De l'article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux*

Jacques Steinlin†, Sylvain Kahane†, Alain Polguère‡, Adil El Ghali†

† LaTTice
UFRL, case 7003
Université Paris 7 — Denis Diderot
75251 Paris Cedex 05, France
{jsteinli,sk,adil}@linguist.jussieu.fr

‡ OLST
Département de linguistique et de traduction
Université de Montréal, CP 6128
Montréal (Québec) H3C 3J7, Canada
alain.polguere@umontreal.ca

Résumé

Cet article s'intéresse à l'architecture d'un dictionnaire centré sur la description formelle des collocations et de la dérivation sémantique. Notre projet s'appuie sur un dictionnaire électronique, le DiCo, qui suit un découpage traditionnel en articles associés à des entrées lexicales. Nous proposons une représentation objet du contenu du DiCo consistant en un découpage en structures élémentaires reliées les unes aux autres. Nous montrons en quoi cette nouvelle ressource, le DiCobjet, permet un accès plus facile aux données et change la vision même du dictionnaire, lequel peut être appréhendé de multiples façons et complètement réorganisé selon des axes originaux tels que les fonctions lexicales.

1 Introduction

Ce travail s'inscrit dans un projet de réalisation d'un dictionnaire électronique centré sur la collocation et la dérivation sémantique. Le dictionnaire relève de la lexicologie explicative et combinatoire, elle-même composante de la Théorie Sens-Texte ((Žolkovskij & Mel'čuk, 1965), (Mel'čuk *et al.*, 1995)). Le projet, initié à l'Université de Montréal par Igor Mel'čuk et Alain Polguère, a pour but de développer un dictionnaire centré sur la modélisation formelle des collocations et de la dérivation sémantique du français[1]. Ce dictionnaire électronique, le DiCo (= Dictionnaire de Combinatoire), est conçu pour servir de description lexicographique générique, description dont peuvent être dérivées i) des bases de données « calculables » telles que celle présentée ici et ii) des descriptions « vulgarisées » grand public du type Lexique Actif du Français ou LAF (Polguère, 2000). Le DiCo et le DEC possèdent une structure assez traditionnelle, inspirée des dictionnaires papier d'usage courant type Petit Robert ou Larousse, et adopté par

*Ce projet est financé par une bourse de recherches Blaise Pascal attribuée à Igor Mel'čuk, gerée par la Fondation de l'Ecole Normale Supérieure et financée par la Région Ile de France.

[1] Le projet a déjà produit la modélisation d'environ 2000 vocables parmi les plus riches en liens lexicaux. Ce nombre est en constante augmentation, de même que la qualité descriptive des articles.

de nombreux dictionnaires électroniques (TLFi (Dendien & Pierrel, 2003), eEH (Arregi *et al.*, 2003)...). Les entrées du dictionnaire sont les vocables (y compris les locutions). Chaque vocable est un regroupement de lexies (unités lexicales) correspondant aux diverses acceptions. Notre objectif est ici de proposer pour le DiCo une architecture plus riche et plus souple afin d'améliorer l'accès aux données. Cette architecture vise d'une part à améliorer les possibilités de consultation du DiCo par des utilisateurs humains et, par là-même, à rendre les informations lexicographiques exploitables par des systèmes de traitement automatique de la langue (notamment en génération de textes). D'autre part, elle ambitionne de faciliter toutes les tâches de mise à jour et de développement du dictionnaire, c'est-à-dire les tâches lexicographiques proprement dites. Pour cette raison, nous avons choisi de réaliser une analyse du dictionnaire existant, le DiCo, et de le transporter dans une base de données relationnelle via une modélisation objet[2]. Nous nous proposons dans cet article de présenter le résultat de cette modélisation, le DiCobjet, et de montrer les avantages d'une telle structure aussi bien au niveau théorique qu'au niveau pratique. Cette structuration peut être comparée à celle qui est à l'œuvre dans WordNet (Fellbaum, 1998). Toutefois, comme nous le verrons, notre réseau ne relie pas seulement les lexies par le biais des relations lexicales, mais aussi par l'ensemble des informations constituant leur description (comme par exemple le régime). Nous introduisons d'abord le DiCo tel qu'il est développé par le lexicographe, puis nous nous intéresserons à sa structure, telle que l'on peut la représenter dans une modélisation objet. Nous terminerons sur les bénéfices théoriques et pratiques obtenus par la modélisation.

