مدينة الملك عبدالعزيز للعلوم والتقنية

المنظمة العربية للتربية والثقافة والعلوم
إدارة العلوم والبحث العلمي

الاجتمـــاع الثانـــي
لخبراء المعجـم الحاسوبي للغة العربيـــة

أبريل 2008

الجوانب التقييسية للمعاجم الحسوبية

أ.د عبد المجيد بن حامدو

الجوانب التقييسية للمعاجم الحاسوبية

أ د. عبد المجيد بن حمادو
مدير مخبر "ميراكل" جامعة صفاقس – تونس
Abdelmajid.benhamadou@isimsf.rnu.tn

يتطرق هذا البحث إلى موضوع تقييس بناء المعاجم الحاسوبية بصفة عامة والمعاجم الحاسوبية العربية بصفة خاصة. وهو جانب حساس ومهم للغاية نظرا للمزايا التي يوفرها لهذه المعاجم من تيسير إثراء المضمون والهيكل وانتشار أوسع وضمان مواكبة التطور السريع للتكنولوجيا.

1. ما هو التقييس ؟ وما هي أهدافه ؟

التقييس هو عمل يهدف إلى الاعتراف رسميا بمواصفات تقنية لمنتج أو لخدمة ما بقرار توافقي تحت إشراف منظمة معترف بها وتتسم بالاستمرارية. هذه المنظمة يمكن أن تكون محلية: MSA, INNORPI, AFNOR, DIN, ANSI, BSI أو عالمية مثل: ISO, IEC, CEN, W3C, OASI.

من أهم الأهداف العامة للتقيس نذكر :

- التوحيد لتسهيل التبادل والإثراء.

- ضمان جودة المنتج أو الخدمة.
- ضمان أكبر قدر ممكن من السلامة.

2. تقييس الموارد المعجمية :

1.2 ما هي الموارد المعجمية التي يمكن تقييسها ؟

يمكن تصنيف الموارد المعجمية التي يمكن تقييسها إلى أربعة أصناف أساسية :

- قواعد البيانات المعجمية،
- قواعد البيانات المصطلحية،
- المعاجم المختلفة للاستعمال البشري (MRD: Machine Readable Dictionaries)
- ومعاجم المعالجة الآلية للغات الطبيعية (NLP: Natural Language Processing).

وتجدر الإشارة إلى أن أهم الموارد المعجمية المتوفرة بالنسبة إلى اللغة العربية تتمثل في المعاجم اللغوية للاستعمال البشري (القاموس المحيط، لسان العرب، الصحاح، المحيط، محيط المحيط، الوسيط، الرائد، الغني،...).البعض من هذه المعاجم متوفر في نسخ إلكترونية على الشابكة أو على أقراص مضغوطة. أغلب هذه النسخ مزودة بآليات بحث بسيطة وغير متطورة بما فيه الكفاية ولا تحافظ على نفس النمط في سرد المعلومات وترتب الكلمات بصفة مختلفة : فمنها من يعتمد ترتيباً ألفبائياً مع الإشارة إلى جذور الكلمات ومنها من يرتب الكلمات بالاعتماد على مخارج الحروف ومنها من يعتمد على ترتيب الجذور.

والسؤال الذي يطرح نفسه هو كيف يمكن استغلال هذا التنوع على مستوى الهيكلة والمضمون في بناء معجم حاسوبي عربي جديد؟ والجواب على هذا السؤال يمر حتما بتوحيد المعاجم المتوفرة هيكلا ومضمونا. ومن هنا تأتي أهمية التقييس كضامن لهذا التوجه.

