Un langage formel d'encodage des fonctions lexicales et son application à la modélisation des collocations

Sylvain Kahane[*] et Alain Polguère[**]

Nous présentons les premiers résultats d'une recherche en cours visant l'élaboration d'un langage formel d'encodage des relations lexicales basé sur le concept de fonction lexicale (Mel'čuk *et al.* 1995:Section 3.5), postulé dans le cadre de la théorie Sens-Texte (Mel'čuk 1997 ; Polguère 1998). Ce langage formel présente l'avantage d'être « calculable », c'est-à-dire de permettre un traitement automatique de l'encodage des liens de fonctions lexicales. De plus, il permet de traiter de façon uniforme les liens paradigmatiques — les dérivations sémantiques — et syntagmatiques — les relations base-collocatif. Cette dernière caractéristique s'avère fondamentale dans la mesure où l'on peut démontrer l'interdépendance des deux phénomènes en question.

Nous ne disposons ici que de très peu d'espace. Nous ferons donc une brève mise en situation du problème, basée sur une présentation de quelques notions importantes liées aux fonctions lexicales. Nous introduirons ensuite, de façon très partielle, la méthode d'encodage des liens lexicaux que nous proposons, en mettant l'accent sur la modélisation des liens base-collocatif.

Notions de base liées aux fonctions lexicales

Le concept de ***fonction lexicale*** [= ***FL***], sur lequel repose entièrement la modélisation que nous proposons ici, permet de rendre compte de deux types de relations lexicales distinctes : les dérivations sémantiques et les relations base-collocatif — c'est-à-dire, les collocations. Nous allons tout d'abord récapituler chacune de ces notions, puis montrer en quoi elles sont en fait intimement liées.

Dérivation sémantique

On dira qu'un lien de ***dérivation sémantique*** existe entre deux lexies dans un des trois cas suivants :

1. Ces deux lexies possèdent (approximativement) le même sens. Il s'agit ici d'une dérivation sémantique vide ou quasi vide, qui correspond à un des cas bien connus de
   - synonymie exacte ou approximative (VÉLO ~ BICYCLETTE),
   - conversion (ACHETER ~ VENDRE),
   - dérivation nominale/verbale/etc. (VENTE ~ VENDRE).
   - lien de généricité (RIVIÈRE ~ COURS D'EAU).
2. Ces deux lexies possèdent des sens opposés (INTERDIRE ~ AUTORISER). Il s'agit ici de l'antonymie exacte ou approximative.
3. Une de ces lexies désigne un élément de la situation désignée par l'autre lexie. Il peut s'agir d'un participant (NAGEUR pour NAGER, DESTINATAIRE pour ENVOYER [une lettre]), d'un circonstant (LIT pour DORMIR, PISCINE pour NAGER) ou d'une caractéristique d'un participant ou circonstant (IRRITABLE pour S'IRRITER, BUVABLE pour BOIRE).

   On remarque que la dérivation **sémantique** doit être distinguée de la dérivation tout court, qui est généralement comprise de façon restrictive comme une dérivation **morphologique**.

* LaTTiCe/TALaNa, UFRL, case 7003, Université Paris 7, 2 place Jussieu, 75251 Paris Cedex 05, France <sk@ccr.jussieu.fr>.

** OLST—Département de linguistique et de traduction, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (Québec) H3C 3J7, Canada <Alain.Polguere@umontreal.ca>.

Relation base-collocatif

Une telle relation existe entre deux lexies quand celles-ci peuvent être combinées pour former une ***collocation***. Dans l'approche Sens-Texte, la collocation n'est pas définie comme étant une cooccurrence « fréquente » de lexies. C'est une combinaison de lexies qui s'oppose à une expression libre : elle est constituée d'une ***base***, qui est choisie « librement » par le locuteur en fonction de son sens, et d'un ***collocatif***, choisi pour exprimer un sens donné **en fonction de la base**. En d'autres termes, une collocation est une expression semi-idiomatique ; elle occupe, vis-à-vis du principe général de compositionalité sémantique, une position intermédiaire : entre l'expression libre et la locution.