2 Présentation du DiCo

Comme nous l'avons dit, le DiCo adopte une structure proche de la structure traditionnelle d'un dictionnaire papier, c'est-à-dire une structuration en articles lexicographiques. Cette structuration est requise par le lexicographe qui veut avoir une vision globale des informations se rapportant à la lexie. La description à l'intérieur de l'article se décompose en rubriques correspondant chacune à un certain type d'information. Les neuf principales sont présentées ici (voir ci-dessous dans le Tableau 1, l'entrée de la lexie ENGUEULADE1)[3] .

Les deux premières rubriques contiennent le nom du vocable et le numéro lexicographique de la lexie. Le champ caractéristiques grammaticales permet de consigner la partie du discours de la lexie ainsi qu'un certain nombre d'autres informations relatives à sa combinatoire : marque d'usage (ex : "fam"), variantes graphiques, contraintes sur la flexion (ex : `pas de pl.`), etc. L'étiquette sémantique inscrit la lexie dans une hiérarchie nécessaire pour décrire la cooccurence lexicale libre des lexies (pour une description détaillée de l'étiquetage sémantique des lexies, voir (Polguère, 2003)). La forme propositionnelle énumère les actants sémantiques de la lexie et leur affecte éventuellement une étiquette sémantique. Le tableau de régime indique par quelles structures syntaxiques régies s'expriment les actants sémantiques. Les champs synonymie et fonctions lexicales décrivent les relations lexicales de la lexie (Mel'čuk *et al.*, 1995). Le nom des fonctions lexicales est noté entre accolades et précède une liste de valeurs. Les fonctions lexicales comme **Magn** ou **Oper**$_i$ permettent de décrire les collocations auxquelles participe l'unité lexicale considérée : *une belle eugueulade*, *une sacrée engueulade* pour l'in-

[2](Fontenelle, 1997) détaille une modélisation de données lexicales similaire, mais concentrée sur les seules fonctions lexicales.

[3]Pour une présentation plus complète de la structure et des formalismes du DiCo, voir (Polguère, 2000) et (Polguère, 2003)).

Champ	Valeur
vocable	ENGUEULADE
acception	1
caractéristiques grammaticales	nom, fém, "fam"
étiquette sémantique	communication langagière
forme propositionnelle	∼ DE L'individu X [VISANT L'individu Y] POUR LE fait Z
tableau de régime	X = I = de N, A-poss Y = II = -- Z = III = Prep-pour N | Prep-pour = {_à propos de_, _au sujet de_, pour}, pour V-inf-passé
synonymes	{QSyn} "fam" savon ; "soutenu" admonestation, remontrance, réprimande ; blâme
fonctions lexicales	{QAnti} compliment, félicitation ; flâtterie {V0} engueuler {Magn} belle, bonne, sacrée | antepos < majeure | postpos {Oper213} essuyer [ART ∼ _de la part de_ N=X Prep-pour N=Z] {Oper23} subir [ART ∼ Prep-pour N=Z]
phraséologie	

FIG. 1 – Description partielle de la lexie ENGUEULADE1

tensification, ou *recevoir une engueulade, subir une engueulade* pour les constructions à verbe support. Enfin le champ phraséologie recense l'ensemble des phrasèmes construits sur la lexie. Il est vide ici, ENGUEULADE1 n'ayant pas de phrasèmes associés.

