Typologie des entités lexicales d'une base de données explicative et combinatoire

Alain Polguère
OLST — Département de linguistique et de traduction
Université de Montréal
alain.polguere@umontreal.ca

1 Exposé du problème

Le premier problème que l'on doit résoudre pour construire une base de données lexicale (BD lexicale), ou pour développer une norme de construction de ce type de base (comme le *Lexical Markup Framework* de l'ISO ISO TC 37/SC 4), est de spécifier quel type d'entité linguistique on cherche à décrire. Le problème peut sembler trivial : puisqu'il s'agit de BD lexicales, il s'agit de décrire les mots de la langue. Or, d'une part, la notion de mot est loin d'être claire ; d'autre part, le type d'architecture que l'on va adopter nous amènera peut-être à considérer que nos bases décrivent en fait des entités de natures diverses, qui ne sont pas toutes ce que l'on pourrait appeler des « mots ».

Le but de cet exposé est de faire le bilan du travail effectué au cours des dernières années sur la BD lexicale du français DiCo (Polguère 2000, 2003a) afin d'identifier quels types d'entités lexicales sont véritablement prises en compte dans cette base. En effet, le travail pratique de construction du DiCo nous a graduellement amené à réviser la façon dont nous envisagions la structure même de la base. Notre approche initiale, très étroitement encadrée par le modèle théorique et descriptif de la lexicologie explicative et combinatoire (Mel'čuk *et al.* 1995), ne semblait pas laisser planer de doutes quant à ce qui était l'objet de notre description : il s'agissait de décrire la combinatoire des unités lexicales — des lexies — du français, telles que postulées par la théorie Sens-Texte et déjà modélisées dans des *Dictionnaires explicatifs et combinatoires* (DEC) — voir, pour le français, Mel'čuk *et al.* (1984, 1988, 1992, 1999). Il s'agit d'une conception très « traditionnelle », que l'on pourrait qualifier de lexicographique, de la modélisation de la structure du lexique. Selon cette conception, le lexique, en tant qu'ensemble de lexies, se décrit de façon linéaire comme une succession d'entrées correspondant aux vocables (mots polysémiques), contenant un article pour chaque acception de ces vocables. Pour remettre en question cette vision traditionnelle de la modélisation du lexique, nous allons examiner les différents types d'entités dont nous devons effectivement rendre compte dans le DiCo. Nous conclurons en expliquant pourquoi la diversité conceptuelle de ces entités doit nous conduire à réévaluer notre conception de ce qu'est l'architecture d'une BD lexicale.

2 Les entités lexicales du DiCo

Dans notre exposé, nous présenterons les différentes entités lexicales mises en jeu dans le DiCo selon un « ordre croissant d'étrangeté ». C'est-à-dire que nous commencerons par des entités lexicales qui sont traditionnellement reconnues comme telles, pour progressivement introduire des entités dont la nature lexicale est plus problématique, voire contestable. Cette hétérogénéité nous intéresse de façon toute particulière dans le cadre de notre présentation à la journée de l'Atala puisque les unités les moins lexicales sont justement celles qui sont révélatrices d'un chevauchement entre le domaine lexical et le domaine grammatical, plus particulièrement syntaxique.

Lexèmes et locutions qui font l'objet de fiches lexicographiques. La nature lexicale de ces entités est clairement établie par la lexicologie explicative et combinatoire, puisqu'il s'agit dans tous les cas de lexies. Il est néanmoins important de rappeler avec force la nature lexicale des locutions. Des entités

