مدينة الملك عبدالعزيز للعلوم والتقنية

المنظمة العربية للتربية والثقافة والعلوم
إدارة العلوم والبحث العلمي

الاجتمـــاع الثانـــي
لخبراء المعجـم الحاسوبي للغة العربيــة

أبريل 2008

المعجم الحاسوبي للغة العربية
الجانب الحاسوبي

الدكتور حسن السيد
الدكتور زكريا الكردي

الجوانب الحاسوبية لمعجم اللغة العربية

الدكتور حسن السيد
Email: hsgroupitc@yahoo.com
الدكتور زكريا الكردي
mzkurdi@yahoo.com

دمشق - 15 شباط 2008 الموافق 8 صفر 1429

مستوى المعجم المطلوب:

يمكن أن يكون مستوى المعجم مرجعيا أو شعبيا(متوسط المستوى) أو طلابيا (شخصيا).

ارتأينا هنا أن يكون العمل موجها لإنجاز معجم متوسط المستوى في البداية (يمكن أن يشتمل على 30000-40000 مفردة).

في تقديرنا أن المشروع بمستواه المقترح يحتاج إلى 3 سنوات؛ ويكون ذلك في مرحلتين:

أ – مرحلة بناء نسخة تجريبية من المعجم Beta Version (18 شهرا).

ب – مرحلة تقييم وتعديل وتحسين النسخة التجريبية لتصبح نسخة معتمدة (18 شهرا).

هناك أربعة قضايا رئيسية في مشروع المعجم الحاسوبي للغة العربية

1. بناء المعجم نفسه.
2. بناء برمجيات مساعدة لبناء المعجم ولتمكين المستثمر من استعماله بسهولة وكفاءة.
3. نشر المعجم وتوزيعه.
4. تحديث المعجم دوريا.