2.2 مزايا تقييس الموارد المعجمية :

مزايا تقييس الموارد المعجمية متعددة نذكر أهمها :

- إمكانية استعمال المعجم لأغراض لغوية أو مصطلحية أو للمعالجة الآلية للغة باعتماد نفس البنية.
- الاستغلال حسب حاجيات المستعمل.
- تيسير عملية تبادل المعاجم بين الأشخاص والمؤسسات قصد الإثراء والاستغلال المشترك.
- دمج المعاجم قصد توليد معاجم متعددة اللغات للترجمة الآلية.
- تبادل الأدوات الحاسوبية مثل :
  - البحث عن المعلومات داخل قواعد البيانات والمعاجم.
  - المدقق الإملائي.
  - المحلل الصرفي.
  - المشكل الآلي.
  - المصنف الآلي.
  - نظم احتساب البيانات الإحصائية.
- الاستفادة المشتركة من المهارات المعجمية الأساسية المتواجدة.

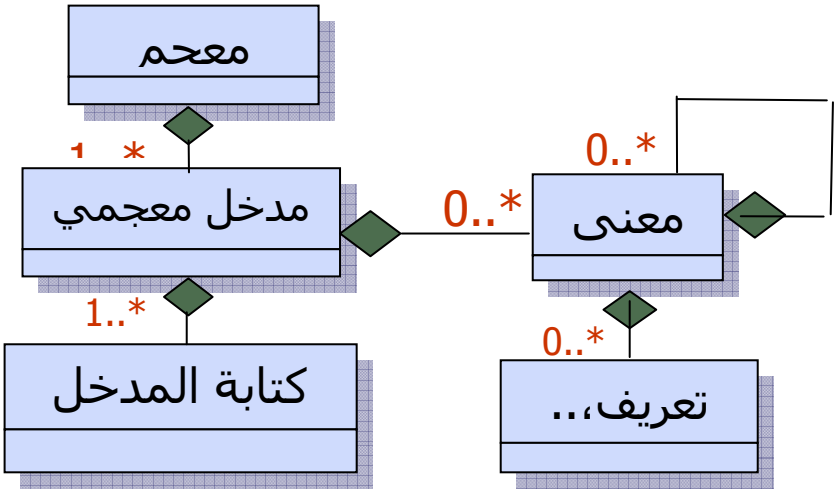
3. تعريف عام لمقياس LMF

LMF (Lexical Markup Framework) يمثل المقياس العالمي "إيزو24613" أرضية لتوصيف قواعد البيانات والمعاجم الحاسوبية أحادية اللغة أو متعددة اللغات لاستعمالات بشرية أو للمعالجة الآلية للغات الطبيعية. بدأ العمل على هذا المقياس منذ صائفة سنة 2003. وفي بداية 2004 قررت مجموعة ISO/37 /SC4 بعث مشروع رسمي يعنى بتطوير هذا المقياس. ومنذ ذلك التاريخ وقع التصويت على 14 نسخة لهذا المقياس والنسخة 15 هي في مرحلة متقدمة وسيقع التصويت عليها الأيام القليلة القادمة، والمتوقع أن تكون الأخيرة قبل تبني المقياس نهائيا. وتشارك في المجموعة 24 دولة من بينها تونس.

يمتاز مقياس LMF بمرونته وبإمكانية تغطيته لمختلف مستويات اللغة (الصرفي، النحوي، الدلالي،...) وكذلك بشموليته اللغوية حيث يمكن اعتماده لكل اللغات بدون تمييز.

يرتكز مقياس LMF على نموذج متكون من : (أنظر الرسم البياني عدد2)

