية (CP1256)، و IBM العربية (CP420)، وMacArabic، و ISO Latin/Arabic (ISO 8859-6)، و ترميزة ASCII.[٤]
- متعدد الصيغ (multi-format) إذ يمكن تشغيل أنواع الملفات مثل: وtxt، وXML ، وHTML، و RTF، و MS-Word.
- قدرته على حفظ الملفات على نمط txt أو HTML[٥] مع مراعاة الترتيب الصحيح للفهرسة.

---

[1] انظر Andrew Roberts, Latifa Al-Sulaiti and Eric Atwell, aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic, Corpora Vol. 1 (1): page 4٥.

[2] Ibid page 47.

[3] ثمة لغات كثيرة في كتابة البرمجة و من أشهرها: Java، و C، و Visual Basic، و C++، و PHP، و Python، و Perl، و Ruby ، C#، و JavaScript. راجع http://www.ojuba.org/wiki/doku.php/docs/pyqt4?s[]=xml

[4] American Standard Code for Information Interchange

[5] Hyper Text Markup Language أي لغة توصيف النصوص المترابطة.

- تحليل نسب شيوع الكلمة و تكرارها.
- يمكن فهرسة الكلمة يمينا أو يسارا.
- إمكانية إعطاء نتائج البحث حسب الكلمة المفتاحية KWIC (key-word in context)، والعبارة (phrase)، و المجاورة (proximity)، و النسبة إلى (boolean)،[1] رمز المتغير (wildcard)،[2] و جذور الكلمة.
- برنامج مجاني (freeware) و مصادره مفتوحة (open-source).
- يفضله اللغويون و معلمو اللغة الذين يرغبون في الحصول على النتائج السريعة في ذخيرة صغيرة الحجم.

٥- آلية الذخيرة المتعددة اللغة (MULTI LANGUAGE CORPUS TOOL)

برنامج متعدد الاستعمال من تصميم Scott Piao[3] بلغة جافا (Java Script) و هو برنامج مجاني من جامعة Lancaster. يتميز البرنامج بأنه يقدم فهرسة يمكن مقارنتها مع فهرسة أخرى تم تشغيلهما في آن واحد وهذا مفيد للذخيرة المتوازية (Parallel Corpus). ويقرأ بعض اللغات و منها الصينية لأنه مدعوم بنظام Unicode.

[1] و ما يصطلح إليه بمتغير منطقي.
[2] رمز يوضع في موضع معين ليدل على أنه يمكن أن يحل محله أي عنصر مناسب لهذا الموقع، كالفراغات التى تحتها خط.
[3] محاضر في جامعة Manchester، قسم علوم الحاسوب، و هو عضو لـ OASIS و هو إئتلاف يسعى إلى التنمية، و الإتفاق والإجماع لبناء المعايير المعتمدة المفتوحة لمجتمع المعلوماتية. للمزيد راجع
http://personalpages.manchester.ac.uk/staff/scott.piao/#professional_activities

يتعامل مع ذخيرة صغيرة الحجم في حدود مليون كلمة كذخيرة Brown.[١] و من الجدير بالذكر أنه يعمل بفاعلية أكبر من بقية المفهرسات التجارية المذكورة آنفا على الرغم من أنه لا يمتع المستخدم بواجهات مريحة و جذابة.

٦- CONCORDANCE[٢]

برنامج كتبه و أصدره R.J.C. Watt[٣]. و يشتغل مع Unicode، إضافة إلى إمكانيته في التعامل مع تعدد نوع ملفات الذخائر. و يمتاز بالبحث أو الفهرس البسيط كما يمكنه أن يبحث عن التعبيرات العامة.[٤]

و إن تكن رخصته أغلى ثمنا لكنه لا يتفوّق على Monoconc Pro أو WordSmith 4.0 من حيث إدارة الموارد. و في محاولة إجراء الفهرسة لكمية كبيرة من النصوص، تحوي ١٠ ملايين كلمة مثلا، يطرأ على البرنامج خلل في الذاكرة إذ لا يقدر على تنفيذ الأمر. و يدل على ذلك أن Concordance صمم ليخدم نصوصا متوسطة الحجم.[٥]

---

[1] أول مخزون نصي (ذخيرة) أعدت عام ١٩٦١، يجمع للغة الإنجليزية الأمريكية المكتوبة و يحتوي على مليون كلمة.

[2] انظر http://www.concordancesoftware.co.uk/. و كان ثمن رخصته ٩٩ دولارا أمريكيا للرخصة الواحدة مع زيادة ٤٠ دولارا أمريكيا لأية رخصة تزيد على الأصل.

[3] المحاضر السابق و منسق برنامج قسم اللغة الإنجليزية، جامعة دوندي (Dundee) الأسترالية.

[4] Daniel Wiechmann and Stefan Fuhs, Corpus Linguistic Theory. Volume 2, Issue 1 (2006), page 113-114, http://www.reference-global.com/doi/abs/10.1515/CLLT.2006.006?cookieSet=1

[5] المرجع نفسه. ويمكن الاستزادة عن البرنامج في موقع: http://www.concordancesoftware.co.uk

٧- CONCAPP 4[١]

و هو من عمل Chris Greaves و هو برنامج مجاني مع واجهات سهلة الاستخدام من حيث التحكم في أي وظيفة من وظائفه المعروضة. على أنه لا يتعامل مع ذخيرة أكبر من ذخيرة Brown مما أدى إلى سرعة الفهرسة للنصوص، و مع أن هذا البرنامج أسرع من البرامج التجارية إلا أنه لا يقدم وظائف كثيرة و متعددة للفهرسة كما هو الحال مع بقية المفهرسات المجانية.

و يعد CONCAPP من البرامج الداعمة لـــ Unicode فهو يتعامل مع الإنجليزية واللغات الأوروبية مع لغات أخرى كالصينية، و اليابانية، و التايلندية و الروسية والعربية كذلك.[٢]

٨- ANTCONC 1.3[٣]

برنامج بناه Lawrence Anthony[٤] و طوره. و قد كان هذا البرنامج في بداية أمره برنامج فهرسة بسيطا، و من ثم تطور شيئا فشيئا ليصبح برنامجا مفيدا لتحليل النصوص كسائر برامج الفهرسة الأخرى. و هو مكتوب بلغة Perl 5.8 مدعما بنظام Komodo. ويمكن تشغيله بعدة نظم تشغيل و هذا يشمل أسرة Windows من Win98 إلى XP، وMacintosh OSX، و Linux.[٥]

---

[1] انظر http://www.edict.com.hk/PUB/concapp/

[2] و ثمة برنامج آخر يسمى Concograme من المصمم نفسه لمن يرغب في التحديث و يباع بـ١٠ دولارات أمريكية. و للمزيد عن البرنامج يراجع موقع المؤلف في: http://www.edict.com.hk/pub/concapp

[3] انظر http://www.antlab.sci.waseda.ac.jp/software/README_antconc3.1.3.txt

[4] محاضر في مركز اللغة الإنجليزية في العلوم و الهندسة، بجامعة وسيدا (Waseda University) اليابانية

[5] راجع http://www.antlab.sci.waseda.ac.jp/software/README_antconc3.2.1.txt

٩- برنامج الفهرسة الميسرة SCP (SIMPLE CONCORDANCE PROGRAM)[١]

صمم SCP لعملية الفهرسة الميسرة. و هو برنامج مجاني من صنع Alan Reed يختص بوظيفة البحث، و قائمة الكلمات (word list)، و الإحصاء. و يمكن تشغيله بلغة غير الإنجليزية لأنه مزود بلوحة الكتابة المرسومة على الشاشة (on-screen-keypad) مع حروف خاصة (special characters)، غير أنه يعجز عن فهرسة الذخيرة كبيرة الحجم. ويمكن تنزيل SCP عبر موقعه:

http://web.bham.ac.uk/a.reed/textworld/scp/

١٠- CW2 (CONCORDANCER FOR WINDOWS 2.0)[٢]

يعد من أوائل برامج الفهرسة و صمم لـــ Windows 3.1 مع واجهة (كلاسيكية) المظهر. و على توقف تطويره منذ السنوات العشر الماضية، فإنه لا تزال هناك أسباب وجيهة للنظر في CW2 و ذلك لسببين: أولهما أنه برنامج مجاني، ثانيهما أنه يقدم خمس طرائق في البحث الدقيق للحصول على النتائج القيمة من البحث. و مع أن CW2 يقبل ملفات الذخيرة المتعددة إلا أنه لابد أن تكون هذه الملفات في نمط txt. و يفضل التعامل مع ذخيرة صغيرة الحجم و يمكن تحميلها عبر موقعه:

http://opinion.nucba.ac.jp/~davidlee/devotedtocorpora/WCONCORD.ZIP

---

[1] انظر http://www.textworld.com/scp
[2] انظر http://www.flwi.ugent.be/nl/upload/courses/bdfrancq/WCONCORD.EXE

١١- TEXTSTAT[1]

برنامج مجاني من تصميم Matthias Hüning مكتوب بلغة Python و هو برنامج بسيط، سريع البحث و الفحص للنتائج. يقرأ نصوصا من ملفات ASCII/ANSI، وHTML، و MS Word، و OpenOffice مع ترميزات متنوعة، و يمكن تحميلها لتكون ذخيرة مفهرسة داخل TextSTAT. و هذا البرنامج متعدد اللغات لأنه يستخدم Unicode فلا حاجة إلى تغيير الترميزات ويمتاز هذا البرنامج بواجهته متعددة اللغات كالإنجليزية، والألمانية، و البرتغالية، والفرنسية.[2] ويعد TextSTAT ٢,٨ أحدث إصدار لهذا البرنامج حتى الآن و هو خاضع للتحديث و التقويم و يمكن تحميله عبر موقعه المذكور سلفا.

و جاء تخير الباحث للبرامج الأربعة السابقة بسبب تقني؛ ذلك أن هذه البرامج تعتمد نظام الكتابة العربية و بخاصة برنامج aConcorde الذي أشير إليه في العرض السابق و وضحت ميزات تفوقه في اختياره برنامجا لتحليل الذخيرة اللغوية المقترحة. و لكن هذا لا يمنع من إمكانية الاستفادة من بقية البرامج المعروضة في التحليل، ذلك أن التطور و التقدم فيها وارد. و قد تصبح ملائمة لتحليل الذخيرة العربية لاحقا. و أرى أن متابعة برامج تحليل الذخيرة هي مطلب مشروع و تطلع مستقبلي لإعداد الذخيرة المنتظرة.

---

[1] انظر http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

[2] انظر Comparative Review:Textstat 2.5, Antconc 3.0, And Compleat Lexical Tutor 4.0, Language Learning & Technology, Vol. 9, No. 3, September 2005, pp. 22-27.

المبحث الرابع

البرامج المنتظرة

ثمة برامج كثيرة في تحليل اللغة العربية قامت بها شركات أجنبية وعربية نذكر منها على سبيل المثال: RDI، و CIMOS، و XEROX، وSYSTRAN، وIBM، وIMAGiNET، و شركة صخر العربية، بالإضافة إلى بعض المؤسسات والجمعيات والجامعات المعنية بمعالجة اللغة العربية الطبيعية مثل LDC، وKACST، و UOB، وIERA، و ELDA.[1] و على أن هذه البرامج قد صممت فرادى غير مسوقة في مشروع يجمعها إلا أنها رافد يعد هدرا تجاهله إزاء بناء الذخيرة اللغوية العربية.

و من المؤسف أن جل هذه البرامج في تحليل العربية غير مجانية مما يحد من حجم الاطلاع عليها. و لابد هنا من التنبيه على الفرق بين برامج الفهرسة و برامج التحليل؛ إذ لا تحوي برامج الفهرسة الخالصة أي برنامج من برامج التحليل اللغوي بخلاف الذخيرة موضوع هذا البحث إذ تسعى إلى أن تشمل فهارسها برامج تحليلية تأسيا بالذخائر المتقدمة كالذخيرة الوطنية البريطانية (BNC)، و الذخيرة الوطنية الأمريكية ما شاكلهما.

[1] راجع Mahtab Nikkhou and Khalid Choukri, Survey on Arabic Language Resources and Tools in the Mediterranean Countries, Revised 7 March 2005, NEMLAR, Center for Sprogteknologi, University of Copenhagen, Denmark.

