Accès unique à des dictionnaires hétérogènes

Mathieu Mangeot-Lerebours

Xerox Research Centre Europe
6, Chemin de maupertuis
F-38240 Meylan
tel. : 04 76 61 51 32
fax : 04 76 61 50 99

GETA-CLIPS-IMAG
Domaine universitaire BP 53
F-38041 Grenoble cedex 9
tel. : 04 76 51 43 80
fax : 04 76 51 44 05

Courriel : Mathieu.Mangeot@imag.fr

1. Introduction

Notre laboratoire utilise pour ses recherches plusieurs ressources lexicales hétérogènes : dictionnaires monolingues, bilingues ou bases lexicales multilingues. Nous avons profité de cette situation pour les utiliser dans le cadre de notre expérience.

Notre but est d'accéder à des ressources hétérogènes à l'aide d'une seule interface en essayant de limiter les développements et de modifier le moins possible les fichiers sources. Cette interface doit être accessible au plus grand nombre d'utilisateurs.

Nous présenterons d'abord les ressources que nous avons utilisées pour notre expérience, puis nous exposerons le système que nous avons élaboré pour répondre à notre problème. Nous présenterons son interface ainsi que son architecture générale, puis nous détaillerons chaque fonctionnalité importante du système. Nous discuterons ensuite de cette solution puis nous la comparerons à d'autres solutions existantes. Enfin, nous conclurons sur les avantages d'un tel système et les extensions éventuelles que nous pourrions envisager.

2. Les ressources

Nous avons à notre disposition, entre autres, cinq dictionnaires. Ils ont été mis à notre disposition tels quels, sans aucun outil de présentation. Les dictionnaires "SGML" n'ont pas non plus de définition de type de document (DTD). Nous avons donc manipulé ces dictionnaires sans connaître leur structure interne.

2.1. DicoSzotar : un dictionnaire pour apprenants du hongrois

C'est un dictionnaire bilingue hongrois-français auquel nous avons ajouté des données multimédia. Certaines entrées disposent d'une image les représentant et d'une prononciation contenue dans un fichier son. L'utilisation des images permet à des apprenants non hungarophones de comprendre la signification de l'entrée sans traduction. Voici l'entrée "`akkor`" (alors) au format d'origine :

```
<entry><headword>akkor</headword><administration><indexer
date="Tue Apr 27 1999">Mathieu Mangeot</indexer> <lesson-
number> 2 </lesson-number> <revisor date="Wed Jun 2 1999">
Ágnes Sandor </revisor> <confidence-rate> trusted </
confidence-rate></administration><syntactic-cat><part-of-
speech> adverbe </part-of-speech> <semantic-cat>
<translation> alors </translation> </semantic-cat>
</syntactic-cat> </ entry>
```

Dans un premier temps, à chaque entrée étaient ajoutés le pluriel et l'accusatif pour un nom et la conjuguaison pour un verbe. Nous avons par la suite supprimé ces données pour finalement associer un générateur de pluriel et d'accusatif pour les noms ainsi qu'un conjugueur pour les verbes. Nous étendons le concept de dictionnaire en associant des actions externes aux entrées.

2.2. Dictionnaire Oxford-Hachette (OHD) [Corréard94]

Il se compose de deux parties. Un dictionnaire anglais-français et un dictionnaire français-anglais. Chaque dictionnaire, codé en SGML, est représenté par un fichier texte de 20 Mo environ. Chaque entrée est représentée par une seule ligne. Voici un extrait de l'entrée "abréger" au format d'origine :

```
<se><hw>abr&eagrave;ger</hw><pr><ph>abKeZe</ph><pr><hg><xt>
15</xt><ps>vtr</ps></hg><s2 num=1>(<ic>rendre court</ic>) t
o shorten [<co>mot, expression</co>]; to summarize [<co>
tex te, discours</co>]; <sl>&hw; &oq;t&eagrave;l&eagrave;
visio n&cq; en &oq;t&eagrave;l&eagrave;&cq;</sl> to shorten
&oq;t elevision&cq; to &oq;TV&cq;; </se>
```

Le mot-vedette "abréger" est suivi de sa prononciation, puis de sa partie du discours "vtr" et de sa traduction en anglais "to shorten" puis "to summarize". Un exemple est donné : "abréger télévision en télé" puis traduit : "to shorten television to TV".