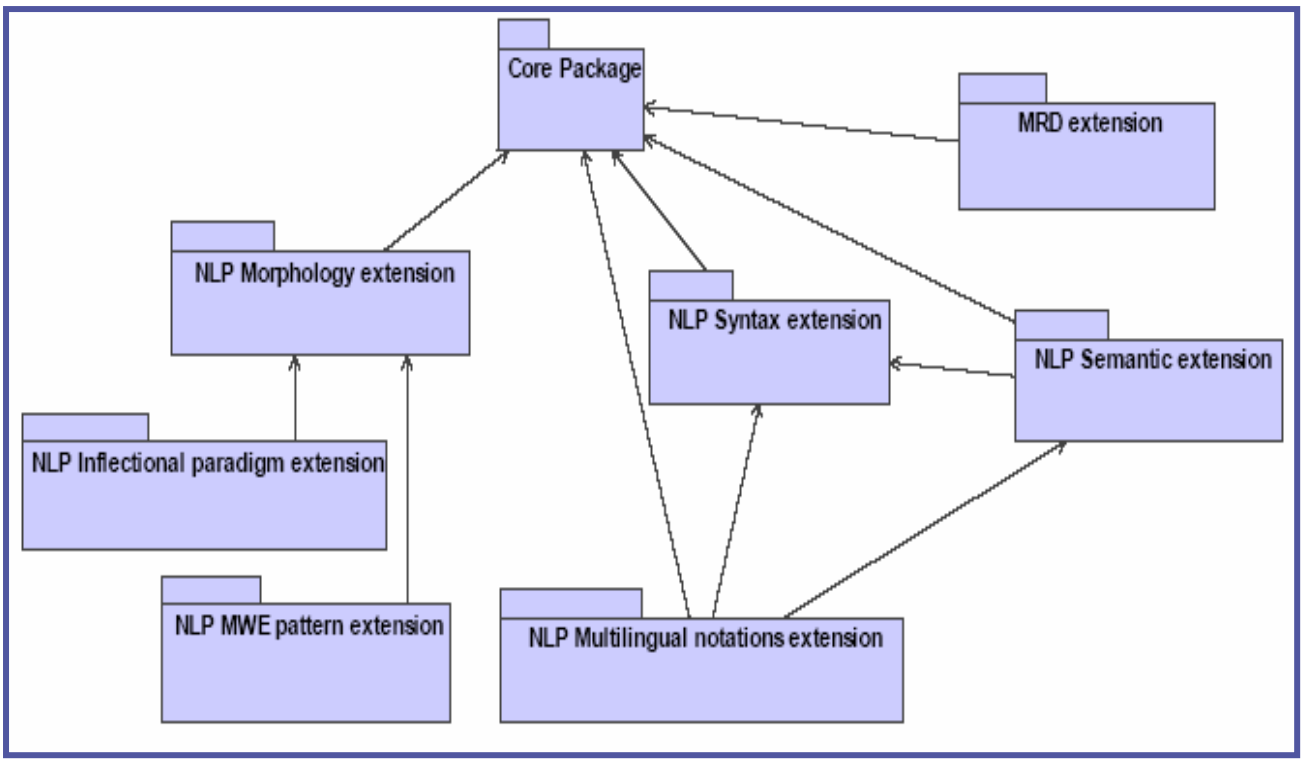
- نواة أساسي (core package) يحتوي على مجموعة مداخل (أنظر الرسم البياني عدد1). يحتوي كل مدخل (بسيط/ مركب/ حرف) على :
  - المعلومات الصرفية النحوية الأساسية (الجنس، قسم الكلام،.)
  - طريقة (أو طرق) كتابة المدخل أو نطقه.
  - معناه أو معانيه (تعريف المدخل، أمثلة وشواهد،..).



الرسم البياني عدد1:نمط النواة الأساسي (core package)

والجدير بالملاحظة أنه وقع الاقتصار على هذا العدد القليل من المعلومات لإمكانية تطبيق النموذج على جميع اللغات.

- وحدات اختيارية متخصّصة (extension packages) يمكن إضافتها للنواة حسب الحاجة إليها:
  - وحدات صيغميّة.
  - وحدات تركيبيّة (نحوية).
  - وحدات دلاليّة،...
- مجموعة أصناف صرفية نحويّة مناسبة للّغة المعتمدة يقع اختيارها من بين القائمة التي يوفرها سجل مرجعي مقيّس : الجنس، قسم الكلام،.... هذه الأصناف تستعمل لتعبئة مختلف مكونات نمط LMF. (Data Category Registry – ISO12620http://syntax.inist.fr/)



الرسم البيانى عدد2: مكونات نمط LMF (LMF Packages)

4. منهجية إعداد معجم حاسوبي (عربي) حسب مقياس LMF

كما ذكرنا أعلاه تتم عملية بناء معجم حاسوبي حسب مقياس LMF بدمج النواة الأساسي بصفر أو عدة وحدات اختيارية متخصصة ومجموعة أصناف نحويّة مناسبة للّغة المعتمدة.

