Informatisation du Dictionnaire Explicatif et Combinatoire

Gilles Sérasset

GETA-CLIPS-IMAG (UJF & CNRS)
BP 53
38041 Grenoble cedex 9
Tél. : 04.76.51.43.80 - Fax : 04.76.51.44.05
Courriel : Gilles.Serasset@imag.fr

Résumé

Nous donnons ici un aperçu du logiciel DECID développé au GETA afin d'informatiser le processus de rédaction du dictionnaire explicatif et combinatoire du français contemporain.

1. Introduction

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmente avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires.

Nous nous sommes donc intéressé à la construction d'outils pour lexicographes et lexicologues. Afin d'avoir une bonne compréhension des problèmes qui se posent, nous avons décidé d'informatiser un dictionnaire complexe, contenant de nombreuses informations structurées, le dictionnaire explicatif et combinatoire du français contemporain (DEC). Le DEC étant un travail de lexicologie, il ne s'agit donc pas à proprement parler d'un dictionnaire, mais plutôt d'un ensemble d'entrées destinées à illustrer une théorie linguistique. Ce ne sont donc pas les données que l'on va informatiser, mais le processus de rédaction de ces données.

Cette action a été menée en collaboration entre le GETA-CLIPS (université Joseph Fourier – Grenoble 1) et le GRESLET (université de Montréal), grâce aux soutiens du réseau LTT de l'AUPELF-UREF et des ministères français et canadiens des affaires étrangères (coopérations franco-québecoises en ingénierie linguistique). Cette action nous a mené à la définition d'une structure informatique reflétant au mieux la structure logique du DEC, tel qu'il est défini par [Mel'ãuk et al. 95], et à la construction d'outils spécialisés pour l'édition du DEC.

Nous présentons ici une brève note de projet qui sera accompagnée d'une démonstration.

2. Aperçu du projet

2.1. Objectifs

Le projet NADIA-DEC est basé sur les travaux de définition d'un système universel de bases de données lexicales multilingues au laboratoire GETA-CLIPS et sur les travaux de définition du dictionnaire explicatif et combinatoire du français contemporain au laboratoire GRESLET.

Ce projet vise la création d'une version informatisée du dictionnaire explicatif et combinatoire du français contemporain. Cette version devra contenir l'ensemble des

informations présentes dans le DEC sous une forme aussi structurée que possible. Elle s'appuie donc sur le système de gestion de bases lexicales multilingues SUBLIM défini au GETA-CLIPS ([Sérasset 94]).

Nous ne visons pas d'application particulière des données ainsi informatisées, afin de ne pas privilégier certains aspects de la structure au détriment des autres. Nous souhaitons que le dictionnaire explicatif et combinatoire soit informatisé sans subir de modification fondamentale par rapport a sa version papier. Ceci garantira l'utilisation des outils développés par les rédacteurs de la version papier, ce qui est une condition de la réussite de ce projet (création de la version informatisée avant la version papier). De plus, nous pensons que l'ensemble des informations trouveront une exploitation dans la communauté TALN. Nous souhaitons donc que l'intégralité des informations présentes sur papier soit disponible sous forme informatique exploitable par chacun.

Ce projet répond à plusieurs motivations de la part de chacun des partenaires. D'une part, il permet de tester le système SUBLIM en l'utilisant pour un dictionnaire mettant en œuvre des structures complexes. D'autre part, les informations contenues dans le dictionnaire explicatif et combinatoire présentent une richesse que l'on ne trouve dans aucun autre dictionnaire informatisé. Enfin, la mise en œuvre de ce projet suppose la création d'outils informatiques simplifiant la gestion d'un tel dictionnaire.

Pour atteindre ces différents objectifs, nous souhaitons non seulement informatiser une version du DEC, mais surtout informatiser sa chaîne de production afin de faire en sorte que le DEC existe d'abord sous forme informatique puis sous une forme imprimée. Ainsi, le travail à effectuer se décompose en différentes étapes :

1. création d'un éditeur informatique spécialisé pour le DEC,
2. réalisation d'un mécanisme d'import des données existantes, actuellement au format R.T.F. (Rich Text Format),
3. réalisation d'un module d'export vers différents formats utilisables informatiquement (SGML/TEI, format LISP…) et pour la publication papier (R.T.F., M.I.F.…).
4. intégrer dictionnaire et éditeur à un système de gestion de données lexicales générique (SUBLIM). Les vérifications des différentes contraintes sur les données seront vérifiées à ce niveau,

Les point 1, 2 et 3 sont en cours de réalisation et sont assez avancés. Le point 4, à plus longue échéance, en est à ses débuts.