Au plan pratique, une base de données relationnelle mono-table sert de support de stockage au DiCo. Chaque entrée constitue une fiche, les rubriques correspondant à autant de champs à l'intérieur de la fiche. Cet aspect n'est pas anecdotique. En effet, le DiCo est développé dans la perspective d'en faire la source de deux dictionnaires différents : un dictionnaire grand public papier (le LAF (Polguère, 2000) et un dictionnaire électronique élaboré pour les applications de Traitement Automatique des Langues (le DiCobjet). À cet effet, au-delà d'un premier niveau de structuration en champs, on trouve un second niveau de structuration, syntaxique celui-ci. Chaque champ est rédigé en respectant un certain nombre de conventions syntaxiques et typographiques. Il devient ainsi possible d'identifier automatiquement les notions-clefs de la description. Par exemple, on observe que dans le champ forme propositionnelle certains éléments apparaissent en minuscules (`individu` ou `fait`). Ceci indique que les chaînes de caractères correspondent aux étiquettes sémantiques des actants sémantiques, eux-mêmes désignés par des capitales (`X`, `Y`, `Z`, `U`, `V`, `W`). Ce dispositif ayant été adopté pour tous les champs, nous avons pu concevoir un compilateur chargé de traduire automatiquement le DiCo en DiCobjet. Un travail similaire a pu être effectué dans le cadre du projet Papillon pour la traduction du DiCo en un format XML (Lapalme & Sérasset, 2003).

Les auteurs du DiCo postulent que la méthodologie lexicographique la plus adaptée à la réalité lexicale est de procéder « traditionnellement », en rédigeant des fiches lexicographiques qui sont de véritables articles de dictionnaire. Cette approche voit la modélisation d'une lexie comme un texte, plutôt qu'un formulaire à remplir. L'approche formulaire présente l'avantage de produire directement des données dans un format approprié aux traitements informatiques.

Elle bride cependant considérablement le lexicographe, le forçant à s'en tenir à une structure de fiche rigide préétablie, alors que chaque lexie peut nécessiter une stratégie particulière. L'approche du DiCo permet de « rédiger » la description tout en se donnant les moyens — par une formalisation de l'encodage — de générer a posteriori une base de type DiCobjet. On peut comparer cette méthodologie à celle de dictionnaires développés directement dans un modèle objet où la collecte des informations est contrainte par le format adopté a priori (comme c'est le cas pour le DAFLES (Selva *et al.*, 2003)). Nous allons maintenant décrire le DiCobjet.

3 Modèle objet du DiCo

3.1 Présentation générale du modèle objet

Chacun des champs décrits à la section précédente contient en fait lui-même des listes d'informations que l'on peut à nouveau découper en champs et ainsi de suite. C'est l'objectif de la compilation. Le contenu du DiCo est transmis au compilateur sous forme d'un texte tabulé où les articles sont à la suite les uns des autres. Le compilateur réalise une analyse syntaxique du fichier et produit une représentation sémantique, le DiCobjet. Un des objectifs que nous nous sommes fixés en réalisant la compilation était d'obtenir une représentation la plus détaillée possible. Mais au fond, la modélisation consiste seulement à mettre au jour (en l'explicitant) la structure du DiCo. Ainsi, en examinant de nouveau la figure 1, on peut s'apercevoir que le champ fonction lexicale est en réalité une liste de fonctions lexicales. Une fonction lexicale est elle-même une liste de valeurs associées à un nom de fonction lexicale (par exemple, $\mathbf{Oper}_{213}$). Chacune de ces valeurs (un verbe support dans le cas présent) peut à son tour être organisée en champs : la valeur proprement dite, sa marque d'usage (`"soutenu"`), son régime, une condition (par exemple `postpos` pour un adjectif), etc. Le travail du compilateur consiste donc à identifier les atomes d'information et à les organiser dans une structure de données adéquate. La structure de données représentant le DiCo est le DiCobjet, dont le schéma est donné dans la figure 2. Dans celle-ci, on a représenté les classes (c'est-à-dire la définition des objets) par des boîtes. La partie haute de la boîte est réservée au nom de la classe. La partie inférieure énumère les attributs qui sont de type primitif (entiers, booléens, chaînes de caractères). Les autres attributs, qui ont pour valeur des classes, sont représentés par les relations entre les classes. Ces relations, bien qu'orientées, sont bi-directionnelles. On les appelle relations d'aggrégation. Une telle relation signifie que la classe se trouvant du côté du losange est propriétaire de la ou des classes à l'autre extrémité du trait. Les valeurs notées à ces extrémités donnent la cardinalité, à savoir le nombre d'objets attendus dans l'aggrégation.