عملية الدمج تتم باتباع المراحل التالية حسب توصيات فريق اللجنة التقنية ISO TC37/SC4 للمنظمة العالمية للتقييس "إيزو" [Francopoulo & George, 2007]:

أ. دراسة المعاجم (العربية) المتداولة (الورقية والحاسوبية) قصد استبيان:

- المكونات الأساسية لمداخلها.
- وطرق ترتيب المداخل (هيكلتها).
- وآليات البحث عن الكلمات التي توفرها للمستعمل.

هذه المرحلة مهمة بالنسبة للمراحل المتبقية لأنها تساعد على تحديد مكونات المعجم باعتماد منهج توحيد البيانات المعجمية المتوفرة. الرسم البياني عدد3 يلخص هذه المراحل.

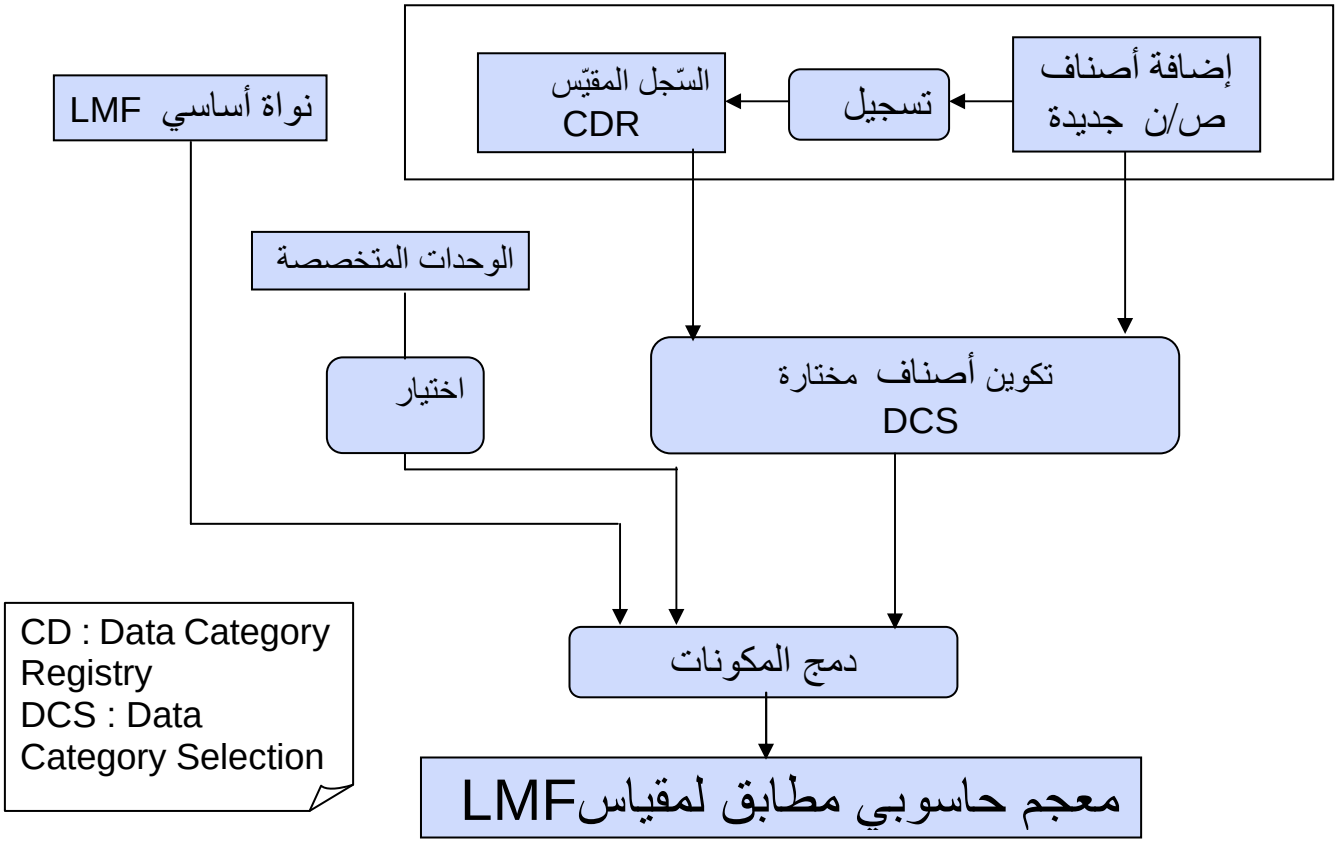
ب. تحديد مكونات النواة الأساسي حسب مقياس LMF.