Les collocations, comme les dérivations sémantiques, peuvent être regroupées en familles. Par exemple :

- modificateurs d'intensification : *colère*$_{\text{base}}$ *noire*$_{\text{collocatif}}$, *chute*$_{\text{base}}$ *brutale*$_{\text{collocatif}}$, *pluie*$_{\text{base}}$ *violente*$_{\text{collocatif}}$, *tomber*$_{\text{base}}$ *les quatre fers en l'air*$_{\text{collocatif}}$, …
- verbes supports : *donner*$_{\text{collocatif}}$ *un coup*$_{\text{base}}$, *recevoir*/*encaisser*$_{\text{collocatif}}$ *un coup*$_{\text{base}}$, *pousser*$_{\text{collocatif}}$ *un soupir*$_{\text{base}}$, …
- verbes de « réalisation » : *suivre*$_{\text{collocatif}}$ *un conseil*$_{\text{base}}$, *concrétiser*/*réaliser*$_{\text{collocatif}}$ *un rêve*$_{\text{base}}$, …

Nous ne nous attarderons pas plus sur la présentation de la modélisation Sens-Texte des collocations, qui a maintes fois été introduite dans la littérature traitant des FL.

Lien conceptuel entre dérivation sémantique et collocation

Les deux phénomènes en cause sont habituellement associés à la distinction entre FL paradigmatiques et FL syntagmatiques. On sait cependant qu'une séparation complète des deux types de FL n'est pas possible, cela pour au moins trois raisons :

1. Les FL paradigmatiques sont liées aux FL syntagmatiques dans le cadre des règles de paraphrasage syntaxique, du type de celles présentées dans Mel'čuk (1988). Par exemple, une expression collocationnelle à verbe support impliquant une lexie nominale L est une paraphrase d'une expression formée à partir du dérivé sémantique verbal de L ; cela se formalise au moyen des FL de la façon suivante :

   $\mathbf{V_0}(\text{L}) \equiv \mathbf{Oper_1}(\text{L})\text{—}2\text{→L}$

   Bob frappe [=$\mathbf{V_0}$(*coup*)] *Marc* ≡ *Bob donne* [=$\mathbf{Oper_1}$(*coup*)] *un coup* [=L] *à Marc.*

2. Les FL paradigmatiques se combinent aux FL syntagmatiques dans le cadre de liens « complexes » entre lexies — cf. *ignorer un conseil*, l'antonyme de *suivre un conseil*.

3. Les valeurs dites ***fusionnées*** de FL syntagmatiques correspondent en fait à des liens paradigmatiques entre lexies ; cf., par exemple, *pluie torrentielle* ≡ *déluge*.

   Il est donc nécessaire, dans le cadre de la modélisation des liens de FL entre lexies, de disposer d'un outil formel permettant de prendre en considération de façon élégante l'interaction existant en langue entre dérivations sémantiques et collocations.

Un encodage « calculable » des liens de fonctions lexicales

Trois niveaux d'encodage

Notre proposition repose sur la prémisse que le concept de FL lui-même, tel que proposé dans la théorie Sens-Texte, est en totale adéquation avec les phénomènes linguistiques dont il doit rendre compte et qu'il n'est pas utile de le réviser en profondeur. Ce qu'il convient de faire, c'est de changer la façon dont les liens de FL sont formellement encodés dans les dictionnaires Sens-Texte — cf., par exemple, Mel'čuk *et al.* (1999) — et les bases de données lexicales structurées grâce aux FL — voir Polguère (2000). L'encodage traditionnellement utilisé dans la littérature Sens-Texte ayant été développé dans le cadre des travaux sur les *Dictionnaires explicatifs et combinatoires* [= DEC], nous y référerons

sous le nom d'***encodage DEC***. En théorie, la base formelle de l'encodage DEC est simple :

1. La relation de FL reçoit un nom conventionnel (**`Syn`**, **`Anti`**, **`Magn`**, **`Oper`**$_{\mathbf{1}}$, etc.).
2. L'origine de la relation encodée par la FL, le ***mot-clé***, est l'argument de la fonction.
3. La cible de la relation est l'image de l'application de la fonction à l'argument en question : une ***valeur***, ou plutôt un ensemble de valeurs.
4. La formule décrivant la relation est **`f`**($L_{\text{mot-clé}}$) = {L_{valeur}}.