La figure 2 doit se lire de la façon suivante : un DiCobjet est une collection (de taille supérieure ou égale à 0) d'objets Vocable. La classe Vocable est caractérisée par un certain nombre d'attributs (seul l'attribut nom étant obligatoire) et une collection d'objets Lexie (la collection devant comprendre au moins un élément). Une Lexie est à son tour décrite comme contenant des collections d'objets complexes : une ÉtiquetteFormule, un ensemble d'actants sémantiques (AsemVar), un ensemble de tableaux de régime (TableauRégime), des Phrasème, une caractérisation grammaticale (Cg) et des RubriqueFL. Une RubriqueFL, en plus d'être définie par un nom, regroupe un ensemble d'objets de type FL, cette dernière classe consistant en la mise en relation d'une FormuleFL (ce que l'on appelle habituellement une Fonction Lexicale), d'une Glose (la paraphrase en langue naturelle de la fonction lexicale) et d'une série de ValeurFL (les valeurs renvoyées par la fonction lexicale appliquée à la lexie décrite). Pour finir, une Formu-

leFL est une structure d'arbre combinant un certain nombre de FLPrimitive. La classe ValeurFL contient nombre d'attributs. Parmi ceux-ci, la valeur à proprement parler (par exemple, FLATTERIE) est une chaîne de caractère. Cette modélisation n'est pas satisfaisante dans la mesure où la valeur est en réalité un objet plus complexe. Dans le cas d'une unité lexicale complexe comme une locution, nous donnons sa composition syntaxique sous la forme d'un arbre de dépendance.