ج. تحديد الوحدات المتخصّصة والاختيارية (extension packages) التي سيقع دمجها مع النواة.

د. وضع نموذج لكل وحدة متخصّصة وقع اختيارها. ويحدد هذا النموذج المكونات الأساسية للوحدة والعلاقات بينها.

ه. اختيار الأصناف الصرفية والنحويّة المناسبة لهذه اللّغة من بين القائمة التي يوفرها سجل مرجعي مقيّس (Data Category Registry – ISO 12620). كما يمكن للمصمم إضافة أصناف جديدة إذا اقتضى الأمر إلى ذلك.

ج. تعمير النموذج المتحصل عليه بهذه الأصناف.



الرسم البياني عدد3: مراحل بناء المعاجم الحاسوبية

5. عرض موجز للمشاريع والإنجازات الدولية

عدة مشاريع وقع إنجازها أو هي بصدد الإنجاز بالنسبة إلى لغات عدة : الفرنسية، الإيطالية، الإنكليزية، العربية، البنغالية، التايلندية واللغات الآسيوية. وسنتعرض فيما يلي إلى أهم هذه المشاريع والإنجازات.

أ. مشروع "مورفالو" MORPHALOU للغة الفرنسية (www.cnrtl.fr). يتمثل هذا المشروع في بناء قاعدة بيانات معجمية للفرنسية حسب مقياس LMF. والاهتمام الحالي بالجانب النحوي. مضمون هذا المعجم يعتمد على مكنز اللغة الفرنسية (Trésor de la Langue Fraçaise).

ب. مشروع "لكسوس" LEXUS (http://www.mpi.nl/lexus) المعجمي والذي يعتمد كليا على مقياسLMF. يمثل هذا المشروع قاعدة متوفرة على

الشابكة تمكن المستعمل من تصميم معاجم جديدة وإدخال البيانات المعجمية التي يختارها بما في ذلك بيانات متعددة الوسائط (صور وأصوات وأفلام الفيديو).

ج. معجم اللغة البنغالية : ICGL2998MichaelMaxwell[1].pdf

د. مشروع إنجاز بيئة مناسبة لبناء معاجم مقيسة للغات الآسيوية [T. Tokunaga et al.]

ه. إنجاز معجم للغة التايلندية يهتم بالجانب الدلالي [T. Charoenporn2007]

و. بناء قاعدة لمعالجة معاجم اللغة الإيطالية حسب مقياس LMF
http://www.senso-comune.it/documents/WorkshopAIIA2007/AIIASensoComune.pdf

6. مساهمة مخبر "ميراكل" في تطوير مقياس LMF للعربية

بمخبر "ميراكل" فريق بحث يشتغل على موضوع تقييس الموارد المعجمية وتصميمها بالنسبة للعربية. ويساهم هذا الفريق بآرائه وخبرته في تعديل مقياس LMF حسب خصائص اللغة العربية. ومن أهم الإنجازات في هذا المجال نذكر :