و حبذا لو تجمع هذه البرامج العديدة[1] التى سنذكرها في برنامج واحد يخدم فهرسة العربية و تحليلاتها اللغوية ابتداء من الوسم/ التمييز بالعلامات (Tokenisation)، فعنونة الكلمات بأقسام الكلمة (Part-of-Speech Tagging)، ثم الإعراب الجزئي (Partial Parsing)، فالتحليل الدلالي (Semantic Analysis)، انتهاء بالحاشية/ التعليقة الخطابية (Discourse Annotation).

و المؤمل من هذا البرنامج أن يكون مجانيا (freeware)، مفتوح المصادر (open source)، سهل الاستخدام. و لعل الأهم من هذا كله أن يعدّ لنا هذا البرنامج نتائج قيمة في الفهرسة دون إغفال ما تواجهه اللسانيات الحديثة من تحديات وإشكالات. أما البرامج المقصودة فنعرضها فيما يلي مع بعض التركيز على برامج "صخر":[2]

١- برولوج[3] المحلل العربي لشعلان (Shaalan's Prolog Arabic Analyzer)

و هو مشروع رسالة ماجستير لشعلان قدمها في عام ١٩٨٩ في جامعة القاهرة معتمدا برنامج برلوج SICStus. لكنه يبدو صعبا بالنسبة إلى اللغويين غير المبرمجين لأنه قديم جاء قبل الترميزات القياسية الجديدة (standard encoding) أي Unicode.

[1] انظر Eric Atwell, Latifa Al-Sulaiti, Saleh Al-Osaimi, Bayan Abu Shawar, , Arabic Language Processing, JEP-TALN 2004. Fez.

[2] انظر http://www.sakhr.com/products_a/Default.aspx?sec=Product

[3] لغة برمجية حاسوبية مصممة لمعالجة الرموز (symbol processing) و بخاصة معالجة اللغة الطبيعية. تستخدم في تطبيقات الذكاء الاصطناعي عموما. راجع: نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ص ٢٩٧.

٢- المحلل الصرفي (Ahmed's Computational Processor of Arabic Morphology)

و هو كذلك مشروع رسالة ماجستير قدمت عام ٢٠٠٠ و يسمى أيضا Morpho3 ما يعني أنه نموذج هجين أو ازدواجي (hybrid model) لتحليل الصرف العربي حيث إنه يجمع معرفتين اثنتين هما القواعد و الإحصاء.

٣- برنامج عنونة الكلمات (Khoja's APT Tagger)

برنامج يأخذ طابع ذخيرة BNC الإنجليزية في عنونة الكلمات و تلقيبها مع بعض تعديلات تناسب قواعد العربية و السبب في ذلك أن العربية لها نظامها الصرفي و النحوي المختلفان عن مجموعة العنونة (tagset) المتعارف عليها في اللغات الأوروبية.

٤- برنامج عنونة الكلمات: النسخة العربية (Freeman's Arabic version of the Brill Tagger)

صمم عام ٢٠٠١-٢٠٠٢ مبنيا على برنامج عنونة Brill Tagger و تصل مجموعة العنونة التى أضافها Freeman إلى ١٤٦ عنونة استنادا إلى جذر

الكلمة. و نظرا إلى أن هذه العنونة صممت لتخدم ذخيرة Brown الإنجليزية فلا مناص من أن تتشابك مع قواعد العربية من حيث تقسيم الكلام.

٥- المحلل Xerox الصرفي متناهي الحالات لـــ Beesley (Beesley's Xerox Finite-State Morphological Analyser)

قدم Beesley ما بين عام ٢٠٠١-٢٠٠٣ المحلل الصرفي للعربية باستخدام آليات النمذجة اللغوية Xerox (Xerox generic finite language tools) متناهي الحالات. و يقصد بذلك الحالات المحدودة لعنصر أساسي في نظم معالجة اللغة الطبيعية و تعرف الكلام.[1] و يعد وسيلة معينة في عملية التدريس بوصفه برنامجا لمعالجة اللغة الطبيعية. و ثمة نسخة مجانية تجريبية يمكن تحميلها من موقع: http://www.xrce.xerox.com/research/mltt/arabic

٦- المحلل الصرفي لـــ Berri، و Zidoum، و Atif.

و في عام ٢٠٠١ ساهم هؤلاء في بناء المحلل الصرفي الآخر و قد قاموا برصد القواعد الصرفية الثابتة، و قائمة من الكلمات أو الرموز للمتغيرات من القواعد ثم قاموا بتعديل الخوارزمية التى تناسب الوسم أو التمييز بالعلامات مع القواعد الصرفية الثابتة و المتغيرة على حد سواء.

[1] انظر نبيل الزهيري، قاموس مصطلحات المعلوماتية و اللغويات الحسابية، ٢٠٠٣، ص ١٣٤.

٧- المحلل الصرفي العربي لـــ Buckwalter

إن نظام هذا المحلل لا يسمح بمزج الأبجديتين اللاتنية و العربية فلا بد من أن تحول النصوص العربية إلى ASCII قبل تحليلها. و أما نتيجة التحليل في نظام ASCII فينبغي أن تحول إلى العربية من جديد حتى تقرأ النتائج. و يمكن تحميل برنامج Buckwalter عبر:

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

٨- برنامج LDC لعنونة الكلمات لمعموري و سيري

و فيه مجموعة من الكلمات المعنونة باستخدام المحلل الصرفي العربي لـــBuckwalter. قام Maeda Kazuai و Hubert Jin بعملية العنونة لصالح اتحاد البيانات اللغوية (Language Data Consortium).

٩- برامج صخر

أصدرت شركة صخر[1] عدة برامج في معالجة اللغة العربية الطبيعة نستعرض منها:

[1] أنشئت شركة صخر للبرمجيات في عام ١٩٨٢ و هي من أكبر شركات البرمجيات العربية. ومقرها في القاهرة لصاحبها السيد محمد عبد الرحمن الشارخ. انظر http://www.sakhr.com/default_a.aspx

- التحليل الصرفي

يتيح المعالج الصرفي متعدد الأطوار Multi-Mode Morphological Processor المعالجة العميقة للكلمة العربية المفردة، ويغطي هذا المحلل نطاق الكلمات العربية بالكامل؛ حديثها وقديمها. ويقوم بتعرف جميع أشكال جذر كلمة، أي أنه يقوم باستخلاص أصل الكلمة بعد تجريدها من اللواصق. ثم استخلاص البيانات الصرفية للكلمة مثل الجذر والميزان الصرفي لها، وقسم الكلم الخاص بها. و يعمل المركِّب بوضع عكسي، إذ يعيد توليد الكلمة بمختلف أشكالها الصرفية. [١]

- التصحيح الآلي

يقوم المصحح صخر الآلي باكتشاف الأخطاء الإملائية العربية والأخطاء العربية الشائعة وتصحيحها إضافة إلى الأخطاء النحوية.[٢]

- استخلاص المداخل

تقوم أداة استخلاص الكلمات المفتاحية بتحليل أي مستند عربي، وتعرف عبارات النص وعناصر البيانات الرئيسة تلقائياً. وتساعد الكلمات المفتاحية على تصنيف المستند في شجرة الموضوعات المحددة من قبل المستخدم لتسهيل استعراض المعلومات والوصول إليها. كما تساعد هذه

[1] راجع http://www.sakhr.com/Technology_a/Keyword/Default.aspx?sec=Technology&item=Keyword
[2] المرجع نفسه

الأداة على ربط المستندات بطريقة ديناميكية، وتقليل الوقت المطلوب للبحث في محتواها.[1]

- التصنيف الآلي

يقوم محرك صخر للتصنيف بتنظيم المعلومات القيمة بدقة وفاعلية وتصنيفها إلى شجرة موضوعات منطقية؛ نظراً لوضوح عملية التصنيف في كل مراحلها، بداية من إنشاء شجرة التصنيف، ثم ملء المحتوى وإدارة الوصول والعرض. و يستخدم هذا المحرك أداة المصحح الآلي لتصحيح النص العربي المدخل تلقائياً؛ لتصحيحه من الأخطاء اللغوية العربية الشائعة.[2]

- الشكل الآلي

نظراً لأن معظم الوثائق العربية لا تحتوي على شكل، قامت شركة صخر بتطوير الشاكل الآلي لمعالجة هذه المشكلة ووضع علامات الشكل على الحروف في النصوص غير المشكولة لمساعدة محرك النطق الآلي على نطق النص العربي بصورة صحيحة. يعتمد الشكل التلقائي على مستويات مختلفة من معالجة اللغة وتحليلها، بدءًا من الصرف وانتهاءً برفع اللبس عن معاني الكلمات. ويتم ذلك بتوظيف أبحاث معالجة اللغة الطبيعية إضافة إلى قواعد البيانات اللغوية الضخمة التى قامت صخر

[1] راجع http://www.sakhr.com/Technology_a/Keyword/Default.aspx?sec=Technology&item=Keyword
[2] المرجع نفسه

بتطويرها على مدار أعوام عديدة. و يقدم محرك الشاكل الآلي نصاً عربياً مشكولاً تصل دقة ضبط المفردات فيه إلى ٩٨%.[١]

- التخاطب

قامت صخر بتطوير محرك النطق الآلي للنصوص TTS (text to speech) و محرك تعرف الكلام الآلي ASR (automatic speech recognition). بينما يقوم محرك النطق الآلي للنصوص بتحويل أي نص عربي إلكتروني إلى صوت طبيعي، بينما يقوم محرك تعرف الكلام الآلي بتعرف الكلمات والأوامر العربية بمختلف الأصوات واللهجات وتحويلها إلى أوامر Command.[٢]

و لعل السؤال عن جدوى هذا المبحث، البرامج المنتظرة، سؤال مشروع. ويجتهد الباحث في الإجابة بأن البرامج التى عرضت في هذا المبحث هي برامج لغوية متخصصة وأن إعداد الذخيرة اللغوية يتكئ على هذه البرامج في مراحل تحليل الذخيرة المختلفة. وبعبارة أخرى نقول إن البرامج التى عرضت و نفذت بشكل جزئي يمكن الإفادة منها إذا ما أصبحت شمولية الطابع متكاملة البناء.

[1] راجع http://www.sakhr.com/Technology_a/Keyword/Default.aspx?sec=Technology&item=Keyword
[2] المرجع نفسه

الفصل الثالث

الخطة المقترحة لذخيرة المجمع اللغوي الماليزي DBP للغة العربية

المبحث الأول

توصيف أنموذج ذخيرة المجمع اللغوي الماليزي DBP

جاءت فكرة بناء الذخيرة للغة الماليزية في عام ١٩٨٣ في Projek Analisis Teks Secara Komputer[1] ، و هو مشروع تحليل النصوص تحليلا حاسوبيا، و كان يهدف أول بدئه إلى تدوين ما ينيف على مليوني كلمة مستندا إلى ذخيرة براون Brown. ونظرا إلى صعوبة عملية تحليل الذخيرة في بداية المشروع و قصور إمكانيات التقنية والآلية، فقد أهملت العملية و استهدف المشروع جمع النصوص و تدوينها فقط. إلى أن جاء فريق من الباحثين من قسم الترجمة الحاسوبية (Unit Terjemahan Melalui Komputer)[٢] في جامعة USM بدأ مشروع بناء ذخيرة DBP بداية جادة و عملية و ذلك في عام ١٩٩٣.[٣]

و تتكون ذخيرة DBP من نصوص ملايوّية قديمة من كتب و روايات، و نصوص أخرى حديثة مأخوذة بطبيعة الحال من الكتب، و نشرات الأخبار، و المجلات. و أما الذخيرة المنطوقة فلم يبلغها الاهتمام حتى الآن لعدة صعوبات و لا سيما ما تعلق منها بتحديات آليات

---

[1] راجع Zaiton Ab. Rahman 1987. Kertas Rancangan Projek Analisis Teks Secara Komputer. Cawangan Penyelidikan, DBP.

[2] راجع http://utmk.cs.usm.my/

[3] انظر
http://dbp.gov.my/lamandbp/main.php?Content=vertsections&SubVertSectionID=551&VertSectionID=25&CurLocation=238&IID=&Page=1

تحليل المنطوق.[١] و قد وصل عدد الكلمات في ذخيرة DBP حتى ٢٥ نوفمبير ٢٠٠٨ ما يناهز ١٣٥ مليون كلمة.[٢]

و أصبحت الذخيرة حينئذ، مصدرا جامعا موثوقا و مستهدفا للباحثين اللغويين في دراساتهم للغة الملايوية. و يتوقع منها أن تعطي نتائج قيمة و دقيقة و حيوية في وصف الملايوية و أن تكون أساسا و مرجعا في وضع المعاجم و علم القواعد على سبيل المثال لا الحصر. كما يمكن اللجوء إلى هذه الذخيرة في الشبكة الإلكترونية[٣] ليفيد منها كثير من الباحثين الآخرين المهتمين بالملايوية.