خطوات إنجاز المعجم

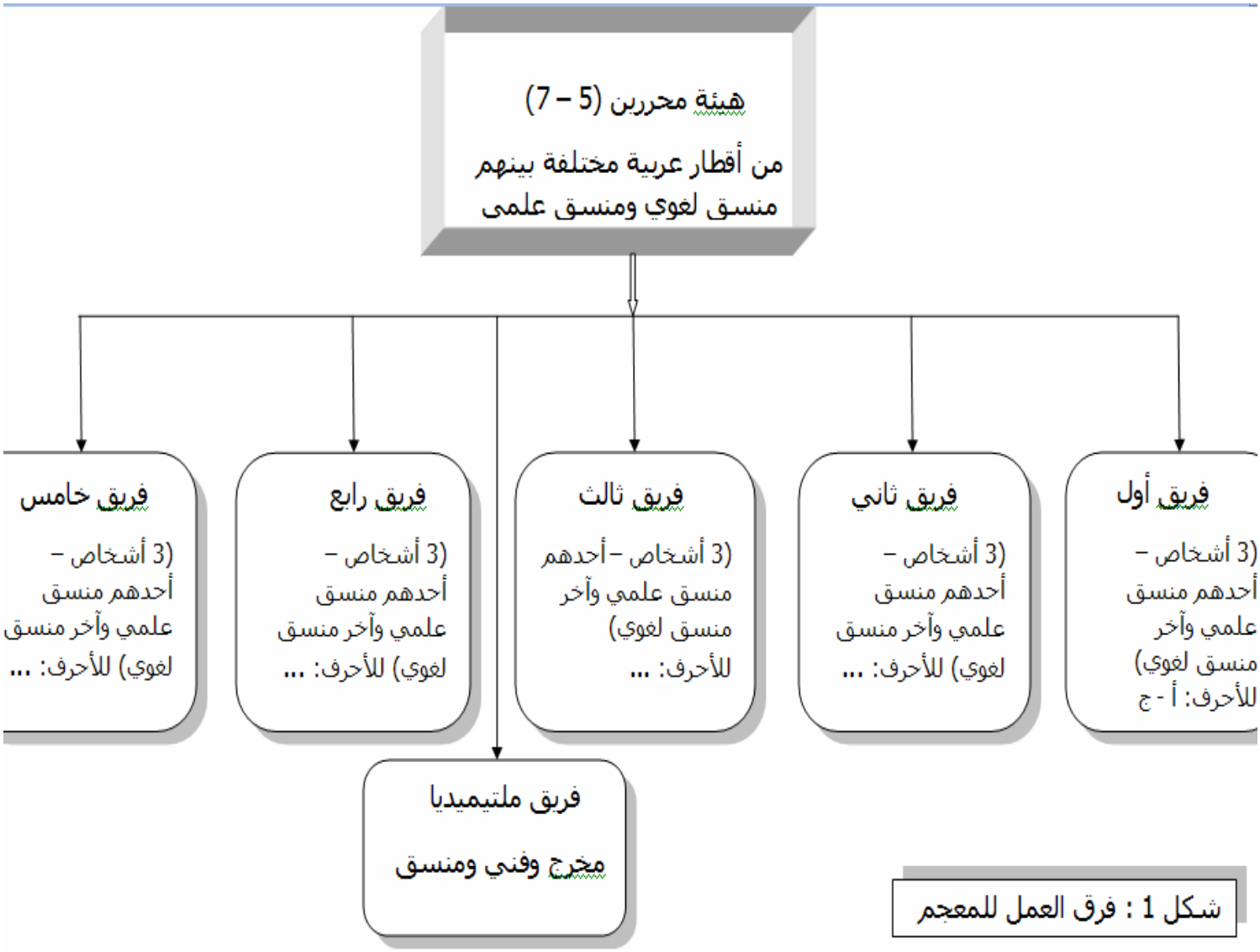
1. انتقاء وتكليف "هيئة محررين" للمعجم، مشكَّلة من خبراء (5 – 7، من أقطار عربية مختلفة) متميزين في اللغة العربية واللسانيات وفي مجالات **علمية** متعددة (مهتمين بثقافة اللغة العربية والترجمة) كما في نماذج المعاجم الانكليزية والفرنسية الشهيرة.
2. تشكيل 5 فرق عمل، يحوي كل فريق 3 أعضاء للقيام بالمهمات التالية:

- إدخال المفردات وفق البطاقة المتفق عليها (كل فريق يعمل على 5 أو 6 أحرف).
- تدقيق المفردات المدخلة وتفرعاتها.

ويتولى أحد أعضاء الفريق التنسيق اللغوي وآخر يتولى التنسيق العلمي مع هيئة المحررين.

يضاف إلى ذلك فريق ملتيميديا (أيضا من 3 أعضاء) من مهماته تجميع قاعدة بيانات ملتيميديا (صور وأصوات وقصاصات أفلام video clips)؛ انظر الشكل 1 المرفق.

3. انتقاء مصادر المادة اللغوية للمعجم (لكل من: المداخل، والتعاريف، والأمثلة، والمترادفات، والتراكيب اللغوية، والفروق اللغوية، والمصطلحات...)؛ يتم تحديد مصادر المادة اللغوية من قبل هيئة المحررين.



شكل 1 : فرق العمل للمعجم

بالنسبة للكلمات الجديدة والمتخصصة (مثل تلك المستعملة في المعلوماتية) يمكن استعمال مشروع "الروبت Robot" ومشروع "نبش المعلومات Data Mining" الذي يعمل عليه فريق جامعة حلب (باشراف الدكتور زكريا الكردي) للبحث في مواقع الانترنيت عن مثل تلك المفردات.