Nous proposons ici un encodage des liens de FL à trois niveaux, chaque niveau répondant à une finalité bien particulière : l'encodage explicite, l'encodage algébrique et l'encodage vulgarisé LAF (pour *Lexique Actif du Français*, voir ci-dessous).

L'***encodage explicite*** est un encodage matriciel qui décrit tout ce qui constitue le lien de FL de façon explicite. Il décrit la valeur en mettant en évidence le contenu sémantique et la patron syntaxique associé à celle-ci dans une matrice ayant la structure :

$$\begin{bmatrix} \texttt{<contenu sémantique>} \\ \texttt{<patron syntaxique>} \end{bmatrix}$$

Il s'agit d'encoder des relations entre une lexie et une valeur donnée et non la fonction lexicale elle-même, qui est une ***lexie généralisée*** —un métasigne désignant un ensemble de lexies potentielles. C'est sur cet encodage que peuvent s'effectuer les calculs automatiques de la façon la plus précise qui soit.

Faute de place, nous devons nous contenter de donner quelques exemples d'encodage, sans justification et sans démontrer toute l'étendue du formalisme. Nous empruntons tous nos exemples à la description des collocations contrôlées par la lexie COLÈRE, qui se modélise sémantiquement comme un prédicat à trois arguments : ‹colère de X envers Y à cause de Z›. Les trois collocations que nous encodons ci-dessous sont : *X* ***éprouve*** *de la colère envers Y à cause de Z*, *Y* ***encourt*** *la colère de X à cause de Z* et [*X est*] ***en*** *colère*.

$$\begin{bmatrix} \texttt{\#} \\ \texttt{V[1,\#,2,3]} \end{bmatrix} \text{ (\textit{colère}) = \textit{éprouver} [ART ~ \textit{envers} } N_Y \text{ \textit{à cause de} } N_Z] \text{ ;}$$

$$\begin{bmatrix} \texttt{\#} \\ \texttt{V[2,\#,3]} \end{bmatrix} \text{ (\textit{colère}) = \textit{encourir} [ART ~ \textit{à cause de} } N_Z] \text{ ;}$$

$$\begin{bmatrix} \texttt{\{1\}\^{}\#} \\ \texttt{A[1\^{},\#]} \end{bmatrix} \text{ (\textit{colère}) = \textit{en} [~]}$$

Dans la première formule, la première ligne de la matrice nous indique que le contenu sémantique de la valeur *éprouver de la colère* est identique au contenu du mot-clé lui-même (*colère*) — dénoté par le symbole #. La seconde ligne nous indique que la partie du discours de la valeur est V[erbe] et que sa diathèse est une structure régie à quatre actants : $X_{[1]}$ *éprouve de la colère*$_{[\#]}$ *envers* $Y_{[2]}$ *à cause de* $Z_{[3]}$.

L'***encodage algébrique*** est un encodage linéaire, proche de l'encodage DEC, qui met en évidence la structure interne des lien de FL, au niveau des opérations qu'ils impliquent. Sa finalité est d'encoder les FL elles-mêmes — c'est-à-dire les lexies généralisées. L'encodage algébrique se définit dans sa sémantique et sa syntaxe relativement à l'encodage explicite. Ainsi, les deux FL **`Magn`** (intensificateur adjectival ou adverbial) et **`Oper`**$_{\mathbf{i}}$ (verbe support) se définissent de la façon suivante :

$$\textbf{\texttt{Magn}} := \begin{bmatrix} \texttt{\{\#\}\^{}Magn} \\ \texttt{A[\#\^{}]} \end{bmatrix} \text{ ou } \begin{bmatrix} \texttt{\{\#\}\^{}Magn} \\ \texttt{Adv[\#\^{}]} \end{bmatrix} \text{ — } \textbf{\texttt{Magn}}(\textit{colère}) = \textit{forte}, \textit{grande}, \textit{terrible}, \ldots \text{ ;}$$

$$\mathbf{Oper_i} := \begin{bmatrix} \# \\ \mathtt{V[i,\#]} \end{bmatrix} \text{ — } \mathbf{Oper_1}(\textit{colère}) = \textit{éprouver}\ [\text{ART} \sim \textit{envers}\ \text{N}_\text{Y}\ \textit{à cause de}\ \text{N}_\text{Z}]$$