و من الطبيعي أن تعتني أي ذخيرة في أول بنائها بجمع نصوص كثيرة و تدوينها دون اللجوء إلى مراحل التحليل اللغوي المعقدة. و قد سمي هذا النوع من الذخيرة بـــ "الذخيرة دون التعاليق" (Unannotated/ Raw Corpus). و ثمة تمايز و تباين في المصطلحات المتعلقة بالذخيرة من حيث مسمياتها يبدأ من مرحلة جمعها حتى تحليلها حاسوبيا؛ إذ تلتبس أحيانا مصطلحات مثل: المدونة، و قائمة البيانات للكلمة، و الأرشيف، و معطيات النصوص، والذخيرة المحوسبة، و الذخيرة المصغرة، و النصوص المنتقاة و ما إلى ذلك.

[1] سبق أن ذكرنا عن مراحل التحليل للذخيرة و هي؛ التمييز بالعلامات، ثم عنونة الكلمات بأقسامها، ثم الإعراب الجزئي، ثم التحليل الدلالي، و أخيرا التعليقة الخطابية. انظر ص ٣٨-٤١ من هذه الرسالة .

[2] انظر
http://dbp.gov.my/lamandbp/main.php?Content=vertsections&SubVertSectionID=551&VertSectionID=25&CurLocation=238&IID=&Page=1

[3] للمزيد يراجع : http://sbmb.dbp.gov.my/knb/nb_konkordans.aspx

و حقيق بنا أن ننبه على أن بعض هذه المصطلحات ضروري لرفع اللبس بينها و ذلك حسب ما جاء في تعريف Sinclair في Preliminary Recommendations on Corpus Typology[1] لهذه المسميات و هي:

- المدونة: و هي مجموعة من النصوص اللغوية كاملة أو غير كاملة، تنتقى و ترتب وفق معايير لسانية دقيقة لتكون أمثلة لظاهرة لغوية ما.
- الذخيرة:[2] هي المدونة التى تم حوسبتها[3] و يمكن أن تحلل تحليلا لسانيا وغالبا ما يطلق مفهوم الذخيرة المحوسبة على corpus.
- الذخيرة المصغرّة/ المجزأة: و هي جزء من الذخيرة الكبيرة الضخمة أو الجزء المنتقى منها.
- المعطيات/ الأرشيف: مجموعة من النصوص غير مرتبة ترتيبا لسانيا وتعرف بـــ ثبت بيانات الكلمة.

و قد جمعت مؤلفات DBP المكتوبة رقميا أو إلكترونيا إلى قائمة البيانات للكلمة، بالإضافة إلى ما اشترت DBP حقوق نشره من مؤلفات أو ما تحصلت عليه بالمجّان. أما بقية النصوص فتكتب كتابة إلكترونية أو تنسخ مباشرة بالماسح الإلكتروني (OCR)[4] ثم تقرأ وتدقّق و تحقّق من جديد لتكون هذه النصوص جمعاء معطيات أو أرشيفا. وتخزّن المعطيات

---

[1] انظر http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

[2] كما اتفقنا على اختيار كلمة الذخيرة في الفصل الأول: الذخيرة اللغوية أصولها و معناها.

[3] يقصد بالحوسبة لها مرحلة من التمييز بالعلامات، و عنونة الكلمات بأقسامها بمعايرهما المتبعة، و مرحلة أخرى متقدمة من تحليلها إعرابا و دلاليا و لها التعاليق الخطابية

[4] Optical Character Recognition

في ثبت البيانات لا في ثبت واحد فقط؛ نظرا لضخامة حجمها و تعرف هذه البيانات بـــ الأرشيفات/ المعطيات المجزأة (sub- data base).[١]

و بعد ذلك تُسلك هذه القوائم المتعددة في تقسيمات مختلفة كالكتب و الجرائد، والمجلات و النصوص التراثية أو الحداثية، شعرا و نثرا. و يجدر الإنباه على أن هذه المسارد ليست في ذاتها الذخيرة اللغوية؛ لأنها لم تحلل بعد تحليلا لغويا حاسوبيا. و يمكن توضيح أنواع النصوص المدونة في ذخيرة DBP كالآتي:[٢]

الجدول ١: أنواع النصوص في ذخيرة DBP

نوع النص	التوضيح	النسبة المئوية
Berita: (1994-2004) Berita Harian, Berita Minggu, Harian Metro, Utusan, Harakah ثمانية عشر معطًى	الأخبار الترفيه الاقتصاد الرياضة	52.65%
Buku: أربعة معطيات	أرشيف واحد: كتب الستينات فما دون أرشيفان إلى أربعة: كتب السبعينات وما بعدها	27.62%

[1] انظر http://dbp.gov.my/korpus/korpus_DBP.pdf

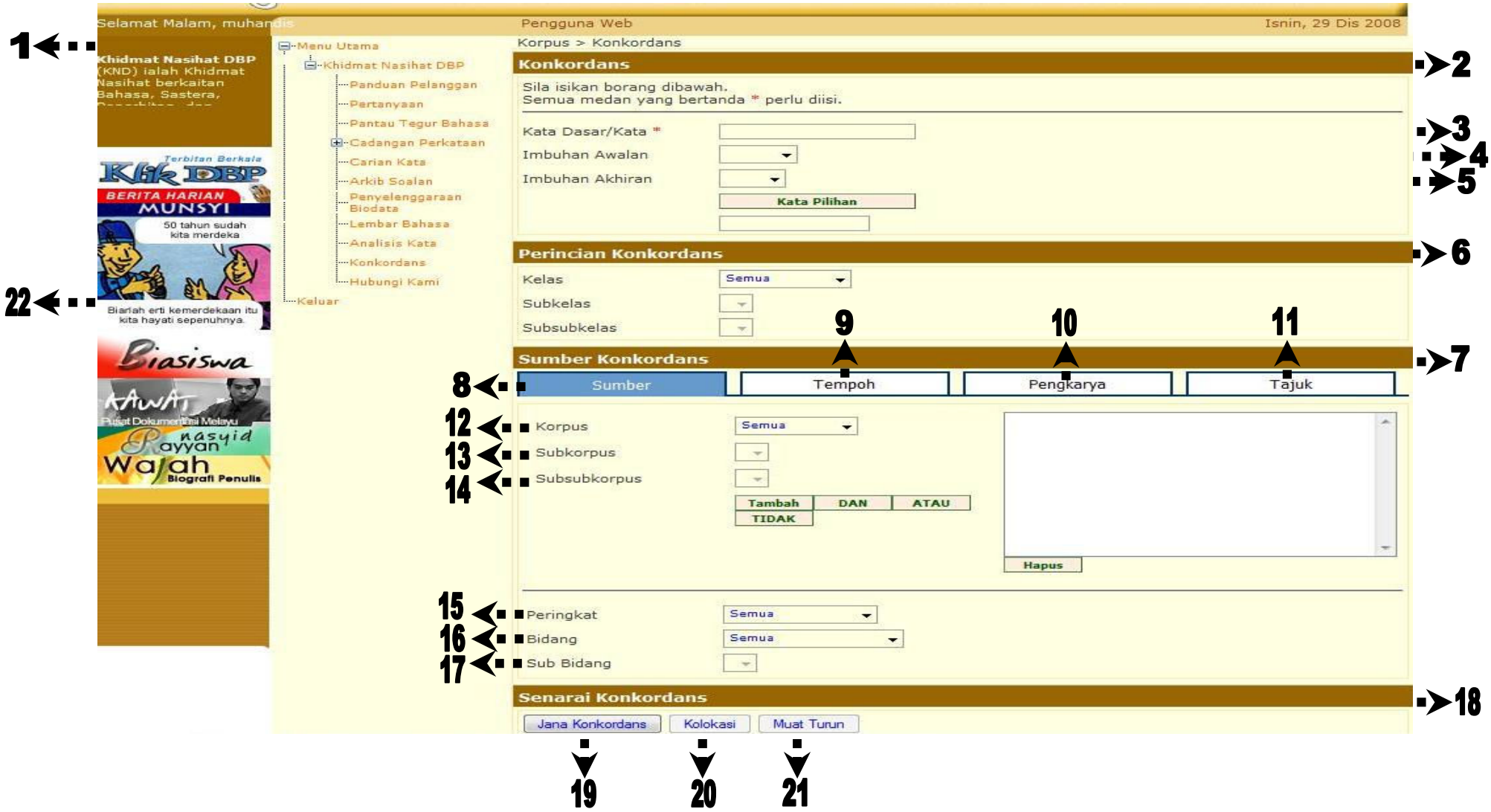
[2] انظر http://dbp.gov.my/korpus/korpus_DBP.pdf

Majalah: أربعة معطيات	المجلات العلمية المجلات غير العلمية المجلات الترفيهية، المجلات النسوية، مجلات الأطفال، المجلات الفكاهية المجلات الرياضة مجلات غيبيات	12.15%
Klasik: معطًى واحد	النصوص التراثية	2.43%
Terjemah: معطًى واحد	النصوص المترجمة إلى الملايوية	1.87%
Sukuan: معطًى واحد	النصوص من لهجات ولايتَي: Sabah & Sarawak	3.28%
Buku Teks: معطًى واحد	الكتب المدرسية المقررة	
Drama: معطًى واحد	المسرحيات	
Efemeral معطًى واحد	الإعلانات و النماذج	
Puisi	النصوص الأدبية	

معطًى واحد		
Kad Bahan معطًى واحد	بطاقات المفردات لتأليف المعجم	

و من المواصفات العامة لذخيرة DBP فهي كالآتية:

الشكل ١: الواجهة الرئيسة لذخيرة DBP



الجدول ٢: توضيح الواجهة الرئيسة لذخيرة DBP

١- **الصفحة الرئيسة:** إرشادات DBP – دليل المستخدم – السؤال – تصحيح لغوي – كلمة مقترحة – البحث عن الكلمة – أرشيف الأسئلة – تنسيق – صفحة لغوية – تحليل الكلمة – الفهرسة – اتصل بنا.

٢- **الفهرسة**
٣- كلمة/ جذر الكلمة
٤- السوابق (be, bel, ber, di, diper, dwi, eka, juru, maha, me, meng, menge, pe, pel, pem, pen, peng, penge, per, se, supra, tata, te, ter)
٥- اللواحق (ah, an, at, i, in, isme, kan, man, nya, wan, wati)

٦- **الفهرسة الدقيقة:** صفة – جذر الكلمة – كلمة أجنبية – مفرد – جمع – فعل – اسم

٧- **مصادر الذخيرة**
٨- المصدر: الأخبار – الكتب – المجلات
٩- زمان النشر: القرن – السنة – التاريخ
١٠- المؤلف/ الكاتب: المؤلف – المترجم – المنسق – المحرر – المُعد – عنوان التأليف – الجنس – القوم
١١- العنوان؛ الناشر: صحيفة Berita Harian – منشورات DBP، نوع التأليف: الأصلية – المقتبسة – المترجمة

العدد: الأول – الثاني – القديم – الجديد

١٢- الذخيرة: الأخبار – الكتب – المجلات – النصوص الأدبية

١٣- الذخيرة المجزأة: نوع الأخبار و الكتب و المجلات و النصوص الأدبية

١٤- مصغرة الذخيرة المجزأة: النوع الدقيق

١٥- فئات المتلقين: الأطفال – الثانوية المتوسطة – الثانوية العالية – الجامعات – العام – الكبار – الصغار

١٦- الموضوع: الآثار – الدين – الاقتصاد – اللغة – الفضاء – التربية – الثقافة – الجغرافيا – الإدارة – الطباعة – التجارة – الزراعة – الصحة – الفلسفة إلخ

١٧- الموضوع المحدد: النوع الدقيق

١٨- ثبت الفهرسة

١٩- تشغيل الفهرسة

٢٠- المتلازمات اللفظية

٢١- التحميل

٢٢- الخروج

و أما لوازم الفهرسة من بحث و فحص في ذخيرة DBP و بعض مواصفاتها للتحليل اللغوي الحاسوبي فيمكن تلخيصها فيما يلي:[١]

١- البحث عن كلمة

عند استصدار كلمة ما تُوسط الذخيرة الكلمة في سياقاتها المتوفرة و هو ما يعرف بـــ KWIC[٢] (الكلمة المفتاحية) داخل السياق. و لنضرب لذلك مثلا بكلمة arab. وتوضح الواجهتان نتائج البحث عن هذه الكلمة.