L'éditeur DECID créé au cours du projet sera mis à disposition de l'ensemble de la communauté TALN.

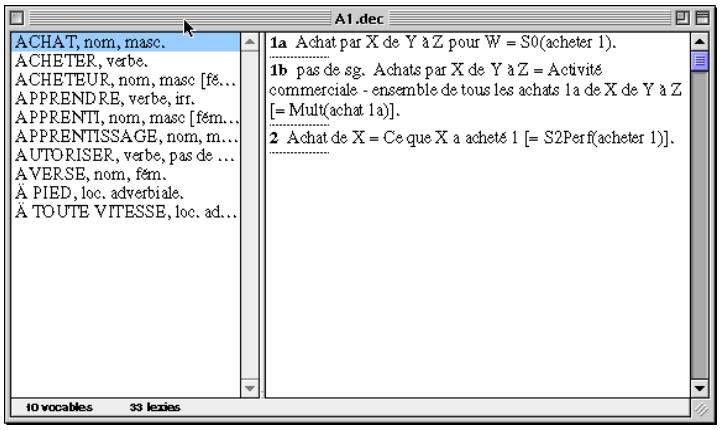
2.2. Originalités

D'autres projets ayant pour but l'informatisation du DEC sont en cours, notamment à l'université de Montréal ([Mel'ãuk & al. 1995], chap. 4) par Alain Polguère. Ces projets utilisent des approches différentes :

- Pour Alain Polguère, l'informatisation du DEC est un moyen de disposer de données pour des traitements automatiques. Ainsi, seules les données complètement formalisables sont retenues lors de cette informatisation. De plus, certaines de ces données sont légèrement modifiées par rapport à la version imprimée et d'autres sont ajoutées.
- le projet NADIA/DEC vise l'informatisation du processus de production du DEC, il doit donc prendre en compte l'ensemble des informations présentes dans le DEC, et ce, quel que soit leur degré de formalisation.

Les deux approches ne sont pas antinomiques et nous travaillons actuellement à l'adaptation de l'éditeur DECID pour la génération (au moins partielles) d'entrées lexicales utilisables par Alain Polguère. Nous envisageons aussi d'augmenter les fonctionnalités de l'éditeur actuel, afin d'intégrer dans le processus de création d'entrées du DEC, la saisie de données propres au formalisme utilisé à l'université de Montréal.

3. L'éditeur spécialisé DECID



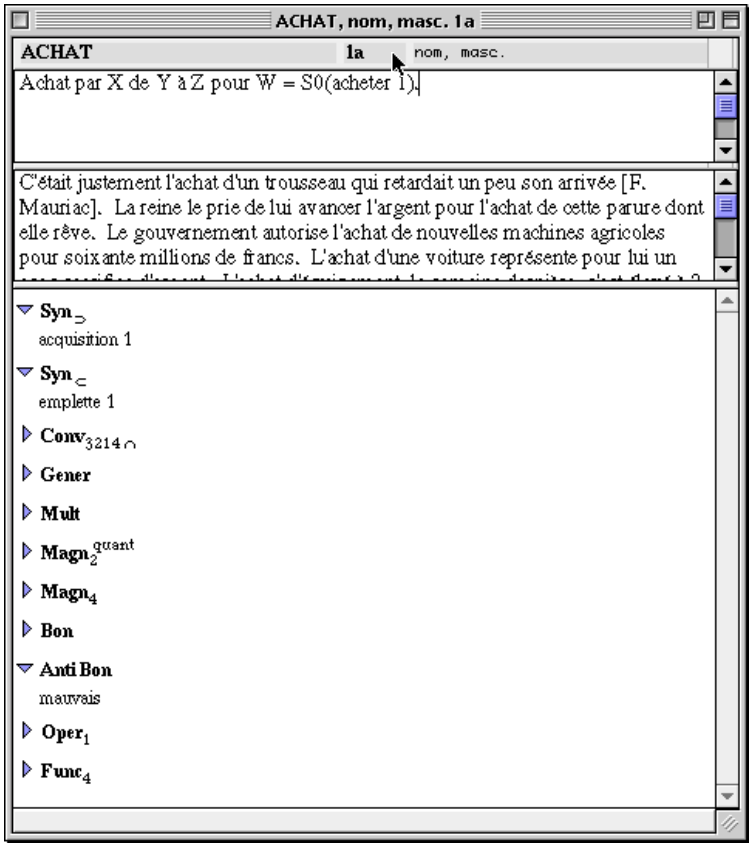
Dès que l'on créé ou que l'on ouvre un dictionnaire, la fenêtre principale du dictionnaire apparaît. Elle se décompose en deux parties. La première (liste de gauche) contient l'ensemble des vocables définis dans ce dictionnaire. Lorsque l'on sélectionne un (ou plusieurs) vocables dans cette liste, les lexies correspondantes apparaissent dans la liste de droite.