4. تشكيل فريق عمل (2-3 أشخاص) لتحديد أماكن تواجد مصادر المادة اللغوية التي حددتها هيئة المحررين، وشكل توافرها (ورقي/الكتروني/ ...)، ووضع ملاحظات عن حقوق النشر والاقتباس للمادة المراد ادراجها في المعجم.
5. حسم أوجه الخلاف بين مصادر المعلومات المتعددة باعتماد نهج محدد يتفق عليه أعضاء هيئة محرري المعجم.
6. تشكيل فريق برمجة (5 مبرمجين بخبرات Java, php, mysql, Python) يقوم بما يلي:

أ. وضع برمجيات مساعدة لتسهيل إنشاء **قاعدة بيانات لمصادر المادة اللغوية** وفق نموذج معياري كذلك الذي اقترحناه في المرفق #1

ب. تصميم بطاقة إدخال لمفردات المعجم (كتلك المقترحة في المرفق #2)، بالتعاون مع اللغويين.

ت. بناء برمجيات تُسَهِّل عملية إدخال المفردات **ومعالجتها وتدقيقها حاسوبيا.**

ث. بناء منتدى تداولي (Blog شبه موقع الكتروني) لتبادل الآراء والمشاكل وحلولها بين العاملين على المعجم.

نوصي بالتأكيد على حيازة كلٍّ من العاملين في المشروع على حاسوب شخصي.

ج- تحديد العتاد الحامل للمعجم (حاسوب مكتبي/مساعد الكتروني شخصي/...)؛ نقترح بدايةً الحواسيب المكتبية (Desktop) والمحمولة (Laptop)، التي ستحدِّد بدورها وسائل الدخل والخرج وخياراته.

ح- تحديد أساليب عرض المادة اللغوية (بالتعاون مع اللغويين) بما في ذلك استعمال الوسائط المتعددة (الصور، الرسوم، الأصوات، مقاطع الفيديو...)

خ- بناء برمجيات تعامل المستثمر مع المعجم بما في ذلك واجهات الاستعمال والبرمجيات الخلفية للبحث والتنقيب والاشتقاق والمحلل الصرفي (إعادة المفردة إلى جذورها) و...

د- اختبار صلاحية البرمجيات، بعد إتمام إدخال أحد أحرف المعجم.

قضايا معلوماتية تفصيلية

برمجيا سيكون هناك 3 طبقات لبرمجيات المعجم (يرجى الاطلاع على الملحق #3 حول توصيف وخصائص بعض البرمجيات الشائعة الاستعمال في بناء تطبيقات الويب):

1. **واجهة التعامل مع المستعمل Graphical user interface**

سنستعمل في ذلك لغة HTML بالاستعانة بلغة Java Script

2. برمجيات بناء المعجم (Web Application Modules) بما في ذلك المستعملة في البحث عن المفردات وجذورها وبرمجيات المحلل الصرفي، وغيرها.

سنستعمل في ذلك لغة Java ولغة PHP5

3. المعطيات نفسها Data Layer

أ- **قاعدة المعطيات** المعتمدة الأفضل لهكذا عمل (لنتذكر بأن المشروع هو حر- مفتوح المصدر). هي قاعدة المعطيات mysql (المفتوحة المصدر) التي أثبتت كفاءة عالية ومقدرة على التعامل مع كم كبير من المعطيات مثل المعجم الذي نحن بصدده.

ب- **النبش في الذخيرة اللغوية** (Data Mining) واستخلاص المفردات منها آليا، وذلك باستعمال بريمجيتين (Modules):

أولا: الروبوت الذي يحتوي على الجزئيات التالية:

- متصفح HTTP Browser للبحث عن المادة اللغوية العربية.
- عرّاف الصفحات العربية Arabic language identifier للتعرف على صفحات الويب المطابقة للغة العربية القياسية.
- مصنف ومخزن نصوص HTML parsing and Text storage

يصنف ويخزن نصوص الذخيرة اللغوية حسب الموضوع (أو الكاتب أو الحقبة الزمنية، ..).