الشكل ٢: واجهة نتائج "البحث عن كلمة" arab

Senarai Konkordans

Jana Konkordans | Kolokasi | Muat Turun

Isih : Konkordans | Isih Senarai

	Ayat Kiri	Kata	Ayat Kanan
1	... maksudnya: "Dialah yang telah mengutuskan dalam kalangan orang	Arab	yang Ummiyin, seorang Rasul (Muhammad saw) dari bangsa ...
2		Arab	malam tadi membuktikan, mereka memang handal baik dari ...
3	... tidak menghiraukan perisytiharan PBB mengenai hubungannya dengan negara	Arab	
4	Ia adalah rekod baru negara	Arab	
5	Ia dibuat dalam enam bahasa rasmi PBB	Arab	Cina, Inggeris, Perancis, Russia dan Sepanyol.
6	... persetujuan mengenai bekalan air yang dijanjikan kepada negara	Arab	itu mengikut perjanjian damai 1994.
7	"Kata khinzir berasal daripada bahasa	Arab	dan merujuk kepada binatang yang sama dan penggunaannya ...
8	... justeru kita diminta menerima amalan dulu daripada bangsa	Arab	
9	... Biro Luar, Hamid Alrawi, dari Parti Sosialis Baath	Arab	Iraq, kagum dengan keharmonian, perpaduan dan sokongan ahli ...
10	... tumpuan masih kepada mata pelajaran agama dan bahasa	Arab	di samping pelajaran akademik.

1 2 3 4 5 6 7 8 9 10

[1] انظر
http://dbp.gov.my/lamandbp/main.php?Content=vertsections&SubVertSectionID=551&VertSectionID=25&CurLocation=238&IID=&Page=1

[2] Key-Word In Context

الشكل ٣: التابع لواجهة نتائج "البحث عن كلمة" arab

Senarai Konkordans

Jana Konkordans | Kolokasi | Muat Turun

Isih : Konkordans | Isih Senarai

	Ayat Kiri	Kata	Ayat Kanan
11	... terbaik sebelum persiapan menghadapi Piala Dunia di Emiriyah	Arab	Bersatu (UAE) November nanti.
12	... Undang-undang (17), Ilmu Wahyu (15), Kejuruteraan (10), Bahasa	Arab	(6), Bahasa Inggeris (6), Perubatan (4) dan Seni ...
13	... Tsai Chi-Huang), Thailand (Thongchai Jaidee, Thammanoon Sriroj), Emiriyah	Arab	Bersatu (Mark Gregson-Walters, Paul Lightbody).
14	"Saya mempunyai ijazah pendidikan dalam bidang Kesusasteraan	Arab	dari sebuah universiti di Indonesia, tetapi tidak berminat ...
15	... bahasa Inggeris tetapi bahasa lain yang penting seperti	Arab	Jepun, Cina, Jerman dan Perancis," katanya ketika dihubungi ...
16	... suci harus patuh dengan peraturan dan undang-undang kerajaan	Arab	Saudi sebagaimana juga kita mengharapkan orang asing yang ...
17	... negara ini menunaikan ibadah haji kerana tindakan kerajaan	Arab	Saudi itu.
18	Dr Abdul Hamid berkata,	Arab	Saudi sentiasa memantau kuota jemaah haji dan lazimnya ...
19	... (300), Sains Kemanusiaan (375), Bahasa Inggeris (80), Bahasa	Arab	(185), Pengajian Berteraskan Ilmu Wahyu dan Warisan Islam ...
20	... itu, kami sependapat bahawa tarian sufi oleh pemuda	Arab	itu lebih menarik dan mengasyikkan daripada tarian gelek.

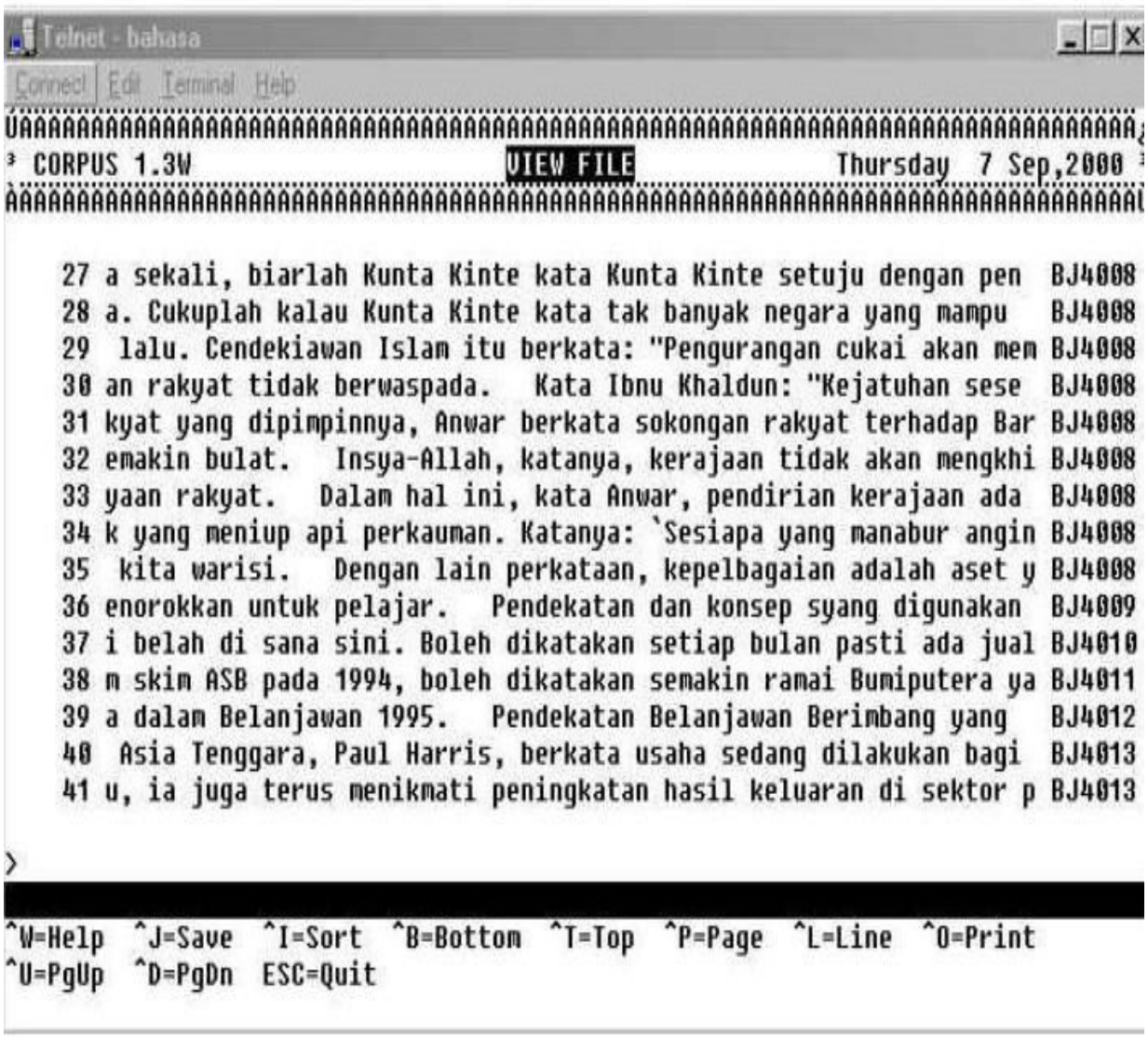
1 2 3 4 5 6 7 8 9 10

٢- البحث برمزي * و ؟

لو أريد البحث عن كلمة ?kata*[1] على سبيل المثال فإن الذخيرة ستعرض هذه الكلمة مجردة كما ستعرضها و قد انضافت إليها السوابق و اللواحق حسب دون عرض الاشتقاقات المختلفة من الجذر kata مثل: kata و perkataan و berkata و ما إلى ذلك.

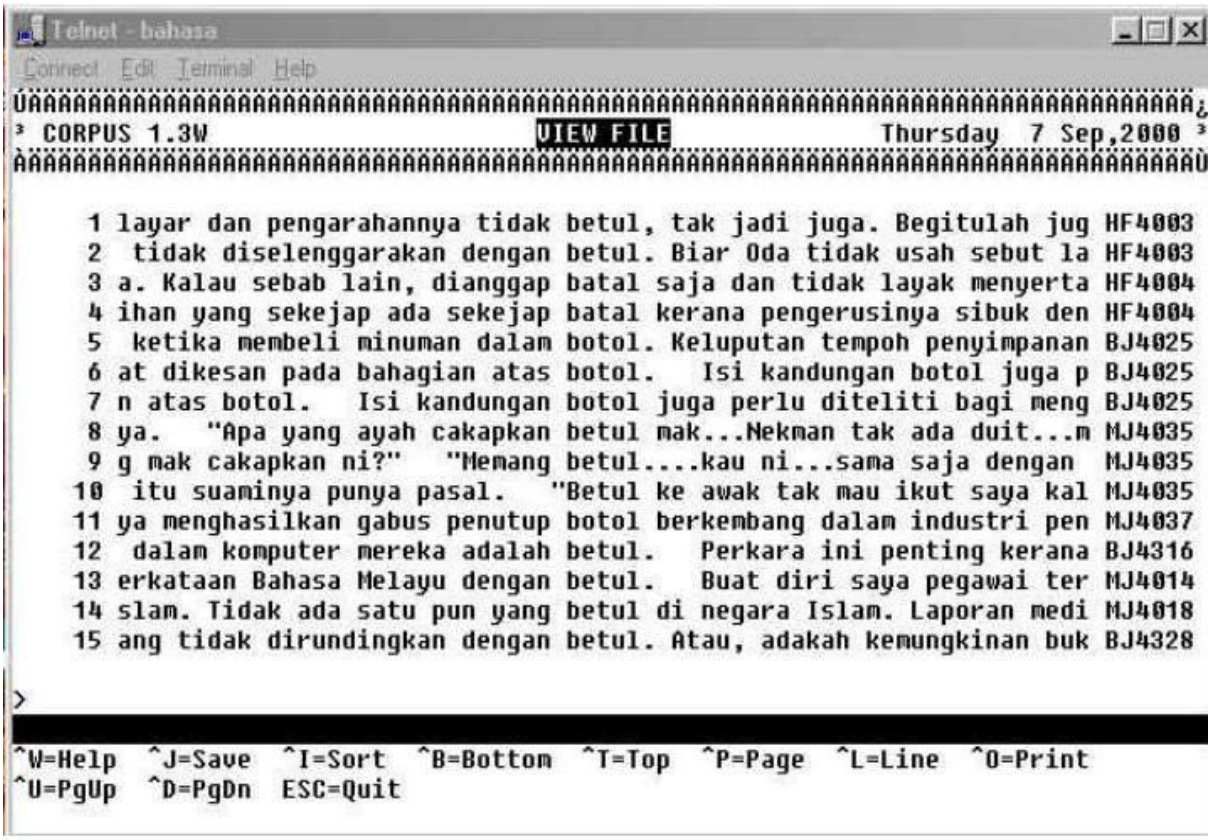
[1] Kata كلمة ملايوية تعني (قول)، perkataan: و معناها كلمة، berkata و تعني يقول/ تقول

الشكل ٤: نتائج البحث برمزي * و ؟ لكلمة kata



و لو أريد البحث عن b?t?l بهذه الصيغة البحثية فإن الذخيرة ستظهر اشتقاقات b t l جميعها عدا السوابق و اللواحق مثل: botol، و batal ، و betul و هلم جرا.