Ces lexies apparaissent sous forme d'un numéro de sens (lorsqu'il y en a) suivi d'un résumé. C'est ce résumé qui est utilisé dans la version papier comme tableau synoptique d'un vocable. Ces résumés sont éditables.

En double-cliquant sur un résumé, on ouvre la fenêtre de la lexie correspondante.

Il faut noter que l'unité du lexique est la lexie et non le vocable. Les informations représentées dans la colonne de gauche sont calculées « à la volée » à partir de l'ensemble des lexies du dictionnaire. Elles ne sont pas stockées dans le dictionnaire.

La fenêtre de lexie comporte quatre parties principales. De haut en bas, on remarque les informations graphiques et morphologiques, puis la définition, les exemples et enfin la liste des fonctions lexicales.

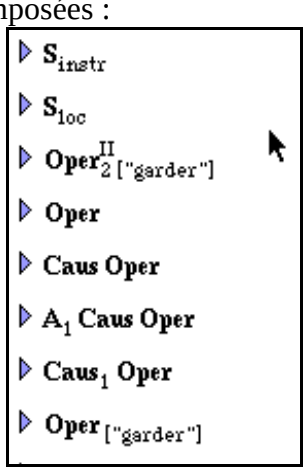


Dans cette fenêtre, chaque élément est éditable. On peut ainsi à tout moment changer la graphie, la définition, le numéro de lexie ou toute autre information.

Les informations graphiques et morphologiques apparaissent comme dans le dictionnaire papier. Il est possible de spécifier la portée de ces informations (propres à une lexie ou commune à toutes les lexies du vocable).

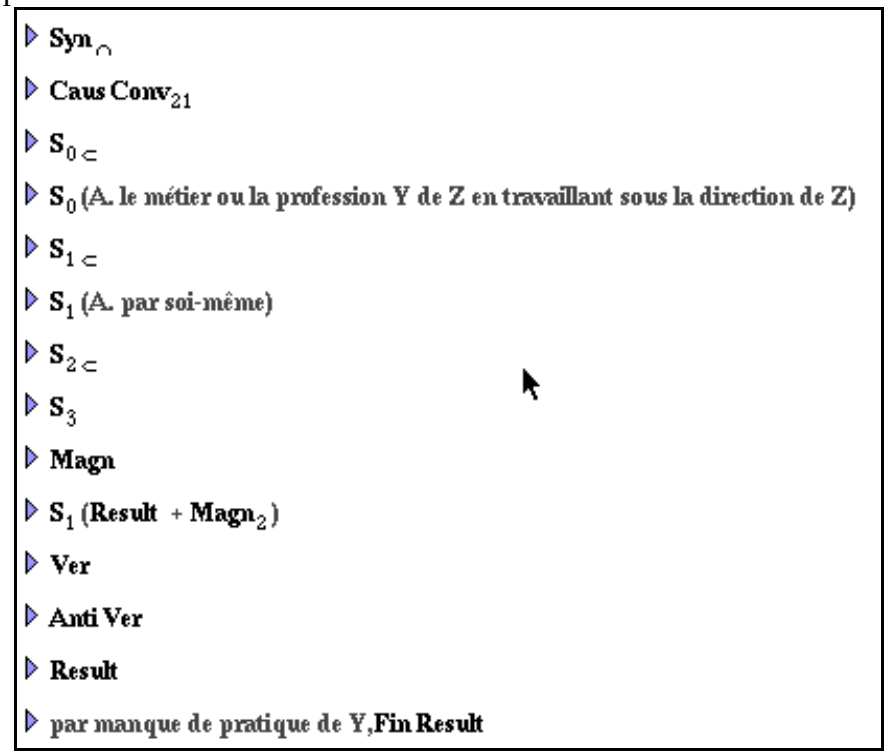
Les fonctions lexicales sont présentées sous forme de liste. Chaque élément de cette liste est éditable. L'éditeur DECID permet d'éditer de manière très simple les fonctions lexicales. Ainsi, il est possible de créer la fonction lexical composée **LiquOper$_1$** en tapant simplement **l**, **i**, **q**, **u**, **o**, **p**, **e**, **r**, **1**.

