# Conception, implémentation et indexation de BaLeM, une base lexicale multilingue

Mathieu Mangeot-Lerebours

GETA-CLIPS - IMAG Campus
BP 53
38041 Grenoble cedex 9
Tél. : 04.76.51.43.80 - Fax : 04.76.51.44.05
Courriel : Mathieu.Mangeot@imag.fr

## 1. Introduction

En Traitement Automatique des Langues Naturelles (TALN), le problème de la gestion des ressources linguistiques est crucial. Le volume des données manipulées, leur grande variété et la vitesse de traduction sont autant de paramètres qui font de la construction de dictionnaires un élément clé de tout système de TALN. La dispersion des outils sur des plates-formes hétérogènes dont les lexicographes ont besoin lors de l'indexation et le coût élevé qui en résulte freinent les avancées dans ce domaine.

Beaucoup d'efforts ont été faits pour essayer de créer une plate-forme unique qui réduirait les coûts de production des dictionnaires mais peu de résultats ont été obtenus. D'autre part, pour le projet Universal Networking Language, nous devons faire face à des besoins très importants. À court terme, des outils d'indexation pour construire les dictionnaires seront nécessaires au projet. Nous pensons qu'il est possible de résoudre les problèmes de dispersion des outils en proposant une application générique multi-outils. Nous pourrons l'expérimenter dans le cadre du projet UNL.

## 2. Problématique

UNL est un projet de communication multilingue interpersonnelle. Il se base sur une langue pivot, l'UNL et 13 autres langues. Il y a au minimum un partenaire pour chaque langue du projet. Il faudra à court terme indexer près de 200 000 entrées avec un coût maximal de 5 F par entrée. Pour satisfaire cette demande, nous avons donc besoin de travailler avec plusieurs indexeurs dispersés en même temps. Il faudra ensuite regrouper les données en construisant une base lexicale pour différents outils et différents partenaires. Nous avons étudié plusieurs solutions :

Le Lexicaliste est un système de gestion de bases de données monolingues. Il ne permet pas d'implémenter une architecture avec pivot interlingue, nécessaire au projet UNL. De plus, son interface multifenêtres devient rapidement compliquée.

Genelex [Genelex 93] permet de décrire une architecture sous forme de graphe. Par contre, aucun outil n'est disponible pour indexer à grande échelle.

Utiliser une base de données sur un PC présente de nombreux avantages mais ne convient pas pour modéliser des structures complexes . Il faudrait en plus pouvoir équiper chaque utilisateur avec le même logiciel et recentraliser les données ensuite.

Construire une super base lexicale centrale sur une puissante station avec des indexeurs reliés par réseau est une solution idéale. C'est malheureusement une solution trop gourmande en équipement. De plus, les liaisons réseau actuelles sont trop coûteuses pour pouvoir travailler à domicile.

Avec les contraintes d'un tel projet, et après analyse des différentes solutions, il n'existe pas de solution satisfaisante. Nous proposerons donc une nouvelle méthode pour indexer une grande base lexicale.

## 3. Solution mise en œuvre

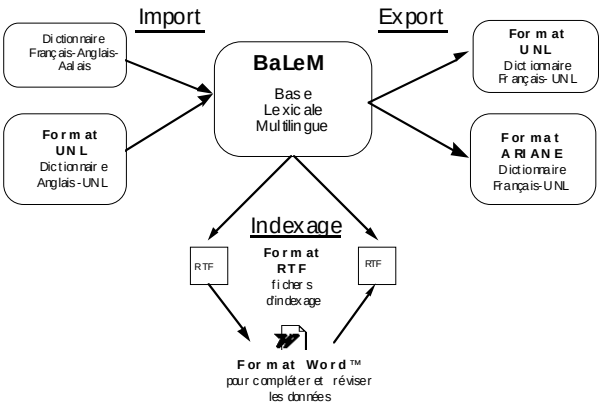


*Figure 4 : solution mise en œuvre*

Les lexicographes indexent la base en remplissant des fichiers Word. Cette technique a permis d'indexer 20 000 entrées équivalent à 50 000 acceptions en 9 mois pour un dictionnaire Français-Anglais-Malais. Nous avons ensuite amélioré cette technique en séparant le travail en deux parties :

La base de données centrale est gérée par un lexicologue. Il récupère d'abord plusieurs dictionnaires qu'il fusionne. Il crée une description des entrées de la base sous forme de grammaire. Grâce à cette description, il prépare le résultat de cette fusion sous forme de fichiers Word™ qui sont envoyés aux indexeurs et récupère les fichiers une fois révisés et complétés pour les intégrer dans la base après filtrage. Il peut renvoyer plusieurs fois les fichiers aux indexeurs, si le résultat n'est pas assez satisfaisant.