2.3. New Oxford Dictionary of English (NODE)

Le NODE [Pearsall98] est un dictionnaire monolingue anglais. Il est composé d'un seul fichier d'environ 2 fois la taille du OHD. Chaque entrée est représentée par une seule ligne. Il est construit selon les mêmes principes que le OHD.

2.4. Dictionnaire Français-Anglais-Malais [Lafourcade96]

Ce dictionnaire propose des traductions de l'entrée française en anglais et en malais. L'anglais a servi d'aide aux lexicographes lors de l'élaboration du dictionnaire. Le fichier complet occupe une dizaine de Mo. Voici un extrait au format d'origine de l'entrée "abréger" :

```
(:fem-entry
(:ENTRY "abréger")
(:FRENCH_PRON "abre-je-")
(:FRENCH_CAT "v.tr.")
(:FRENCH_GLOSS "un texte")
(:ENGLISH_EQU "to shorten")
(:ENGLISH_EQU "to abridge")
(:MALAY_EQU "memendekkan")
(:MALAY_EQU "meringkaskan")
)
```

2.5 Base ELRA

Cette base est composée de 6 dictionnaires. Un dictionnaire de concepts où chaque ligne représente un concept avec son numéro puis une définition du concept en français. Ensuite, pour l'anglais, le français, l'italien, l'espagnol et l'allemand ; chaque ligne correspond à un numéro de concept suivi de sa traduction et éventuellement d'une catégorie grammaticale. Chaque fichier occupe environ 12 Mo. Voici le concept n° 92 et ses traductions anglaises et françaises.

```
92;E;abbreviate;v_trans
92;F;abréger;v_trans
```