linguistiques du type ⌜COUP DE FOUDRE⌝, ⌜EN MIETTES⌝, ⌜LEVÉE DE BOUCLIERS⌝, etc.[1] sont des lexies de la langue à part entière, même si leur signifiant dans la phrase est une construction linguistique et non un mot-forme. Les dictionnaires de langue traditionnels ne les introduisent certes pas encore de façon systématique dans leur nomenclature, préférant enchâsser leur description à l'intérieur des articles de lexèmes. Il s'agit cependant d'une mauvaise pratique, qui pousse l'utilisateur du dictionnaire à se poser de mauvaises questions. Ainsi, on ne devrait pas avoir à se demander s'il faut chercher ⌜COUP DE FOUDRE⌝ sous COUP ou sous FOUDRE. On devrait directement y accéder dans la nomenclature générale du dictionnaire, dans une entrée spécifique. Par contre, il est vrai que les locutions ne sont pas des entités lexicales comme les autres puisque leurs signifiants ne sont pas des entités morphologiques mais des entités représentables syntaxiquement (groupes verbaux, nominaux, prépositionnels, etc.). La prise en compte des locutions au plus haut niveau de la structuration d'une BD lexicale, celui de sa nomenclature, met clairement en évidence la pénétration de la syntaxe et, plus généralement, de la grammaire de la langue dans la modélisation de la structure du lexique.

Lexèmes et locutions ciblés par des liens de fonctions lexicales. Nous faisons référence ici à des entités telles que celles apparaissant en gras dans l'exemple ci-dessous.

```
GRATITUDE
     /*[X] manifester de la G. à Y*/
{Real12/Caus1Manif} exprimer, manifester, marquer,
                    montrer, témoigner [ART ~ à N=Y],
                    _faire preuve_ [de ~ envers N=Y]
```

Remarque. Pour une présentation des formalismes du DiCo, voir (Polguère 2000, 2003a). Noter que les traits de soulignement remplacent dans le DiCo l'usage de ⌜…⌝ pour indiquer la nature locutionnelle d'une expression.

Comme dans le cas précédent, il s'agit de lexies, au sens classique. Ces lexies feront elles-mêmes l'objet d'une description dans la BD lexicale, au moyen de fiches lexicographiques standard. Il convient de noter que les lexies ciblées par certains types de liens de fonctions lexicales standard peuvent être des acceptions très particulières des vocables correspondants, acceptions qui doivent être décrites dans les dictionnaires ou les BD lexicales de façon spéciale. Tel est le cas des collocatifs en gras ci-dessous.

```
INDIGNATION
     /*[X] qui éprouve de l'I.*/
{A1} gonflé, plein, rempli [d'~] //indigné
```

La description de l'acception de REMPLI ciblée ici devra être minimale. Elle devra simplement spécifier que la lexie en question est une valeur de $\mathbf{A_1}$ pour un ensemble de lexies dénotant des sentiments, ensemble qui ne peut être spécifié en compréhension.

Remarque. Faute de place, nous allons supposer, dans ce court résumé, que le lecteur est familier avec le système des fonctions lexicales. Pour une présentation des fonctions lexicales, voir Mel'čuk (1997:41-57).

Dans une BD lexicale intégrant des définitions, telle que la BDéf (Altman et Polguère 2003), il sera quasiment impossible de produire une définition analytique[2] de cette lexie, définition qui pourrait rendre compte du fait que l'on dit *rempli d'indignation* mais pas *rempli de mécontentement*. Des lexies de ce type sont des entités lexicales sans véritable autonomie sémantique et fonctionnelle ; leur comportement en langue est modélisé plus par l'ensemble des pointeurs de liens lexicaux qui y mènent que par leur propre article lexicographique.

Collocations et expressions libres ciblées par des liens de fonctions lexicales. Pour qui connaît un peu le système des fonctions lexicales, il pourrait sembler étrange de considérer que ces dernières

1. Nous écrivons les noms de lexèmes en petites majuscules ; les noms de locutions s'écrivent aussi en petites majuscules, mais encadrés par ⌜…⌝, pour bien mettre en évidence leur nature syntagmatique.
2. Pour la notion de définition analytique (définition paraphrastique par genre prochain et différences spécifiques), voir Polguère (2003b:150-151).