Les indexeurs travaillent à domicile sur leur ordinateur personnel. Ils n'ont besoin que du logiciel Word™ sur Mac ou PC. Pour faciliter le travail des lexicographes, nous avons ajouté des outils d'aide à l'indexation sous forme de macros Word™.

## 4. Le serveur BaLeM

La base centrale est implémentée en MCL (Macintosh Common LISP). La structure de la base n'est pas figée, ce qui nous permet de l'adapter aux changements. Nous pouvons à tout moment intégrer des dictionnaires existants ou générer automatiquement des dictionnaires pour différents systèmes de traduction comme ARIANE [Boitet 97] ou le système Deco utilisé à l'Université des Nations Unies par le centre UNL pour le japonais et l'anglais.

## 5. Les postes des lexicographes

Le lexicographe dispose d'une vue globale du dictionnaire avec lequel il travaille. Il peut corriger très rapidement les erreurs qu'il détecte et s'inspirer des articles précédents ou suivants, qu'il voit en totalité sans avoir à ouvrir de

Chaque unité d'information est donnée sous forme de paragraphe dans un style particulier ([Gaschler et al. 94]). À l'aide des macros, le lexicographe peut sélectionner la catégorie dans une liste (ce qui évite les erreurs dans les abréviations), vérifier la validité d'une entrée ou calculer l'ensemble des styles valides à la suite d'un élément d'information afin d'insérer un nouvel élément d'information ([Mangeot 97]).

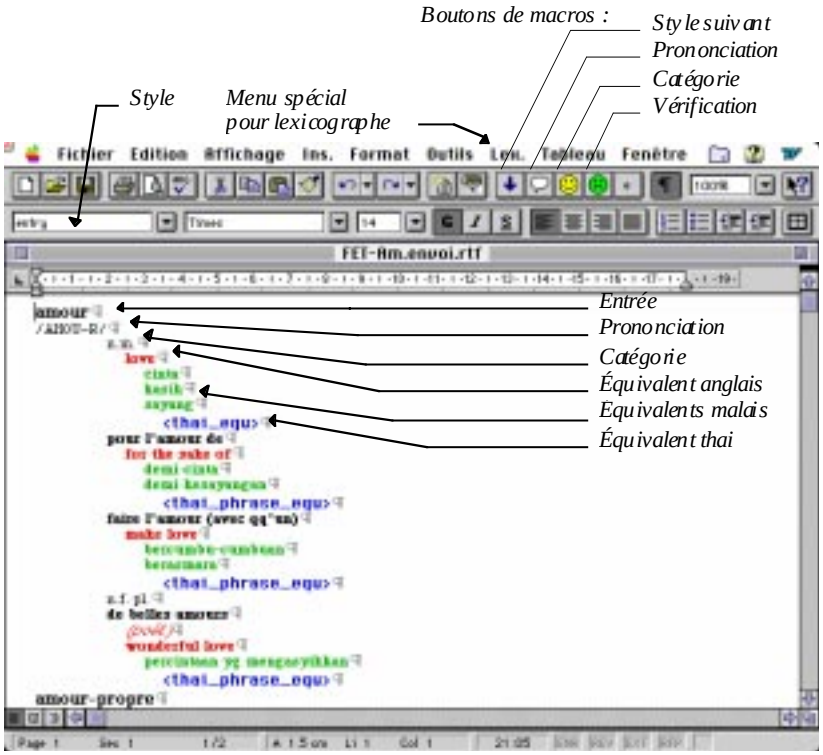


*Figure 6 : Exemple de fichier d'édition du dictionnaire Français Anglais Thaï*

## 6. Conclusion

La technique utilisée pour remplir notre base nous a permis d'indexer 20000 acceptions UNL en 7 mois. Elle nous semble satisfaisante et générique. Nous continuerons donc à l'utiliser pour indexer le reste des termes UNL (UW). Elle pourra facilement être reprise pour la fabrication de grandes BDLM.

Nous avons depuis implémenté une interface Web directement reliée à notre base grâce à MCL pour la consultation de notre base par les utilisateurs. Nous mettons actuellement en œuvre une interface pour le lexicologue administrateur codée en MCL et intégrée à la base pour permettre des modifications en temps réel.

## Références

Boitet C., (1997) GETA's methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects. Proc. PACLING, Ohme, Tokyo, Japan, PACLING, vol. 1/1, pp 23-

Gaschler J., and Lafourcade M., (1994) Manipulating Human-Oriented Dictionaries with Very Simple Tools. Proc. COLING'94, Kyoto, Japon, vol. 1/2, pp 283-286

Genelex (1993) Projet Eureka Genelex, modèle sémantique. Rapport Technique, Projet Eureka, Genelex, 4 mars 1994.

Mangeot-Lerebours M., (1997) Outils pour lexicographes : aspects informatiques. DEA Informatique Systèmes et Communications, Université Joseph Fourier Gren

Sérasset G., (1994) SUBLIM: un système universel de bases lexicales multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions. Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier GRENOBLE