Comme on peut le voir en comparant les exemples ci-dessus pour la collocation *encourir la colère*, l'encodage algébrique est moins précis que l'encodage explicite ; on dira qu'il possède une moins grande ***granularité***. Il permet cependant de construire des formules complexes entièrement représentables dans l'encodage explicite à partir 1) d'un ensemble fini de FL simples et 2) d'un petit ensemble d'opérations définies, comme les FL simples, sur la base de l'encodage explicite. Nous illustrerons ici ce fait à partir de l'opération de ***produit***. Le produit de deux FL **f** et **g**, noté **g.f**, se définit de la façon suivante :

$$\mathbf{g.f} := \begin{bmatrix} \mathtt{c(f):\# \rightarrow c(g)} \\ \mathtt{pdd(f)[d(f)]:\# \rightarrow d(g)} \end{bmatrix}.$$

où :

1. `c(`**f**`)` et `c(`**g**`)` désignent la représentation du contenu sémantique de **f** et **g**,
2. `pdd(`**f**`)` est la partie du discours de **f**,
3. `d(`**f**`)` et `d(`**g**`)` sont les représentations des diathèses de **f** et **g**,
4. $\mathit{Expr}_1\mathtt{:\#}\rightarrow\mathit{Expr}_2$ signifie que, dans l'expression Expr_1, `#` (qui dénote l'argument de la fonction) doit être remplacé par l'expression Expr_2.

Soit la FL **Incep**, qui signifie ‹commencer à› et se définit par :

$$\mathbf{Incep} := \begin{bmatrix} \mathtt{Incep[\#]} \\ \mathtt{pos(\#)[\#]} \end{bmatrix}.$$

À partir de cette définition, de la définition déjà donnée pour $\mathbf{Oper_1}$ et de la définition du produit, nous pouvons dériver « algébriquement » la définition de $\mathbf{Incep.Oper_i}$:

$$\mathbf{Incep.Oper_i} := \begin{bmatrix} \mathtt{Incep[\#]} \\ \mathtt{V[i,\#]} \end{bmatrix} \text{ — par ex., } \mathbf{Incep.Oper_1}(\textit{colère}) = \textit{se mettre}\ [\textit{en} \sim].$$

Le produit de FL est une opération naturelle de « combinaison » correspondant à l'opération d'***union linguistique***, qui s'applique aux signes linguistiques en général. Une deuxième opération qui joue un rôle important dans l'encodage algébrique est la ***fusion***, qui permet notamment de modéliser les valeurs dites fusionnées des FL syntagmatiques.

L'***encodage LAF*** est en quelque sorte une vulgarisation de l'encodage algébrique. Sa finalité est d'interfacer la modélisation algébrique en vue d'une utilisation des données par le grand public. Il tire son nom du projet de *Lexique Actif du Français* [=LAF], un dictionnaire de dérivations sémantiques et de collocations du français en cours de construction. Nous ne pouvons entrer plus en détail dans ce niveau d'encodage. On trouvera une présentation du LAF et de l'encodage LAF dans Polguère (2000).

Bibliographie

Mel'čuk, I. A. (1988) Paraphrase et lexique : la théorie Sens-Texte et le Dictionnaire explicatif et combinatoire. In Mel'čuk *et al.* : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques II*, Montréal : Les Presses de l'Université de Montréal, 9-58.

Mel'čuk, I. A. (1997) *Vers une linguistique Sens-Texte. Leçon inaugurale (faite le Vendredi 10 janvier 1997)*, Collège de France, Chaire internationale.

Mel'čuk, I. A., N. Arbatchewsky-Jumarie, L. Iordanskaja, S. Mantha et A. Polguère (1999) *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques IV*, Montréal : Les Presses de l'Université de Montréal.

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

Polguère, A. (1998) La théorie Sens-Texte. *Dialangue*, Vol. 8-9, Université du Québec à Chicoutimi, 9-30.

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