الشكل ٥: نتائج البحث برمزي ? و ? لكلمة b?t?l

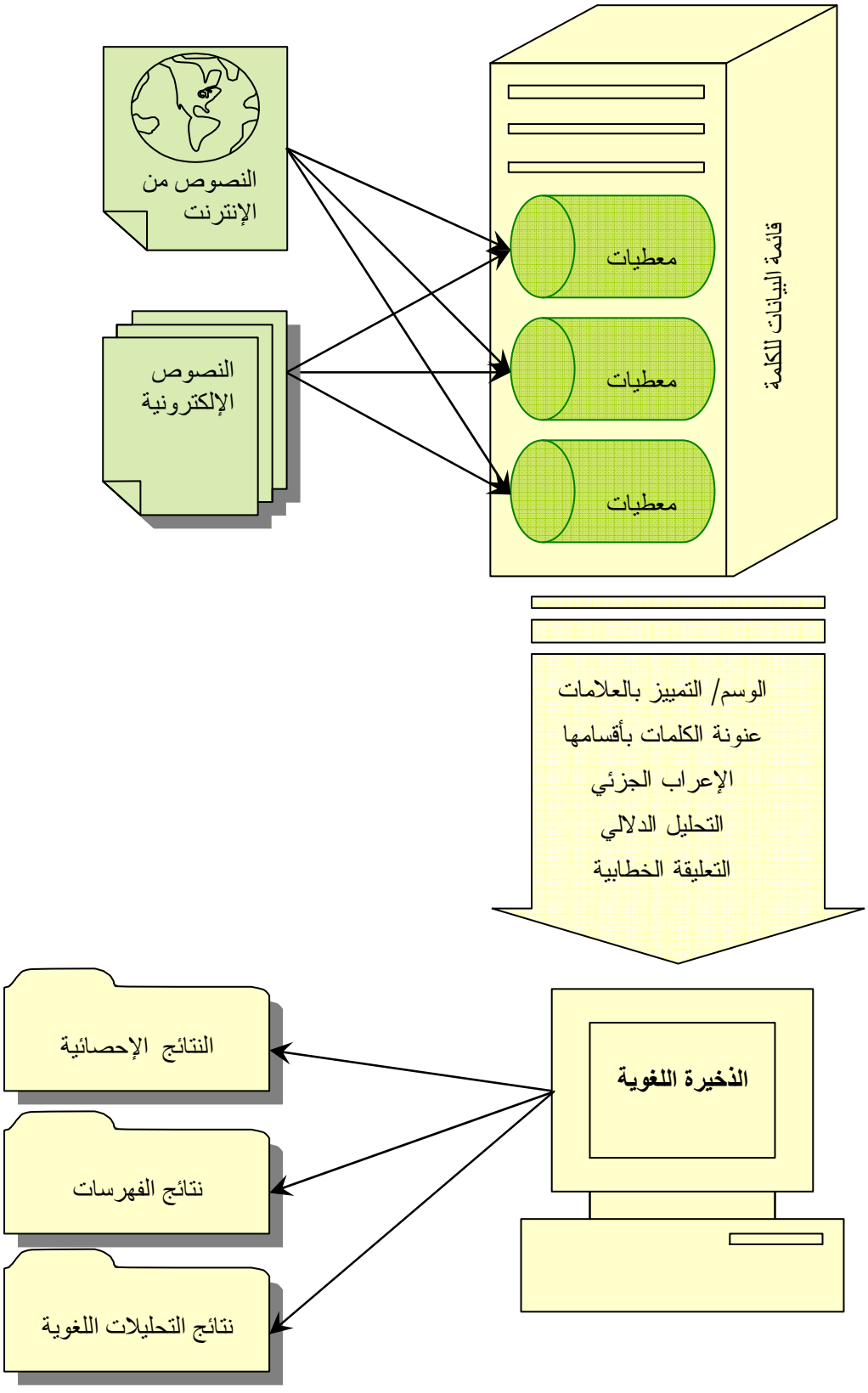


٣- نظام تحليل النص الملايوي MATA (Malay Text Analysis)

يقدم هذا النظام:

- عدد الكلمات (bilangan perkataan)
- نسب ترددها (kekerapan perkataan)
- جذور الألفاظ و ثبت ترددها (bilangan dan senarai kata akar)
- الكلمات الجديدة و ثبت ترددها (bilangan dan senarai kata baru)
- عدد الكلمات الخطأ و ثبتها (bilangan dan senarai kata tak sahih)

الشكل ٦: مخطط بناء ذخيرة DBP للعربية



المبحث الثاني

الأنموذج المقترح

عند البَدء بإعداد ذخيرة لغوية معينة يتوجب على الباحث الاتكاء على مجموعة عامة من المبادئ المتسلسلة و المنظمة عَبْر مراحل محددة أبرزها: جمع المادة و حفظها بصفتها معطيات في ثبت البيانات للكلمة، مع مراعاة مسألة السماح بحقوق النشر والطبع. ثم مرحلة التحليل البدائي التى تعتني بتغيير نوع الملفات الحاسوبية جميعها – في عينة الدراسة المقترحة في الذخيرة – إلى الترميز الموحد أو الشفرة الموحدة أي اليونيكود.

و يأتي التحليل اللساني لاحقا للمرحلة السابقة فيعتني بتحليل النصوص وفق المراحل الخمس المعروفة لتحليل الذخيرة؛ ومنها تمييز الكلمات بالعلامات لميز الأصلِ من الزائد أي تجريد الكلمة من سوابقها و لواحقها، و عنونة الكلمات بأقسامها من حيث عددها وجنسها وصيغها و صولا إلى المرحلة المتقدمة في نتائج تحليل الإحصاء و الفهرسة والتحليلات اللغوية.

و لبناء الذخيرة اللغوية العربية في ماليزيا لا بد من مراعاة العناصر الآتية:

أ- هدف الذخيرة

إن العمل في إطار المعالجة الحاسوبية للغات الطبيعية يرتبط ارتباطا عضويا بالذخيرة اللغوية بصفتها موارد لغوية جمة. و ليس من قبيل المبالغة القول إن من أهم الركائز في مجال معالجة اللغات الطبيعية وجود هذه الذخيرة نفسها لتعطي صورا متكاملة عن اللغة الطبيعية كونها مصدرا شاملا في العملية التعليميّة التعلّميّة. و لأن العربية لغة الدين في ماليزيا فمن المتوقع أن يُقبِل الناس هناك على أي مشروع جديد من شأنه أن يقيم هذه اللغة على ألسنتهم و ينشرها بينهم.

و لا ننكر ما تسعى الذخيرة إلى تحقيقه من غرض اقتصادي يتمثّل في إتاحة تعلم اللغة العربية في ماليزيا ضمن تقنيات متطورة توائم بنية العقل الماليزي و تنسجم معها. ناهيك عن الكلفة المادية الزهيدة مقابل الوصول إلى المعلومات و الحصول عليها.

و ليس خافيا ما يمكن أن تبلغه الذخيرة من مرام أبعد من ذلك على المستوى الاقتصادي؛ فإذا ما انتشرت العربية في ماليزيا جراء تطوير أساليب عرضها و تقديمها إلى المجتمع الماليزي فإن ذلك، لا شك، سيسهم في انفتاح العلاقات الاقتصادية بين ماليزيا والعرب بوجه عام.

أما على صعيد علم اللسانيات و تقدم تكنولوجيا اللغة و هندستها، فإن وجود الذخيرة العربية في ماليزيا سيسهم بشكل أو بآخر، في تطوير ما يلي:

١- مصادر المعرفة

- إنشاء الذخيرة المتوازية (Parallel Corpus) بين الملايوية والعربية
- إنشاء الذخيرة التربوية (Pedagogic Corpus) التى تخدم المتعلمين بوجه خاص لتحسين معارفهم في اللغة و إغناء مهاراتهم التعلمية و برامجهم التعليمية.
- إنشاء قواميس إلكترونية، أحادية اللغة (العربية)، أو متعددة اللغات (الملايوية – الإنجليزية – العربية).
- تأليف القواميس القابلة للقراءة الآلية MRD (Machine-Readable Dictionary).
- بناء مصادر القواعد أحادية اللغة (العربية) أو متعددة اللغات (الملايوية والعربية والإنجليزية)

٢- آليات اللغة (Language Tools)

- نظام معالجة النصوص كالمفهرسات
- نظام المصحح الإملائي
- المحلل الصرفي
- المعرب النحوي

٣- أنظمة الترجمة

- نظام الترجمة الآلية
- نظام الموارد اللغوية

٤- التخاطب مع الآلة

- نظام OCR[1] (نظام تعرف الحروف ضوئيا)
- استنطاق النصوص
- تحويل المكتوب منطوقا (text to speech)
- نظام التعلم المبني على الشبكة الإلكترونية

ب- الفئة المستهدفة من المستخدمين

يقوم بناء الذخيرة العربية في ماليزيا على أساس تعليم اللغة بمساعدة الحاسوب (CALL)[2] بالدرجة الأولى. و لا يجمل بنا تجاهل ما ستقدمه هذه الذخيرة من فوائد جمة على صعيد تسهيل البحث و سرعة الوصول إلى المعلومة. و غني عن الذكر بيان أهمية CALL في عملية التعليم و التعلم.[3] و لا شك في أن الفرص سانحة الآن في دولة مثل ماليزيا التى

---

[1] Optical Character Recognition

٢- تعليم اللغة بمساعدة الحاسب الآلي (Computer-Assisted Language Learning, CALL) هو أحدُ فروع حقل التعليم بمساعدة الحاسب الآلي (Computer-Assisted Learning). ويعد التدريس بمساعدة هذا الحاسب وكذلك تعليم اللغات بمساعدته تطبيقين من تطبيقات علم الذكاء الاصطناعي (Artificial Intelligence)، وهو العلم الذي يسعى لجعل الآلة تقوم بما يقوم به البشر.

[3] راجع

- Michael Fullan, 1992, The Meaning of Educational Change, Published by OISE Press/Ontario Institute for Studies in Education.
- Arif Karkhi Abu Khudairi, 1993, Teknik Moden Dalam Pengajaran Bahasa Arab Kepada Bukan Arab
- James B. Ellsworth, 2000, Surviving Change: A Survey of Educational Change Models, Educational Research Information Center (U.S)

تشجع أي مشروع يسهم في تطوير التعليم لاسيما تلك المشاريع التى يؤمّل منها أن تُحدث نقلة في تفعيل مهارات الطلبة اللغوية و تعزيزها.

و من المتوقع أن يعي كل من الطلبة – في جميع مراحلهم الدراسية – و المثقفين المهتمين باللغة: تعليمها و تعلمها، أن الذخيرة اللغوية مادة لا تتسم بالأهمية و الإفادة فحسب بل تتجاوزهما إلى حد الإمتاع أيضا؛ فهي تمكّن المستخدم من اعتماد صحة الكلمة بناء على شيوعها وفق السياقات التى توفرها الذخيرة، كما تمكنه من اختيار المقالات المناسبة في المقامات المختلفة. و لا يغيب عن ذهنهم أيضا كيف يمكن للسياق أن يغير معنى الكلمة المتعارف عليه بما يعرف بتلازم اللفظ.

و في المستوى المتقدم يمكن للمستخدم تعرّف قواعد اللغة بتنوعها و تشعبها وكيفيات استخدامها في الجملة. و من اللافت للنظر أن الذخيرة اللغوية تضع العملية التعليمية التعلمية في متناول المستخدمين بين أصابعهم وأمام أعينهم على شاشات الحاسوب.

ج- اختيار نصوص معينة

إذا كان لنا أن نتصور ذخيرة كصحيفة مثلا، فيمكننا القول إن أية عبارة مسطّرة على ورق الصحيفة هي نصوص لغوية مكتوبة. ثم نقوم باختيار النصوص وفق مسوغات خاصة

- Mary Ann Fitzgerald, Michael Orey and Robert Maribe Branch, 2002, Educational Media and Technology Yearbook 2002.
- JM Cooper, 2005, Classroom Teaching Skills (8th ed., pp. 151-184). Boston: Houghton-Mifflin. Tomlinson.
- Mona & Nor Azilah, 2007. Kajian Keperluan Latihan dalam Pengajaran ICT di Kalangan Guru-Guru.

تبعا لأهدافنا في الانتقاء و الجمع؛ ذلك أنّ حجم الذخيرة يعتمد على نوعية النصوص التى استخدمناها، و كيفية اختيارنا لها، وفق النسب التى خصصناها. و إذا ما قسمنا هذه النصوص إلى عدة أقسام مثل الأخبار السياسية والمقالات الأدبية والنشرات الجوية فيمكننا تحديد القسم الذي نريد استخدامه في التحليل.