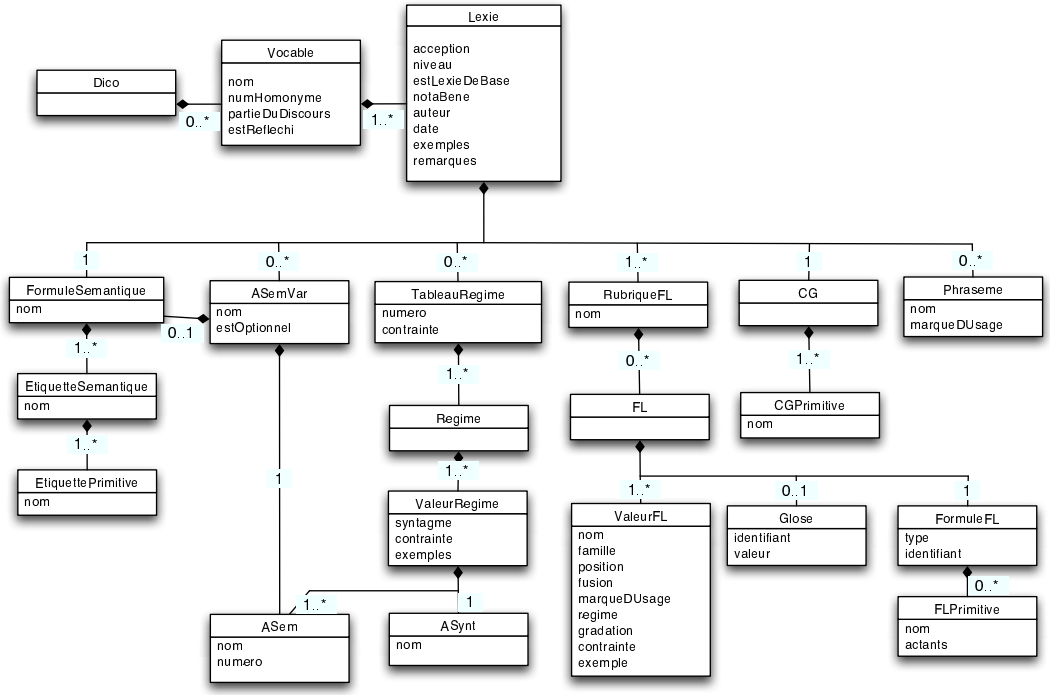

FIG. 2 – Modèle objet du DiCo

Il faut souligner que la relation d'aggrégation ne préjuge pas de l'orientation du parcours entre les objets. Si le TableauRégime est clairement un attribut d'une Lexie, il est aussi possible de remonter d'un TableauRégime donné aux objets Lexie qui le possèdent. Nous reviendrons dans la conclusion sur les conséquences que cela peut avoir sur la vision même du dictionnaire.

3.2 Exploitation

Les intérêts d'une modélisation objet sont multiples. Il devient possible d'effectuer des requêtes sophistiquées[4] et de filtrer le résultat de celles-ci afin de ne conserver que l'information pertinente (au lieu des fiches complètes). On peut par exemple récupérer la liste des lexies ayant au moins deux actants dont le premier se réalise par le régime `de N, A-poss` ou encore les lexies ayant à la fois deux actants sémantiques et un **Real**$_{@}$[5].

À cette augmentation de la puissance d'interrogation, il faut ajouter la possibilité d'effectuer des recherches sous d'autres angles que celui imposé par la structuration en vocables. On peut par exemple considérer le dictionnaire du point de vue des fonctions lexicales. Ainsi la version

[4]En réalité, rien dans le format initial du dictionnaire ne s'oppose à la réalisation de telles requêtes, à condition de disposer d'un dialecte d'expressions rationnelles. La représentation que nous proposons mâche le travail de l'utilisateur en segmentant l'information qui devient immédiatement disponible (sans traitement particulier).

[5]@ note un actant externe, c'est-à-dire ne figurant pas dans la forme propositionnelle.

actuelle du DiCo donne 143 valeurs différentes pour la fonction lexicale **Oper**$_1$ (la figure 3 en donne un échantillon). On récupère également 48 fonctions lexicales différentes renvoyant la valeur *donner* (voir la figure 4).

nb	Valeur	nb	Valeur	nb	Valeur	nb	Valeur
64	avoir	14	pousser	1	aller	1	être pris
42	être	11	émettre	1	battre	1	jauger
27	faire	10	constituer	1	célébrer	1	occuper
20	posséder	9	_faire preuve_	1	comprendre	1	partir
17	ressentir	8	se trouver	1	couvrir	1	peupler
15	éprouver	(...)	(...)	1	être (présent)	1	proférer

FIG. 3 – Valeurs d'**Oper**$_1$ par ordre de fréquence

nb	FL	nb	FL	nb	FL
10	{CausFunc1}	3	{Liqu1Oper1}	2	{Labreal21}
7	{Oper12}	3	{Son}	2	{Liqu1Real1}
6	{Oper1}	2	{Caus1Func0}	2	{Real31–I}
5	{Oper2}	2	{Caus de nouveau Func1}	2	{Real31–II}
4	{Oper13}	2	{Caus1Func0}	2	{Realo–II}
4	{Real12}	2	{Fact1}	1	{PredAble2}
3	{Caus3Func0}	1	{Fact21-production}	...	...