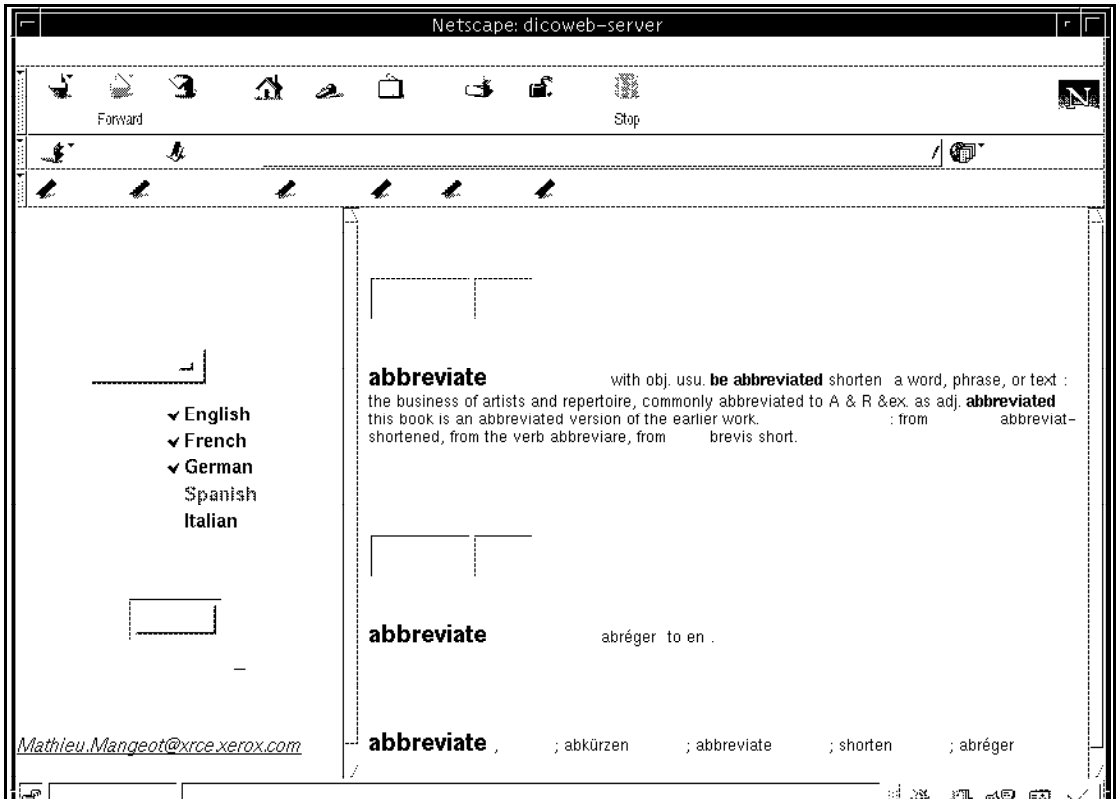
3. Le serveur

Ce serveur de dictionnaires est conçu pour un usage humain. Il sert pour des expérimentations. Pour des raisons légales, il n'est pas accessible au public. Je présenterai son interface, son architecture et quelques points importants.

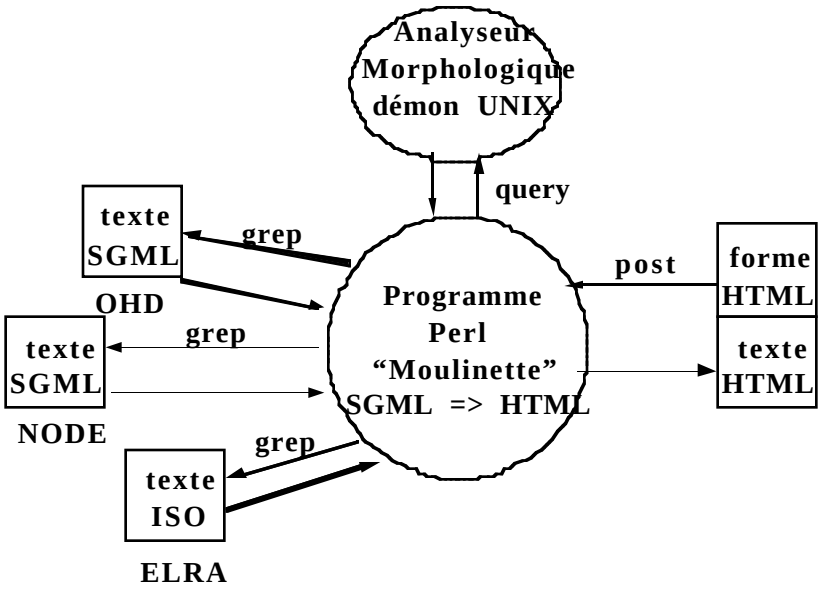
3.1. Interface

L'utilisateur sélectionne la langue source, dans laquelle il va taper l'entrée, puis les langues cibles qu'il désire. Il n'est possible de sélectionner qu'une seule langue source à la fois. Par contre, l'utilisateur peut choisir autant de langues cibles qu'il le souhaite. Il peut, avant de consulter les dictionnaires, envoyer le mot qu'il vient de taper à un analyseur morphologique en cochant la case correspondante. S'il clique sur les boutons "`previous`" ou "`next`" des parties "OHD" ou "NODE", il pourra consulter les entrées précédentes et suivantes correspondant, dans l'ordre alphabétique, à celles affichées.

Dans un souci de clarté, nous fixons au départ une seule couleur ainsi qu'une police différente pour chaque langue, qu'elle soit source ou cible, et cela pour tous les dictionnaires. L'utilisateur s'habituera ainsi à ce mode de représentation.