د- تحديد نوع الذخيرة و حجمها

ينبغي مراعاة هذا العنصر في بناء الذخيرة العربية في ماليزيا، و يرتكز هذا العنصر على ثلاثة أسئلة هي: ما نوع النصوص التى نريدها؟، وما حجمها؟، و كم يتوافر منها؟ ولقد رسمنا سلفا بضعة أهداف لوجود هذه الذخيرة في ماليزيا، كما أننا حددنا الفئات المستهدفة من المستخدمين، و كذلك اخترنا نصوصا متوافرة – إلى الآن – ينبغي الالتزام بها؛ و لذلك نرى من الضروري أن تتسم الذخيرة بما يلي:

- الاعتماد على النصوص المكتوبة حسب دون المنطوقة.
- الاعتماد على النصوص الحديثة (ما بعد القرن التاسع عشر الميلادي) إلا كمّا يسيرا من نصوص قديمة أو مخطوطات.
- ذخيرة دون تعاليق (raw corpus/ unannotated).
- متاحة أمام مستخدمي الإنترنت.

هـ – إدارة معطيات الذخيرة

من الممكن الإفادة من ذخيرة DBP الماليزية في تنظيم المعطيات وترتيبها. ولعلّ الخطة الصورية يمكن رسمها هنا كما يلي:

الجدول ٣: إدارة معطيات ذخيرة DBP العربية

- **الفهرسة:** الكلمة
- السوابق
- اللواحق

- **الفهرسة الدقيقة:** اسم – فعل – حرف – مفرد – مثنى – جمع – مذكر – مؤنث
- مصادر الذخيرة: صحف و مجلات – إجابات الطلبة عن الاختبارات – مواقع إنترنت – كتب مدرسية – رسائل أو بحوث جامعية – قصص شعبية مترجمة – نصوص تراثية مخطوطة – قواميس
- زمان النشر: ما قبل القرن التاسع عشر الميلادي – ما بعد القرن التاسع عشر الميلادي
- المؤلف: الكاتب – المترجم – المنسق – المحرر – المعد – الجنسية – الجنس – الطلبة – الأساتذة
- أنواع التآليف: أصلية – مقتبسة – مترجمة

- **فئات المتلقين:** الأطفال – الثانوية المتوسطة – الثانوية العالية – الجامعات – العام – الكبار – الصغار

- **الموضوع:** الآثار – الدين – الاقتصاد – اللغة – الفضاء – التربية – الثقافة – الجغرافيا – الإدارة – الطباعة – التجارة – الزراعة – الصحة – الفلسفة

- **ثبت الفهرسة**
- تشغيل الفهرسة
- المتلازمات اللفظية
- الإحصاء
- ترتيب الكلمة وفق شيوعها
- التحميل

و- مواجهة مشكلة حقوق الطبع و النشر

لا تواجه عملية بناء الذخيرة اللغوية العربية في ماليزيا أية إشكالات مع الجهات المعنية من مؤسسات ماليزية حكومية و غير حكومية على صعيد إتاحة ما تملكه من مواد الذخيرة المطلوبة أمام هذا المشروع.

ز- التحليل البدائي للنصوص (تحويل رمز/ امتداد أنواع الملفات جميعها إلى Unicode)

إن انضمام الذخيرة العربية في ماليزيا إلى ذخيرة DBP الماليزية من الوجهة التقنية ممكن جدا؛ كونها ذخيرة وطنية تجمع مصادر شتى بحيث يضاف إلى ثبت بياناتها معطيات جديدة هي نصوص اللغة العربية.[1] و ذلك لعدة أسباب منها: الإفادة من التقنية المطورة لذخيرة DBP، ومطاوعة نظام هذه الذخيرة للتعامل مع معظم أنظمة اللغات في العالم باعتماده نظام الترميز العالمي Unicode. وإذا كانت الذخيرة العربية في ماليزيا في بدئها مستقلة كونها محاولة تجريبية رائدة إلا أنها تبقى جهدا فرديا متواضعا يتطلع بأمل و طموح عريضين إلى دعم مادي و تقني فلا بد أن تحتضن من مؤسسات كبيرة مثل المجمع اللغوي الماليزي.

و لا ننسى وفرة البرامج المعدَّة لخدمة الذخيرة العربية مثل Monoconc، وWordsmith Pro، و Xiara، و على رأسها aConcorde و على ما في هذه البرامج من قصور كنا قد أشرنا إليه[2] إلا أن المأمول أن تخضع هذه البرامج لعمليات التطوير والتحديث بما يؤدي إلى بناء الذخيرة اللغوية العربية في شكلها الأمثل.

و ستقتصر إمكانية هذه الذخيرة على فهرسة الكلمات العربية وفق ترتيب قياسي للذخيرة (KWIC)، كما يمكن أن تخضع لعملية البحث عن الكلمة، و البحث بالرموز. وكذلك يمكن سرد الكلمة المبحوث عنها كما هو في السياق. إضافة إلى نتائج إحصائية أساسية في

[1] و قد وصفنا نموذج الذخيرة DBP الماليزية في المبحث السابق. انظر صفحة ٦١-٧٤ من البحث.
[2] انظر ص ٤٢-٤٩ من هذا البحث

الذخيرة هي مجموعة الكلمات للنصوص المراد تحليلها، وتردد نسبها المئوية، ورصد أكثر الكلمات أو الحرف شيوعا.

ه- التحليل اللساني للذخيرة في مراحلها الخمس[١]

المأمول أن تدرس برامج تحليل الذخيرة دراسة تقنية دقيقة بدعم مؤسسي ماليزي لا سيما من المجمع اللغوي الماليزي (DBP) و نذكر منها على سبيل المثال لا الحصر البرامج التى تقدمها الشركات: RDI، و CIMOS، و XEROX، وSYSTRAN، وIBM، وIMAGiNET، و شركة صخر العربية. و من المؤكد أن تقدّم هذه البرامج وتطورها سيجيب عن أسئلة لسانية حاسوبية كالنقل الحرفي (Diacritization)، و عنونة الكلمات (Part of Speech Tagging)، و التحليل الصرفي، و المعرب، و التصحيح والشكل الآليين، و استنطاق النصوص و إلخ.

كما يمكن استثمار تجارب بضع مؤسسات علمية و تجارية في تحليل الذخيرة العربية؛ نذكرها هنا على سبيل المثال لا الحصر وهي: CLARA (Corpus Linguae Arabicae)،[٢] و The Penn Arabic Treebank،[٣] و Prague Arabic Dependency Treebank.[٤] فالمؤسسات الثلاث السابقة أصدرت كل منها ذخيرة عربية مطورة و متقدمة

[1] سبق أن ذكرنا عن هذه المراحل. انظر ص ٣٨-٤١ من هذا البحث

[2] راجع http://www.ilc.pi.cnr.it/

[3] راجع www.ircs.upenn.edu/arabic/

[4] راجع http://ufal.mff.cuni.cz/padt/PADT_1.0/docs/index.html

في عنونة الكلمات و التحليل الصرفي و النحوي. و من المنطقي أن هذا الاستثمار يتطلب التفاهم العميق في منظورين اثنين: المنظور اللساني أو اللغوي، والمنظور الحاسوبي.

المبحث الثالث

معطيات الذخيرة

استقصى الباحث مصادر المواد اللغوية العربية في ماليزيا فوجدها تقتصر على ما يلي:

الجدول ٤: معطيات الذخيرة

نوع النص	التوضيح
الصحيفة • صحيفة إلكترونية لوكالة برناما (Bernama)[١] • صحيفة "أهلا"[٢]	• لديها أرشيف أسبوعي في موقع إلكتروني يمكن تصفحه و قد بدأت في تقديم خدماتها باللغة العربية منذ عام ٢٠٠٤. • غير متوافرة إلا أن نسخها تتوافر إلكترونيا و يتيسر الحصول عليها من إدارة الصحيفة.
المجلة • الأسواق[١]	• نشرت أول عدد لها في ديسمبير ٢٠٠٨، و قد أطلعني مدير المجلة على مسودة النسخة الإلكترونية بهذا العدد.

[1] أنشأت وكالة الأنباء الوطنية الماليزية (برناما) أكتوبر في ٢٠٠٤ شعبة الخدمة العربية و تعنى هذه الشعبة بترجمة أبرز الأخبار و الأحداث المهمة محلية و دولية إلى اللغة العربية، بالإضافة إلى المقالات و المعلومات عن نظم البنوك و التأمينات. انظر http://www.bernama.com/arabic/v2/

[2] صحيفة نصف شهرية تأسست في اليوم الأول من شهر يوليو ٢٠٠٥، وكانت في بادئ أمرها تهتم بالشؤون السياحية في ماليزيا ثم توسعت لتشمل الأخبار السياسية وأحداث الشرق الأوسط، وقضايا العالم الإسلامي والحياة، إضافة إلى المنتدى الثقافي لرواد الفكر والأدب، و محبي اللغة العربية، فضلا عن شؤون الأسرة، والرياضة، والفن.

إنشاء الطلبة • أوراق الامتحانات في الجامعات و المعاهد	• تدرّس العربية بوصفها مادة تخصصية في عدة جامعات و منها: جامعة ملايا (UM)، و الجامعة الإسلامية العالمية الماليزية (UIAM)، و الجامعة الوطنية (UKM)، و جامعة Putra (UPM)، و جامعة العلوم الإسلامية الماليزية (USIM)، و غيرها. و يقترح أن تجمع إنشاءات الطلبة ثم تكتب إلكترونيا.
مواقع الإنترنت • موقع JAKIM [2] • مواقع المناسبات	• تُرجمت أهم محتوياته إلى العربية • صممت بعض المواقع الرسمية المعنية بالمناسبات العامة كالندوات و المؤتمرات و الملتقيات العالمية مثل: المؤتمر العالمي للحضارة الإسلامية[٣]، و الدورة العاشرة لمؤتمر القمة الإسلامي[٤]
كتب المدرسة • الابتدائية: في ستة كتب • الثانوية: في خمسة كتب	• العربية: في ستة مستويات • اللغة العربية: الاتصالية في ثلاثة مستويات، و اللغة

[1] مجلة شهرية اقتصادية صدَر أول عدد منها في ديسمبير ٢٠٠٨ و تُعنى بالأخبار و الإعلانات الاقتصادية محررة باللغة العربية، حتى يتسنى للمستثمرين العرب الاطلاع عليها. و قد طبع العدد الأول منها و نشر في كل من كوالا لمبور، و دبي، و قطر، و بحرين.

[2] مصلحة الشؤون الإسلامية الماليزية (JAKIM) هي هيئة حكومية ماليزية تهتم بالشؤون و المقدسات الإسلامية. و مهمتها تطوير المجتمع و ترقيته أخلاقيا على أسس إسلامية متينة. وهي تتفق مع تصور الحكومة الماليزية في بناء مؤسسة تقوم بإدارة العمل الإسلامي على وجه فعال ومؤثر. انظر http://www.islam.gov.my/arabic/index.html

[3] انظر http://www.kiperak.edu.my/ISCIS/ISCIS%20AR/mukadimah_ar.html

[4] انظر http://www.oic-oci.org/oicnew/arabic/conf/is/10/IS10-main-ar.htm

• الثانوية العالية: في كتاب واحد	العربية العالية في مستويين • الأدب و البلاغة لعثمان خالد ويمكن الحصول على النسخة الإلكترونية هذه الكتب جميعها بالتعاون مع وزارة التربية الماليزية.
رسائل و بحوث الجامعية	• رسائل جامعية في مرحلتي الماجستير و الدكتورة
الترجمة	• ترجمت إلى العربية أكثر من ٢١ سلسلة من القصص الشعبية الماليزية (cerita rakyat) نشرها معهد الترجمة الوطنية بماليزيا ITNMB[1] سنة ٢٠٠٦. أعدت لتكون من ضمن المواد المساعدة في تعليم اللغة العربية بالمدارس الحكومية و خاصة بعد إنشاء برنامج JQAF.[2]
النصوص التراثية	• سجلت في المكتبة الوطنية الماليزية (PNM)[3] ١٠٦ مخطوطات عربية عثر عليها في ماليزيا جلها في العلوم الإسلامية و بعضها في اللغة العربية ككتب القواعد و القواميس.
القواميس	• قاموس اللغة الملايوية – اللغة العربية – اللغة الملايوية من تأليف عبد الرؤوف، و عبد الحليم و خير الأمين.