puissent encoder des liens entre des lexies et des expressions qui ne sont pas à proprement parler des lexies. Considérons cependant les entités en gras dans l'exemple ci-dessous.

```
PRISE DE SANG
     /*[Y] subir une P. D. S.*/
{Oper2} avoir, se faire faire, subir [ART ~] //donner A-poss=Y/du sang
```

On pointe ici, par un lien de fonction lexicale standard (**Oper$_2$**), vers des expressions qui ne sont pas des locutions. La première est en fait une expression libre, calculable à partir d'un des **Oper$_1$** de ⌜PRISE DE SANG⌝ (*faire*) et d'une expression causative du français (*se faire* [V_{inf}]). La seconde est une collocation de la lexie SANG. La cohabitation de ces expressions avec des lexies véritables de la langue (AVOIR et SUBIR) au sein de la valeur de **Oper$_2$** de ⌜PRISE DE SANG⌝ ne fait pas d'elles des lexies, bien entendu, mais elle leur confère une certaine valeur lexicale. **Elles deviennent des nœuds du réseau lexical de la langue.**

Notons que ce phénomène d'intégration dans le réseau lexical de la langue d'expressions qui ne sont pas des lexies se produit aussi dans le cas de liens de fonction lexicale non standard, comme ci-dessous.

```
SUICIDER [SE]
{SE S. d'une façon
 spécifique} s'ouvrir les veines; "fam" _se brûler la cervelle_,
             "fam" _se faire sauter le caisson_, se tirer une balle
             dans la tête; se pendre; absorber une forte dose de N,
             prendre du poison, s'empoisonner; se jeter dans ART N
             | N est un cours d'eau ou une étendue d'eau; se jeter
             sous (les roues de) N | N est un véhicule; se jeter dans
             le vide, se jeter par la fenêtre; s'immoler par le feu
```

Nous n'avons malheureusement pas la place de justifier ici la différence de statut lexical que nous attribuons à des expressions comme ⌜*se brûler la cervelle*⌝ — une locution — vs *se tirer un balle dans la tête* — une collocation de la lexie BALLE. Cela pourra être fait au cours de l'exposé.

Le phénomène qui vient d'être examiné ne doit pas surprendre. En fait, il est légitime de se demander si, notamment, tous les liens de fonctions lexicales syntagmatiques — c'est-à-dire ceux pointant vers des collocatifs — ne créent pas automatiquement une « lexicalisation » des collocations dont ils rendent compte. Ainsi, en introduisant la formule suivante dans l'article de DiCo de ⌜PRISE DE SANG⌝ :

```
     /*[X] faire une P. D. S.*/
{Oper12} faire [ART ~ à N=Y], pratiquer [ART ~ sur N=Y]
```

ce sont en fait les deux expressions au complet *faire une prise de sang* et *pratiquer une prise de sang* que l'on insère dans le réseau lexical de la langue. Nous croyons, en fait, que les collocations sont des entités lexicales, qui sont intégrées en tant que telles dans la connaissance linguistique du locuteur. Il peut accéder à ces entités de multiples façons. Notamment, il peut les construire (ajout du collocatif à la base) ou y accéder quasiment en bloc.

Fonctions lexicales elles-mêmes. Dans Polguère (2003c), nous avons expliqué pourquoi les fonctions lexicales elles-mêmes (**Syn**, **Anti**, **S$_i$**, **Magn**, **Oper$_i$**, etc.) devaient être considérées comme des sortes de « metalexies ». Nous n'avons pas ici la possibilité de développer cette idée, mais il nous semble très important de prendre en compte, dans le design d'une BD lexicale du type du DiCo, le fait nous utilisons de telles metalexies pour encoder les liens lexicaux et, donc, tisser le réseau lexical de la langue.