3.2. Architecture générale



Un script cgi écrit en perl fait la liaison entre l'utilisateur, les analyseurs morphologiques et les dictionnaires. Lorsque l'utilisateur a choisi ses langues source et cibles puis tapé son entrée, le résultat est envoyé au script. Si l'analyse morphologique est sélectionnée, le dit script envoie l'entrée à l'analyseur morphologique correspondant à la langue source. La réponse est ensuite décodée.

Les dictionnaires sont alors sélectionnés en fonction des langues cibles et les fichiers texte originaux sont parcourus par le script qui cherche l'entrée décrite par une expression régulière perl. Les lignes vérifiant l'expression régulière sont alors sélectionnées puis passées à travers une "moulinette" qui transforme le texte source en HTML. Le tout est renvoyé sous forme de page HTML à l'utilisateur.

3.3. Analyse de l'entrée

Lorsque l'utilisateur a sélectionné l'analyse morphologique, l'entrée est d'abord envoyée par le script à l'analyseur morphologique correspondant à la langue source. Le résultat est ensuite décodé de façon à fournir une liste d'entrées plausibles. Ainsi, si l'utilisateur tape l'entrée "`cochons`", la liste des nouvelles entrées sera "`cocher`" et "`cochon`". Les analyseurs morphologiques sont des démons UNIX qui tournent en permanence. Ils répondent à des requêtes de différentes applications et étaient déjà utilisés avant que nous ne programmions cette interface.

Le but ici n'est pas de fournir une véritable recherche aidée par le contexte, mais de proposer une petite aide supplémentaire. En effet, il existe des outils spécialisés dans la recherche à l'aide du contexte. Ces outils évitent par exemple que, lorsque l'utilisateur tape "`cochons`", il obtienne l'entrée "`cocher, nom commun`" qui n'a rien à voir avec sa première demande. Notre système n'est pas conçu pour résoudre ce genre de problème. Cependant, l'analyse morphologique de l'entrée peut s'avérer utile lorsqu'on ne maîtrise pas la langue source. La liste des nouvelles entrées est ensuite utilisée par le script pour consulter les dictionnaires.

3.4. Recherche de l'entrée

Selon les langues sélectionnées, le script consulte les dictionnaires correspondants. Par exemple, si l'utilisateur ne sélectionne que l'anglais comme langue source et cible, le script consultera le dictionnaire NODE monolingue anglais et la base ELRA. S'il choisit le français comme langue source et l'anglais comme langue cible, le script consultera le dictionnaire OHD français-anglais et la base ELRA. S'il choisit l'espagnol comme langue source, le script ne consultera que la base ELRA. Les dictionnaires ne subissent aucune modification, ils sont consultés directement dans leur format d'origine.

Perl dispose d'un puissant langage d'expressions régulières. À chaque dictionnaire correspond une expression régulière. Pour chercher une entrée du OHD, par exemple, on utilisera le patron : `/<[hc]w>$entry</` où `$entry` représente l'entrée demandée.

Le dictionnaire FeM est unidirectionnel, du français vers l'anglais et le malais. Cependant, grâce aux expressions régulières, nous pouvons chercher la traduction d'un mot malais en français ou plus exactement, dans quelles entrées françaises apparaît ce mot malais. L'utilisateur pourra alors se faire une idée de la traduction française de celui-ci.