[1] Institut Terjemahan Negara Malaysia Berhad

[2] و هو برنامج تربوي يهدف إلى تمكين الطلبة الإعدادية لأربع مواد و هي J: لكتابة الجاوية (الكتابة العربية للغة الملايووية)، Q: القرآن، A: العربية، و F: فرض عين.

[3] Perpustakaan Negara Malaysia

الخاتمة

سعت هذه الدراسة إلى رسم نموذج لبناء الذخيرة اللغوية العربية في ماليزيا، وذلك بتمثل بعض الذخائر اللغوية المتداولة عربية و غير عربية و قد اعتمدت الدراسة ذخيرة DBP الماليزية نموذجا.

و من هذا المنطلق، جاءت الدراسة في فصول ثلاثة؛ تحدث الفصل الأول عن الذخيرة اللغوية، معناها و أصولها، و أنواعها، و نماذجها عربية و غير عربية. و أما الفصل الثاني فعرض إلى بعض تحديات التقنية في برامج الذخيرة و آليتها (الفهرسة)، و ما هو المنتظر منها. و أما الفصل الثالث، فجاء عمّا يمكن الاستفادة منه من ذخيرة DBP الماليزية في رسم خطة الذخيرة العربية في ماليزيا.

و ثمة إشكالات كثيرة صادفت هذه الدراسة في سعيها إلى رسم النموذج المقترح للذخيرة العربية في ماليزيا و هي:

- أن جل المصادر و المراجع في لسانيات الذخيرة مكتوبة بالإنجليزية فمن الطبيعي جدا أنها تخدم اللغة الإنجليزية أولا من حيث النظرية و التطبيق. وهذا واضح تمام الوضوح من خلال إمكانية برامج الفهرسة من تحليل العربية.

- عانت الدراسة كثيرا من مصطلحات دقيقة في الإنجليزية عن علم اللسانيات الذخيرية إذ قابلتها مصطلحات عربية متعددة مترجمة و معربة.
- أن التعامل مع برامج فهرسة الذخيرة نظريا و تطبيقيا لَعَمَلٌ رائق على ما فيه من المشقة، إذ هي محتاجة على الدوام إلى جهد بينيّ مشترك (لغوي–حاسوبي) في سبيل تطوير برامج الفهرسة اللغوية توصيفا و تحليلا و تفسيرا.

و قد تبنت هذه الدراسة بعض النتائج و التوصيات أبرزها:

- ضرورة تدوين النصوص العربية المنشورة في ماليزيا لتصبح مصدرا إلكترونيا جامعا أي ذخيرة لغوية يسهل تناولها لدى المهتمين بالعربية على اختلاف مستوياتهم من متعلمين و متخصصين.
- يمكن أن تنضم هذه الذخيرة إلى ذخيرة DBP الماليزية في المستقبل بصفتها معطيات واحدة تنضاف إلى قوائم المعطيات الماليزية الموجودة.
- الحاجة إلى برنامج مطوّر للذخيرة العربية ليخدم عملية تحليل العربية في مستوياتها المختلفة و لا شك في حجم الإفادة المتهيئة من التقنيات الماليزية المتقدمة .
- الحاجة إلى منهج جديد معتمد على الذخيرة اللغوية في الدراسات اللغوية لحل مشكلة تعلم العربية لدى المتعلمين الماليزيين، و المأمول أن تعتمد هذه الذخيرة في تأليف كتب المدارس و المعاهد و الجامعات المقررة أو تأليف القواميس و المعاجم ذات الصلة.

- الحاجة إلى تبني مشروع الذخيرة من جهة حكومية معتمدة؛ ذلك أن هذا المشروع هو مشروع ريادي طَموح يتجاوز الأعمال الفردية إلى ضرورة تكاتف فريق عمل كبير قد يحتاج لإنجازه إلى أكثر من جنسية و بطبيعة الحال إلى ميزانية تليق بمستوى هذا المشروع و هذا التحدي.

و لا يخفى أن هذه الخاتمة تطل على أفق منداح لا يوازيه في بعد المرامي غير ما فيه من العراقيل و العقبات. و ليس ما جاء في هذه الدراسة من توصيف نظري إلا المشروع في شِقّه الأول؛ إذ تترقب هذه الدراسة تمامها في الجانب التطبيقي بالإفادة من نظم البرمجيات المتقدمة في ماليزيا لبناء الذخيرة اللغوية العربية و هو المشروع الذي أعود إلى بلدي حاملا إنجازه على عاتقي، إن شاء الله .

و آخر دعوانا أنِ الحمد لله رب العالمين.

المصادر و المراجع

الكتب

- أحمد مختار عمر و آخرون، (٢٠٠٠)، **المكنز الكبير: معجم لغوي مهني متخصص للمترادفات**، شركة سطوو، ط ١.
- حداد، إ. و، (١٩٨٨)، **المعجم الحديث لمصطلحات الكمبيوتر والمعلوماتية**، (إنجليزي – عربي)، مكتبة لبنان.
- الزهيري، نبيل، (٢٠٠٣)، **قاموس مصطلحات المعلوماتية و اللغويات الحسابية**، مكتبة لبنان ناشرون.
- صيني محمود إسماعيل، و عبد العزيزناصف مصطفى، و أحمد سليمان مصطفى، (١٩٩٣)، **المكنز العربي المعاصر**، مكتبة لبنان ناشرون، ط ١.
- العناتي، وليد و الجبر، خالد، (٢٠٠٧)، **دليل الباحث إلى اللسانيات الحاسوبية العربية**، دار جرير، عمان.
- العناتي، وليد، و برهومة، عيسى، (٢٠٠٧)، **اللغة العربية و أسئلة العصر**، ط ١، دار الشروق للنشر و التوزيع.
- قنديلجي، عامر إبراهيم، (٢٠٠٣)، **المعجم الموسوعي لتكنولوجيا المعلومات و الإنترنت**، دار المسيرة للنشر و الطباعة، ط ١.
- كوزيك، (١٩٨٨)، **معجم مصطلحات المعلوماتية و الحاسبات، الإلكترونية و الآلية** (إنجليزي – عربي)، مكتبة لبنان.
- الكيلاني تيسير و مازن، (١٩٨٧)، **معجم الكيلاني لمصطلحات الحاسب الإلكتروني** (إنجليزي – عربي)، مكتبة لبنان.
- مركز الحاسب الآلي، مجمع اللغة العربية بالقاهرة، (١٩٩٥)، **معجم الحاسبات**، الهيئة العامة لشؤون المطابع الأميرية.
- الموسى، نهاد، (٢٠٠٠)، **العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية**، ط ١، المؤسسة العربية للدراسات و النشر، بيروت.

- Charles F. Meyer, (2002), **English Corpus Linguistic: An Introduction**, Cambridge University Press, First Published.
- David Crystal, (2003). **The Cambridge Encyclopedia Of The English Language**, 2nd Edition. Cambridge University Press.

- Igor A. Bolshakov and Alaxender Gelbukh, (2004). **Computational Linguistics, Models, Resources, Applications**, D.R. Instituto Politecnico Nacional.
- Jan Svartvik, (1992). **Directions in Corpus Linguistics**: **Proceedings of Nobel Symposium 82 Stockholm**, Mouton de Gruyter.
- Jeanne McCarten, (2007). **Teaching Vocabulary, Lesson from the Corpus, Lesson for the Classsroom**, Cambridge University Press, First Published.
- John Sinclair, (1991). **Corpus, Concordance, Collocation**, Oxford University Press, First Published.
- Karin Aijmer and Altenberg, (1991). **English Corpus Linguistic, Studies in Honour of Jan Svartvik,** Longman London Group UK Limited, First Published.
- Michael Barlow**, (**2000). **Concordancing with MonoConc Pro 2.0**, Athelstan.
- Michael McCarthy, (2006). **Explorations in Corpus Linguistics**, Cambridge University Press, First Published.
- Michael Pearce, (2007). **The Routledge Dictionary of English Language Studies**, Published by Routledge, 2007.
- Mike Scott, (1998). **WordSmith Tools Manual version *3.0***, Oxford University Press.
- Mitkov and Ruslan, (2003). **The Oxford Handbook of Computational Linguistic**, Oxford Press Limited, First Published.
- Paul Baker, Andrew Hardie and Tony McEnery, (2006). **A Glossary of Corpus Linguistics**, Edinburgh University Press.
- Tony McEnery and Andrew Wilson, (1996). **Corpus Linguistics**, Edinburgh University Press.
- Tony McEnery, Richard Xiao and Yukio Tono, (2006). **Corpus-Based Language Studies, an Advanced Resource Book**, Routledge Taylor & Francis Group.
- Wolfgang Teubert & Ramesh Krishnamurthy, (2007). **Corpus Linguistics: Critical Concepts in Linguistics,** Routledge Taylor & Francis Group, First Published.

البحوث و المقالات و الدوريات

- بوشعيب راغين، (٢٠٠٧) الحوسبة التوليدية للصر العربي، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- الثبيتي، عبد المحسن بن عبيد، (٢٠٠٧) استخدام ذخائر النصوص لاستخلاص المصطلحات المتخصصة، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- الزامل، عبد الله بن عبد الرحمن ، (٢٠٠٧) العلاقة الصرفية بين الجذور و الأوزان (تصنيف جديد لجذور اللغة العربية)، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- سويف منصر، (٢٠٠٦) من مشاريع الخطاب اللساني المغاربي الذخيرة اللغوية العربية أنموذجا، **مجلة اللسانيات و اللغة العربية.**
- عبد الرحمن الحاج صالح، نماذج من البحث العلمي الخاص باللغة العربية لمواجهة تحديات العصر.
- **http://www.isesco.org.ma/arabe/publications/Langue_arabe/p12.php**
- العناتي، وليد، (٢٠٠٥) الدليل نحو بناء قاعدة بيانات للسانيات الحاسوبية العربية، **ندوة تقنية المعلومات و العلوم الشرعية و العربية.**
- العناتي، وليد، (٢٠٠٥) اللسانيات الحاسوبية العربية (المفهوم، التطبيقات، الجدوى)، **مجلة الزرقاء للبحوث و الدراسات، المجلد السابع، العدد الثاني.**
- غازي، عز الدين، (٢٠٠٧) قواعد المعطيات المعرفية للمصطلحات العربية، مشروع مقترح، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- الغامدي، منصور بن محمد ، (2006) تصميم رموز حاسوبية لتمثل ألفبائية صوتية دولية تعتمد على الحرف العربي، ٦٤ - **مجلة جامعة الملك عبد العزيز :العلوم الهندسية.**
- الغامدي، منصور بن محمد، و آخرون، (2006) نظام حاسوبي لتشكيل النص العربي، التقرير الفني النهائي، **http://www.mghamdi.com/ATD.pdf**
- الغامدي، منصور بن محمد، (٢٠٠٧) مساهمة اللغويين العرب في مشاريع معهد بحوث الحاسب والإلكترونيات، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- الغامدي، منصور بن محمد، إبراهيم بن عبد الله الخراشي و عماد بن عبد الرحمن الصغير، (٢٠٠٧) جهود معهد بحوث الحاسب و الإلكترونيات في بحوث اللغة العربية، **الندوة الدولية الأولى عن الحاسب و اللغة العربية.**
- القحطاني سعد بن هادي، (٢٠٠٥) تحليل اللغة العربية بواسطة الحاسوب، **مجلة مجمع اللغة العربية الأردني، العدد ٦٨.**

- مكتب تنسيق التعريب، **مجلة اللسان العربي، (٢٠٠٠) المنظمة العربية للتربية و الثقافة والعلوم، العدد ٥٢.**