Régimes lexicaux. Toute lexie prédicative se voit associée, dans une base type DiCo, un régime décrivant sa valence active (les dépendances syntaxiques qu'elle contrôle). Voici comme exemple deux régimes distincts associés aux deux acceptions du vocable MUTISME décrites dans le DiCo.

```
MUTISME#a
[Ex. “Il boudait, s’enfonçait pendant des heures dans un mutisme total.”]
X = I = de N, A-poss

MUTISME#b
[Ex. “Le gouvernement se confine dans un prudent mutisme.“]
X = I = de N, A-poss
Y = II = _à propos de_ N, sur N, _quant à_ N
```

Les deux régimes en gras ci-dessus doivent considérés comme des entités lexicales, à mi-chemin entre le lexique et la grammaire de la langue. En effet, chaque régime, pris en bloc, est caractéristique d’une possible expression au niveau syntaxique d’une structure actancielle donnée, ce qui donne parfois l’illusion que le régime (la configuration syntaxique) a par elle-même un sens. Nous pensons plutôt qu’elle est doit être vue comme outil d’expression d’un type ou d’une famille de sens et, de ce fait, est en connexion directe avec le réseau lexical de la langue. Nous n’avons malheureusement pas le loisir de développer ce point dans le présent texte.

3 Conclusion : incidence sur la structuration et l’utilisation des BD lexicales

L’énumération que nous venons de faire des différents types d’entités lexicales manipulées par dans une BD lexicale de type DiCo est malheureusement trop succincte. Il faudrait commenter en détail chaque type identifié et montrer comment il se positionne dans le réseau lexical de la langue. Nous espérons cependant avoir apporté suffisamment d’informations pour montrer l’importance théorique et pratique de la prise en compte d’un ensemble hétérogène d’entités lexicales dans la conception d’une BD lexicale. En effet, la focalisation sur les seules lexies (regroupées au sein de vocables polysémiques), si elle doit mener à une simple structuration de la base par fiches, ne fera que refléter le point de vue du lexicographe (le concepteur de la base). Or, trois perspectives doivent être considérées simultanément dans la structuration d’une BD lexicale :

1. la perspective du lexicographe, concepteur de la base ;
2. la perspective de l’usager qui, avant tout, navigue dans le réseau lexical ;
3. la perspective de la modélisation de l’activité linguistique, notamment dans un cadre de traitement automatique de la langue.

Nous croyons qu’il convient de repenser entièrement l’architecture conceptuelle et informatique des BD lexicales. Pour une tentative en ce sens, on pourra consulter Steinlin *et al.* (2004), qui propose une architecture « éclatée » de la base DiCo.

Références

Altman J., Polguère A. (2003) La BDéf : base de définitions dérivée du Dictionnaire explicatif et combinatoire, *Proceedings of the First International Conference on Meaning-Text Theory* (*MTT-03*), Paris, 43-54.

Mel’čuk I. (1997) *Vers une linguistique Sens-Texte*, leçon inaugurale, Paris : Collège de France.

Mel’čuk I. *et al.* (1984, 1988, 1992, 1999) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, vol. I-IV, Montréal : Les Presses de l’Université de Montréal.

Mel’čuk I., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve : Duculot.

Polguère A. (2000) Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French, *Proceedings of EURALEX’2000*, Stuttgart, 517-527.

Polguère A. (2003a) Étiquetage sémantique des lexies dans la base de données DiCo, *Traitement Automatique des Langues* (*T.A.L.*), 44:2, 39-68.

Polguère A. (2003b) *Lexicologie et sémantique lexicale. Notions fondamentales*, coll. « Paramètres », Montréal : Les Presses de l’Université de Montréal.

Polguère A. (2003c) Collocations et fonctions lexicales : pour un modèle d’apprentissage. In F. Grossmann et A. Tutin (dir.) : *Les Collocations. Analyse et traitement*, coll. « Travaux et Recherches en Linguistique Appliquée » — collection associée à la *Revue Française de Linguistique Appliquée*, série E, n° 1, Amsterdam : De Werelt, 117-133.

Steinlin J., Kahane S., Polguère A., El Ghali A. (2004) De l’article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux, *Proceedings of EURALEX’2004*, Lorient, 177-186.