في هذه المهمة سنستعمل لغة البرمجة Java لوجود مكتبة واسعة فيها تدعم بروتوكول HTTP وتدعم مصنف النصوص HTML parsing.
كما سنستعمل لغة الترميز السريع Python التي تتميز بكونها غرضية التوجه، وتدعم بروتوكول HTTP، وتدعم مصنفي النصوص HTML parsing، XML.

وكذلك يمكن أن نستعمل تجميعة من خوارزميات التعلم الآلي:
WEKA (Waikato Environment for Knowledge Analysis) المكتوبة بلغة Java لأتمتة وتسهيل عمليات نبش المعطيات (Data Mining).
هذه تجميعة الخوارزميات هذه مفتوحة المصدر أيضا وتحتوي على أدوات للتبويب، وايجاد الارتباط (Regression) بين المفردات، والعنقدة/البحث العنقودي (Clustering)، وغير ذلك.

ثانيا. مستخلص آلي للتوصيف المعجمي للمفردات.

4. **برمجيات دعم العمل التشاركي Groupware** من أجل:

- مساعدة اللغويين (والعلميين) في الاطلاع على أعمال بعضهم البعض (من يعمل ماذا) وتبادل الآراء في قضية محددة والوصول إلى توافق عليها.

- مساعدة اللغويين (والعلميين) في التعاون على حل الاشكاليات الصعبة.

ت - برمجيات أخرى مساندة:

1. برمجيات البحث والتنقيب.
2. برمجيات التحليل الصرفي.
3. برمجيات الاشتقاق والتصريف.
4. برمجيات احصائيات الزائرين لمواقع المعجم.

نشر المعجم وتوزيعه

بعد انتهاء بناء المعجم (المرحلة الأولى من المشروع)، توضع النسخة التجريبية (Beta version) من المعجم على عدة مواقع الكترونية (موقع الأليكسو، موقع مدينة الملك عبد العزيز، ...)، وتوضع اعلانات عنه في مواقع رائجة مثل google وغيره، للحصول على تغذية راجعة وتقييم للمعجم من مراجع متنوعة.

آلية تحديث المعجم

يتعاون أعضاء هيئة المحررين وغيرهم من الاختصاصيين اللغويين مع المعلوماتيين واختصاصيي العلوم الأخرى في وضع مثل هكذا آلية ومقوماتها وشروطها.

موازنة تقديرية للمشروع

الموارد البشرية المطلوبة للمشروع (تقديرا)

أ – يحتاج المشروع كما ذكرنا أعلاه إلى عدد من اللغويين وإلى اختصاصيين في العلوم يُقَدر عددهم 7 أشخاص يشكلون هيئة المحررين الدائمين.

ب - يحتاج المشروع، في المرحلة الأولى، إلى 5 مبرمجين، ويمكن اختصار العدد إلى 3 في المرحلة الثانية.

ت - يحتاج المشروع، في المرحلة الأولى، إلى عدد من الفنيين (6×3 =18) العاملين في الفرق الستة المقترحة أعلاه من أجل اقتباس وادخال وتحضير مواد المعجم وإخراجها، ويمكن اختصار العدد إلى 12 في المرحلة الثانية.

ث - يحتاج المشروع في كلا المرحلتين إلى عدد من المشرفين والمنسقين المؤقتين بعمل جزئي (اشراف علمي، اشراف لغوي، تنسيق لغوي، تنسيق علمي، إدارة، سكرتارية...)

التكلفة التقديرية للمشروع

آ – تكلفة الموارد البشرية:

المرحلة الأولى: 30 شخص×18 شهرا×1000 دولارا = 540000 دولارا أمريكيا

المرحلة الثانية: 22 شخص×18 شهرا×1000 دولارا = 396000 دولارا أمريكيا

يضاف إلى ذلك:

ب- تكلفة أجهزة ومعدات	30,000 دولارا أمريكيا
ت- تكلفة خبراء مؤقتين	36,000 دولارا أمريكيا
ث- تكلفة اجتماعات وندوات 3x3مرات في السنة× 10,000 =	90,000 دولارا أمريكيا
ج- تكاليف متنوعة	10,000 دولارا أمريكيا
ح- تكاليف طارئة	30,000 دولارا أمريكيا