- تغيير بعض الروابط بين مكوّنات نمط LMF لاحتواء خصائص اللغة العربية.
- إضافة وحدة صوتمية (Phonological package)
- إضافة أصناف صرفية نحويّة خاصة بالعربيّة للسّجل المقيّس الذي يعتمد عليه نمط LMF :المرفوع، المنصوب، المجْزوم، جمع التكسير،...
- إنجاز أوّل قاعدة بيانات معجميّة للّغة العربيّة ArabicLDB مطابقة كليا لمقياسLMF مزودة بوحدة صرفية.
- تصميم وحدة تركيبيّة (نحوية) لإضافتها إلى القاعدة المعجمية.
- إنجاز آلية متطورة للبحث عن البيانات المعجميّة داخل القاعدة تعمل على ثلاثة مستويات:

- البحث البسيط عن الكلمة داخل المعجم يمكن المستعمل من عرض الخصائص الصرفية والنحوية للكلمة، معناها (أو معانيها)، الأمثلة والشواهد،... (أنظر الرسم البياني عدد4).
- البحث الموجه يتمتع فيه المستعمل بالمساندة الكافية للوصول إلى غايته مثل تصحيح الكلمة المدخلة،..
- البحث المتطور يمكن المستعمل من استرجاع بيانات متطورة تهم مجموعة من الكلمات أوكل المعجم مثل معلومات إحصائية، قائمة الكلمات التي لها نفس الجذر، والإبحار داخل المعجم باستغلال مختلف الروابط بين الكلمات،... BEN ABDERRAHMEN 2007

الرسم لبياني عدد4 : واجهة البحث البسيط عن الكلمة داخل المعجم

شكر

أريد أن أشكر فريق البحث المتكون من : بلال القرقوري، عبدالحميد عبد الواحد، محمد جميل ، قيس الهدار، عائدة خماخم فاتن بكور، ومحمد بن عبدالرحمان على مساهمتهم المباشرة أو غير المباشرة في إثراء هذا التقرير.

المراجع

BEN ABDERRAHMEN M., CHAARI. F, GARGOURI B., JMAIEL M. (2006). Des services orientés besoin pour l'exploitation des bases lexicales normalisées. *10th Maghrebian Conference on Software Engineering and Artificial Intelligence MCSEAI'06, 07-09 Décembre 2006, Agadir, Maroc.*

BEN ABDERRAHMEN M., GARGOURI B., JMAIEL M. (2007). LMF-QL: A graphical Tool to Query LMF databases. *Third Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics October 5-7 2007, Poznań, Poland.*

FRANCOPOULO G., GEORGE M. (2007). ISO/TC 37/SC 4 Rev.14. Language resource management − Lexical markup framework http://www.tagmatica.fr/(LMF).

IDE N., ROMARY L. (2004). A Registry of Standard Data Categories for Linguistic Annotation. Fourth International Conference on Language Ressources and Evaluation, 135-138.

IDE N., ROMARY L. (2004). International standard for a linguistic annotation framework. International Journal of Natural Language Engineering, 10 Numéro 3-4, 211-225.

KHEMAKHEM A., GARGOURI B., ABDELWAHED A., FRANCOPOULO G (2007). Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613. Traitement

Automatique des Langues Naturelles : du 5 au 8 juin 2007 à Toulouse.

BACCAR F., KHEMAKHEM A., GARGOURI B., HADDAR K., BEN HAMADOUA. (2008) Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe. Submitted to TAL'2008 conference.

T. Tokunaga, V. Sornlertlamvanich, T. Charoenporn, N. Calzolari, M. Monachini, C. Sonia, C.R. Huang, Y. J. Xia. 2006. Infrastructure for Standardization of Asian Language Resources. Proceedings of the COLING/ACL 2006, pp. 827-834.

T. Charoenporn, S. Thoongsup, V. Sornlertlamvanich, and H .Isahara.2007. Thai Thai Lexicon. In The 17th Annual Conference of the Southeast Asian Linguistics Society (SEALS) University of Maryland, College Park, August 31st to September 2nd, 2007