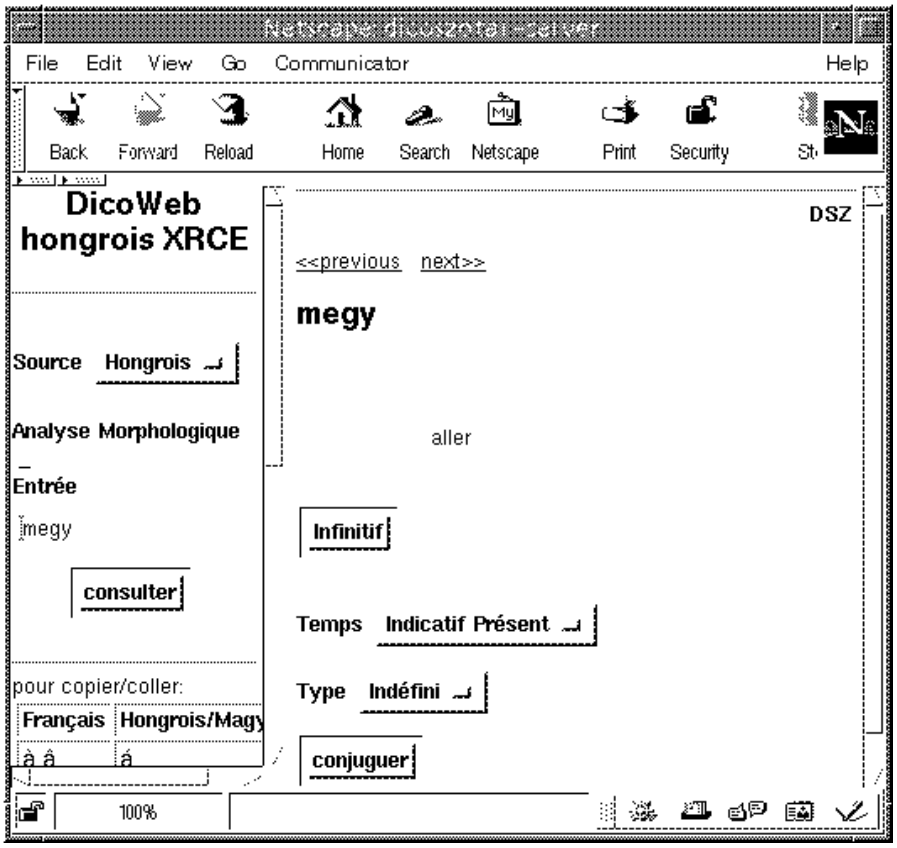
Pour la recherche dans la base ELRA, le script cherche d'abord les numéros de concept dans le dictionnaire correspondant à la langue source, puis cherche dans les dictionnaires correspondant aux langues cibles les traductions correspondant aux numéros de concept.

L'utilisateur peut profiter directement du langage d'expressions régulières. En effet, s'il tape une entrée sous forme d'expression régulière, celle-ci sera interprétée telle quelle par le script. Par exemple, si l'utilisateur tape “`b.ll`” (ici, le point correspond à n'importe quel caractère) et sélectionne l'anglais comme langue source, il obtiendra les entrées “`ball`”, “`bell`”, “`bill`”, “`boll`” et “`bull`”.

3.5. Entrée précédente et suivante

Pour les dictionnaires classés par ordre alphabétique (ici le OHD et le NODE), il est possible de consulter les entrées précédant et suivant celles affichées. Pour cela, lorsque le script consulte un dictionnaire à la recherche d'une entrée, il compte les lignes. Lorque l'utilisateur demande l'entrée précédente ou suivante, le script utilise ce numéro de ligne pour faire sa recherche. Elle s'effectue donc plus rapidement que lorsque le script effectue une recherche à l'aide d’une expression régulière. L'utilisateur se retrouve partiellement dans le contexte de la lecture d'un dictionnaire papier où le contexte de l'entrée est directement sous ses yeux.

3.6. Actions associées aux entrées



Nous avons associé des action principalement aux entrées de DicoSzotar[1], pour faciliter l'apprentissage du hongrois. L'accusatif ou le pluriel des noms hongrois n'étant pas réguliers, nous avons associé un générateur d'accusatif et de pluriel aux entrées correspondantes. Pour les verbes, nous avons associé un conjugueur. L'utilisateur recherche d'abord une entrée dans l'interface de départ. L'entrée s'affiche alors dans la partie droite du navigateur. Au bas de l'entrée, une nouvelle interface est affichée. Un bouton pour l'accusatif ou le pluriel des noms ou une liste pour sélectionner le temps et mode de conjuguaison que l'on désire. Les résultats s'affichent encore dans la partie droite de la fenêtre.