FIG. 4 – Fonctions lexicales renvoyant *donner* comme une des valeurs

Il devient possible également de concevoir une aide au lexicographe en proposant des patrons de fiches. Par exemple, on peut projeter des fiches à partir des valeurs de fonctions lexicales. L'entrée de SAVONII (SAVON dans le sens de 'réprimande' — *Quel savon il s'est fait passer par le proviseur*) se construit naturellement à partir de ENGUEULADE1, puisque les deux lexies sont en relation de quasi-synonymie (fonction lexicale **QSyn**). On peut donc proposer automatiquement au lexicographe une fiche pour la lexie comportant une étiquette sémantique, une forme propositionnelle, un tableau de régime et les fonctions lexicales **Syn**, **QSyn**, **Gener** et **Anti** ainsi que les valeurs correspondantes. D'une façon générale, toutes les valeurs renvoyées par une fonction lexicale sont des lexies et doivent donner lieu à la création d'une fiche.

Ainsi, *subir* qui est renvoyée par la fonction lexicale **Oper**$_{23}$(ENGUEULADE1) — *Il a encore subi une méchante engueulade pour son retard d'hier* — doit figurer dans le DiCo comme ayant au moins trois actants syntaxiques dont le premier est de type `individu`, le deuxième de type `communication langagière` et le troisième de type `fait`. Ce dernier se réalise par un groupe prépositionnel introduit par les prépositions de la famille spécifiée (`[ART Prép-pour N=Z]`[6]). Il apparaît par ailleurs que, lorsque l'on s'attache à la description de lexies apparentées, il est intéressant de croiser les informations déjà disponibles afin d'obtenir un squelette d'article. Cette procédure est déjà appliquée, mais à la main. Notre modélisation permet de comparer précisément les articles et d'en extraire l'intersection.

[6]Profitons-en pour dire que `Prep-pour` renvoie à une liste de prépositions (ici `_à propos de_`, `_au sujet de_`, `pour`) qui est utilisée à plusieurs endroits de l'article et reçoit pour cette raison un nom.

3.3 Apports théoriques et pratiques du DiCobjet, et état d'avancement

Le DiCobjet a été extrait automatiquement du DiCo. Ceci était envisageable grâce à l'usage de conventions strictes dans la rédaction du DiCo (seuls quelques ajustements mineurs ont été nécessaires pour réduire les ambiguïtés existantes). À noter qu'une des exigences principales des lexicographes est de pouvoir continuer à développer le DiCo sous son format initial (ou un format équivalent). En effet, tel que nous l'avons dit plus haut, la rédaction d'un dictionnaire ne peut se résumer à un remplissage à l'aveugle de champs prédéfinis et le lexicographe doit pouvoir appréhender un article de dictionnaire dans sa totalité.

La modélisation objet du dictionnaire accroît comme attendu l'accès aux données du DiCo, en permettant des requêtes très variées. Mais surtout le résultat renvoyé par ces requêtes n'est plus nécessairement l'article, ce peut être une valeur de fonction lexicale ou bien une collection de régimes. Outre son interrogation et son exploitation, la modélisation facilite la révision du DiCo en autorisant une vérification aisée de l'homogénéité de chaque type d'information. Elle facilite aussi son développement par les calculs sur les données et en permettant par exemple de croiser les informations de plusieurs entrées pour construire le squelette d'une nouvelle entrée.

La modélisation objet du dictionnaire n'accroît pas seulement les possiblités d'exploitation des données. Le résultat le plus important et le moins attendu du projet est que notre vision même du dictionnaire a changé. Par exemple, au niveau de surface, le dictionnaire n'est plus nécessairement une liste de vocables. Le modèle objet peut en fait être attrapé par n'importe quel bout et le DiCo peut donc être retourné comme une vulgaire chaussette. On peut par exemple attraper le Dico par l'objet FL et produire pour chaque fonction lexicale une entrée donnant notamment les listes des valeurs pour chaque mot-clé. On a alors une vision beaucoup plus sémantique du dictionnaire, où l'on voit comment des sens comme l'intensification, la causation, la réalisation, etc. se réalisent de façons très diverses[7]. On peut aussi entrer dans le dictionnaire par les valeurs de fonctions lexicales et enrichir automatiquement l'article de dictionnaire de ces entitées lexicales un peu particulières. On n'a plus affaire alors à un dictionnaire de 2000 entrées comme au départ, mais à un dictionnaire de plusieurs dizaines de milliers d'entrées.