- Abduelbaset Goweder, Anne De Roeck, **Assessment of a Significant Arabic Corpus,** Department of Computer Science, University of Essex, Wivenhoe Park, Colchester CO4 3SQ.
- Ahmed Alabdali and Jim Cowie, (2005). **Regional Corpus of Modern Arabic Standard**, Conference sur les Technologies du Langage dans la Societr de l'Information.
- Andrew Roberts, Latifa Al-Sulaiti and Eric Atwell**, aConCorde: Towards an Open-Source, Extendable Concordancer for Arabic, Corpora Vol. 1 (1).**
- Andrius Utka, (2004). Phases Of Translation Corpus Compilation And Analysis,. **International Journal of Corpus,** John Benjamins Publishing Company.
- Azhar Bin Muhammad, (2005). Beberapa Aspek Keunikan Dan Keistimewaan Bahasa Arab Sebagai Bahasa Al-Quran, **Jurnal Teknologi, Universiti Teknologi Malaysia, 42(E) Jun.**
- Bassam Haddad, (2007). 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.**
- Bente Maegaard, Khalid Choukri, Chafik Mokbel and Mustafa Yaseen, (2005). **Language Technology for Arabic, NEMLAR**, **Center for Sprogteknologi**. University of Copenhagen, Denmark.
- Daniel Wiechmann and Stefan Fuhs, (2006). **Concordancing Software Corpus**, Linguistics and Linguistic Theory 2-1, Walter de Gruyter.
- Daniel Wiechmann and Stefan Fuhs, (2006). **Corpus Linguistic Theory. Volume 2, Issue 1**.
- David Coniam, (2004). Concordancing Oneself, Constructing Individual Textual Profiles, **International Journal of Corpus Linguistics 9:2, John Benjamins Publishing Company**.
- Eric Atwell, Latifa Al-Sulaiti, Saleh Al-Osaimi and Bayan Abu Shawar, (2004). **A Review of Arabic Corpus Analysis Tools**, JEP-TALN 2004, Arabic Language Processing, Fez.
- Eric Atwell, Latifa Al-Sulaiti, Saleh Al-Osaimi, Bayan Abu Shawar, (2004). **Arabic Language Processing**, JEP-TALN 2004, Fez.
- Frank Richter, **Introduction to Computational Linguistics**, (2005). Seminar fur Sprachwissenschaft, Eberhard-Karls-Universitat Tubingen, Germany,
- Gerry Knowles and Zuraidah Mohd Don, **The Totion of a "lemma"**
- Ghadeer Khalil, Graham Tranfield and Tony Allen, (2007). Arabic speech Recognition Using English Based Engines, **The 1st International Symposium on Computers and Arabic Language & Exhibition.**

- Headwords, Toots and Lexical Sets, (2004)**. International Journal of Corpus Linguistics,** John Benjamins Publishing Company.
- Jacqueline Léon, (2005). Claimed and Unclaimed Sources of Corpus Linguistics, **Henry Sweet Society Bulletin, Issue No. 44.**
- Jan Aarts and Willem Meijs, (1990). **Theory and Practice in Corpus Linguistics,** Rodopi.
- Karim Bouzouba and Adil Kabbaj, (2007). An Intergrated Development Platform for Arabic Language Processing, **The 1st International Symposium on Computers and Arabic Language & Exhibition.**
- Khaled Shaalan, Habib Talhami and Ibrahim Kamel, (2007). Automatic Morphological Generation for the Indexing Of Arabic Speech Recordings**, International Journal Of Computer Processing Of Oriental Languages, Vol. 20, No. 1.**
- Latifa Al-Sulaiti and Eric Atwell, (2006). The design of a corpus of Contemporary Arabic, **International Journal of Corpus Linguistics 11:1, John Benjamins Publishing Company**.
- Laurence Anthony, (*2005)*. AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom, **IEEE International Professional Communication Conference Proceedings.**
- Mahtab Nikkhou and Khalid Choukri, (2005). Survey on Arabic Language Resources and Tools in the Mediterranean Countries, Revised 7 March (2005), **NEMLAR, Center for Sprogteknologi**, University of Copenhagen, Denmark.
- Mark Van Mol and Hans Paulussen, (2004). **Natural Language Processing and Arabic: the Leuven Tandem Approach**, JEP-TALN 2004, Arabic Language Processing, Fez.
- Mark Van Mol, Exploring annotated Arabic Corpora, Preliminary Results. **http://ilt.kuleuven.be/arabic/ENG/onderzoek/index.php**
- Mark Van Mol, **The Semi-Automatic Tagging Of Arabic Corpora.**
- Maryam Mohammadi, (2007). Specialized Monolingual Corpora in Translation, **Translation Journal, Volume 11, No. 2.**
- Muhammad Atiyya, Khalid Choukri and Mustafa Yaseen, (2005). Specifications of the Arabic Written Corpus, **NEMLAR, Center for Sprogteknologi,** University of Copenhagen, Denmark.
- Muhammed Attiyya Mohamed Elaraby Ahmed, 2000. **A Large-Scale Computational Processor Of the Arabic Morphology, and Applications**, Msc Thesis, Cairo University.
- Nizar Habash and Owen Rambow, (2007). Arabic Diacritization through Full Morphological Tagging, **Proceedings of NAACL HLT, Rochester, NY.**
- Nizar Habash and Owen Rambow, (2007). Morphophonemic and Orthographic Rules in a Multidialectal Morphological Analyzer and Generator for Arabic Verbs, **The 1st International Symposium on Computers and Arabic Language & Exhibition**.

- Ramzi Abbès, Joseph Dichy And Mohamed Hassoun, **The Architecture Of A Standard Arabic Lexical Database: Some Figures, Ratios And Categories From The Diinar.1 Source Program**.
- Randi Reppen, (2001). Northern Arizona University: Review Of Monoconc Pro And Wordsmith Tools**, Language Learning & Technology, Vol. 5, No. 3.**
- Rusli Abdul Ghani, Norhafizah Mohamed Husin dan Chin Lee Yim, (2004). **Pangkalan Data Korpus DBP: Perancangan, Pembinaan dan Pemanfaatan**, Dewan Bahasa Dan Pustaka Malaysia.
- Sameh Alansary, Magdy Nagi and Noha Adly. **Building an International Corpus of Arabic (ICA): Progress of Compilation Stage.**
- Sameh Alansary, Magdy Nagi and Noha Adly (2007), **Building An International Corpus Of Arabic (ICA): Progress Of Compilation Stage**.
- Sameh Alansary, Magdy Nagi and Noha Adly, (2008), **Towards Analyzing the International Corpus of Arabic (ICA): Progress of Morphological Stage.**
- Sandra Kubler: **Introduction to Corpus Linguistic, Seminar fur Sprachwissenschaft**, University of Tubingen.
- Serge Sharoff, Open-Source Corpora, Using The Net To Fish For Linguistic Data, (2006). **International Journal of Corpus Linguistics 11:4, John Benjamins Publishing Company**.
- Siamak Rezaei, **Tokenizing an Arabic Script Language**
- Steven Krauwer, Bente Maegaard, Khalid Choukri, Lise Damsgaard Jørgensen, (2004). Report on BLARK for Arabic, **NEMLAR, Center for Sprogteknologi**, University of Copenhagen, Denmark.
- Toufik Sari and Mokhtar Sellami, State-of-the-art of Off-line Arabic Handwriting Segmentation, (2007). **International Journal of Computer Processing of Oriental Languages, Chinese Language Computer Society & World Scientific Publishing Co**.
- Victoria Ros´En, Paul Meurer And Koenraad De Smedt, (2005). Constructing A Parsed Corpus With A Large LFG Grammar, **Proceedings Of The LFG'05 Conference**, University Of Bergen, CSLI Publications.
- William Black, Sabri Elkateb, Horacio Rodriguez and Musa Alkhalifa, (2006). Introducing the Arabic WordNet Project, **GWC Proceedings**.
- Wolfgang Teubert, (2005). My Version of Corpus Linguistics, **International Journal of Corpus Linguistics**, John Benjamins Publishing Company.
- Xunlei Rose Hu and Eric Atwell, **Survey of Machine Learning Approaches to Analysis of Large Corpora**.
- Yousif a. El-imam and Zuraida Mohammed Don, (2000). Text-To-Speech Conversion Of Standard Malay, **International Journal Of Speech Technology 3.**
- Zaiton Ab. Rahman (1983). **Kertas Rancangan Projek Analisis Teks Secara Komputer**. Cawangan Penyelidikan, Dewan Bahasa Dan Pustaka Malaysia.

مواقع الإنترنت

- المجلة العربية لعلوم و هندسة الحاسوب
/http://www.phillips-publishing.com/jcsea
- شركة صخر
http://www.sakhr.com/default_a.aspx
- مجتمع مطوري الموقع
http://www.almashroo.com/articles
- مدينة الملك عبد العزيز للعلوم و التقنية (KACST)
http://www.kacst.edu.sa/ar/default_ar.aspx

- aConcorde
http://www.andy-roberts.net/software/aConCorde/
- Acronym Finder
http://www.acronymfinder.com/
- Arab Center for Arabization, Transaltion, Authorship & Publication
http://www.acatap.htmlplanet.com/journal.htm
- Arabic Corpus Tool
http://arabicorpus.byu.edu/index.php
- Buckwalter Arabic Corpus
http://www.qamus.org/
- Cambridge and Nottingham Corpus of Discourse In English
http://www.cambridge.org/elt/corpus/cancode.htm
- Center for Corpus Research, UOB
http://www.corpus.bham.ac.uk/pclc/#corpora
- Concordance
http://www.concordancesoftware.co.uk/
- Corpus Linguistic: Bibliography
http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/bibliography/bibliography.html#title
- Dewan Bahasa Dan Pustaka
http://dbp.gov.my/lamandbp/main.php
- European Language Resources Association
http://www.elra.info/
- European Parliament Proceedings Parallel Corpus
http://www.statmt.org/europarl/index.html
- Gateway to Corpus Linguistics
http://www.corpus-linguistics.de/
- Google Book Search
http://books.google.com/
- ICAME 2001 Future Challenges in Corpus Linguistics, Conference Programme

http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Events/icamepr.htm

- Interesting Sites about Natural Language Processing
  http://coleweb.dc.fi.udc.es/cole/sites_cl.html
- International Corpus of English
  http://www.ucl.ac.uk/english-usage/ice/
- International Corpus of Leaner English
  http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm
- Journal of Computer Science and Engineering, in Arabic
  http://www.phillips-publishing.com/jcsea/
- Katholieke Universitiet Leuven, Arabic Corpus
  http://ilt.kuleuven.be/arabic/ENG/onderzoek/index.php#onderzoek_b
- King Saud University Libraries Catalog
  http://catalog.library.ksu.edu.sa/
- Leuvan Corpus
  http://ilt.kuleuven.be/arabic/ARAB/indexARAB.php
- Linguistic Data Consortium
  http://www.ldc.upenn.edu/
- MEDAR
  http://www.medar.info/Archive/index.php
- Monoconc Pro
  http://www.athel.com/mono.html
- Penn Arabic Treebank
  http://www.ircs.upenn.edu/arabic/
- Preliminary Recommendations on Text Typology
  http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html
- Prokamus
  http://prokamus.wordpress.com/
- Science Development Network
  http://www.sciencedev.net/fe/GetPage.aspx?pid=22
- Software, Tools, Lists, Resources
  http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/software.htm
- Statistical Natural Language Processing and Corpus-Based Computational Linguistics: An Annotated List of Resources
  http://nlp.stanford.edu/links/statnlp.html
- The Association for Computational Linguistic
  http://www.aclweb.org/
- The Linguist List, East Michigan University
  http://linguistlist.org/
- The Louvain Corpus of Native English Essays
  http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/research%20learner%20corpora.html
- The Michigan of Academic Spoken English
  http://www.lw.lsa.umich.edu/eli/micase/index.htm
- Type of Corpus
  http://donelaitis.vdu.lt/publikacijos/SDoCL1.htm#CL

- University of Cambridge, Faculty of Modern & Medieval Languages
  http://www.mml.cam.ac.uk/call/cert/14/
- Wmatrix Corpus Analysis and Comparison Tool
  http://ucrel.lancs.ac.uk/wmatrix/
- Wordsmith Tools 4
  http://www.lexically.net/wordsmith/index.html
- World Scientific Journals Online
  http://db0.worldscinet.com/
- Xaira
  http://www.oucs.ox.ac.uk/rts/xaira/

TOWARDS BUILDING A MODEL OF ARABIC CORPUS IN MALAYSIA

By

Aswandi Bin Laman

Supervisor

Dr. Nihad Al-Mousa

ABSTRACT

This study is a draft proposal of an individual which may be the basis for a major institutional project in the area of building an Arabic Corpus in Malaysia. Despite the importance of subjects related to corpus linguistics, as far as this researcher is concerned, relevant studies regarding how to define corpus linguistics, its nature and types as well as the software and tools required to clarify it, are scarce. As a result, so are attempts to propose building it both at individual and institutional levels. This study not only delves into theoretical descriptions of corpus linguistics but suggest a model which with the benefit of advanced Malaysian technological know-how, may well be realized for practical applications in the development of an arabic corpus in Malaysia.