Pour la prononciation, nous pouvons associer un phonétiseur à chaque entrée. Pour le hongrois, nous disposions seulement des fichiers son. L'utilisateur peut donc en cliquant sur le bouton de son écouter la prononciation de l'entrée.

3.7. Pages fabriquées à la volée

Pour éviter de convertir à chaque fois le texte source en HTML, nous aurions pu convertir en une seule fois tous les dictionnaires source cependant, même si cette solution réduit le temps d'attente lors de la recherche d'une entrée, elle présente deux inconvénients importants. En effet, la fabrication à la volée des pages HTML permet d'une part de respecter le copyright en interdisant aux utilisateurs de récupérer entièrement le dictionnaire en une seule fois et d'autre part de retoucher le rendu final directement en modifiant le script perl.

3.8. Ajout d'une nouvelle ressource

---

[1] **Error! Bookmark not defined.**

Les critères que doivent satisfaire les nouvelles ressources pour être ajoutées au système sont simples : l'entrée doit soit être disposée sur une seule ligne, soit pouvoir être extraite à l'aide d'un outil simple comme sggrep (grep pour SGML). Il suffit alors de formuler l'expression régulière adéquate pour trouver l'entrée du dictionnaire, puis associer une feuille de style au texte pour le rendu final.

4. Discussion

Avec cet outil, nous sommes en mesure de proposer un système très simple qui permet de visualiser des sources hétérogènes. Cela permet entre autres de comparer les définitions et/ou traductions entre les différents dictionnaires. L'ajout d'une nouvelle ressource au système est lui aussi très simple puisque le texte d'origine ne subit aucune modification et les fichiers sont utilisés tels quels par le système. De plus, l'interface web permet évidemment une utilisation du système multiplate-forme et multi-utilisateurs.

Comme nous n'avons pas besoin d'analyser les ressources avant de les afficher, nous pouvons facilement intégrer dans ce système des ressources mal structurées du point de vue SGML ou contenant des erreurs. Il est même envisageable d'afficher des ressources incomplètes ou en cours d'élaboration. Nous privilégions cependant le format XML lors de la récupération d'un dictionnaire [correard98].

Cet outil a été développé pour un usage humain. Cependant, nous pourrions envisager de l'utiliser pour construire un dictionnaire "qui n'existe pas", résultant de la synthèse de toutes les ressources disponibles. Il serait par exemple possible de construire un dictionnaire en prenant la définition dans une source, les traductions dans d'autres, l'étymologie dans une troisième et les exemples dans un dernier. Cependant, l'intérêt d'un tel dictionnaire nous paraît discutable pour un usage humain. Une information brute, sans indications de provenance ou de contexte, est difficile à utiliser pour un humain. Il faudrait alors garder la provenance de chaque information.

5. Comparaison avec d'autres approches

5.1. Les travaux de M. Haï [Haï98]

Les outils développés par M. Haï permettent de récupérer des données non structurées puis de construire de nouveaux ensembles lexicaux. Notre système, lui, utilise des dictionnaires déjà structurés même si ceux-ci peuvent être incomplets. L'étape de la récupération a déjà été effectuée.

5.2. Le projet Blak [Fischer98]

Blak est un assistant de découverte des caractères chinois. Il utilise des ressources accessibles par Internet, mais dans un format compilé. Il faut donc y accéder à partir d'une interface spécialisée. Toutes nos ressources sont disponibles localement et dans un format textuel. Nous pouvons donc accéder directement aux données sans passer par une interface spécialisée.

5.3. Le projet INTERLEX [INTERLEX]

Le but de ce projet est de convertir des dictionnaires bilingues ou multilingues (généraux et terminologiques), disponibles sur cédérom ou au format papier, en ressources électroniques accessibles par Internet. Ce projet européen (MLIS) est multipartenaire (université et industriels) et est prévu pour durer 18 mois. Toutes les ressources seront

converties dans un format standard (GENETER). Par la suite,les partenaires pourront accéder par une interface ad hoc à une base de données regroupant toutes les ressources.