Le travail d'analyse et de modélisation des données constitue aussi un test des qualités formelles de présentation des données. Ce travail a ainsi permis de mettre à l'épreuve un certain nombre de concepts et de nous contraindre à une utilisation très peu permissive de ceux-ci. De cette façon, puisque l'interface est stable, on a la garantie qu'il sera possible de connecter de nouvelles composantes développées de façon indépendante (par exemple, la BDéf, un dictionnaire de définitions formalisées actuellement en cours de développement ; (Altman & Polguère, 2003)).

Nous avons débuté le projet par une phase de prototypage avec un modèle partiel du DiCo. Le prototype s'étant avéré concluant, une deuxième phase a consisté à restructurer le modèle afin de représenter toute la sémantique du DiCo. Nous sommes entrés dans la phase d'implantation consistant en la programmation des interfaces du modèle objet, le développement d'un analyseur complet et la définition de la structure de la base de données[8]. Cette phase sera suivie du

[7] Il est possible à partir de l'encodage traditionnel des fonctions lexicales d'extraire une formule sémantique (ainsi qu'un patron syntaxique). Cette idée est à la base d'un encodage plus explicite (Sylvain & Alain, 2001) que nous projetons d'ajouter automatiqment au DiCo.

[8] Les API (Application Programming Interface) sont réalisées au moyen du langage de programmation Java qui est portable sous n'importe quel système d'exploitation. Les programmes développés ainsi que l'ensemble des outils utilisés sont placées sous licence GPL, c'est-à-dire que leur utilisation est libre, selon les termes de cette licence, et peuvent donc être employés par tout chercheur désireux de développer des bases de données lexicales sur le modèle du DiCo.

développement d'une interface utilisateur afin de rendre la ressource accessible à l'ensemble de la communauté scientifique.

Références

ALTMAN J. & POLGUÈRE A. (2003). La BDéf, base de définitions dérivée du dictionnaire explicatif et combinatoire. In *Actes MTT 2003*.

ARREGI X., ARRIOLA J., ARTOLA X., DIAZ DE ILARRAZA A., GARCÍA E., V. L., SARASOLA K., SOROA A. & URIA L. (2003). Semiautomatic conversion of the euskal hitzegia basque dictionary to a queryable electronic form. *Traitement Automatique des langues, TAL, 44 n°2, 107-124*.

DENDIEN J. & PIERREL J. M. (2003). Le trésor de la langue française informatisé. un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des langues, TAL, 44 n°2, 11-37*.

C. FELLBAUM, Ed. (1998). *WordNet, An Electronic Lexical Database*. MIT Press.

FONTENELLE T. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Niemeyer.

LAPALME G. & SÉRASSET G. (2003). Batch creation of Papillon entries from DiCo. In *Workshop Papillon*.

MEL'ČUK I., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot.

POLGUÈRE A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for french. In *Proceedings of EURALEX 2000*.

POLGUÈRE A. (2003). Étiquetage sémantique des lexies du DiCo. *Traitement Automatique des langues, TAL, 44 n°2*.

SELVA T., VERLINDE S. & BINON J. (2003). Vers une deuxième génération de dictionnaires électroniques. *Traitement Automatique des langues, TAL, 44 n°2, 177-197*.

SYLVAIN K. & ALAIN P. (2001). Formal foundations of lexical functions. *Workshop on Collocation, ACL*.

ŽOLKOVSKIJ A. & MEL'ČUK I. (1965). O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija*, **6**, 23–28.