فتكون التكلفة الاجمالية المقدرة للمشروع **1,132,000 دولارا أمريكيا**

ملحق #1

الحقول المطلوبة لبناء قاعدة بيانات مصادر المعجم

اسم المصدر	المؤلف	الناشر	شكل المادة اللغوية E, PP, Ph,O	حجم المادة (صفحات/مفردات)	جودة المادة لغويا	كيفية الاتصال بالناشر	كيفية الاتصال بالمؤلف	ملاحظات حول حقوق الاقتباس والنشر
الاصطلاحات:								
E= وثيقة الكترونية	PP= ورقية مطبوعة	PH= ورقية بخط اليد	O: غير ذلك					

ملحق # 2

نموذج مقترح لبطاقة إدخال مفردات المعجم

تصريفات المفردة	
الصيغة الصوتية للمفردة	
معنى أول (صيغ وأمثلة)	
معنى ثاني (صيغ وأمثلة)	
معنى ثالث (صيغ وأمثلة)	
تاريخ المفردة وتطورها	
تكرار المفردة في النصوص المعاصرة	
تكرار المفردة في النصوص القديمة	
مشتقات المفردة	
مرادفات المفردة	
اضضاد المفردة	

ملحق #3

خصائص بعض لغات البرمجة المقترح استعمالها

ظهرت في نهايات القرن الماضي نزعة لدى مبرمجي الحواسيب إلى التحول عن الطرق التقليدية للبرمجة المتبعة في "لغات البرمجة المعيارية System Programming languages"، مثل Pascal, C, C++, Java، إلى مايسمى "لغات الترميز السريع Scripting Languages" مثل JavaScript, Perl, Python, Tcl.

ظهرت لغات البرمجة التقليدية (رفيعة المستوى Higher level Languages) في نهاية الخمسينات من القرن الماضي كتسهيل وترقية للغة التجميع (Assembly Language) أو لغة الآلة التي كانت تخاطب الآلة (الحاسوب) مباشرة، بحيث كان يتوجب على المبرمج مثلا أن يكتب عشرات الأسطر من الشيفرة ليتعامل مع أماكن وكيفية تخزين المعطيات ومعالجتها.

بينما تستعمل لغات البرمجة التقليدية (رفيعة المستوى High Level) مركبة "المجمِّع Compiler " لتقوم بمهمة تحويل الشيفرة إلى تعليمات يلغة الآلة وبأعمال أخرى تسهِّل تنفيذ البرنامج.

بالمقابل فإن لغات الترميز السريع (الحديثة نسبيا) هي مجرد تصميغ أو تطوير لتطبيقات برمجية مكتوبة على الغالب بلغات البرمجة (المعيارية). لذلك يطلق على لغات الترميز السريع في بعض الأحيان اسم لغات التصميغ Gluing Languages، أو لغات تكامل المنظومة البرمجية System Integration Languages.

وعلى سبيل المثال فإننا نحتاج في لغة **Tcl** إلى سطر برمجي واحد للتحكم بحجم الخط في طباعة كلمة "Hello"، بينما نحتاج إلى 7 أسطر برمجية في لغة Java، وفي لغة C++ نحتاج إلى أكثر من 20 سطرا.

ولغات البرمجة سريعة الترميز موجودة منذ زمن بعيد، إلا أن هناك عدة عوامل تساهم في انتشارها وتطويرها بقوة من جديد؛ وأهم هذه العوامل:

- تزايد امكانات الحواسيب وسرعتها بشكل سريع جدا، يسمح بتصميغ عدد من التطبيقات الكبيرة وتشغيلها معا بكفاءة. وكلما ازدادت سرعة الحواسيب كلما مال المبرمجون إلى لغات الترميز السريع.