Notre technique se distingue de ce projet par sa simplicité. En effet, il nous aura fallu moins d'une personne/mois pour la développer et une demi-journée pour inclure le dictionnaire FeM. De plus, nous gardons le format d'origine des dictionnaires sans les modifier.

6. Conclusion

Le but de cette expérience a été atteint : nous pouvons, à l'aide d'un outil simple, accéder par une interface unique à un grand nombre de dictionnaires hétérogènes et ajouter une ressource au système avec un minimum de développement. Suite au succès de cet outil, nous avons réutilisé cette méthode pour développer un serveur accessible au public[2].

Nous n'envisageons pas de développement ultérieur de cet outil puisque ce n'est pas un produit finalisé. Cependant, cette étude a dégagé quelques perspectives : il serait par exemple intéressant de pouvoir fournir à l'utilisateur un fichier de préférences pour qu'il puisse fabriquer des entrées sur mesure. Par exemple, certains utilisateurs préfèrent les exemples d'un dictionnaire et l'étymologie d'un autre, etc. D'autre part, pour éviter le passage du texte d'origine dans une "moulinette" pour produire du HTML, il serait peut être intéressant d'utiliser le langage XML [Connolly98]et d'associer une feuille de style à chaque dictionnaire. Le texte d'origine ne subirait alors aucune transformation, pourvu que celui-ci soit représentable en XML.

Références

Atkins, B. T. S. and Zampolli A. (1994) *Computational Approaches to the Lexicon.* Oxford University Press, 480 p.

Connolly, D. (1997) *XML Principles, Tools and Techniques* World Wide Web Journal, Volume 2, Issue 4, Fall 1997, O'REILLY & Associates, 250 p.

Corréard, M-H. & Grundy, V. (1994) *Le dictionnaire Hachette-Oxford.* Oxford University Press & Hachette, 1950 p.

Corréard, M-H. & Mangeot-Lerebours M. (1999) *XML- A Solution For LDBs, Eds and MRDs?,* Proc. COMPLEX'99, Pécs, Hungary, 6 p.

Haï, D. (1998) *Conception, implémentation et expérimentation de techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes.* Thèse de nouveau doctorat, Spécialité Informatique, Institut National Polytechnique de Grenoble, 168 p.

ELRA http://www.icp.inpg.fr/ELRA/

Fischer, L. et al. (1998) *BLAK, un assistant de découverte des caractères chinois fonctionnant par accès dynamique à des ressources lexicales.* Proc. NLP+IA '98, Moncton, N.B., Canada, 18 au 21 aout 1998, vol. 1/2, pp. 13-17.

Heid et al. (1992) *Extracting linguistic information from machine-readable versions of traditional dictionaries, a metalexicographic method and some tools.* Proc.

[2] **Error! Bookmark not defined.**

COMPLEX'92, Conference on Computational Lexicography an Text Research, Budapest, Hongrie, Linguistics Institute, Hungarian Academy of Sciences, Budapest, pp. 161-174.

Ide, N. and Veronis, J. (1995) *Text Encoding Initiative, background and context.* Kluwer Academic Publishers, 242 p.

INTERLEX `http://interlex.uax.es/`

Lafourcade, M. (1996) *Serveurs de dictionnaires - Etude de cas avec l'outil ALEX et le projet de dictionnaire français-anglais-malais.* Proc. Séminaire LEXIQUE, Grenoble, 13 et 14 novembre 1996, CLIPS-IMAG, Pôles langage naturel et parole du GDR-PRC CHM., vol. 1/1, pp. 185-192.

Mangeot-Lerebours, M. (1998) *Conception, implémentation et indexation de BaLeM, une base lexicale multilingue.* Proc. TALN'98, Paris, pp. 215-218.

Pearsall, J. (1998) *The New Oxford Dictionary of English,* Clarendon Press, Oxford, 2154 p.