Actuellement, l'interface de création des fonctions lexicales ne permet que la création de fonctions standards simples ou composées :



Fonctions lexicales simples et composées

Elle ne permet pas encore la création de fonctions lexicales non standard ou complexes. Par contre, le noyau de représentation d'une fonction lexicale, ainsi que le module d'affichage sont suffisamment avancés pour gérer ce type de fonction. Ainsi, il est possible de récupérer et de visualiser des fonctions lexicales non standard ou complexes à partir des fichier word du dictionnaire publié.



Fonctions lexicales non standard et complexes

4. Discussions et perspectives

L'éditeur présenté ici, n'est qu'une première étape vers un outil véritablement utilisables par les lexicographes souhaitant éditer un dictionnaire explicatif et combinatoire. Pour cet objectif, il nous faudra notamment le doter d'un système de vérification de contraintes de cohérence et l'intégrer à des outils plus généraux utilisés par tout lexicographe (recherches de collocations dans un corpus, outils d'acquisition d'information…).

Il nous a néanmoins permis de mieux comprendre les besoins des lexicographes en terme d'outils. Certes, l'intégration à des outils plus généraux est importante. Mais de très nombreuses fonctionnalités plus simples, que nous n'envisagions pas a priori, sont tout aussi importantes. On nous a ainsi demandé des filtres permettant d'extraire différentes informations, telle qu'une liste de toutes les abréviations utilisées dans le dictionnaire (afin de détecter des incohérences dans les notations de différents lexicographes).

Les requêtes étant très nombreuses, il nous apparaît qu'un outil pour lexicographe doit fournir un langage de description de filtres permettant vérifications et extractions de données diverses.

Une autre difficulté provient de la nature même du dictionnaire explicatif et combinatoire qui n'est pas un travail de lexicographie, mais un travail de lexicologie. La structure des entrées est donc en permanente évolution. On peut ainsi observer des différences dans les informations des quatre volumes actuellement publiés. Cette constante évolution de la structure logique (qui apparaît dans une moindre mesure dans d'autres projets de dictionnaires) rend impossible la maintenance d'un outil ad-hoc.

Nous souhaitons donc maintenant orienter nos développement vers l'intégration des aspects d'interface dans le cadre d'un outil générique de gestion de bases de données lexicales multilingues, tel que nous l'avons défini dans [Sérasset 1994]. Une telle approche nous permettra de construire une plate-forme générique où l'on pourra paramétrer à la fois les structures de données utilisées mais aussi les différentes "vues" (interfaces) de ces structures.

5. Bibliographie

Mel'ãuk I., Arbatchewsky-Jumarie N., Eltnisky L., Iordanskaja L. & Lessard A. (1984) *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques I.* Presses de l'université de Montréal, Montréal (Québec), Canada, 172 p.

Mel'ãuk I., Arbatchewsky-Jumarie N., Dagenais L., Elnitsky L., Iordanskaja L., Lefebvre M.-N. & al. (1988) *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques II.* Presses de l'université de Montréal, Montréal (Québec), Canada, 332 p.

Mel'ãuk I., Arbatchewsky-Jumarie N., Iordanskaja L. & Mantha S. (1992) *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques III.* Presses de l'université de Montréal, Montréal (Québec), Canada, 323 p.

Mel'ãuk I., Clas A. & Polguère A. (1995) *Introduction a la lexicologie explicative et combinatoire.* Universités francophones et champs linguistiques, AUPELF-UREF et Duculot, Louvain la Neuve, 256 p.

Sérasset G. (1994) *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions.* Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1, 194 p.