- تطور الانترنيت، فالانترنيت في الواقع هي البيئة المثلى للغات سريعة الترميز. ولذلك نجد مثلا أن لغة Perl هي الأكثر استعملا في كتابة برامج CGI، و Javascript هي الأكثر استعمالا في برمجة صفحات الويب.

- ازدياد أهمية تطوير برمجيات واجهات التعامل مع المستعمل Graphical User Interface.

  فهذه البرمجيات ماهي في الواقع إلا تجميعة من التطبيقات المصمَّغة مع بعضها. وهذا بالضبط عمل اللغات سريعة الترميز، التي يمكن أن تقوم بالمطلوب بسرعة وكفاءة.
  ومعظم بيئات التطوير السريع لواجهات المستعمل rapid-development GUI ترتكز على لغات البرمجة سريعة الترميز .

- سهولة تعلم اللغات سريعة الترميز بالمقارنة مع اللغات المعيارية التي تتعامل مع الكائنات Objects والحزم Threads.

يبين الجدول أدناه مقارنة سريعة بين بعض لغات البرمجة المقترح استعمالها

لغة البرمجة	استعمالاتها	ميزاتها
Java	بناء التطبيقات بشكل عام	مرنة/قوية/متعددة البيئات
JavaScript	برمجة صفحات الويب لخدمة المستعمل	
Perl	معالجة نصوص الانترنيت	الاختصار وجودة التمثيل
PHP	تطبيقات الويب	القوة والبساطة
Prolog	ذكاء اصطناعي/حل مسائل	برمجة تعريفية
Python	تطبيقات/تعليمية/دعم لمهمات الويب	البساطة/سهلة الفهم/قوة النمذجة Modularity
Visual Basic	بناء تطبيقات	تطوير سريع للتطبيقات/بساطة

المراجع

- A study of the Script-Oriented Programming (SOP) suitability of selected languages – from The Scriptometer.
- A Slightly Skeptical View on Scripting Languages by Dr. Nikolai Bezroukov
- Scripting: Higher Level Programming for the 21st Century by John K. Ousterhout
- Are Scripting Languages Any Good? A Validation of Perl, Python, Rexx, and Tcl against C, C++, and Java (PDF) — 2003 study
- The Computer Language Benchmarks Game at Alioth
- Language Study — Syntax across languages.
- Programming Language Comparison — A comparison of nine programming languages and related information.
- Computer Language Shootout Scorecard — Comparison of benchmark results for dozens of languages.
- Scriptometer scores — Multiple comparisons of 26 programming languages.
- Are Scripting Languages Any Good? A Validation of Perl, Python, Rexx, and Tcl against C, C++, and Java — PDF — 2003 study
- An empirical comparison of C, C++, Java, Perl, Python, Rexx, and Tcl — PDF — March 2000 refereed journal paper
- An empirical comparison of C, C++, Java, Perl, Python, Rexx, and Tcl for a search/string-processing program — PDF — March 2000 technical report (same author, experiment, and data as above, but has additional analysis and charts)
- ABAP2Java.com Comparision and Translation of ABAP and Java
- Comparing Web Languages in Theory and Practice — PDF — Research to fulfill Kristofer J. Carlson's master's degree requirements.
- The Encyclopedia of Computer Languages — As of May 2006, the encyclopedia lists 8512 computer languages with 17837 bibliographic records featuring 11064 extracts.
- PLEAC Programming Language Examples Alike Cookbook.
- The hundred-year language by Paul Graham. Keynote from PyCon2003 (about Python): how languages evolve and what increase in CPU speed might bring us.
- TIOBE Programming Community Index The TIOBE Programming Community index gives an indication of the popularity of programming languages.
- OHLOH Language Statistics The programming languages page on Ohloh gives an actively updated indication of the popularity of programming languages in open-source projects.
- Comparison Cheat Sheet between Languages --seems to be down